

UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Identificación de Subclases de Enanas Blancas en Datos Espectroscópicos mediante Espacios Latentes y Clustering

Memoria para optar al título de
Ingeniera Civil Electrónica

Alumno
Javiera Valentina Peña Véliz

Profesor Guía
Dr. Mauricio Araya

Profesor Correferente
Dra. Odette Toloza

SEPTIEMBRE 2026

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): X Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Identificación de Subclases de Enanas Blancas en Datos Espectroescópicos mediante Espacios Latentes y Clustering

Nombre del candidato(a): Javiera Valentina Peña Véliz

Carrera / Grado: Ingeniería Civil Electrónica

Campus: Casa Central **Departamento:** Electrónica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Mauricio Araya, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

X El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

_____

_____

_____

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 30/10/2026. **Firma:** _  _____

Estudiante o Candidato(a):

Fecha: 30/10/2026 **Firma:**  _____

AGRADECIMIENTOS

QUISIERA expresar mi profunda gratitud a todas las personas que contribuyeron a mi formación académica y, muy especialmente, a mi profesor guía Mauricio Araya, que me acompañó a lo largo de toda esta memoria. Su orientación constante y la generosidad con que compartió su conocimiento fueron determinantes en cada etapa del trabajo, y por ello le estoy sinceramente agradecida.

Agradezco asimismo a la profesora Odette Toloza, cuyos conocimientos dieron fundamento físico a este trabajo. Su ayuda en el análisis e interpretación de los espectros sintetizados resultó esencial para dotar de sentido astrofísico a los resultados obtenidos.

Agradezco también a mi familia y amigos por su acompañamiento incondicional a lo largo de esta etapa.

Javiera Valentina Peña Véliz

ÍNDICE

AGRADECIMIENTOS **I**

ÍNDICE DE FIGURAS **V**

ÍNDICE DE TABLAS **X**

RESUMEN **XI**

ABSTRACT **XIII**

1. INTRODUCCIÓN **1**

2. TRABAJO A DESARROLLAR Y RESULTADOS ESPERADOS **3**

2.1. Objetivos del trabajo 3

2.2. Etapas del trabajo 4

2.2.1. Preprocesamiento de datos 4

2.2.2. Extracción de embeddings en espacio latente 4

2.2.3. Clustering no supervisado 4

2.2.4. Evaluación comparativa 5

2.2.5. Síntesis de espectros representativos y validación astronómica . . 5

2.3. Evaluaciones a realizar 5

2.4. Resultados esperados 6

3. ESTADO DEL ARTE **7**

3.1. Limitaciones de la clasificación tradicional 7

3.1.1. Formación del espectro de una enana blanca DA 7

3.1.2. Evolución espectral: difusión gravitacional, convección y acreción . 9

3.1.3. Magnetismo 10

3.2. Aprendizaje supervisado vs no supervisado 11

3.3. Redes neuronales profundas 11

3.3.1. Conceptos fundamentales . . . 11
3.3.2. Espacio latente (embedding) . . . 12
3.3.3. El fenómeno del sesgo funcional . . . 12
3.3.4. Métodos de reducción de dimensionalidad . . . 13
3.4. Clustering en aprendizaje no supervisado . . . 13
3.4.1. Algoritmos de clustering relevantes . . . 14
3.4.2. Selección del número de clusters mediante el eigengap . . . 14
3.5. Métricas de validación interna . . . 15
3.6. Técnicas de Machine Learning usadas comúnmente en astronomía . . . 17

4. ALTERNATIVAS POSIBLES DE SOLUCIÓN DEL TRABAJO 19
4.1. Alternativas bibliográficas . . . 19
4.2. Alternativas de reducción de dimensionalidad . . . 20
4.3. Alternativas de algoritmos de clustering . . . 20
4.4. Alternativas de métricas de validación y selección de K . . . 21
4.5. Criterios de estabilidad y robustez . . . 22
4.6. Resumen de alternativas consideradas . . . 22
4.7. Selección de alternativas . . . 22
4.7.1. Criterios de selección . . . 22
4.7.2. Reducción de dimensionalidad seleccionada . . . 23
4.7.3. Algoritmos de clustering seleccionados . . . 23
4.7.4. Métricas de validación y criterios de estabilidad seleccionados . . . 24

5. DESARROLLO Y EVALUACIÓN METODOLÓGICA DEL PIPELINE 26
5.1. Pipeline implementado . . . 26
5.2. Ambiente y herramientas utilizadas . . . 27
5.3. Datos utilizados . . . 27
5.3.1. Malla espectral . . . 28
5.4. Preprocesamiento de los espectros . . . 29
5.4.1. Procedimiento aplicado . . . 29
5.4.2. Calidad de los datos y limitaciones declaradas . . . 30
5.5. Red neuronal y extracción de embeddings . . . 30
5.6. Reducción de dimensionalidad . . . 32
5.7. Metodología de clustering y métricas . . . 32
5.8. Exploración visual de los espacios latentes . . . 33
5.9. Selección de k y primeros experimentos con K-Means, Spectral Clustering y DBSCAN para WDA. . . . 34
5.10. Síntesis de espectros representativos . . . 47
5.11. Experimentos con Spectral Clustering y síntesis espectral . . . 47

5.11.1. Exploradores interactivos 49
5.11.2. Matrices de migración 49
5.11.3. Capa `flatten` 51
5.11.4. Capa `dense` 57
5.11.5. Capa `dense_1` 62
5.11.6. Capa `dense_2` 67
5.11.7. Métricas internas de las particiones 70
5.12. Comportamiento de DBSCAN sobre la clase WDA 72
5.13. Aplicación del pipeline a la clase WDB 73
5.13.1. Exploración de los espacios latentes 73
5.13.2. Selección de k mediante el eigengap 74
5.13.3. Capa `flatten` 76
5.13.4. Capa `dense` 79
5.13.5. Capa `dense_1` 82
5.13.6. Capa `dense_2` 85
5.13.7. Comportamiento de DBSCAN sobre la clase WDB 89
5.13.8. Métricas internas de las particiones 90
5.14. Cierre del capítulo 91

6. RESULTADOS E INTERPRETACIÓN ASTROFÍSICA 93
6.1. Interpretación física de las particiones de la clase WDA 93
6.1.1. Particiones de DBSCAN: separación por temperatura efectiva . . 93
6.1.2. Capa `flatten`: la temperatura como único eje de separación . . 95
6.1.3. Capa `dense`: anchura de las líneas y gravedad superficial 95
6.1.4. El grupo frío de `dense` con $k = 5$ 96
6.2. Interpretación física de las particiones de la clase WDB 97
6.2.1. Capa `flatten`: separación por presencia de hidrógeno 97
6.2.2. Capa `dense`: un rasgo en emisión próximo a los 7.300 Å 98
6.2.3. Profundidad de la representación y naturaleza de los rasgos detectados 99
6.3. Discusión 99

7. CONCLUSIONES 102
7.1. Resumen 102
7.2. Conclusiones 102
7.3. Trabajo Futuro 104

BIBLIOGRAFÍA 106

Índice de figuras

5.1. Proyecciones 2D de los espacios latentes de las cuatro capas para la clase WDA. Fila superior: PCA. Fila inferior: t-SNE. En ninguna capa ni método de proyección se observan subgrupos claramente separados a simple vista. . 34
5.2. Curva de Silhouette Score en función del número de clusters k para la capa `flatten`. 35
5.3. Curva de Silhouette Score en función del número de clusters k para la capa `dense`. 35
5.4. Curva de Silhouette Score en función del número de clusters k para la capa `dense_1`. 35
5.5. Curva de Silhouette Score en función del número de clusters k para la capa `dense_2`. 36
5.6. Número de clusters indicado por el eigengap del Laplaciano en función de la profundidad de la capa, para las clases WDA y WDB. La capa más profunda requiere en ambos casos el mayor número de grupos, si bien la progresión no es monótona. 38
5.7. Particiones obtenidas por K-Means, DBSCAN y Spectral Clustering sobre la capa `flatten` de WDA (n = 1,100). En cada figura los tres algoritmos se disponen verticalmente en ese orden. 41
5.8. Particiones obtenidas por K-Means, DBSCAN y Spectral Clustering sobre la capa `dense` de WDA (n = 1,100). En cada figura los tres algoritmos se disponen verticalmente en ese orden. 42
5.9. Particiones obtenidas por K-Means, DBSCAN y Spectral Clustering sobre la capa `dense_1` de WDA (n = 1,100). En cada figura los tres algoritmos se disponen verticalmente en ese orden. 43
5.10. Particiones obtenidas por K-Means, DBSCAN y Spectral Clustering sobre la capa `dense_2` de WDA (n = 1,100). En cada figura los tres algoritmos se disponen verticalmente en ese orden. 44

5.11. Particiones de Spectral Clustering sobre la capa `flatten` de WDA (n = 1,100), para los valores de k evaluados. Columna izquierda: proyección PCA. Columna derecha: proyección t-SNE. 51
5.12. Matrices de migración entre las particiones de la capa `flatten` de WDA. Cada celda indica cuántas muestras del grupo de la fila pasan al grupo de la columna. 52
5.13. Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 3$) sobre la capa `flatten` de WDA. 54
5.14. Comparación de los 3 espectros representativos obtenidos con Spectral Clustering ($k = 3$) sobre la capa `flatten`, superpuestos al espectro promedio de la clase WDA. 55
5.15. Espectro promedio de cada grupo obtenido con DBSCAN (ε = 1,5 y `min_samples` = 10) sobre la capa `flatten` de WDA, incluido el conjunto de puntos etiquetados como ruido. 55
5.16. Comparación de los espectros representativos de los grupos de DBSCAN sobre la capa `flatten`, incluido el ruido, superpuestos al espectro promedio de la clase WDA. 56
5.17. Particiones de Spectral Clustering sobre la capa `dense` de WDA (n = 1,100), para los valores de k evaluados. Columna izquierda: proyección PCA. Columna derecha: proyección t-SNE. 57
5.18. Matrices de migración entre las particiones de la capa `dense` de WDA. Cada celda indica cuántas muestras del grupo de la fila pasan al grupo de la columna. 58
5.19. Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 5$) sobre la capa `dense` de WDA. 59
5.20. Comparación de los 5 espectros representativos obtenidos con Spectral Clustering ($k = 5$) sobre la capa `dense`, superpuestos al espectro promedio de la clase WDA. 60
5.21. Espectro promedio de cada grupo obtenido con DBSCAN (ε = 1,5 y `min_samples` = 10) sobre la capa `dense` de WDA, incluido el conjunto de puntos etiquetados como ruido. 60
5.22. Comparación de los espectros representativos de los grupos de DBSCAN sobre la capa `dense`, incluido el ruido, superpuestos al espectro promedio de la clase WDA. 61
5.23. Particiones de Spectral Clustering sobre la capa `dense_1` de WDA (n = 1,100), para los valores de k evaluados. Columna izquierda: proyección PCA. Columna derecha: proyección t-SNE. 62

5.24. Matrices de migración entre las particiones de la capa `dense_1` de WDA. Cada celda indica cuántas muestras del grupo de la fila pasan al grupo de la columna. 63
5.25. Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 2$) sobre la capa `dense_1` de WDA. 64
5.26. Comparación de los 2 espectros representativos obtenidos con Spectral Clustering ($k = 2$) sobre la capa `dense_1`, superpuestos al espectro promedio de la clase WDA. 65
5.27. Espectro promedio de cada grupo obtenido con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$) sobre la capa `dense_1` de WDA, incluido el conjunto de puntos etiquetados como ruido. 65
5.28. Comparación de los espectros representativos de los grupos de DBSCAN sobre la capa `dense_1`, incluido el ruido, superpuestos al espectro promedio de la clase WDA. 66
5.29. Particiones de Spectral Clustering sobre la capa `dense_2` de WDA ($n = 1{,}100$), para los valores de k evaluados. Columna izquierda: proyección PCA. Columna derecha: proyección t-SNE. 67
5.30. Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 8$) sobre la capa `dense_2` de WDA. 68
5.31. Comparación de los 8 espectros representativos obtenidos con Spectral Clustering ($k = 8$) sobre la capa `dense_2`, superpuestos al espectro promedio de la clase WDA. 69
5.32. Espectro promedio de cada grupo obtenido con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$) sobre la capa `dense_2` de WDA, incluido el conjunto de puntos etiquetados como ruido. 69
5.33. Comparación de los espectros representativos de los grupos de DBSCAN sobre la capa `dense_2`, incluido el ruido, superpuestos al espectro promedio de la clase WDA. 70
5.34. Proyecciones 2D de los espacios latentes de las cuatro capas para la clase WDB ($n = 1{,}100$). Fila superior: PCA. Fila inferior: t-SNE. Al igual que en WDA, no se aprecian agrupamientos evidentes en ninguna capa ni proyección. 74
5.35. Autovalores del Laplaciano normalizado para la capa `flatten` de WDB ($n = 1{,}100$). El mayor salto entre autovalores consecutivos indica $k = 2$. . 75
5.36. Autovalores del Laplaciano normalizado para la capa `dense` de WDB ($n = 1{,}100$). El mayor salto entre autovalores consecutivos indica $k = 2$. . . . 75
5.37. Autovalores del Laplaciano normalizado para la capa `dense_1` de WDB ($n = 1{,}100$). El mayor salto entre autovalores consecutivos indica $k = 2$. . 75

5.38. Autovalores del Laplaciano normalizado para la capa `dense_2` de WDB ($n = 1{,}100$). El mayor salto entre autovalores consecutivos indica $k = 8$. . 76
5.39. Particiones de Spectral Clustering sobre la capa `flatten` de WDB ($n = 1{,}100$). Columna izquierda: proyección PCA. Columna derecha: t-SNE. . 77
5.40. Partición obtenida con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$) sobre la capa `flatten` de WDB, en ambas proyecciones. 77
5.41. Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 2$) sobre la capa `flatten` de WDB. 78
5.42. Comparación de los 2 espectros representativos ($k = 2$) sobre la capa `flatten`, superpuestos al espectro promedio de la clase WDB. 78
5.43. Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 3$) sobre la capa `flatten` de WDB. 79
5.44. Comparación de los 3 espectros representativos ($k = 3$) sobre la capa `flatten`, superpuestos al espectro promedio de la clase WDB. 79
5.45. Particiones de Spectral Clustering sobre la capa `dense` de WDB ($n = 1{,}100$). Columna izquierda: proyección PCA. Columna derecha: t-SNE. 80
5.46. Partición obtenida con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$) sobre la capa `dense` de WDB, en ambas proyecciones. 80
5.47. Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 2$) sobre la capa `dense` de WDB. 81
5.48. Comparación de los 2 espectros representativos ($k = 2$) sobre la capa `dense`, superpuestos al espectro promedio de la clase WDB. 81
5.49. Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 3$) sobre la capa `dense` de WDB. 82
5.50. Comparación de los 3 espectros representativos ($k = 3$) sobre la capa `dense`, superpuestos al espectro promedio de la clase WDB. 82
5.51. Particiones de Spectral Clustering sobre la capa `dense_1` de WDB ($n = 1{,}100$). Columna izquierda: proyección PCA. Columna derecha: t-SNE. . 83
5.52. Partición obtenida con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$) sobre la capa `dense_1` de WDB, en ambas proyecciones. 83
5.53. Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 2$) sobre la capa `dense_1` de WDB. 84
5.54. Comparación de los 2 espectros representativos ($k = 2$) sobre la capa `dense_1`, superpuestos al espectro promedio de la clase WDB. 84
5.55. Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 3$) sobre la capa `dense_1` de WDB. 85
5.56. Comparación de los 3 espectros representativos ($k = 3$) sobre la capa `dense_1`, superpuestos al espectro promedio de la clase WDB. 85

5.57. Particiones de Spectral Clustering sobre la capa `dense_2` de WDB (n = 1,100). Columna izquierda: proyección PCA. Columna derecha: t-SNE. . 86
5.58. Partición obtenida con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` = 10) sobre la capa `dense_2` de WDB, en ambas proyecciones. 86
5.59. Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 8$) sobre la capa `dense_2` de WDB. 87
5.60. Comparación de los 8 espectros representativos ($k = 8$) sobre la capa `dense_2`, superpuestos al espectro promedio de la clase WDB. 88
5.61. Espectro promedio de cada grupo obtenido con DBSCAN sobre la capa `dense_2` de WDB. 88
5.62. Comparación de los espectros de la partición de DBSCAN sobre la capa `dense_2`, superpuestos al espectro promedio de la clase WDB. 89

Índice de tablas

4.1. Resumen de alternativas metodológicas consideradas en el proyecto. . . 22

5.1. Componentes seleccionados para el pipeline e implementados en este capítulo. 27
5.2. Librerías utilizadas para la implementación de la solución. 27
5.3. Clases de enanas blancas consideradas y tamaños muestrales. 28
5.4. Arquitectura completa de la red preentrenada `model_4C_3FC.h5`. Se destacan las cuatro capas seleccionadas como espacios latentes candidatos. . . . 31
5.5. Capas de la red utilizadas como espacios latentes y sus dimensiones. . . 31
5.6. Número de clusters seleccionado para cada capa según el algoritmo (clase WDA, $n = 1{,}100$). Para K-Means se emplea el máximo del Silhouette Score. Para Spectral Clustering, la heurística del eigengap del Laplaciano. 37
5.7. Número de clusters formados por DBSCAN en función de ε para cada capa de la clase WDA. Entre paréntesis, el número de puntos etiquetados como ruido ($n = 1{,}100$, `min_samples` $= 10$). 39
5.8. Métricas internas de las particiones de Spectral Clustering evaluadas sobre la clase WDA ($n = 1{,}100$). Silhouette y Calinski–Harabasz: valores más altos indican mejor partición. Davies–Bouldin: valores más bajos. 71
5.9. Número de clusters indicado por el eigengap del Laplaciano en cada capa de la clase WDB ($n = 1{,}100$), con la magnitud del salto entre autovalores consecutivos en que se sustenta. 76
5.10. Métricas internas de las particiones de Spectral Clustering evaluadas sobre la clase WDB ($n = 1{,}100$). Silhouette y Calinski–Harabasz: valores más altos indican mejor partición. Davies–Bouldin: valores más bajos. 90

6.1. Flujo normalizado en el extremo azul del espectro promedio de cada grupo obtenido con DBSCAN sobre la clase WDA ($n = 1{,}100$). El valor equivale a la razón entre el flujo azul y el flujo en la longitud de onda de normalización, y constituye por tanto una medida de la pendiente del continuo. . . . 94

RESUMEN

Las enanas blancas son remanentes estelares compactos cuya clasificación espectral actual (clases como WDA, WDB, entre otras) ha permitido grandes avances en su estudio. Sin embargo, existen indicios de que dentro de estas clases principales podrían existir subestructuras no reconocidas, originadas por diferencias en la composición química atmosférica, fenómenos de microfísica del plasma, magnetismo o acreción, que modulan la forma e intensidad de las líneas espectrales sin cambiar la etiqueta química gruesa.

El presente trabajo propone y construye un marco metodológico basado en aprendizaje no supervisado para descubrir y validar subclases de enanas blancas a partir de sus espectros. La metodología utiliza una red neuronal profunda preentrenada como extractor de características: se extraen representaciones intermedias (espacios latentes) de cuatro capas de la red, se reducen mediante Análisis de Componentes Principales (PCA) y se aplican sobre ellas algoritmos de clustering de distinta naturaleza (K-Means, Spectral Clustering y DBSCAN). La calidad de los agrupamientos se evalúa mediante métricas internas (Silhouette, Davies–Bouldin, Calinski–Harabasz y DBCV), su estructura se examina mediante matrices de migración entre configuraciones con distinto número de clusters y, para cada grupo, se generan espectros representativos que se contrastan con la fenomenología astrofísica de las enanas blancas.

Los experimentos se realizan sobre espectros del catálogo Sloan Digital Sky Survey (SDSS) para las clases WDA y WDB. Los resultados muestran que la capa de la que se extrae el embedding no determina únicamente la calidad de los agrupamientos, sino el tipo de estructura que puede recuperarse: las capas superficiales separan por rasgos directamente perceptibles del espectro, mientras que las capas intermedias construyen descriptores derivados capaces de aislar características que ninguna propiedad local expresa por sí sola. En la clase WDA las particiones recuperan ejes de variación físicos ya establecidos, la temperatura efectiva y la anchura de las líneas de Balmer ligada a la gravedad superficial. En la clase WDB se obtiene el hallazgo principal del trabajo: la partición ordena los espectros según la presencia e intensidad de las líneas de hidrógeno, reproduciendo sin supervisión la distinción entre objetos DB y DBA que en la práctica astronómica se establece mediante

inspección visual, lo que sugiere una vía de automatización de dicha tarea. Se identifica además una fuerte sensibilidad de DBSCAN a su parámetro de densidad sobre estos espacios latentes, cuya densidad varía de forma continua, lo que vuelve poco informativos sus resultados con un radio fijo y sugiere, para progresar por esta vía, una selección adaptativa de dicho parámetro o el empleo de métodos como HDBSCAN. Se muestra, por último, que es el análisis de los espectros representativos, y no las métricas internas, lo que permite interpretar físicamente los grupos obtenidos.

ABSTRACT

White dwarfs are compact stellar remnants whose current spectral classification (classes such as WDA, WDB, among others) has enabled major advances in their study. Nevertheless, there is evidence that unrecognized substructures may exist within these main classes, arising from differences in atmospheric chemical composition, plasma microphysics, magnetism or accretion, which modulate the shape and intensity of spectral lines without changing the coarse chemical label.

This work proposes and builds a methodological framework based on unsupervised learning to discover and validate white dwarf subclasses from their spectra. The methodology uses a pretrained deep neural network as a feature extractor: intermediate representations (latent spaces) are extracted from four network layers, reduced via Principal Component Analysis (PCA), and clustered using algorithms of different nature (K-Means, Spectral Clustering and DBSCAN). Cluster quality is assessed through internal metrics (Silhouette, Davies–Bouldin, Calinski–Harabasz and DBCV), their structure is examined by means of migration matrices between configurations with different numbers of clusters, and representative spectra are synthesized for each group and compared against the astrophysical phenomenology of white dwarfs.

Experiments are carried out on spectra from the Sloan Digital Sky Survey (SDSS) for the WDA and WDB classes. Results show that the layer from which the embedding is extracted does not merely affect cluster quality, but determines the kind of structure that can be recovered: shallow layers separate spectra by directly perceptible features, whereas intermediate layers build derived descriptors capable of isolating characteristics that no single local property expresses on its own. For the WDA class, the partitions recover already established physical axes of variation, namely effective temperature and the width of the Balmer lines, which is tied to the surface gravity of the star. For the WDB class the main finding of this work is obtained: the partition orders spectra according to the presence and strength of hydrogen lines, reproducing without supervision the distinction between DB and DBA objects that in astronomical practice is established through visual inspection, which suggests a path towards automating that task. A strong sensitivity of DBSCAN to its density

parameter is further identified over these latent spaces, whose density varies continuously. This renders its results uninformative under a fixed radius and points towards an adaptive selection of that parameter, or towards methods such as HDBSCAN, as the way to make progress with density-based approaches. Finally, it is the analysis of representative spectra, rather than the internal metrics, that makes it possible to interpret the resulting groups in physical terms.

Capítulo 1

INTRODUCCIÓN

Las enanas blancas son los remanentes estelares compactos que quedan cuando estrellas de baja y mediana masa agotan su combustible nuclear. Son objetos pequeños y muy densos, cuya luz visible está dominada por la atmósfera externa: una capa delgada que, pese a su poca masa, determina gran parte de las características espectrales observables. Por esta razón, la clasificación espectral de enanas blancas se centra en qué elementos dominan su atmósfera y en cómo se manifiestan sus líneas.

Las principales clases espectrales y sus características asociadas son [1]:

- DA: atmósferas dominadas por hidrógeno. Presentan líneas de Balmer en su espectro.
- DB: atmósferas dominadas por helio neutro. Presentan líneas de He I en su espectro.
- DC: espectros sin líneas prominentes (espectro continuo).
- DO: atmósferas dominadas por helio ionizado. Presentan líneas de He II en su espectro.
- DZ: presentan líneas metálicas en su espectro.
- DQ: atmósferas dominadas por carbono.
- WD: clase genérica para enanas blancas que no pueden clasificarse en las clases anteriores.

En la red neuronal preentrenada utilizada en este trabajo, las clases se denotan como WDA, WDB, etc. Variantes "híbridas" o con sufijos (como DAZ o DBA) indican mezclas o rasgos adicionales. Estas etiquetas existen porque las atmósferas de las enanas blancas pueden diferir en composición, temperatura, presión, magnetismo y procesos de acreción, lo

que altera qué líneas aparecen, su intensidad y su forma. En otras palabras, las diferencias atmosféricas se traducen en patrones espectrales distintos, y eso motiva la existencia de múltiples clases y subclases.

Además de la composición química dominante, el aspecto espectral de una enana blanca está condicionado por diversos procesos electrohidrodinámicos y de interacción que tienen lugar en su atmósfera superficial. Estos procesos pueden modular la forma e intensidad de las líneas sin cambiar la etiqueta química gruesa (DA, DB, etc.), generando subestructuras dentro de las clases tradicionales que los métodos de clasificación clásicos, basados en la inspección de unas pocas líneas, tienden a pasar por alto.

Ante esta situación, surge la necesidad de un enfoque que vaya más allá de la clasificación estándar y permita descubrir patrones ocultos en grandes volúmenes de datos. Este proyecto propone el uso de una red neuronal profunda preentrenada para obtener representaciones intermedias de los espectros estelares y aplicar sobre ellas técnicas de clustering, con el objetivo de identificar de manera objetiva, reproducible y verificable posibles subclases de enanas blancas, así como generar espectros representativos útiles para la comunidad astronómica.

La presente memoria se organiza de la siguiente manera. El Capítulo 2 describe el trabajo a desarrollar, los objetivos, las etapas, las evaluaciones a realizar y los resultados esperados. El Capítulo 3 presenta el estado del arte, abarcando las limitaciones de la clasificación tradicional, los fundamentos de las redes neuronales profundas, el espacio latente, la reducción de dimensionalidad, los algoritmos de clustering y sus métricas de validación. El Capítulo 4 evalúa las distintas alternativas de solución consideradas. El Capítulo 5 expone la propuesta de solución seleccionada. Finalmente, el Capítulo 6 detalla el desarrollo de la solución y los resultados experimentales obtenidos para las clases WDA y WDB.

Capítulo 2

TRABAJO A DESARROLLAR Y RESULTADOS ESPERADOS

Este capítulo describe el problema abordado, los objetivos del trabajo, las etapas metodológicas que componen el marco propuesto, las evaluaciones que se realizarán para verificar su correcto funcionamiento y los resultados que se espera obtener.

2.1. Objetivos del trabajo

El objetivo general es desarrollar un marco metodológico basado en aprendizaje no supervisado que permita descubrir y validar subclases de enanas blancas a partir de sus espectros, utilizando representaciones en espacio latente obtenidas de una red neuronal preentrenada y técnicas de clustering, con el fin de generar espectros representativos útiles para la comunidad astronómica.

Los objetivos específicos de este trabajo son:

- **Verificar la calidad de los espectros disponibles** y su compatibilidad con el preprocesamiento esperado por la red preentrenada.
- **Procesar los espectros estelares** y utilizarlos como entrada en una red neuronal profunda preentrenada.
- **Extraer representaciones en distintas capas intermedias de la red.**
- **Aplicar y comparar distintos algoritmos de clustering** de distinta naturaleza sobre las representaciones obtenidas.

- **Definir y aplicar métricas de evaluación** (por ejemplo, Silhouette) para medir la calidad de los agrupamientos.
- **Identificar y caracterizar posibles subgrupos** dentro de las clases principales de enanas blancas.
- **Seleccionar la capa y los métodos de clustering más adecuados** para la identificación de subgrupos.
- **Generar espectros estelares representativos de cada cluster** y compararlos con los originales para validar su coherencia con la fenomenología astrofísica de las enanas blancas.

2.2. Etapas del trabajo

Para alcanzar los objetivos planteados, se desarrollará un marco de trabajo organizado en las siguientes etapas:

2.2.1. Preprocesamiento de datos

- Descarga de espectros de enanas blancas desde bases de datos disponibles (SDSS).
- Normalización y escalado de los datos para adaptarlos a los requisitos de entrada de la red neuronal preentrenada.

2.2.2. Extracción de embeddings en espacio latente

- Procesamiento de los espectros en la red neuronal preentrenada.
- Extracción y almacenamiento de las representaciones latentes de las capas intermedias seleccionadas para utilizarlas como embeddings.

2.2.3. Clustering no supervisado

- Aplicación de técnicas de reducción de dimensionalidad para visualización y clustering preliminar.
- Aplicación de algoritmos de clustering de distinta naturaleza sobre los embeddings.
- Ajuste de los hiperparámetros clave de cada algoritmo (por ejemplo, el número de clusters k, o los parámetros ε y `min_samples` de DBSCAN).

2.2.4. Evaluación comparativa

- Inspección visual de la distribución de clusters en espacios 2D mediante reducción de dimensionalidad.
- Medición de la calidad de los clusters usando métricas internas (por ejemplo, Silhouette, Davies–Bouldin, Calinski–Harabasz, entre otros) para cada combinación capa/algoritmo/hiperparámetros, permitiendo una comparación objetiva.
- Análisis de la consistencia de los agrupamientos en experimentos repetidos.
- Estudio del efecto de la capa utilizada sobre la calidad de los agrupamientos.

2.2.5. Síntesis de espectros representativos y validación astronómica

- Generación de espectros representativos para cada cluster.
- Comparación de estos espectros sintetizados con espectros originales para su validación.
- Discusión sobre la relevancia de los subgrupos identificados y su potencial como nuevas subclases, evaluando si pueden explicarse por diferencias físicas o químicas plausibles.

2.3. Evaluaciones a realizar

Verificación de datos:

- Validar la compatibilidad dimensional y de rango espectral de los datos con los requisitos de entrada de la red preentrenada, remuestreando todos los espectros a una escala de longitud de onda común, normalizando los flujos y estandarizando la longitud de los vectores espectrales.

Evaluación del desempeño del clustering:

- Comparación entre al menos tres algoritmos de distinta naturaleza aplicados sobre los embeddings de cada capa de interés.
- Evaluación de la calidad de los agrupamientos utilizando un conjunto de métricas internas complementarias, para evitar depender de un único criterio.
- Estudio del efecto de la capa utilizada en la red sobre la calidad de los agrupamientos.

Evaluación de la síntesis espectral:

- Generación de espectros promedio o representativos de cada cluster.
- Comparación con espectros originales para verificar su fidelidad.

- Validación de la coherencia con la fenomenología astrofísica de las enanas blancas (composición atmosférica, presencia de líneas particulares, campos magnéticos, acompañantes, etc.).

2.4. Resultados esperados

- Determinar qué capa de la red preentrenada ofrece las representaciones más útiles para el clustering, identificando en qué nivel de abstracción se capturan las características más relevantes para discriminar subclases.
- Obtener una evaluación comparativa concluyente que determine el algoritmo de clustering (y su configuración) más adecuado para este tipo de datos astrofísicos, proporcionando una recomendación metodológica para estudios futuros.
- Identificar posibles subgrupos dentro de las clases principales de enanas blancas, obtenidos de manera objetiva mediante clustering.
- Obtener un conjunto de espectros sintéticos representativos de alta calidad para cada subclase identificada, que funcionen como prototipos de referencia.
- Desarrollar un pipeline de software reproducible (implementado en Python), documentado y estructurado en etapas secuenciales, que abarque el preprocesamiento, la extracción de embeddings, el clustering y la síntesis espectral, facilitando su ejecución paso a paso, su uso y su extensión.

Capítulo 3

ESTADO DEL ARTE

3.1. Limitaciones de la clasificación tradicional

La clasificación espectral de las enanas blancas se ha desarrollado históricamente mediante criterios empíricos basados en la inspección visual de líneas características o en mediciones de intensidad y anchura de ciertas transiciones atómicas (por ejemplo, las líneas de la serie de Balmer o de helio). Estos esquemas, que definen familias como DA, DB, DC, DQ, DZ o DO, resultan muy útiles para distinguir composiciones atmosféricas dominantes, pero presentan limitaciones importantes frente a los grandes volúmenes de datos y la diversidad observacional que ofrecen los catálogos modernos (por ejemplo, SDSS y Gaia). En particular, el aspecto espectral no depende solo de "qué elemento domina", sino también de procesos de plasma y de interacción que modulan la forma, anchura e intensidad de las líneas sin necesariamente cambiar la etiqueta química gruesa.

3.1.1. Formación del espectro de una enana blanca DA

Interpretar agrupamientos obtenidos sobre espectros exige precisar primero qué determina su forma. Conviene partir de una observación sobre el propio esquema de clasificación: a diferencia de lo que ocurre en la secuencia principal, las clases espectrales de las enanas blancas no indican temperatura efectiva sino **composición fotosférica**, de modo que las DA se definieron por sus líneas de Balmer intensas y las DB por sus líneas de He I [2].

El espectro emergente se calcula mediante modelos de atmósfera que reciben como parámetros la temperatura efectiva $T_{\rm eff}$, la gravedad superficial $\log g$ y las abundancias de los elementos. Los dos primeros actúan como subrogados de los valores típicos de temperatura y presión de la atmósfera [2]. Bajo el supuesto de equilibrio termodinámico local, la excitación

y la ionización del gas quedan gobernadas por los factores de Boltzmann y la ecuación de Saha, y la ley de Kirchhoff permite igualar la función fuente a la función de Planck a la temperatura local de cada capa, aunque el campo de radiación en sí mismo no se supone planckiano [2]. La forma del continuo depende además de la opacidad del gas, cuyas fuentes dominantes en una atmósfera de hidrógeno son las absorciones ligado-libre y libre-libre del hidrógeno neutro y del ion H^- [2].

Puesto que la temperatura efectiva fija el flujo total transportado a través de la atmósfera [2], la forma del continuo varía de manera sistemática con ella. De ahí que, tras normalizar un espectro en una longitud de onda fija, la pendiente del continuo entre el extremo azul y el rojo constituya el rasgo más directamente ligado a la temperatura, criterio que se emplea más adelante en este trabajo.

La visibilidad de las líneas de hidrógeno, en cambio, no acompaña a la temperatura de forma monótona, y la propia clasificación lo evidencia. Una enana blanca de atmósfera de hidrógeno puro se clasifica como DA a lo largo de la mayor parte de la secuencia de enfriamiento, pero pasa a ser DC, es decir, sin rasgos espectrales, por debajo de los 5,000 K, porque a esas temperaturas las transiciones del hidrógeno no se excitan [1]. La profundidad de las líneas resulta por tanto un indicador de temperatura menos robusto que la pendiente del continuo.

El perfil de las líneas aporta una información distinta. Su ensanchamiento está dominado por el *efecto Stark*, único mecanismo para el que existen teorías y datos satisfactorios en el caso del hidrógeno neutro, las llamadas teorías unificadas, que describen el perfil completo desde el núcleo hasta el ala lejana [2]. No es el único en juego a lo largo de toda la secuencia, pues por debajo de los 8,000 K en atmósferas ricas en hidrógeno cobra importancia el ensanchamiento por partículas neutras [2]. Los perfiles teóricos resultan muy sensibles a variaciones de los parámetros atmosféricos, razón por la cual comparar los perfiles observados con los predichos constituye el método más exitoso para determinar $T_{\rm eff}$ y $\log g$ en enanas blancas con líneas de hidrógeno [3]. Tremblay y Bergeron ilustran esa sensibilidad mostrando los perfiles de cinco líneas de Balmer calculados para valores de $\log g$ entre 7,0 y 9,0 [3].

La precisión alcanzable depende del tratamiento de la microfísica del plasma. Incorporar los efectos no ideales de forma consistente dentro del cálculo de los perfiles Stark, en lugar de mediante un parámetro *ad hoc*, modificó de manera apreciable los parámetros atmosféricos derivados para las DA, con un desplazamiento de la masa media de $+0{,}034\,M_\odot$ respecto de los análisis previos [3], y su modelado sigue siendo una de las cuestiones abiertas en el estudio de estos objetos [4].

En conjunto, dos espectros clasificados como DA pueden diferir de manera sistemática en la pendiente de su continuo y en el perfil de sus líneas sin que ninguna de esas diferencias altere la etiqueta asignada. Dado que cerca del 80 % de la población de enanas blancas es de tipo DA [3], se trata de información descartada en la mayor parte del catálogo.

3.1.2. Evolución espectral: difusión gravitacional, convección y acreción

La composición de la atmósfera de una enana blanca tampoco es un atributo fijo. Resulta de varios mecanismos de transporte de elementos que compiten entre sí [1], y que conviene distinguir porque actúan en sentidos opuestos.

La **difusión gravitacional estratifica**. En ausencia de movimientos macroscópicos que compitan con ella, los elementos más pesados difunden hacia abajo y dejan flotando en la superficie al más liviano de los presentes, lo que explica que la composición observada esté dominada por un solo elemento, en la mayoría de los casos por varios órdenes de magnitud [2]. Las observaciones, sin embargo, contradicen esa predicción de primer orden. Al inicio de la secuencia de enfriamiento entre un 20 y un 30 % de las enanas blancas presenta una atmósfera rica en helio, proporción que varía de forma sistemática con la temperatura efectiva a lo largo de la secuencia [1]. Es de esa variación estadística, y no de la observación directa de un objeto individual, de donde se deduce la existencia del fenómeno [1].

La **convección mezcla**. Conforme la estrella se enfría, el constituyente principal de la envoltura se recombina y se vuelve muy opaco, con lo que se desarrolla una zona convectiva que mezcla por completo las especies presentes en su interior. Su efecto es el opuesto al de la difusión, pues borra la segregación y produce un perfil de composición plano en toda la extensión de la zona [1]. Puede además arrastrar hacia la superficie material de capas más profundas, mecanismo con que se explica el carbono observado en los objetos de tipo DQ [2].

La **acreción de material externo aporta**. Su particularidad es que, a diferencia de los demás mecanismos, no se limita a redistribuir lo que ya está presente sino que incorpora especies nuevas al inventario químico de la estrella [1]. El material puede proceder del medio interestelar o de los restos de un cuerpo planetario disrumpido por efecto de marea, y da origen a los tipos con trazas de metales [2, 5].

La competencia entre estos procesos, conocida como **evolución espectral**, implica que la clase espectral de una enana blanca no es una propiedad permanente del objeto sino un estado que puede cambiar durante su vida [1]. Dos objetos que comparten etiqueta pueden hallarse en puntos distintos de esa competencia, de modo que dentro de una sola clase cabe esperar gradientes continuos que la etiqueta gruesa no registra.

El caso más pertinente para este trabajo es la designación **DBA**, que corresponde a enanas blancas de atmósfera dominada por helio en las que el hidrógeno resulta detectable. Las DBA clásicas, con $30{,}000 \gtrsim T_{\mathrm{eff}} \gtrsim 11{,}000$ K, llegan a representar entre el 60 y el 75 % de la población dominada por helio en esa parte de la secuencia, y sus abundancias de hidrógeno se extienden en el rango $-6{,}5 \lesssim \log N_{\mathrm{H}}/N_{\mathrm{He}} \lesssim -3{,}0$ [1]. Resulta especialmente relevante que el límite inferior de ese rango no sea físico sino observacional, pues depende de la **visibilidad de las líneas de hidrógeno**, que cambia con la temperatura efectiva [1]. La frontera entre una DB y una DBA queda así determinada en parte por la capacidad de detectar un rasgo débil.

Cabe añadir un efecto que no proviene de la composición. Algunas decenas de objetos cuyo espectro presentaba rasgos de tipo DAB o DBA resultaron ser, al examinarlos con mayor detalle, sistemas binarios no resueltos formados por una enana blanca DA y una DB [1].

3.1.3. Magnetismo

Los campos magnéticos detectados en enanas blancas aisladas se distribuyen en el rango 10^3–10^9 G [6], y su efecto sobre el espectro es el desdoblamiento Zeeman de las líneas de absorción. Cuando el campo es suficientemente intenso ese desdoblamiento resulta reconocible y el objeto recibe una clasificación magnética específica, de modo que el fenómeno no queda oculto sino que define la etiqueta.

El problema aparece en el régimen opuesto, y la literatura lo señala de forma explícita. Por debajo de aproximadamente 1 MG los desdoblamientos Zeeman **no son resolubles en el espectro de flujo** y su detección exige espectropolarimetría circular a través de las líneas [6]. Un procedimiento que opere únicamente sobre el flujo, como la clasificación visual convencional y como el desarrollado en este trabajo, carece por construcción de la información necesaria para identificar esos campos. Ello se refleja en el estado del conocimiento. Mientras el corte superior de la distribución de intensidades parece ser real, el límite inferior está determinado hoy únicamente por la sensibilidad de las campañas espectropolarimétricas disponibles, y establecer la importancia del grupo de campo débil, por debajo de 0,1 MG, requiere observaciones mucho más sensibles en telescopios de la clase de los 8 m [6]. Las mediciones actuales no permiten siquiera determinar si existe un campo mínimo o si, por debajo de cierta intensidad, todas las enanas blancas son magnéticas [6].

A la intensidad se suma la geometría, pues las enanas blancas fuertemente magnéticas tienden a exhibir estructuras de campo complejas y no dipolares, con presencia de multipolos de orden superior en algunos objetos [6]. La consecuencia conjunta es que un campo débil, una geometría complicada o un espectro ruidoso pueden introducir distorsiones sistemáticas en el perfil de las líneas que una clasificación convencional no registra. El objeto conserva una etiqueta no magnética mientras su espectro porta una deformación real, que es exactamente el tipo de subestructura intraclase que este trabajo busca. Cabe añadir que el grupo de campo intenso difiere del de campo débil en masa media, con unas $0{,}85\,M_\odot$ frente a unas $0{,}66\,M_\odot$ de las débilmente magnéticas o no magnéticas [6], lo que vincula este fenómeno con la gravedad superficial discutida en la Sección 3.1.1.

En conjunto, los tres apartados anteriores describen procesos que generan un continuo de características espectrales en lugar de categorías discretas. Limitarse a observar “qué línea está o no está” ignora la información global del espectro y oculta subestructuras potencialmente relevantes, lo que justifica el uso de redes neuronales profundas para obtener representaciones latentes del espacio espectral junto con técnicas de clustering no

supervisado.

3.2. Aprendizaje supervisado vs no supervisado

Los métodos de aprendizaje automático pueden dividirse, de forma general, en dos grandes familias: el aprendizaje supervisado y el no supervisado [7]. En el aprendizaje supervisado, el modelo se entrena a partir de ejemplos etiquetados, donde cada muestra de entrada posee una clase o valor conocido. Este enfoque ha demostrado gran eficacia en tareas como la clasificación de galaxias o la detección de tránsitos exoplanetarios, donde las categorías están claramente delimitadas mediante etiquetas generadas por humanos, como en el proyecto Galaxy Zoo [8]. Sin embargo, en el caso de las enanas blancas, las etiquetas disponibles (DA, DB, etc.) son amplias y no describen la complejidad interna de cada clase. Las posibles subestructuras finas no cuentan con etiquetas previas, lo que limita la aplicabilidad del enfoque supervisado.

El aprendizaje no supervisado utiliza conjuntos de datos que carecen de etiquetas predefinidas. El algoritmo busca estructuras internas o patrones ocultos en la distribución de los datos, agrupando objetos que comparten características inherentes. En astronomía, este enfoque se ha empleado con éxito para clasificar espectros estelares y detectar poblaciones subyacentes en grandes catálogos [9]. En el contexto de este proyecto, el aprendizaje no supervisado permite explorar el espacio espectral de las enanas blancas sin depender de categorías predefinidas, revelando posibles agrupamientos latentes que podrían corresponder a subclases físicas aún no reconocidas. Se requiere, por tanto, una metodología que derive una clasificación *a posteriori* (basada en el análisis de los datos) en lugar de imponer una clasificación *a priori*.

3.3. Redes neuronales profundas

La complejidad de los espectros de enanas blancas requiere una herramienta de extracción de características capaz de manejar relaciones no lineales en datos de alta dimensionalidad. Las redes neuronales profundas (*Deep Learning*) cumplen este rol.

3.3.1. Conceptos fundamentales

Una red neuronal es un modelo computacional inspirado en la estructura cerebral, compuesta por capas de nodos o neuronas. Una red profunda construye capas adicionales sobre la señal de entrada, donde cada capa captura una representación más abstracta de las características presentes en la capa inferior. Formalmente, una red neuronal es una función paramétrica compuesta que transforma un vector de entrada x (en este caso, un espectro) en otro vector de salida $\hat{y}$ mediante capas encadenadas. Cada capa l aplica una transformación

afín seguida de una función de activación no lineal ϕ (ReLU, tanh, Sigmoid, etc.) [7]:

$$z^{(l)} = W^{(l)} h^{(l-1)} + b^{(l)}, \qquad h^{(l)} = \phi^{(l)}\left(z^{(l)}\right), \tag{3.1}$$

con $h^{(0)} = x$ y $h^{(L)} = \hat{y}$.

Las representaciones internas $h^{(l)}$ pueden interpretarse como características progresivamente más abstractas del espectro original. En lugar de un único "cuello de botella" como en un autoencoder, en este trabajo se consideran múltiples capas de la red profunda preentrenada como posibles espacios latentes. Cada espacio latente se reduce luego de dimensionalidad mediante PCA, mitigando parcialmente los efectos de la maldición de la dimensionalidad antes del clustering.

3.3.2. Espacio latente (embedding)

Llamaremos espacio latente al espacio vectorial interno que produce una capa de la red profunda preentrenada. Cada $h^{(l)}$ constituye una representación latente del espectro de entrada, de dimensión d_l, que resume las características extraídas por la red hasta la capa l. La importancia del espacio latente radica en que una red profunda opera mediante una jerarquía de abstracción:

- **Capas iniciales:** cercanas a la entrada, capturan características de bajo nivel o patrones generales, como formas de líneas o relaciones locales entre longitudes de onda.
- **Capas intermedias y profundas:** capturan patrones más abstractos y específicos de la tarea de entrenamiento.

3.3.3. El fenómeno del sesgo funcional

La elección de la capa de la cual se extrae el embedding es un aspecto crítico. La red utilizada en este trabajo no proviene de un trabajo publicado: fue desarrollada previamente en el grupo de investigación y se emplea aquí como extractor de características ya entrenado. Su entrenamiento tuvo por objetivo clasificar espectros de enanas blancas en las categorías establecidas (DA, DB, etc.), en la misma línea de los clasificadores neuronales supervisados de enanas blancas reportados en la literatura reciente [10, 11], de modo que sus capas quedaron ajustadas a ese propósito, lo que introduce un sesgo funcional hacia la tarea de clasificación. Las capas cercanas a la salida se encuentran sesgadas por la tarea final de la red, ya que el embedding estará orientado a reforzar las fronteras de las clases ya conocidas. La consecuencia metodológica es que un embedding sesgado por la salida final podría enmascarar o distorsionar agrupaciones de subclases que no estaban presentes en el entrenamiento original. Este comportamiento es consistente con lo documentado

en la literatura de transferencia de aprendizaje: las capas iniciales de una red profunda aprenden características generales, transferibles entre tareas, mientras que las capas finales se especializan progresivamente en la tarea con que fue entrenada [12]. Por eso se deben evaluar múltiples capas.

3.3.4. Métodos de reducción de dimensionalidad

La reducción de dimensionalidad transforma el conjunto de datos de un espacio de alta dimensión a uno de baja dimensión, manteniendo la mayor cantidad de información relevante posible. En este trabajo se usa para (i) atenuar la maldición de la dimensionalidad antes de agrupar, (ii) visualizar estructuras en 2D/3D y (iii) acelerar el cómputo.

- **Análisis de Componentes Principales (PCA):** encuentra una nueva base ortogonal para el espacio de datos a partir de la diagonalización de la matriz de covarianza. Los autovectores (componentes principales) definen las nuevas direcciones y los autovalores indican la varianza explicada. Es eficiente y robusto, pero solo captura relaciones lineales y es sensible al ruido.
- **Kernel PCA (KPCA):** extiende PCA mediante el truco del kernel para capturar estructuras no lineales, a costa de un alto costo computacional (matriz de Gram $N \times N$).
- **t-SNE:** algoritmo no lineal diseñado para visualización [13]. Preserva la estructura local, pero distorsiona las distancias y densidades globales, la separación y el tamaño aparente de los grupos en el map 2D no reflejan necesariamente la disimilitud real en el espacio original. Su hiperparámetro principal es la perplejidad, que controla el número efectivo de vecinos considerado al modelar las vecindades locales, con valores típicos entre 5 y 50.
- **Isomap:** aprendizaje de *manifolds* que preserva distancias geodésicas, sensible al parámetro k y costoso ($O(N^3)$).
- **UMAP:** algoritmo no lineal que preserva estructura local y global, y se utiliza tanto para visualización como para extracción de embeddings, con buen rendimiento en datos altamente estructurados, como las líneas de absorción en espectros estelares.

3.4. Clustering en aprendizaje no supervisado

El clustering es un método clave del aprendizaje no supervisado cuyo objetivo es agrupar objetos similares sin usar etiquetas predefinidas. En este proyecto se aplica directamente sobre los embeddings extraídos de los espacios latentes candidatos. El clustering busca maximizar la similitud (cohesión) entre objetos del mismo grupo y maximizar la disimilitud

(separación) entre grupos diferentes. El resultado es un conjunto de clases *a posteriori* que, en este caso, representan las posibles subclases de enanas blancas.

3.4.1. Algoritmos de clustering relevantes

- **Particionales (K-Means, Kernel K-Means):** buscan una partición en k grupos optimizando un criterio global (suma de distancias punto-centroide). K-Means [14] es simple y rápido, pero requiere fijar k a priori y asume clusters convexos. Kernel K-Means captura formas no lineales mediante el truco del kernel.
- **Basados en densidad (DBSCAN, HDBSCAN):** agrupan puntos densamente conectados y detectan ruido, siendo efectivos para clusters de forma arbitraria. DBSCAN [15] agrupa según la densidad de puntos vecinos y etiqueta como ruido los puntos que no pertenecen a ninguna región densa, sin fijar el número de clusters a priori, requiere definir ε (distancia máxima para considerar vecinas a dos muestras) y `min_samples` (número mínimo de puntos para formar una región densa). HDBSCAN [16] elimina la dependencia de ε extendiendo DBSCAN a múltiples escalas de densidad, manejando densidades variables.
- **Jerárquicos y de propagación (Agglomerative, Affinity Propagation):** construyen una jerarquía (dendrograma) o eligen *exemplars* representativos, no requieren fijar k a priori, pero suelen ser costosos.
- **Basados en modelos (GMM):** asumen que los datos provienen de una mixtura de distribuciones gaussianas y proporcionan asignación suave (probabilidades de pertenencia).
- **Basados en grafos (Spectral Clustering):** reformulan el clustering como una partición de grafos: los puntos de datos son los nodos y las aristas codifican la similaridad entre espectros. Spectral Clustering utiliza los autovectores de la matriz Laplaciana del grafo de similaridad para proyectar los datos a un espacio de baja dimensión donde la estructura de grupos se vuelve separable, aplicando allí un algoritmo simple como K-Means [17, 18]. Es útil para descubrir clusters de formas complejas y arbitrarias, y con frecuencia supera a los algoritmos tradicionales de clustering [17], aunque es costoso para grandes conjuntos de datos. Requiere fijar el número de clusters k a priori, para lo cual dispone de un criterio propio derivado de su misma formulación: la heurística del *eigengap*, descrita a continuación.

3.4.2. Selección del número de clusters mediante el eigengap

En los algoritmos que requieren fijar k a priori, este valor suele escogerse optimizando alguna métrica de validación interna, como el Silhouette Score. Spectral Clustering, sin

embargo, admite un criterio adicional derivado directamente de la teoría espectral de grafos: la heurística del *eigengap* [17].

Sea L el Laplaciano normalizado del grafo de similitud y sean $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ sus autovalores ordenados de menor a mayor. La heurística indica escoger el número de clusters k tal que los primeros k autovalores sean pequeños y λ_{k+1} sea comparativamente grande, es decir, el índice que maximiza el salto entre autovalores consecutivos:

$$k^* = \arg\max_k \left(\lambda_{k+1} - \lambda_k\right). \tag{3.2}$$

La justificación es directa: en el caso ideal de k componentes completamente desconectadas en el grafo, el Laplaciano tiene exactamente k autovalores nulos, y el salto hasta λ_{k+1} es máximo. Cuando los grupos están conectados pero débilmente, esta estructura se preserva de forma aproximada, de modo que la magnitud del salto informa sobre cuántos cortes de baja conductancia admite el grafo.

La diferencia conceptual con el Silhouette Score es relevante: mientras este último mide compacidad y separación en términos de distancias euclidianas, el eigengap evalúa la conectividad del grafo. Por ello constituye un criterio independiente y particularmente pertinente para Spectral Clustering, cuya partición se construye precisamente sobre esa estructura de grafo.

Entre estas familias, DBSCAN y Spectral Clustering resultan particularmente atractivas para la búsqueda de subclases espectrales, por motivos complementarios: DBSCAN no impone el número de grupos y puede aislar espectros peculiares como ruido, mientras que Spectral Clustering captura geometrías no convexas del espacio latente que los métodos basados en centroides no logran representar. Ambas familias cuentan además con aplicaciones astronómicas recientes que avalan su uso sobre espacios embebidos de espectros, como se revisa en la Sección 3.6.

3.5. Métricas de validación interna

La validación interna es necesaria en el aprendizaje no supervisado porque no existen etiquetas de "verdad fundamental" para comparar los clusters descubiertos. Las métricas internas evalúan la calidad de los agrupamientos basándose únicamente en la estructura inherente de los datos.

- **Silhouette Score [19]:** mide, para cada punto, la diferencia entre la cohesión con su propio cluster $a(i)$ y la separación respecto del cluster vecino más cercano $b(i)$, mediante

$$S(i) = \frac{b(i) - a(i)}{\max\left(a(i),\, b(i)\right)}. \tag{3.3}$$

  Varía entre -1 y $+1$. Valores cercanos a 1 indican buen agrupamiento. Permite además estimar el número óptimo de clusters.

- **Índice Calinski–Harabasz (CH) [20]:** evalúa la razón entre la dispersión inter-cluster y la intra-cluster usando la varianza, mediante

$$CH = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \cdot \frac{N-k}{k-1}, \tag{3.4}$$

donde $B_k = \sum_{i=1}^{k} n_i (c_i - c)(c_i - c)^T$ y $W_k = \sum_{i=1}^{k} \sum_{x \in C_i} (x - c_i)(x - c_i)^T$ son las matrices de dispersión inter-cluster e intra-cluster, c_i y c son los centroides de cada cluster y de todo el conjunto, $n_i = |C_i|$, N es el número total de puntos y k el número de clusters. Valores más altos indican clusters densos y bien separados, favorece clusters esféricos.

- **Índice Davies–Bouldin (DB) [21]:** mide la similitud promedio entre cada cluster y el más cercano, penalizando la mala separación y la alta dispersión interna, mediante

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{s_i + s_j}{d(c_i, c_j)} \right), \tag{3.5}$$

donde $s_i = \frac{1}{n_i} \sum_{x \in C_i} \|x - c_i\|$ es la dispersión promedio del cluster i respecto de su centroide, y $d(c_i, c_j)$ es la distancia entre los centroides de los clusters i y j. Valores más bajos indican mejor clustering.

- **DBCV (Density-Based Cluster Validation) [22]:** métrica diseñada para algoritmos basados en densidad (DBSCAN, HDBSCAN), preferible a Silhouette cuando la densidad es heterogénea, mediante

$$DBCV(\mathcal{C}) = \sum_{i=1}^{k} \frac{|C_i|}{N} V_C(C_i), \qquad V_C(C_i) = \frac{DSPC(C_i) - DSC(C_i)}{\max\left(DSPC(C_i),\, DSC(C_i)\right)}, \tag{3.6}$$

donde $DSC(C_i)$ es la dispersión interna del cluster (arista máxima de su árbol de expansión mínima sobre distancias de alcanzabilidad mutua) y $DSPC(C_i)$ es su separación respecto del cluster más cercano. El puntaje varía en $[-1, 1]$, donde valores más altos indican clusters más densos y mejor separados.

- **SSE / WCSS (Within-Cluster Sum of Squares):** mide la compacidad interna de los clusters, mediante

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2, \tag{3.7}$$

la suma de las distancias al cuadrado de cada punto al centroide de su cluster. Es la función que K-Means minimiza y la base del método del codo. Disminuye monótonamente al aumentar k y no mide separación.

- **Índice Dunn [23]:** razón entre la distancia mínima de separación entre clusters y el máximo diámetro (dispersión) de cualquier cluster, mediante

$$D = \frac{\min\limits_{1\le i<j\le k} \delta(C_i, C_j)}{\max\limits_{1\le l\le k} \Delta(C_l)}, \tag{3.8}$$

  donde $\delta(C_i, C_j)$ es la distancia mínima entre puntos de clusters distintos y $\Delta(C_l)$ es el diámetro (distancia máxima entre puntos) del cluster C_l. Valores más altos indican mejor clustering, aunque es sensible al ruido y a los outliers.

- **Índice $\mathbf{S}_{Dbw}$ [24]:** combina la dispersión promedio de los clusters respecto de la dispersión total con una medida de separación entre centroides normalizada, mediante

$$S_{Dbw}(k) = \alpha \cdot Scat(k) + Dis(k), \tag{3.9}$$

  donde $Scat(k)$ mide la dispersión promedio de los clusters relativa a la dispersión total de los datos, $Dis(k)$ mide la separación entre centroides normalizada por la desviación estándar de los clusters, y α pondera la separación frente a la dispersión. Valores más bajos indican mejor clustering.

- **Índice Xie–Beni [25]:** razón entre la cohesión (SSE) y la separación mínima al cuadrado entre centroides, mediante (en su especialización a asignación dura, sin pertenencia difusa)

$$XB = \frac{\sum\limits_{i=1}^{k} \sum\limits_{x\in C_i} \|x - c_i\|^2}{N \cdot \min\limits_{i\neq j} \|c_i - c_j\|^2}, \tag{3.10}$$

  valores más bajos indican mejor clustering, aunque asume clusters de forma esférica y es sensible a outliers.

3.6. Técnicas de Machine Learning usadas comúnmente en astronomía

En la literatura astronómica, K-Means se usa como base reproducible cuando los grupos esperados son compactos en un espacio ya reducido (por ejemplo, tras aplicar PCA sobre espectros), como en la clasificación de más de 170.000 espectros estelares de SDSS/SEGUE [26] y en el agrupamiento de 153.847 espectros de alta resolución de APOGEE [27]. Enfoques no supervisados emparentados, basados en mezclas gaussianas sobre un subespacio discriminante, se han aplicado a la clasificación de cientos de miles de espectros de galaxias del SDSS [28]. Cuando las subpoblaciones se solapan y se requiere probabilidad de pertenencia, GMM es estándar, por ejemplo en quimio-cinemática estelar con Gaia y APOGEE [29]. DBSCAN y HDBSCAN son muy populares en Gaia para el hallazgo de cúmulos abiertos, por su tolerancia al ruido y a formas no convexas. Trabajando

sobre el espacio astrométrico de Gaia EDR3 se han descubierto cientos de nuevos cúmulos con estos algoritmos [30].

En cuanto a las métricas, Silhouette se usa para ajustar hiperparámetros en embeddings de espectros y fotometría, por ejemplo, seleccionando la configuración de UMAP + K-Means que maximiza Silhouette en fuentes con exceso Hα [31], Calinski–Harabasz se emplea en el *chemical tagging* del sondeo Gaia-ESO [32], Davies–Bouldin para valorar agrupamientos en estudios de remanentes de supernova [33], y DBCV como métrica de validación para clustering basado en densidad en espacios embebidos [34]. Cabe destacar que este último trabajo [34] emplea precisamente Spectral Clustering combinado con validación por densidad para la caracterización agnóstica de espectros de exoplanetas, es decir, el mismo esquema metodológico (agrupamiento sobre representaciones embebidas de espectros, validado con métricas internas y DBCV) que se adopta en esta memoria.

En el dominio específico de las enanas blancas, los esfuerzos recientes de aprendizaje automático se han concentrado en la clasificación supervisada: se han entrenado redes neuronales sobre decenas de miles de espectros del SDSS, alcanzando precisiones cercanas al 100 % para las clases DA, DB y DZ [10], y se han clasificado y parametrizado más de cien mil candidatas a partir de los espectros XP de Gaia DR3 [11]. Estos trabajos consolidan la viabilidad de clasificadores supervisados precisos para enanas blancas, enfoque al que pertenece también la red preentrenada que este proyecto reutiliza como extractor de características, pero dejan comparativamente inexplorada la búsqueda no supervisada de subestructura dentro de las clases ya establecidas, que es precisamente el vacío que este trabajo aborda.

Capítulo 4

ALTERNATIVAS POSIBLES DE SOLUCIÓN DEL TRABAJO

La identificación de subclases de enanas blancas a partir de sus espectros es un problema intrínsecamente no supervisado: las fronteras entre tipos espectrales son difusas, existen poblaciones raras o poco representadas y, en muchos casos, no se dispone de etiquetas confiables a gran escala. En este capítulo se realiza una evaluación sistemática de distintas alternativas de solución, considerando herramientas bibliográficas consolidadas y métodos específicos de reducción de dimensionalidad, algoritmos de clustering y métricas de validación interna. Cada alternativa se analiza en función de su pertinencia al problema, su costo computacional, sus requerimientos de datos y su grado de alineación con las restricciones prácticas del proyecto.

4.1. Alternativas bibliográficas

- **DRACULA** [35] (Dimensionality Reduction And Clustering for Unsupervised Learning in Astronomy): paquete de Python desarrollado para la exploración de la diversidad espectral de Supernovas Tipo Ia. Implementa la secuencia completa de reducción de dimensionalidad, clustering y visualización, en particular, su módulo de visualización mediante mapas autoorganizados (SOM) reporta, para cada celda de la grilla, el espectro promedio y la desviación estándar de los espectros asignados a ella. Dado que en este proyecto ya se cuenta con una red preentrenada, la fase de entrenamiento de autoencoders de DRACULA no aporta a la solución final, por lo que se descarta como alternativa directa para la extracción de embeddings. No obstante, se rescata y adapta su procedimiento de síntesis de espectros representativos a partir

de los clusters encontrados, el cual se incorpora en la etapa de síntesis espectral de esta memoria (Sección 5.10).

- **pyUPMASK:** algoritmo no supervisado diseñado para la decontaminación de cúmulos estelares. Separa los datos en dos clases ("cúmulo" y "campo"), mientras que en este proyecto se busca descubrir nuevas clases, no separar en dos. Se descarta.
- **ASPECT:** herramienta de exploración basada en un Self-Organizing Map (SOM) de muy alta resolución. Aunque es un referente metodológico directo, su implementación excede los recursos computacionales del proyecto (tiempos del orden de 100 días por entrenamiento). Se descarta como solución directa.
- **Fisher-EM [28]:** modelo de mezcla gaussiana en un subespacio discriminante aplicado a 702.248 espectros del SDSS, obteniendo 86 grupos espectrales. Constituye un referente relevante, pero su implementación directa excede los recursos del proyecto. Se descarta como herramienta práctica.

4.2. Alternativas de reducción de dimensionalidad

Se evaluaron métodos lineales y no lineales. PCA se selecciona como método prioritario, y el efectivamente utilizado en el pipeline, por su bajo costo, estabilidad y amplio uso en pipelines astronómicos. Kernel PCA, Random Projection, Isomap, Spectral Embedding, Diffusion Maps y SOM se descartan por falta de escalabilidad, sensibilidad a hiperparámetros o requerimientos de muestreo denso. UMAP se identifica como la principal alternativa no lineal, reservada para trabajo futuro por su capacidad de capturar estructuras no lineales al combinarse con algoritmos de densidad. t-SNE se emplea exclusivamente como herramienta de visualización y análisis exploratorio del espacio latente. LLE se considera solo de forma condicional, a revaluar si aumenta significativamente el número de espectros disponibles.

4.3. Alternativas de algoritmos de clustering

Como se describió en el Capítulo 3, los algoritmos de clustering se agrupan en familias con supuestos geométricos distintos (particionales, basados en densidad, basados en grafos, jerárquicos y basados en modelos). Dado que el objetivo de este trabajo es desarrollar un pipeline reproducible, y no un benchmark exhaustivo de algoritmos, el criterio de selección adoptado no es probar la mayor cantidad de métodos posible, sino **cubrir las principales familias con un representante de cada una**. Esta estrategia permite evaluar paradigmas de agrupamiento genuinamente distintos con un número acotado de experimentos, reservando las variantes adicionales de cada familia para trabajo futuro. Bajo este criterio se evaluaron los siguientes algoritmos:

- **K-Means (familia particional):** ampliamente usado en análisis espectrales por su simplicidad, rapidez y escalabilidad. Se selecciona como representante de los métodos basados en centroides.
- **DBSCAN [15] (familia basada en densidad):** identifica clusters de forma arbitraria y detecta ruido sin fijar el número de grupos a priori. Su capacidad de aislar espectros peculiares como outliers es astrofísicamente valiosa. Se selecciona como representante de los métodos de densidad.
- **Spectral Clustering [17] (familia basada en grafos):** utiliza la descomposición espectral del grafo de similitud para capturar estructuras no convexas en espacios comprimidos, aunque es costoso para grandes conjuntos. Se selecciona como representante de los métodos basados en grafos.

Las siguientes alternativas se identifican como **viables y se reservan para trabajo futuro**, por representar familias o variantes que amplían, sin duplicar, la cobertura de los tres representantes seleccionados:

- **GMM (familia basada en modelos):** modela los clusters mediante una mixtura de gaussianas con asignación suave (probabilidad de pertenencia), lo que resulta especialmente pertinente para subtipos que se solapan, como las variantes híbridas DAB/DBA. Aporta además la selección de k por criterios BIC/AIC.
- **HDBSCAN (variante jerárquica de densidad):** extiende DBSCAN eliminando la dependencia del parámetro ε y manejando densidades variables, por lo que aborda directamente la sensibilidad a hiperparámetros que exhibe DBSCAN. Se valida de forma natural mediante DBCV.
- **Agglomerative Clustering (familia jerárquica):** no asume geometría esférica y entrega un dendrograma interpretable que permite visualizar la jerarquía de subgrupos y seleccionar el número de clusters cortando el árbol a distintas alturas.

Finalmente, se descartan **Kernel K-Means** (costo cuadrático que no compensa su ganancia sobre Spectral Clustering para capturar no linealidad) y **Affinity Propagation** (alto costo por la matriz $N \times N$ y fuerte sensibilidad al parámetro `preference`).

4.4. Alternativas de métricas de validación y selección de K

Las métricas de validación interna consideradas son Silhouette Score, Calinski–Harabasz, Davies–Bouldin y DBCV, cada una adecuada para distintos tipos de algoritmo.

Para la selección del número de clusters k en los algoritmos que lo requieren a priori se dispone de criterios de distinta naturaleza. El más general es el máximo del Silhouette Score,

aplicable a cualquier algoritmo, que privilegia particiones compactas y bien separadas en términos de distancias euclidianas. Para Spectral Clustering existe además la heurística del *eigengap* (Sección 3.4.2), que deriva de la misma teoría espectral en que se funda el algoritmo y evalúa la conectividad del grafo de similitud en lugar de la compacidad geométrica. Finalmente, otras familias disponen de criterios propios: AIC/BIC en el caso de GMM y el análisis del dendrograma en el de Agglomerative Clustering.

4.5. Criterios de estabilidad y robustez

- **Proporción de ruido:** cantidad de puntos etiquetados como ruido, indicador directo del desempeño de DBSCAN.
- **Consistencia del dendrograma y estabilidad jerárquica:** criterios asociados a las alternativas reservadas para trabajo futuro (Agglomerative y HDBSCAN), que evalúan si la estructura jerárquica se mantiene al variar el criterio de enlace o el nivel de corte.

4.6. Resumen de alternativas consideradas

La Tabla 4.1 resume las alternativas metodológicas consideradas en el proyecto, distinguiendo los algoritmos de clustering seleccionados como representantes de cada familia de aquellos reservados para trabajo futuro.

Tabla 4.1: Resumen de alternativas metodológicas consideradas en el proyecto.

Reducción de dim.	**Algoritmos de clustering**	**Métricas de validación**
PCA	K-Means (particional)	Silhouette Score
UMAP	Spectral Clustering (grafos)	Calinski–Harabasz
t-SNE (visualización)	DBSCAN (densidad)	Davies–Bouldin
LLE (condicional)	*GMM (futuro)*	DBCV
—	*HDBSCAN, Agglomerative (futuro)*	—

4.7. Selección de alternativas

Sobre la base de las alternativas revisadas en las secciones anteriores, esta sección fija las que integran el pipeline del trabajo y expone el criterio con que se escogieron.

4.7.1. Criterios de selección

La elección entre alternativas se guía por cuatro criterios prácticos, derivados de las restricciones del proyecto.

- **Escalabilidad:** capacidad de manejar un aumento en el número de espectros sin degradar el rendimiento ni la calidad de los resultados.
- **Simplicidad de implementación:** facilidad de integrar y ajustar la alternativa dentro del ecosistema de librerías disponible (`scikit-learn`, `umap-learn`, `hdbscan`).
- **Costo:** costo computacional en tiempo de ejecución, consumo de memoria y necesidad de hardware especializado.
- **Estabilidad:** robustez de los resultados frente a variaciones en la inicialización, perturbaciones de los datos o ajustes moderados de hiperparámetros.

A estos se suma un criterio metodológico determinante para la etapa de clustering: dado que el objetivo del trabajo es desarrollar un pipeline reproducible y no un benchmark exhaustivo de algoritmos, la selección no busca probar la mayor cantidad de métodos posible, sino **cubrir las principales familias de clustering con un representante de cada una**. Esta estrategia permite contrastar paradigmas de agrupamiento genuinamente distintos con un número acotado de experimentos.

4.7.2. Reducción de dimensionalidad seleccionada

Se selecciona **PCA** como método de reducción de dimensionalidad del pipeline. La elección se fundamenta en su bajo costo computacional, su carácter determinista, la disponibilidad de implementaciones maduras y su amplio uso en pipelines astronómicos, factores que satisfacen simultáneamente los criterios de escalabilidad, simplicidad y costo. PCA proporciona además un mapeo explícito y reutilizable, lo que permite homogeneizar a una dimensión común las representaciones de capas de distinto tamaño.

Adicionalmente, se selecciona **t-SNE** como herramienta exclusiva de visualización y análisis exploratorio del espacio latente, sin emplearlo como paso previo al clustering, dada su distorsión de la geometría global. Las restantes alternativas no se incorporan al pipeline: UMAP se reserva como principal extensión no lineal para trabajo futuro, mientras que LLE, Kernel PCA, Isomap, Random Projection, Spectral Embedding, Diffusion Maps y SOM se descartan por limitaciones de escalabilidad, sensibilidad a hiperparámetros o requerimientos de muestreo denso.

4.7.3. Algoritmos de clustering seleccionados

Aplicando el criterio de cobertura de familias se seleccionan tres algoritmos, uno por cada una de las familias con supuestos geométricos más contrastantes descritas en la Sección 4.3.

- **K-Means**, por la familia particional, en calidad de línea base reproducible.

- **Spectral Clustering**, por la familia basada en grafos, por su capacidad de capturar estructuras no convexas.
- **DBSCAN**, por la familia basada en densidad, por no requerir fijar el número de grupos y por aislar espectros peculiares como ruido, característica astrofísicamente valiosa.

La inclusión conjunta de los tres responde a una lógica de complementariedad: la suposición de clusters aproximadamente esféricos que hace K-Means en el espacio reducido puede no cumplirse, y son precisamente Spectral Clustering y DBSCAN los que permiten capturar las estructuras no convexas o basadas en densidad que K-Means no representa.

Por otra parte, **GMM**, **HDBSCAN** y **Agglomerative Clustering** se identifican como alternativas aptas, pero se reservan para trabajo futuro (Capítulo 7), por representar familias o variantes que amplían la cobertura ya alcanzada sin duplicarla. Su incorporación queda fuera del alcance temporal de este trabajo, pero no de la arquitectura del pipeline, diseñada para admitirlas sin modificaciones estructurales.

4.7.4. Métricas de validación y criterios de estabilidad seleccionados

Para evaluar la calidad de las particiones se seleccionan el **Silhouette Score**, el índice **Calinski–Harabasz** y el índice **Davies–Bouldin**, aplicables a los tres algoritmos, complementados con el **DBCV** en el caso de DBSCAN, por estar diseñado específicamente para agrupamientos basados en densidad. El uso de un conjunto de métricas complementarias, y no de un único criterio, permite detectar particiones ficticias que optimizan una métrica en particular.

Para los dos algoritmos que requieren fijar el número de clusters a priori se selecciona, en cada caso, el criterio más apropiado a su naturaleza. En **K-Means** se emplea el **máximo del Silhouette Score**, evaluado sobre un barrido de valores de k, que es el criterio estándar para un algoritmo cuya partición se define por distancias a centroides. En **Spectral Clustering** se emplea la **heurística del eigengap** (Sección 3.4.2), derivada de la misma teoría espectral de grafos en que se funda el algoritmo.

Como criterio de robustez se selecciona el **análisis de migración** entre configuraciones con distinto número de clusters, que permite determinar si los grupos se conservan, se fusionan o se subdividen de forma ordenada al modificar el hiperparámetro k, complementado con el seguimiento de la cantidad de puntos etiquetados como ruido en el caso de DBSCAN, indicador directo de qué fracción de la muestra queda efectivamente particionada.

En síntesis, el pipeline experimental se construye en torno a tres algoritmos: K-Means, Spectral Clustering y DBSCAN, uno por cada una de las tres familias con supuestos geométricos más contrastantes, con PCA como reductor de dimensionalidad y t-SNE como

herramienta de visualización, mientras que GMM, HDBSCAN y Agglomerative Clustering quedan documentados para trabajo futuro. Su integración en etapas secuenciales y su validación experimental se desarrollan en el Capítulo 5.

Capítulo 5

DESARROLLO Y EVALUACIÓN METODOLÓGICA DEL PIPELINE

Este capítulo describe la implementación del pipeline y su evaluación metodológica sobre datos reales de enanas blancas del catálogo SDSS. Se detallan la integración de las etapas seleccionadas en el Capítulo 4, el ambiente de desarrollo, los datos utilizados y su preprocesamiento, la extracción de embeddings desde las capas de la red, la reducción de dimensionalidad, la metodología de clustering y la síntesis espectral. Sobre esa base se obtienen las particiones de las cuatro capas latentes para las clases WDA y WDB y se evalúan mediante métricas internas y matrices de migración. La interpretación astrofísica de esas particiones se aborda en el Capítulo 6.

5.1. Pipeline implementado

El pipeline integra las alternativas seleccionadas en la Sección 4.7 en una secuencia de etapas, resumida en la Tabla 5.1. Los espectros preprocesados se procesan mediante la red neuronal preentrenada para extraer representaciones latentes de distintas capas, y estas se reducen a una dimensión común mediante PCA. Sobre las representaciones resultantes se aplican los tres algoritmos de clustering seleccionados. La calidad y la robustez de las particiones se evalúan con el conjunto de criterios definido y, finalmente, se sintetizan espectros representativos por cluster, adaptando el procedimiento de espectro promedio y banda de desviación estándar identificado en DRACULA (Sección 4.1). Las secciones

siguientes detallan la implementación de cada etapa en el orden en que se ejecutan.

Tabla 5.1: Componentes seleccionados para el pipeline e implementados en este capítulo.

Etapa	**Alternativa seleccionada**
Extracción de características	Red neuronal preentrenada (capas intermedias)
Reducción de dimensionalidad	PCA
Visualización	PCA y t-SNE
Clustering	K-Means, Spectral Clustering y DBSCAN
Validación de calidad	Silhouette, Calinski–Harabasz, Davies–Bouldin, DBCV
Validación de robustez	Matrices de migración y recuento de ruido
Síntesis espectral	Espectro promedio y desviación estándar por cluster

5.2. Ambiente y herramientas utilizadas

Las soluciones se desarrollaron en lenguaje Python, utilizando las librerías principales resumidas en la Tabla 5.2.

Tabla 5.2: Librerías utilizadas para la implementación de la solución.

Propósito general	**Manejo de datos**	**Visualización**	**ML / Deep Learning**
NumPy	pandas	Matplotlib	scikit-learn (PCA, K-Means, DBSCAN, Spectral)
os / pathlib	—	seaborn	TensorFlow / Keras (extracción de capas)
SciPy	—	ipywidgets (interact)	—

El desarrollo se realizó en cuadernos Jupyter (se utilizaron tres cuadernos para la construcción incremental del pipeline, y `exp_wda` y `exp_wdb` para los experimentos por clase), lo que permitió iterar de forma reproducible sobre cada etapa.

5.3. Datos utilizados

Los datos corresponden a espectros obtenidos del Sloan Digital Sky Survey. Cada objeto se almacena en un archivo de texto `.dat` con tres columnas: longitud de onda, flujo y error del flujo.

Las etiquetas espectrales no provienen de un algoritmo, sino de la clasificación manual realizada por astrónomos en el marco del esfuerzo de clasificación espectroscópica del proyecto, y se distribuyen en archivos CSV que contienen, entre otros campos, el nombre del archivo, la fecha juliana modificada (MJD), el identificador del objeto, la clasificación asignada y un indicador de calidad del dato. El cruce entre espectro y etiqueta se realiza mediante la combinación de MJD e identificador de objeto, que aparece tanto en el nombre

del archivo `.dat` como en los CSV. La distribución de las clases más relevantes para este trabajo se resume en la Tabla 5.3.

Tabla 5.3: Clases de enanas blancas consideradas y tamaños muestrales.

Clase	**N.º de espectros etiquetados**	**Con archivo .dat asociado**	**Usados en experimentos**
WDA	13.748	13.723	1.100
WDB	1.186	1.180	1.100

Las clases WDA y WDB se eligieron por ser las dos clases principales con mayor cantidad de muestras. La primera agrupa los objetos cuyo espectro presenta **líneas de hidrógeno**, en particular las de la serie de Balmer, y la segunda aquellos en que se observan **líneas de helio**. En ambos casos el criterio de clasificación es observacional, basado en qué líneas resultan visibles en el espectro, y de él se infiere la composición dominante de la atmósfera. Para ambas clases se fijó un tamaño muestral de $n = 1{,}100$ espectros, seleccionados con una semilla fija (`random_state` $= 42$) para garantizar reproducibilidad.

Conviene precisar el papel que cumplen estas etiquetas, dado que el método desarrollado es no supervisado. Las etiquetas se emplean **exclusivamente para delimitar la muestra**: definen el conjunto de espectros sobre el cual se busca subestructura, de modo que la búsqueda quede acotada al interior de una clase ya establecida y cualquier agrupamiento encontrado sea necesariamente una subdivisión de ella. En ningún punto posterior del procedimiento se entrega información de etiqueta a los algoritmos: la red recibe únicamente el vector de flujo, y el clustering opera sobre las representaciones latentes sin acceso a la clase, al identificador del objeto ni a ninguna otra variable externa. Tampoco se evalúa la concordancia entre las particiones obtenidas y las etiquetas, puesto que lo que se busca no es reproducir la clasificación conocida sino encontrar divisiones dentro de ella.

La etiqueta genérica WD, que agrupa las enanas blancas sin subclase determinada, resulta de interés por su carácter heterogéneo, pero su análisis se plantea como trabajo futuro (Capítulo 7).

5.3.1. Malla espectral

Todos los espectros empleados comparten la malla de longitudes de onda estándar del espectrógrafo, lo que resulta determinante para el procedimiento descrito más adelante. La malla consta de 4,648 puntos y cubre el rango que va de 3,565,96 a 10,396,35 Å.

El muestreo es **logarítmico** y no lineal: lo que permanece constante de un píxel al siguiente es el incremento del logaritmo de la longitud de onda, con $\Delta \log_{10} \lambda = 1{,}0000 \times 10^{-4}$. El paso lineal equivalente, en consecuencia, no es constante: crece desde 0,82 Å en el extremo azul hasta 2,40 Å en el rojo, un factor cercano a 2,9. Esta elección de malla es la habitual

en espectroscopía estelar, porque un paso constante en $\log \lambda$ equivale a un paso constante en velocidad, de 69,0 $\mathrm{km\,s^{-1}}$ por píxel en este caso, lo que simplifica el tratamiento de los desplazamientos de las líneas.

Puesto que la malla es común a todos los archivos, **no se aplica ninguna interpolación** a una malla de referencia: la correspondencia entre el índice de un vector y su longitud de onda es idéntica para todos los espectros y basta con operar sobre índices. Esta propiedad es la que permite que el recorte y la normalización descritos en la Sección 5.4 actúen sobre la misma región espectral en todos los objetos.

5.4. Preprocesamiento de los espectros

5.4.1. Procedimiento aplicado

Cada espectro se carga desde su archivo `.dat` y se procesa para ser compatible con la entrada esperada por la red neuronal. De las tres columnas del archivo se utiliza únicamente la del **flujo**. La columna de longitud de onda se emplea solo para verificar la malla, dado que, al ser común a todos los archivos, el procedimiento puede operar sobre índices (Sección 5.3.1). El procedimiento, implementado en la función `process_spectrum`, consta de tres pasos.

Recorte de los extremos. Se descartan los primeros 48 y los últimos 400 puntos del vector de flujo, con lo que la longitud pasa de 4,648 a 4,200 puntos y el rango retenido queda comprendido entre 3,605,6 y 9,481,6 Å. El objetivo es eliminar los bordes del rango instrumental, donde la sensibilidad del espectrógrafo decae y la relación señal-ruido es considerablemente menor.

Ajuste a la longitud de entrada de la red. La red espera vectores de 3,750 puntos, de modo que se recortan 225 puntos adicionales en cada extremo, de forma centrada. El rango espectral que finalmente ingresa a la red queda así comprendido entre 3,797,3 y 9,002,9 Å.

Este paso incluye una rama alternativa para el caso de un espectro más corto de lo esperado: el *padding* consiste en completar el vector con ceros hasta alcanzar los 3,750 puntos, añadidos al final. Con el conjunto de datos empleado esa rama no llega a ejecutarse, puesto que todos los archivos comparten la malla de 4,648 puntos y tras el primer recorte quedan 4,200, siempre por encima del tamaño requerido, de modo que funciona como resguardo defensivo. Conviene advertir, de todos modos, que si llegara a activarse rompería la correspondencia entre índice y longitud de onda, porque desplazaría el espectro respecto de los demás en lugar de alinearlo: un archivo con una malla distinta exige interpolación, no relleno.

Normalización. El flujo se divide por su valor en un índice de referencia fijo, el 2,100 del vector resultante, que corresponde a 6,158,6 Å. Si ese valor es nulo o no numérico se emplea el flujo medio del espectro, y si tampoco este resulta utilizable el espectro se descarta. Esta operación elimina la escala absoluta de flujo, que depende de la distancia y del radio del

objeto, y deja los espectros en una escala comparable entre sí. Es importante notar que *no* elimina la pendiente del continuo: tras la normalización, el valor que toma el espectro en el extremo azul expresa la razón entre el flujo azul y el flujo en la longitud de onda de normalización, que según lo expuesto en la Sección 3.1.1 constituye un indicador de la temperatura efectiva.

El preprocesamiento no incorpora ninguna corrección de longitud de onda. Conviene precisar que, al tratarse de objetos de nuestra propia galaxia, el desplazamiento pertinente no es el cosmológico sino el **corrimiento gravitacional** (*gravitational redshift*), consecuencia de la intensa gravedad superficial de la enana blanca, al que se suma el desplazamiento Doppler producido por la velocidad radial del objeto. Ambos son del orden de unas decenas de $\mathrm{km\,s^{-1}}$, muy por debajo de la anchura que el efecto Stark confiere a las líneas de Balmer en estos objetos, lo que puede constatarse al verificar que las líneas de la serie en longitud de onda de reposo coinciden con los rasgos de absorción observados. Por ese motivo el agrupamiento puede operar sobre la malla sin corregir, aunque la identificación precisa de un rasgo estrecho por su longitud de onda sí requeriría incorporar estas correcciones.

5.4.2. Calidad de los datos y limitaciones declaradas

El preprocesamiento es deliberadamente simple, por estar determinado por el formato de entrada de una red ya entrenada. Conviene por ello declarar qué tratamientos habituales no se aplican.

- **Errores del flujo.** La tercera columna del archivo contiene el error del flujo y no se utiliza, de modo que no hay ponderación por varianza inversa ni interpolación de los píxeles que la reducción marca como defectuosos. Estos se concentran en los extremos del rango, que el recorte descarta.
- **Relación señal-ruido.** No se aplica ningún corte. Los únicos criterios de inclusión son la clase asignada y la existencia del archivo asociado, con un descarte posterior de los espectros cuyo valor de normalización resulta inutilizable.
- **Extinción interestelar.** No se corrige el enrojecimiento. Esta omisión merece atención particular, porque la normalización elimina la escala absoluta de flujo pero no la pendiente del continuo, que es la cantidad empleada como indicador de temperatura. Un objeto más enrojecido resulta por tanto indistinguible de uno intrínsecamente más frío.

5.5. Red neuronal y extracción de embeddings

Se utiliza una red neuronal convolucional preentrenada (modelo `model_4C_3FC.h5`), compuesta por cuatro bloques convolucionales seguidos de tres capas totalmente conectadas,

entrenada originalmente para clasificar espectros de enanas blancas en las doce categorías presentes en el catálogo de etiquetas. La Tabla 5.4 detalla su arquitectura completa, tal como la reporta el modelo cargado.

Tabla 5.4: Arquitectura completa de la red preentrenada `model_4C_3FC.h5`. Se destacan las cuatro capas seleccionadas como espacios latentes candidatos.

Capa (tipo)	**Forma de salida**
`conv1d_input` (InputLayer)	(None, 3750, 1)
`conv1d` (Conv1D)	(None, 1874, 128)
`max_pooling1d` (MaxPooling1D)	(None, 937, 128)
`conv1d_1` (Conv1D)	(None, 467, 64)
`max_pooling1d_1` (MaxPooling1D)	(None, 233, 64)
`conv1d_2` (Conv1D)	(None, 115, 32)
`max_pooling1d_2` (MaxPooling1D)	(None, 57, 32)
`conv1d_3` (Conv1D)	(None, 27, 16)
`max_pooling1d_3` (MaxPooling1D)	(None, 13, 16)

Capa (tipo)	**Forma de salida**
`flatten` (Flatten)	(None, 208)
`dense` (Dense)	(None, 128)
`dropout` (Dropout)	(None, 128)
`dense_1` (Dense)	(None, 64)
`dropout_1` (Dropout)	(None, 64)
`dense_2` (Dense)	(None, 32)
`dense_3` (Dense)	(None, 12)

Los nombres de capa empleados a lo largo de este capítulo (`flatten`, `dense`, `dense_1` y `dense_2`) son los identificadores que la propia biblioteca asigna a cada capa al construir el modelo, y son los que utilizan las funciones de extracción para interceptar la salida correspondiente. La arquitectura muestra con claridad el mecanismo de compresión progresiva que motiva el análisis por capas: el espectro de entrada, un vector de 3,750 valores de flujo correspondientes a igual número de muestras en longitud de onda, entre 3,797,3 y 9,002,9 Å según lo descrito en la Sección 5.4.1, se reduce sucesivamente hasta los 208 valores de `flatten`, y de ahí a 128, 64 y 32 en las capas densas, antes de proyectarse sobre las 12 clases de salida. Cada una de esas etapas descarta información que la red juzga irrelevante para clasificar, lo que explica por qué la elección de la capa condiciona de forma decisiva qué subestructura permanece disponible para el clustering.

A partir del modelo completo se construyen extractores que devuelven la salida de capas intermedias específicas. Se seleccionaron cuatro capas como espacios latentes candidatos, que cubren distintos niveles de abstracción (Tabla 5.5).

Capa	**Dimensión de salida**	**Nivel de abstracción**
`flatten`	208	Bajo (salida aplanada de los bloques convolucionales)
`dense`	128	Intermedio-bajo
`dense_1`	64	Intermedio-alto
`dense_2`	32	Alto (última capa latente, previa a la salida)

Tabla 5.5: Capas de la red utilizadas como espacios latentes y sus dimensiones.

Para cada conjunto de espectros, se genera una función que procesa las muestras a través

de cada extractor y almacena las representaciones latentes correspondientes.

5.6. Reducción de dimensionalidad

Las representaciones de las capas tienen dimensiones heterogéneas (208, 128, 64 y 32). Para homogeneizarlas y mitigar la maldición de la dimensionalidad antes del clustering, cada embedding se reduce a un espacio común de 32 dimensiones mediante PCA. La capa `dense_2` ya posee 32 dimensiones, por lo que se conserva directamente. Para la visualización en 2D se emplean PCA y t-SNE, este último exclusivamente con fines ilustrativos.

5.7. Metodología de clustering y métricas

Sobre las representaciones latentes se aplican los tres algoritmos seleccionados en la Sección 4.7.3, cuyo funcionamiento y supuestos geométricos se describieron en la Sección 4.3. Esta sección no los vuelve a presentar, sino que especifica con qué entrada y con qué configuración se ejecuta cada uno, puesto que es ahí donde se concentran las decisiones de implementación.

La entrada es común a los tres y se construye en dos pasos a partir del embedding de cada capa. Primero se reduce a 32 dimensiones mediante PCA, salvo en `dense_2`, que ya posee esa dimensión y se conserva sin transformar (Sección 5.6). Después se estandariza cada componente a media nula y varianza unitaria mediante `StandardScaler`, paso necesario porque las componentes principales presentan varianzas de magnitud muy distinta y los tres algoritmos operan sobre distancias euclidianas, de modo que sin estandarizar las primeras componentes dominarían la métrica. La matriz que reciben los algoritmos es, por tanto, de $n \times 32$, con $n = 1{,}100$ espectros por clase.

A partir de esa matriz común, cada algoritmo recibe lo siguiente:

- **K-Means** recibe únicamente el número de grupos k, que se hace variar en un barrido para seleccionarlo por el máximo del Silhouette Score. El resto de los parámetros conserva los valores por omisión de la implementación empleada, entre ellos la inicialización `k-means++` y las diez repeticiones con distinta semilla de las que se retiene la de menor inercia.

- **Spectral Clustering** recibe el número de grupos k y la especificación de la afinidad con que se construye el grafo. Se emplea afinidad por vecinos más cercanos (`affinity = "nearest_neighbors"`), de modo que el algoritmo no opera sobre la matriz de distancias completa sino sobre un grafo de vecindad, y el valor de k se selecciona mediante la heurística del eigengap sobre ese mismo grafo (Sección 5.9).

- **DBSCAN** no recibe el número de grupos, sino los dos parámetros que definen qué se considera una región densa: el radio de vecindad ε, fijado en 1,5, y el número mínimo

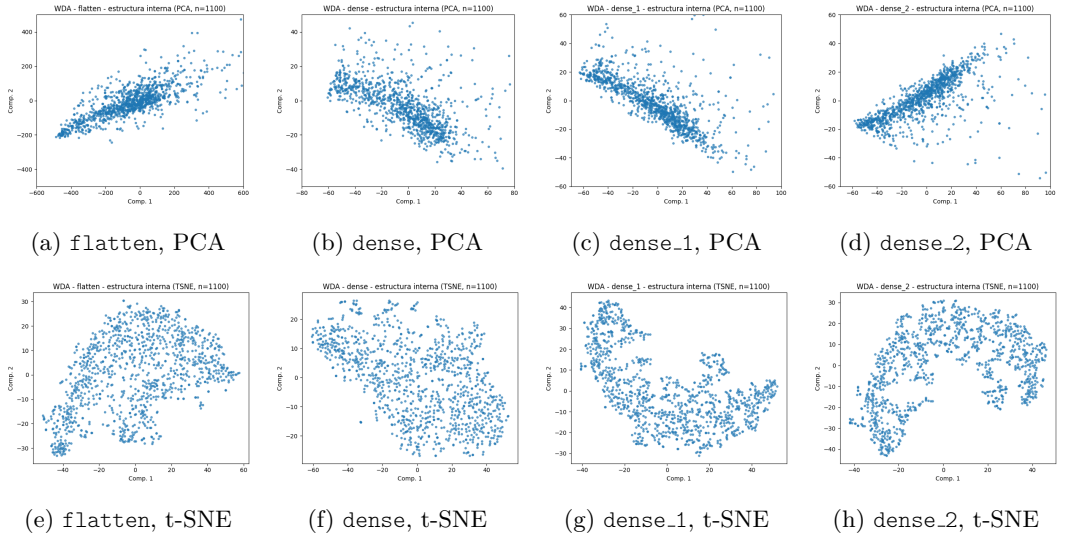
de puntos `min_samples`, fijado en 10. La justificación de ambos valores se desarrolla en la Sección 5.9. A diferencia de los otros dos, su salida incluye una etiqueta adicional, -1, para los puntos que no quedan asignados a ninguna región densa.

La calidad de cada partición se evalúa mediante el Silhouette Score (sil), el índice Davies–Bouldin (DB) y el índice Calinski–Harabasz (CH), y en el caso de DBSCAN se añade el DBCV, por tratarse de la métrica específica para agrupamientos basados en densidad. Las tres primeras se calculan sobre la misma matriz estandarizada que recibió el algoritmo, de modo que los valores son comparables entre capas. Para los puntos etiquetados como ruido se procede según se detalla en la Sección 5.11.1. El análisis de robustez se apoya en las matrices de migración entre particiones con distinto número de clusters, cuya construcción se describe en la Sección 5.11.2.

5.8. Exploración visual de los espacios latentes

Antes de aplicar cualquier algoritmo de clustering, en el cuaderno `task1` se verificó si los espacios latentes exhibían agrupamientos observables a simple vista. Para ello, las representaciones de las salidas de las cuatro capas (`flatten`, `dense`, `dense_1` y `dense_2`) se proyectaron a dos dimensiones mediante PCA y t-SNE. Como convención metodológica, el título de cada gráfico indica explícitamente el método de proyección utilizado (PCA o t-SNE), dado que ambos producen geometrías distintas y su interpretación no es intercambiable: PCA preserva las direcciones de máxima varianza (estructura global, lineal), mientras que t-SNE preserva vecindades locales.

La Figura 5.1 reúne las ocho proyecciones (cuatro capas $\times$ dos métodos), organizadas en una grilla donde la fila superior corresponde a PCA y la inferior a t-SNE, y las columnas recorren las capas de la más superficial (`flatten`) a la más profunda (`dense_2`).



(a) `flatten`, PCA (b) `dense`, PCA (c) `dense_1`, PCA (d) `dense_2`, PCA

(e) `flatten`, t-SNE (f) `dense`, t-SNE (g) `dense_1`, t-SNE (h) `dense_2`, t-SNE

Figura 5.1: Proyecciones 2D de los espacios latentes de las cuatro capas para la clase WDA. Fila superior: PCA. Fila inferior: t-SNE. En ninguna capa ni método de proyección se observan subgrupos claramente separados a simple vista.

El resultado de esta exploración (Figura 5.1) es concluyente: no se observan subgrupos formados a simple vista, en ninguna de las capas ni bajo ninguno de los dos métodos de proyección. Esto confirma que, si existe subestructura dentro de las clases, no es separable por inspección visual de las proyecciones, y justifica el uso de algoritmos de clustering que operen sobre el espacio completo de 32 dimensiones, junto con métricas internas que cuantifiquen la calidad de las particiones más allá de lo que la proyección 2D permite apreciar.

5.9. Selección de k y primeros experimentos con K-Means, Spectral Clustering y DBSCAN para WDA.

En el cuaderno `task2` se realizaron los primeros experimentos de clustering utilizando K-Means, Spectral Clustering y DBSCAN para la clase de enanas blancas WDA. Dado que los dos primeros algoritmos requieren fijar el número de clusters k a priori, se implementó la función `evaluate_kmeans_k()`, que ejecuta K-Means para un rango de valores de k sobre cada capa y grafica la curva del Silhouette Score en función de k.

Las Figuras 5.2 a 5.5 presentan las curvas obtenidas para cada una de las cuatro capas de la clase WDA.

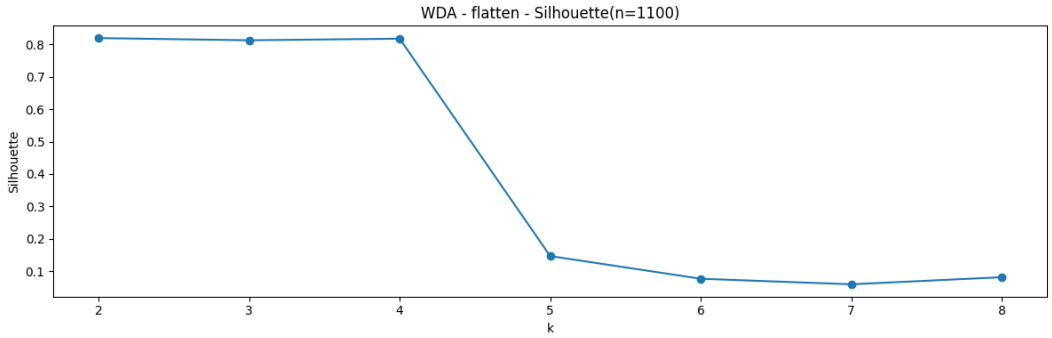


Figura 5.2: Curva de Silhouette Score en función del número de clusters k para la capa `flatten`.

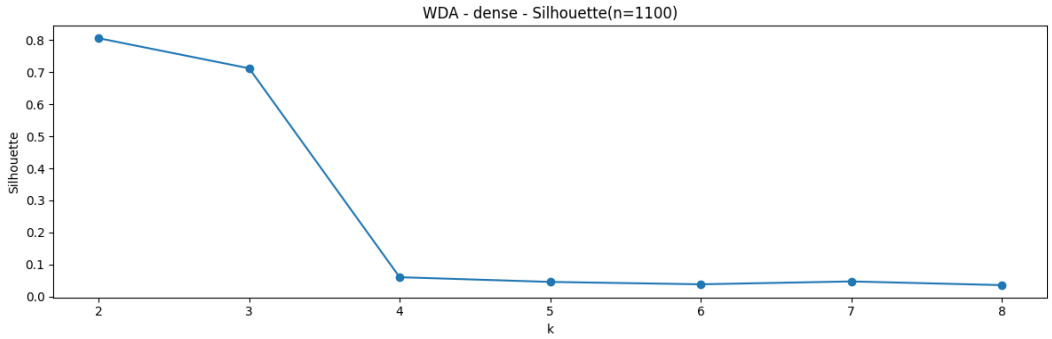


Figura 5.3: Curva de Silhouette Score en función del número de clusters k para la capa `dense`.

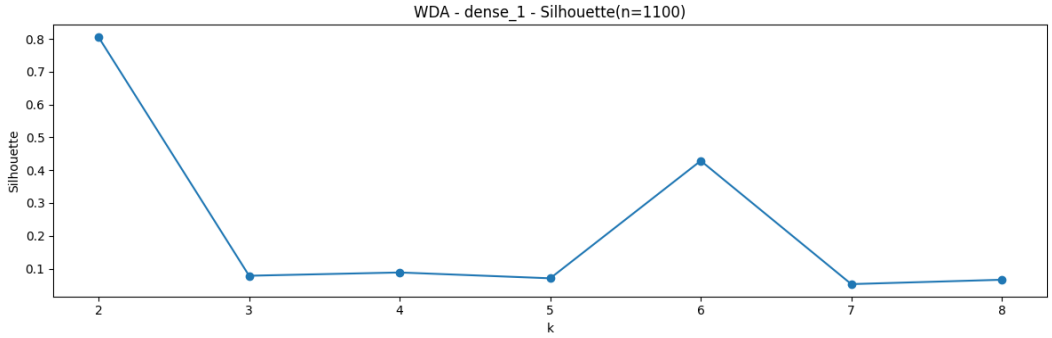


Figura 5.4: Curva de Silhouette Score en función del número de clusters k para la capa `dense_1`.

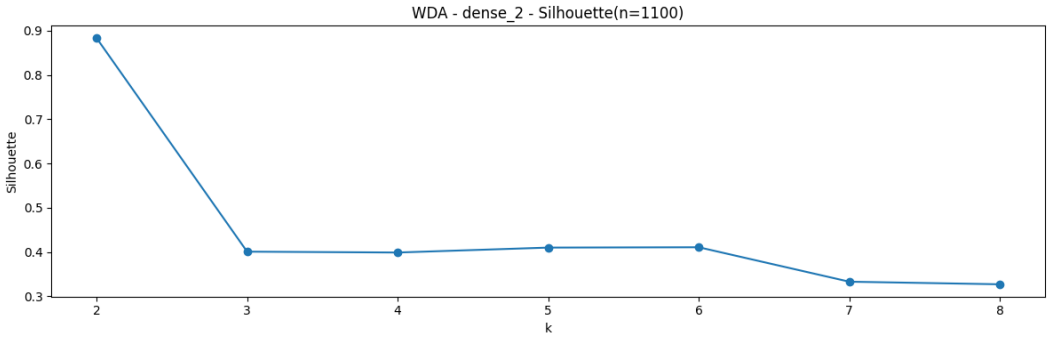


Figura 5.5: Curva de Silhouette Score en función del número de clusters k para la capa `dense_2`.

A partir de estas curvas se seleccionó, para **K-Means**, el valor de k que maximiza el Silhouette Score en cada capa. Con este criterio las cuatro capas coinciden en $k = 2$, que es el valor con el que se ejecutaron los experimentos reportados más adelante. Conviene precisar, no obstante, que la nitidez de ese máximo varía entre capas: en `dense_1` y `dense_2` la caída posterior a $k = 2$ es abrupta, lo que hace del máximo una preferencia inequívoca, mientras que en `flatten` la curva se mantiene prácticamente plana entre $k = 2$ y $k = 4$ antes de desplomarse en $k = 5$, de modo que en esa capa el criterio no discrimina entre esos tres valores.

Para **Spectral Clustering** se empleó un criterio distinto, propio del algoritmo: la heurística del *eigengap* descrita en la Sección 3.4.2, que escoge k a partir del mayor salto entre autovalores consecutivos del Laplaciano normalizado del grafo de similitud (Ecuación 3.2). La razón de emplear un criterio diferenciado es que la partición de Spectral Clustering no se construye sobre distancias a centroides, sino sobre la conectividad del grafo, de modo que resulta coherente seleccionar k a partir de la estructura de autovalores de ese mismo grafo y no de una medida de compacidad euclidiana. Aplicado sobre las capas de WDA, este criterio indica $k = 3$ para `flatten`, $k = 5$ para `dense`, $k = 2$ para `dense_1` y $k = 8$ para `dense_2`.

La Tabla 5.6 reúne los valores obtenidos por ambos criterios, y es la referencia a la que remiten los experimentos de las secciones siguientes.

Tabla 5.6: Número de clusters seleccionado para cada capa según el algoritmo (clase WDA, $n = 1{,}100$). Para K-Means se emplea el máximo del Silhouette Score. Para Spectral Clustering, la heurística del eigengap del Laplaciano.

Capa	***k* K-Means (Silhouette)**	***k* Spectral (eigengap)**
`flatten`	2	3
`dense`	2	5
`dense_1`	2	2
`dense_2`	2	8

El contraste entre ambas columnas es en sí mismo informativo. El Silhouette converge al mismo valor en las cuatro capas, lo que en este caso no expresa una preferencia por dos grupos sino la distorsión descrita más adelante: sobre una nube sin separaciones claras, la partición más compacta que K-Means puede construir es la que aísla un punto extremo del resto. El eigengap, en cambio, sí distingue entre capas, porque mide una propiedad distinta (la conectividad del grafo de vecindad) y no la compacidad respecto de centroides. La mayor discrepancia se produce en `dense_2`, donde el eigengap indica $k = 8$ frente al $k = 2$ del Silhouette. Ninguno de los dos criterios queda invalidado por ello: leídos en conjunto sugieren una organización jerárquica en esa capa, con una división principal en dos grupos sobre la cual existiría subestructura adicional de menor contraste.

Resulta relevante examinar cómo evoluciona el k indicado por el eigengap al avanzar hacia capas más profundas, dado que es el criterio con el que efectivamente se fija el número de clusters de Spectral Clustering. Los valores obtenidos son $k = 3$, 5, 2 y 8 para `flatten`, `dense`, `dense_1` y `dense_2` en la clase WDA, y $k = 2$, 2, 2 y 8 en la clase WDB. En ambos casos la capa más profunda es la que requiere, con diferencia, el mayor número de grupos, mientras que las capas superficiales se mantienen en valores reducidos. La Figura 5.6 resume este comportamiento.

La interpretación se apoya en el sesgo funcional discutido en el Capítulo 3. La capa `dense_2` es la última representación previa a la salida de clasificación, y está orientada a separar las doce categorías con que la red fue entrenada. Ese sesgo organiza el espacio en regiones localmente conectadas y de frontera nítida, y es precisamente ese tipo de estructura la que detecta el eigengap, cuyo criterio se basa en la existencia de cortes de baja conductancia en el grafo de vecindad. Al acercarse a la salida, la representación no se vuelve más homogénea sino más fragmentada, de modo que el número de cortes que el criterio identifica aumenta.

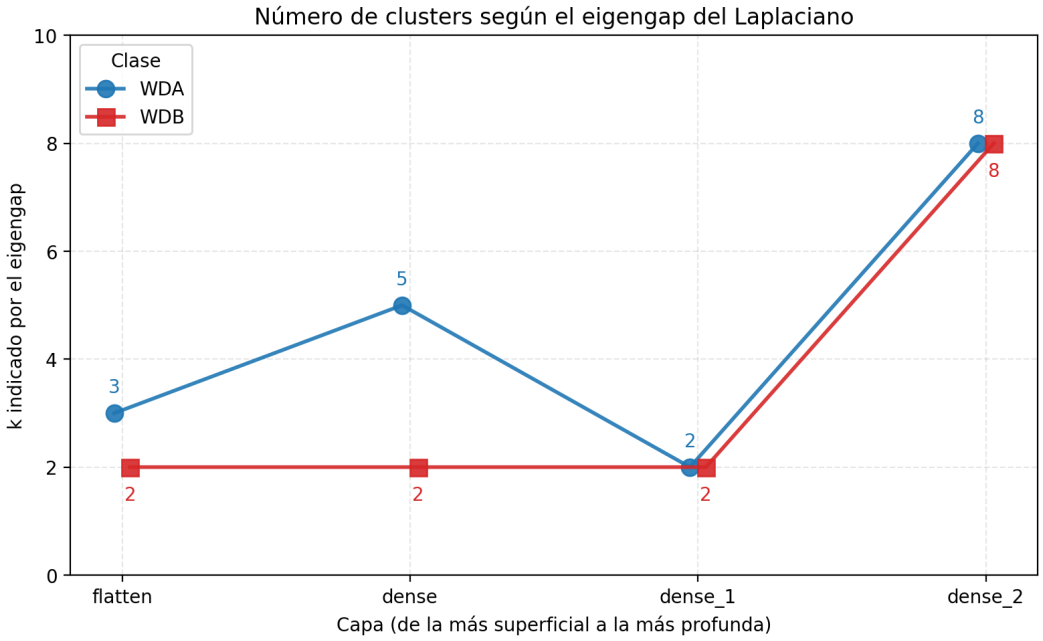


Figura 5.6: Número de clusters indicado por el eigengap del Laplaciano en función de la profundidad de la capa, para las clases WDA y WDB. La capa más profunda requiere en ambos casos el mayor número de grupos, si bien la progresión no es monótona.

Considerando los criterios de selección de k para K-Means y Spectral Clustering, y dado que DBSCAN no requiere fijar el número de grupos a priori, resta especificar la configuración con la que se ejecuta este último algoritmo. DBSCAN sustituye la elección de k por dos hiperparámetros que definen de forma conjunta qué se considera una región densa: el radio de vecindad ε y el número mínimo de puntos `min_samples` que dicha vecindad debe contener para que un punto sea calificado como punto núcleo.

Para `min_samples` se adoptó un valor de 10 en todos los experimentos. Esta elección responde a un compromiso propio del objetivo del trabajo: dado que las subclases buscadas podrían corresponder a poblaciones poco frecuentes dentro de cada clase espectral, un umbral demasiado alto impediría que un subgrupo minoritario llegue siquiera a formarse, mientras que uno demasiado bajo permitiría que un puñado de espectros próximos por azar constituyan un cluster espurio.

Para ε, en cambio, no existe un valor evidente a priori, dado que su magnitud apropiada depende de la escala de distancias de cada espacio latente, la cual varía entre capas. Como punto de partida se adoptó $\varepsilon = 1{,}5$.

Para verificar que este valor no fuese una elección arbitraria, se realizó un barrido sobre los valores 0,5, 1,0, 1,5, 2,0, 2,5 y 3,0 de ε en las cuatro capas de la clase WDA, registrando

en cada caso el número de grupos formados y la cantidad de puntos etiquetados como ruido. Los resultados se presentan en la Tabla 5.7.

Tabla 5.7: Número de clusters formados por DBSCAN en función de ε para cada capa de la clase WDA. Entre paréntesis, el número de puntos etiquetados como ruido (n = 1,100, `min_samples` = 10).

Capa	**0,5**	**1,0**	**1,5**	**2,0**	**2,5**	**3,0**
`flatten`	0 (1100)	1 (1084)	1 (790)	1 (581)	1 (426)	1 (302)
`dense`	0 (1100)	0 (1100)	**3** (1030)	1 (737)	1 (501)	1 (352)
`dense_1`	0 (1100)	0 (1100)	**2** (1005)	1 (639)	1 (460)	1 (313)
`dense_2`	3 (720)	1 (221)	1 (115)	1 (82)	1 (62)	1 (52)

El barrido respalda la elección. En las capas `dense` y `dense_1`, $\varepsilon = 1{,}5$ es el único valor de todo el rango explorado que produce más de un grupo: por debajo de ese radio la totalidad de la muestra queda etiquetada como ruido, y por encima los grupos se funden en un único cluster que absorbe progresivamente la nube. El valor adoptado se sitúa, por tanto, en la ventana estrecha en que la estructura de densidad de estos dos espacios resulta detectable, lo que confirma que se trata de una configuración adecuada para el propósito de la exploración y no de una elección indiferente.

En `dense_2` el comportamiento es distinto: allí $\varepsilon = 1{,}5$ entrega un único cluster, mientras que $\varepsilon = 0{,}5$ produce tres grupos, si bien a costa de etiquetar como ruido un 65 % de la muestra. Esta diferencia es coherente con el carácter más compacto de la representación en la última capa latente, donde las distancias entre puntos son sistemáticamente menores y un radio equivalente abarca una porción mayor del espacio.

El valor se mantuvo constante en todos los experimentos, decisión que obedece al mismo principio que guía la comparación con k constante entre capas: si los hiperparámetros varían junto con la representación, las diferencias observadas dejan de ser atribuibles al espacio latente. Dado que el propósito de esta etapa es comparar capas y clases entre sí, y no obtener la mejor partición posible para cada una, se privilegió la comparabilidad por sobre el ajuste individual de ε en cada espacio.

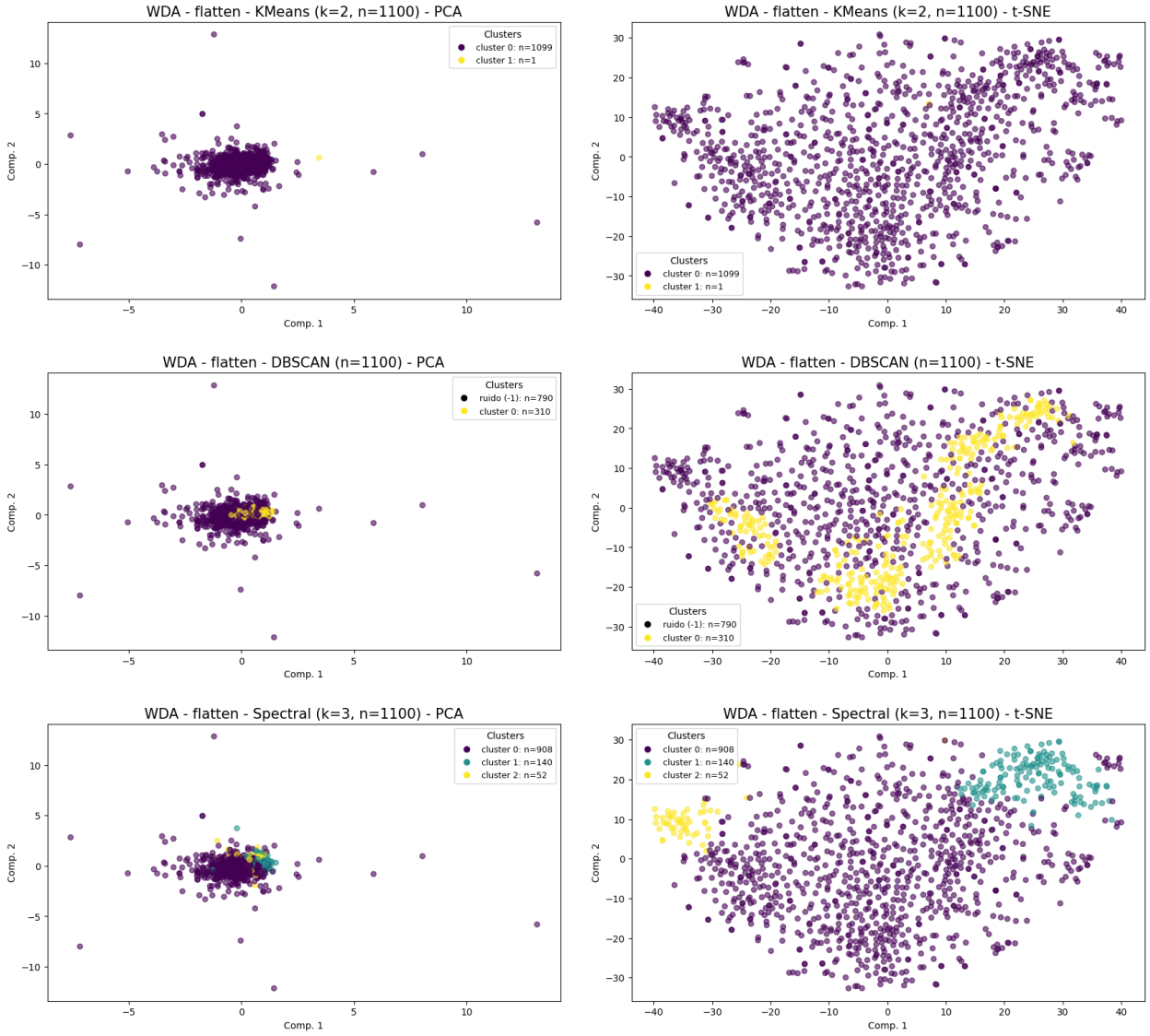
Establecidos los criterios de selección de k para K-Means y Spectral Clustering, y fijados los hiperparámetros de DBSCAN, se dispone de la configuración necesaria para ejecutar los tres algoritmos sobre los espacios latentes. Esta primera ronda de experimentos se realizó íntegramente sobre la clase WDA, que cumple el rol de caso de desarrollo: al ser la clase con mayor número de espectros disponibles, permite ajustar y depurar el procedimiento sobre una base muestral amplia. Una vez consolidado, el mismo pipeline se aplica sin modificaciones a la clase WDB, cuyos resultados se reportan en la Sección 5.13.

Los resultados de cada capa se presentan en un par de figuras dispuestas una junto

a la otra. Ambas contienen las particiones producidas por los tres algoritmos sobre el mismo espacio latente, apiladas verticalmente en el orden K-Means, DBSCAN y Spectral Clustering. Lo único que las distingue es el método empleado para proyectar el espacio de 32 dimensiones al plano, PCA en la primera y t-SNE en la segunda. La leyenda de cada panel indica el número de muestras asignadas a cada grupo, dato necesario para juzgar si una partición reparte los datos de forma equilibrada o si concentra la mayoría en un solo cluster.

Mostrar ambas proyecciones de forma simultánea responde a una necesidad interpretativa ya señalada en la Sección 5.8: PCA preserva las direcciones de máxima varianza, esto es, la estructura global y lineal del espacio, mientras que t-SNE preserva vecindades locales. Un agrupamiento visible bajo una sola de las dos proyecciones puede ser consecuencia de la geometría particular de esa transformación, mientras que uno que se sostiene bajo ambas constituye evidencia más sólida de que la partición refleja estructura del espacio latente y no del método de visualización. Disponerlas lado a lado permite realizar esa comparación de forma directa.

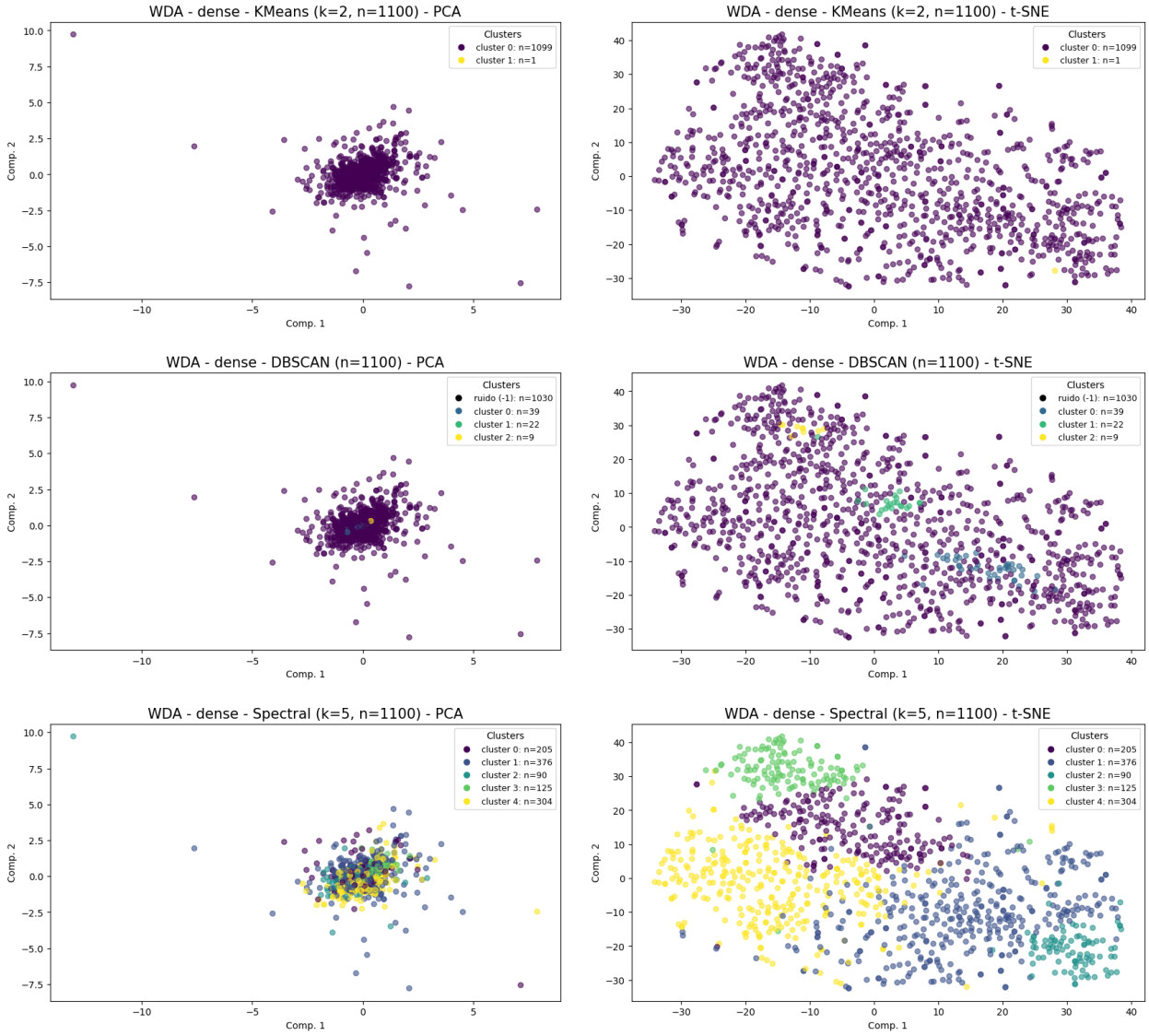
Para facilitar la trazabilidad de cada experimento, el título de cada panel sigue una convención que reúne, en este orden, la clase espectral analizada, la capa de la que proviene el embedding, el algoritmo de clustering junto con sus hiperparámetros y el tamaño muestral, y el método de proyección empleado. Así, una figura de título `WDA - flatten - KMeans (k=2, n=1100) - PCA` corresponde a la partición obtenida por K-Means con $k = 2$ sobre las representaciones de la capa `flatten` de 1,100 espectros de la clase WDA, visualizada mediante PCA.



(a) Proyección PCA

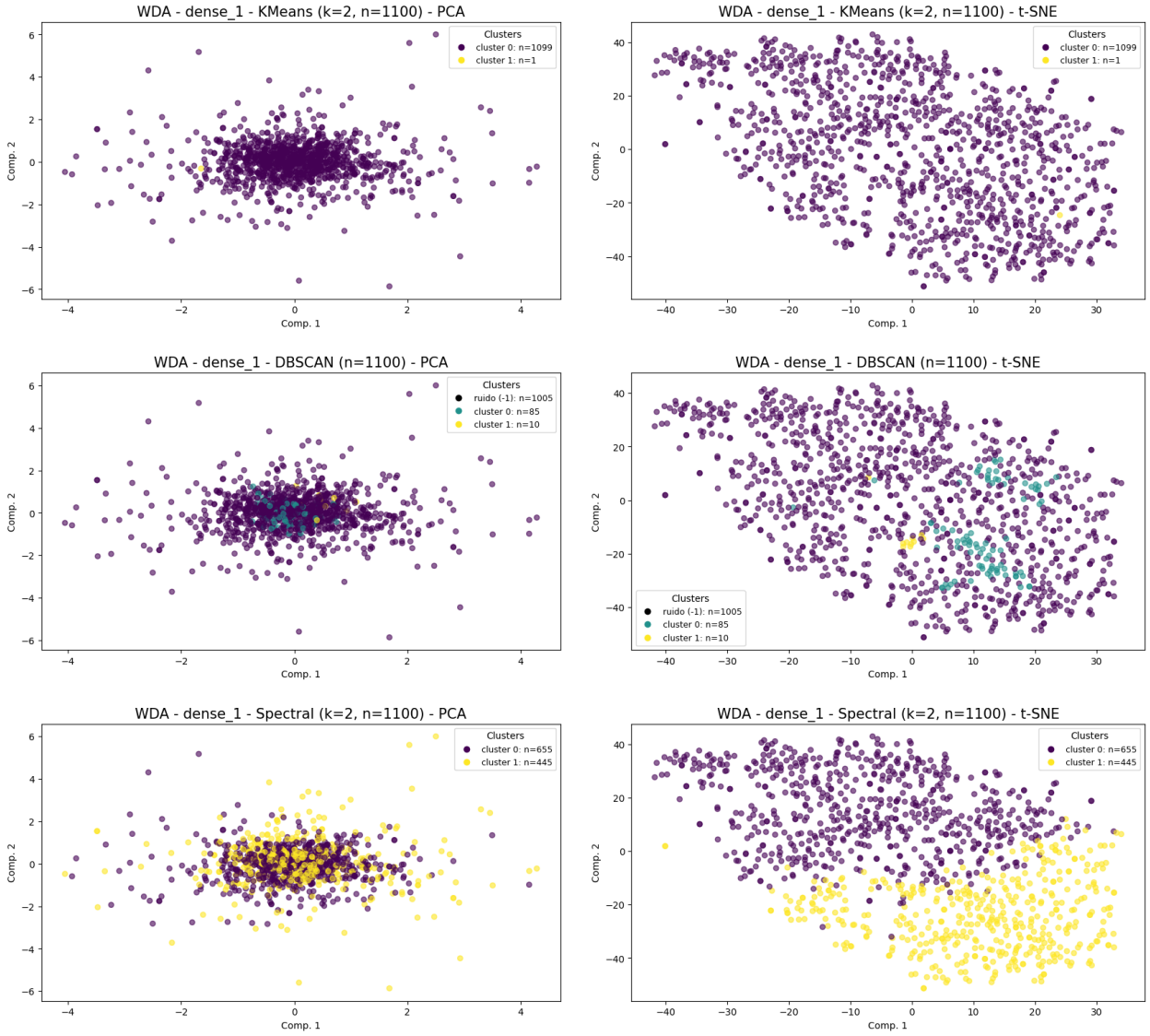
(b) Proyección t-SNE

Figura 5.7: Particiones obtenidas por K-Means, DBSCAN y Spectral Clustering sobre la capa `flatten` de WDA ($n = 1{,}100$). En cada figura los tres algoritmos se disponen verticalmente en ese orden.



(a) Proyección PCA (b) Proyección t-SNE

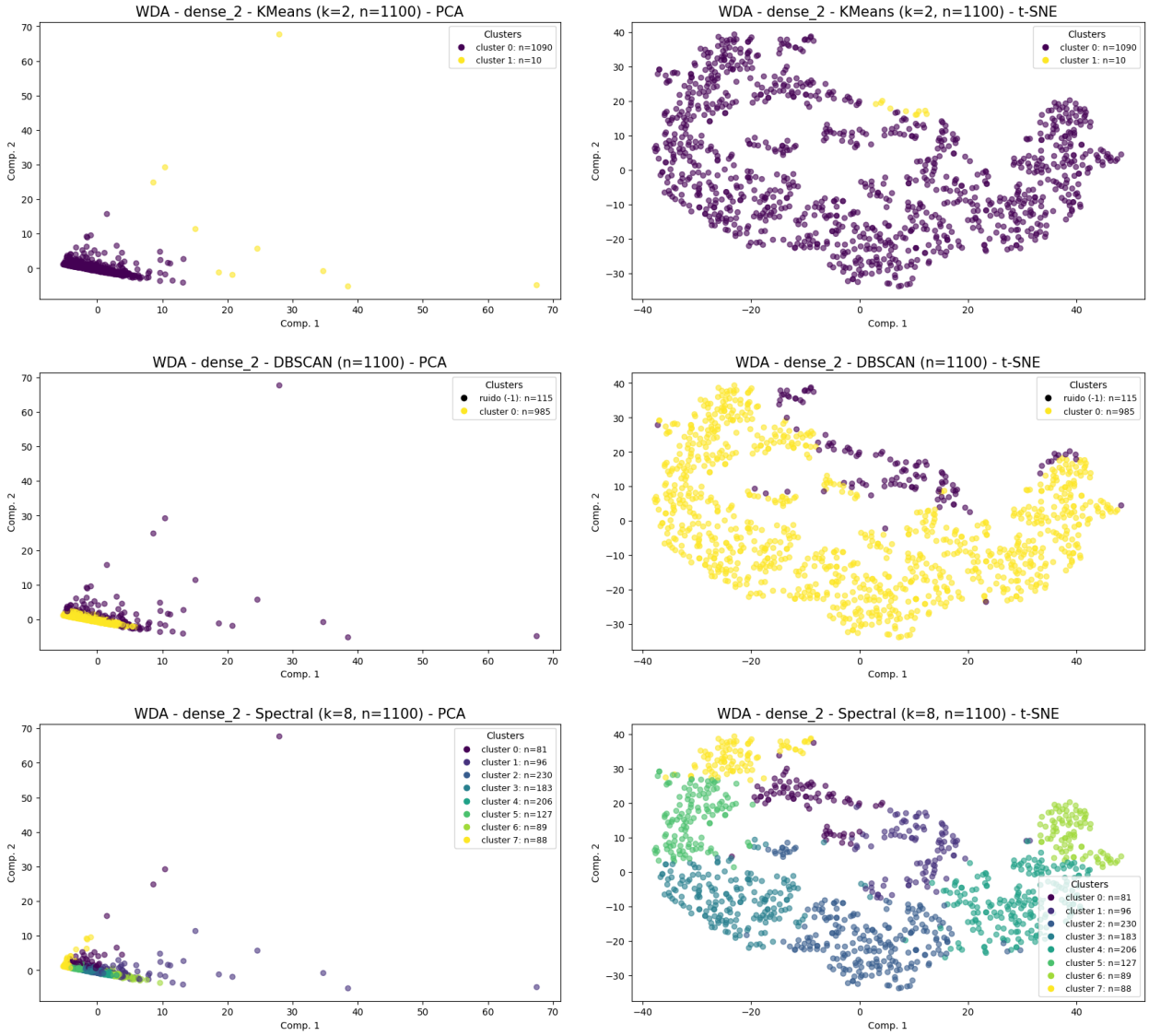
Figura 5.8: Particiones obtenidas por K-Means, DBSCAN y Spectral Clustering sobre la capa `dense` de WDA ($n = 1{,}100$). En cada figura los tres algoritmos se disponen verticalmente en ese orden.



(a) Proyección PCA

(b) Proyección t-SNE

Figura 5.9: Particiones obtenidas por K-Means, DBSCAN y Spectral Clustering sobre la capa `dense_1` de WDA ($n = 1{,}100$). En cada figura los tres algoritmos se disponen verticalmente en ese orden.



(a) Proyección PCA (b) Proyección t-SNE

Figura 5.10: Particiones obtenidas por K-Means, DBSCAN y Spectral Clustering sobre la capa `dense_2` de WDA ($n = 1{,}100$). En cada figura los tres algoritmos se disponen verticalmente en ese orden.

Las Figuras 5.7 a 5.10 admiten una primera lectura cualitativa, previa a cualquier evaluación métrica, que orientó el resto del trabajo.

El resultado más inmediato concierne a K-Means, cuyas particiones resultan degeneradas en las cuatro capas. Con $k = 2$, el algoritmo no divide la muestra en dos poblaciones

comparables, sino que asigna la práctica totalidad de los espectros a un único grupo y reserva el segundo para un puñado de puntos extremos: en `flatten`, `dense` y `dense_1` el reparto es de 1,099 contra 1 espectro, y en `dense_2` de 1,090 contra 10. Estas particiones no constituyen agrupamientos en ningún sentido útil, lo que el algoritmo hace es aislar los valores atípicos más alejados del centro de la nube.

La causa es geométrica. K-Means minimiza la suma de distancias al centroide, criterio que solo produce grupos significativos cuando la nube de puntos presenta zonas de mayor concentración separadas entre sí. En estos espacios latentes la distribución es esencialmente unimodal: una masa central compacta rodeada de puntos dispersos, sin discontinuidades que el algoritmo pueda aprovechar. Ante esa geometría, la partición que minimiza el criterio es precisamente la que separa el punto más lejano del resto, ya que cualquier corte que atraviese la masa central incrementa la dispersión intra-cluster sin compensación. El fenómeno se agudiza en `dense_2`, donde la presencia de una decena de espectros con valores extremos en la primera componente principal basta para capturar por completo el segundo centroide.

Este comportamiento es consistente con la suposición de convexidad del algoritmo, que no se satisface en la geometría de los embeddings, y motivó su descarte para la fase experimental sistemática.

La revisión de estos primeros resultados dejó planteadas, además, cuatro cuestiones que definieron las etapas siguientes:

- **Flexibilización del criterio de selección de k.** Conviene no tratar el valor indicado por cada criterio como una determinación cerrada, sino como el punto de partida de una exploración acotada. En el caso de K-Means, la lectura de las curvas de Silhouette no exige necesariamente tomar el máximo global: cuando varios k consecutivos alcanzan una calidad equivalente, el máximo no expresa una preferencia real y resulta igualmente defendible situarse en el último valor de la meseta.

  El mismo razonamiento se aplica a Spectral Clustering, con la salvedad de que allí k no se selecciona mediante el Silhouette sino con el eigengap del Laplaciano. Dado que ese criterio entrega un único valor por capa, sin una curva que permita apreciar cuán marcada es la preferencia, resulta necesario evaluar también los valores vecinos al indicado. Solo comparando particiones con k contiguos puede establecerse si los grupos se conservan, se fusionan o se subdividen de forma ordenada, que es lo que distingue una estructura jerárquica genuina de una partición sensible al hiperparámetro.

- **La estructura observada en `dense` con DBSCAN.** Sobre esta capa, la proyección t-SNE muestra tres grupos delimitados con una nitidez llamativa frente al resto de la nube. Que la separación sea tan clara plantea una pregunta que excede lo metodológico: si responde a una diferencia desde el punto de vista de la física entre los objetos o si es un efecto de la representación. Responderla exige salir del espacio latente y volver

al espectro, lo que motivó la síntesis de espectros representativos por cluster y su comparación con el espectro promedio de la clase.

Resulta de interés extender ese mismo análisis a la partición que Spectral Clustering produce sobre esta capa con $k = 5$, cuyos grupos son de naturaleza distinta a los de DBSCAN. Mientras DBSCAN aísla tres núcleos densos de 39, 22 y 9 espectros y relega el resto a la categoría de ruido, Spectral Clustering reparte la totalidad de la muestra en cinco grupos espacialmente coherentes y de tamaños comparables. Ambas particiones responden, por tanto, a preguntas diferentes: una identifica concentraciones locales excepcionales, la otra describe cómo se organiza la nube en su conjunto. Comparar los espectros representativos de una y otra permite establecer si los núcleos densos que detecta DBSCAN corresponden a poblaciones con rasgos espectrales propios o si son subconjuntos de los grupos mayores identificados por Spectral Clustering.

- **La separación obtenida por Spectral Clustering en `flatten`.** Con $k = 3$ se forman tres grupos que se sostienen tanto en la proyección PCA como en la t-SNE. Dado que ambas transformaciones tienen sesgos geométricos distintos, esa coincidencia sugiere que la partición no es un artefacto de visualización y justifica analizar sus espectros.

- **Una complementariedad entre algoritmo y profundidad de capa.** A partir de las figuras anteriores se observa que DBSCAN entrega sus resultados más informativos en las capas cercanas a la salida, donde la compactación del espacio genera regiones densas bien delimitadas, mientras que Spectral Clustering resulta más productivo en las capas cercanas a `flatten`, de estructura más difusa y geometría no convexa.

De la tercera cuestión se desprende, por último, la práctica metodológica que estructura la etapa siguiente: comparar experimentos con la misma cantidad de clusters entre capas. Si un valor de k produce un resultado interesante en una capa, ese mismo k se evalúa en las demás, de modo que las diferencias observadas sean atribuibles a la representación y no a la elección del hiperparámetro. Sobre esa base, los experimentos de la Sección 5.11 someten el resultado a dos comprobaciones complementarias: la primera mantiene k fijo y cambia de capa, tal como se acaba de describir. La segunda mantiene la capa y varía k, repitiendo el agrupamiento con $k = 2$ y $k = 4$ para observar qué ocurre con los tres grupos hallados originalmente. Si al reducir k uno de ellos es absorbido por otro, y al aumentarlo alguno se subdivide sin que el resto se altere, los grupos responden a una jerarquía real de la estructura, si en cambio la partición se reorganiza por completo ante cualquiera de los dos cambios, el resultado inicial era un artefacto de la elección de k.

Con independencia de este descarte, cada partición se evalúa también con el conjunto de métricas internas definidas en la Sección 4.7.4. El cálculo está integrado en las propias funciones de agrupamiento: al ejecutar `cluster_plot_Spect()` o `cluster_plot_DBSCAN()`

sobre una capa, cada una genera la proyección con las etiquetas obtenidas y, junto a ella, reporta el número de clusters formados, la cantidad de puntos etiquetados como ruido, el Silhouette Score y, los índices Davies–Bouldin y Calinski–Harabasz. En el caso de DBSCAN se añade el DBCV, por tratarse de la métrica específica para agrupamientos basados en densidad.

Esta integración tiene dos consecuencias prácticas. La primera es que la lectura visual y la cuantitativa quedan disponibles en un mismo paso, lo que permite contrastar de inmediato la impresión que produce la proyección con lo que indican las métricas, evitando conclusiones apoyadas solo en la geometría del gráfico. La segunda concierne al tratamiento del ruido: las métricas se calculan excluyendo los puntos etiquetados como -1, dado que no constituyen un grupo y su inclusión distorsionaría los índices. Por la misma razón, cuando una configuración produce menos de dos clusters las métricas no se reportan, ya que carecen de definición en ese caso. Estas métricas cumplen aquí una función de verificación: permiten contrastar la calidad de la partición obtenida con el k indicado por el eigengap y comparar capas entre sí bajo un mismo criterio. Los resultados por capa y configuración se presentan en las tablas de las Secciones 5.11.7 y 5.13.8.

5.10. Síntesis de espectros representativos

Una vez obtenidas las particiones, se sintetizan espectros representativos por cluster calculando el espectro promedio de las muestras asignadas a cada grupo. Para cada experimento se generan: (i) un gráfico por cluster con su espectro promedio (y, opcionalmente, su banda de desviación estándar $\pm 1\sigma$), (ii) un gráfico de diferencia entre el promedio de cada cluster y el promedio total de la clase, suavizado para resaltar rasgos sistemáticos, y (iii) un gráfico combinado que superpone todos los promedios junto al espectro medio de la clase. Sobre estos gráficos se pueden superponer las líneas de la serie de Balmer (Hε a 3.970 Å, Hδ a 4.101 Å, Hγ a 4.340 Å, Hβ a 4.861 Å y Hα a 6.563 Å) para evaluar si las diferencias entre clusters se concentran en transiciones físicamente relevantes. Esta etapa es la que permite contrastar los subgrupos descubiertos con la fenomenología astrofísica esperada.

5.11. Experimentos con Spectral Clustering y síntesis espectral

Las conclusiones de la sección anterior definen el programa de esta etapa. Descartado K-Means y establecido que la partición con $k = 3$ sobre `flatten` se sostiene bajo ambas proyecciones, el experimento central consiste en evaluar ese mismo valor en las cuatro capas. A cada partición se le calculan sus métricas internas y se le sintetizan los espectros representativos, que son los que permiten interpretar los grupos en términos físicos y no solo geométricos.

Los valores de k evaluados en cada capa responden a dos referencias. La primera es el

k que indica el eigengap del Laplaciano, criterio propio de Spectral Clustering adoptado en la Sección 5.9, que señala $k = 3$ para `flatten`, $k = 5$ para `dense`, $k = 2$ para `dense_1` y $k = 8$ para `dense_2`. La segunda es $k = 3$, valor que las conclusiones de la sección anterior identificaron como de interés por la partición que produce sobre `flatten`, y que se replica en las demás capas para que las diferencias observadas resulten atribuibles a la representación y no a la elección del hiperparámetro.

A estas dos referencias se añaden valores contiguos, siguiendo la flexibilización descrita en la Sección 5.9. La razón es que el eigengap entrega un único valor por capa, sin una curva que permita apreciar cuán marcada es esa preferencia frente a las alternativas próximas, evaluar k vecinos es lo que permite establecer si la partición indicada se sostiene o si se reorganiza ante una perturbación mínima del hiperparámetro. En concreto se evalúan $k = 2$, 3 y 4 en `flatten`, $k = 3$, 4 y 5 en `dense`, $k = 2$, 3 y 4 en `dense_1`, y $k = 3$ y 8 en `dense_2`.

Ambas decisiones obedecen al mismo principio, propio de los tests de ablación: si una partición refleja estructura real y no un artefacto de la configuración, debe resistir perturbaciones controladas. Replicar un mismo k entre capas permite además localizar en qué punto de la jerarquía se pierde la información que la sustenta, y privilegiar así la representación que menos descarta, dado que la compactación hacia la salida reduce de forma progresiva la variabilidad disponible.

Cabe precisar que el Silhouette Score y los demás índices internos se reportan para todas estas particiones, pero cumplen aquí una función de evaluación *a posteriori* y no de selección: permiten comparar la calidad de las particiones una vez obtenidas, mientras que la elección de k descansa en el eigengap y en el análisis de migración entre configuraciones.

La síntesis de espectros representativos, en cambio, no se realiza para todas las configuraciones evaluadas sino únicamente para la que el eigengap selecciona en cada capa: $k = 3$ en `flatten`, $k = 5$ en `dense`, $k = 2$ en `dense_1` y $k = 8$ en `dense_2`. El motivo es doble. Por una parte, generar los espectros de todas las particiones multiplicaría el número de figuras sin un aumento proporcional de información, dado que las configuraciones con k contiguos comparten la mayor parte de sus grupos, según muestran las matrices de migración analizadas más adelante. Por otra, la partición indicada por el eigengap es la que el criterio propio de Spectral Clustering señala como la más acorde a la estructura del grafo de similitud, de modo que constituye la referencia natural para el análisis físico. El caso de `dense` resulta además de particular interés, por motivos que se discuten en la subsección correspondiente.

Esta restricción responde a criterios de presentación y no limita el alcance del análisis: el explorador `spectra_viewer()` descrito a continuación permite generar los espectros representativos de cualquier otra configuración a demanda, de manera que quien consulte los cuadernos puede reproducir la síntesis para el valor de k y la capa que estime convenientes.

5.11.1. Exploradores interactivos

El número de combinaciones de capa, algoritmo y número de clusters que exige este análisis hace impracticable escribir una celda de código por experimento. Para agilizarlo se desarrollaron tres exploradores basados en `ipywidgets`:

- `spectral_explorer()`: ejecuta y almacena en caché experimentos de Spectral Clustering variando la capa y k mediante menús desplegables, y muestra las proyecciones t-SNE y PCA junto a las métricas de la partición. Las etiquetas quedan guardadas en `spectral_store` para su reutilización posterior.
- `dbscan_explorer()`: análogo al anterior, con control de capa, ε y `min_samples`, e incorporando el DBCV entre las métricas reportadas.
- `spectra_viewer()`: a partir de los experimentos ya cacheados, muestra los espectros promedio de cada grupo con controles para superponer el espectro medio de la clase, las líneas de la serie de Balmer y la banda de dispersión $\pm 1\sigma$.

El uso de una caché es lo que hace viable el procedimiento: una configuración ya calculada no vuelve a ejecutarse, de modo que comparar particiones es inmediato una vez obtenidas. Su aporte principal, sin embargo, es el de `spectra_viewer()`: poder activar y desactivar las líneas de Balmer sobre el mismo gráfico permite juzgar de un vistazo si las diferencias entre grupos se concentran en transiciones atómicas concretas o se reparten a lo largo del continuo, que es la distinción que separa una subclase con sentido físico de un artefacto numérico.

Todos los experimentos ejecutados a través de los exploradores quedan almacenados en estructuras de caché con una convención de nombres uniforme, de modo que sus etiquetas pueden recuperarse después para la síntesis espectral, lo que garantiza la trazabilidad de los resultados reportados.

5.11.2. Matrices de migración

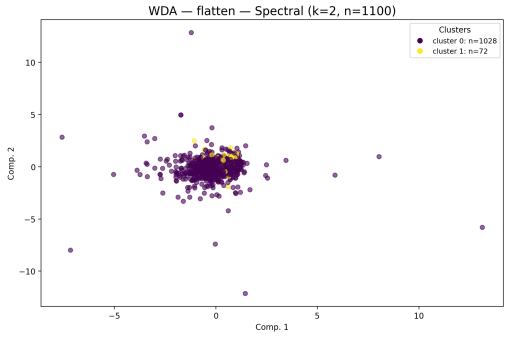
Para determinar si los grupos se absorben o se subdividen al variar k no basta con comparar las proyecciones visualmente, dado que las etiquetas que asigna el algoritmo son arbitrarias y no guardan correspondencia entre una ejecución y otra. Se emplea para ello una *matriz de migración*, construida con la función `plot_migration_matrix()`: es una tabla de contingencia en la que las filas corresponden a los grupos de la partición con $k = a$, las columnas a los de la partición con $k = b$, y cada celda registra cuántas muestras pertenecen simultáneamente a ambos.

Su lectura es directa. Si una fila concentra prácticamente todas sus muestras en una sola columna, ese grupo se mantiene intacto al cambiar k. Si una fila reparte sus muestras entre dos columnas de forma equilibrada, el grupo se ha subdividido. Si dos filas apuntan a la

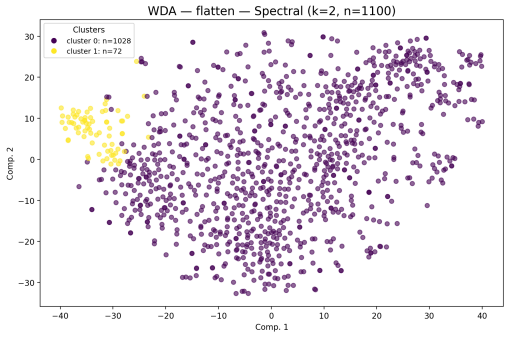
misma columna, ambos grupos se han fusionado. Una partición que describe una jerarquía real produce matrices con estructura de bloques, en las que cada grupo se conserva o se divide de forma limpia. Una partición inestable produce matrices dispersas, con las muestras repartidas sin patrón entre casi todas las celdas.

5.11.3. Capa `flatten`

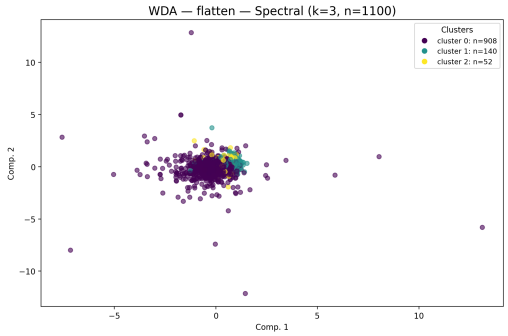
Las figuras en 5.11 presentan las particiones obtenidas sobre esta capa para $k \in \{2, 3, 4\}$, en ambas proyecciones:



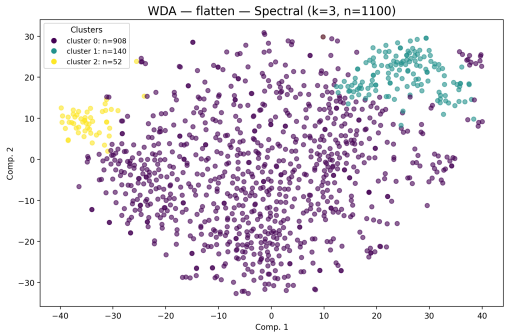
(a) $k = 2$, PCA



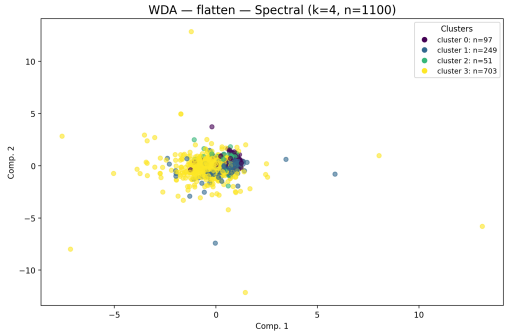
(b) $k = 2$, t-SNE



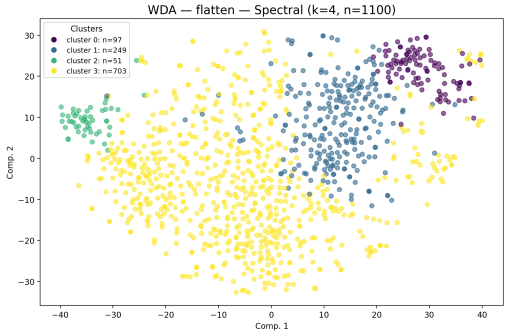
(c) $k = 3$, PCA



(d) $k = 3$, t-SNE



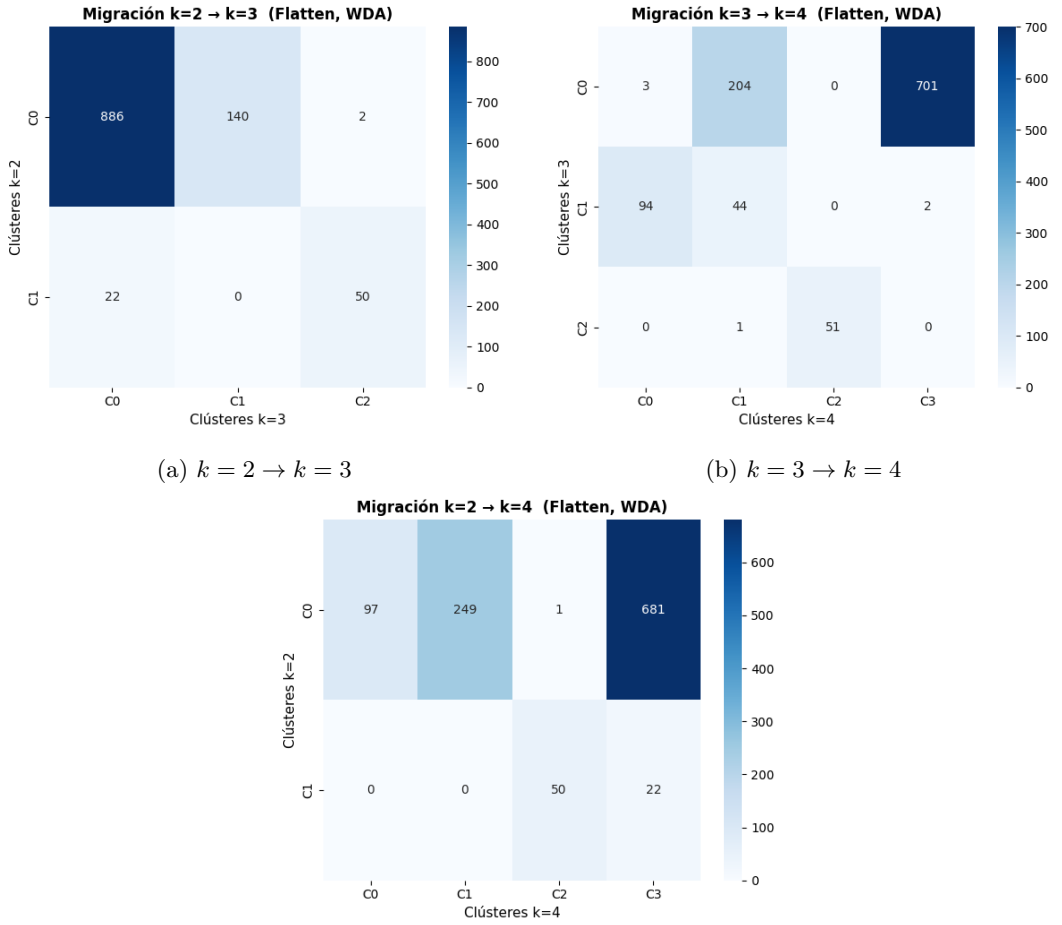
(e) $k = 4$, PCA



(f) $k = 4$, t-SNE

Figura 5.11: Particiones de Spectral Clustering sobre la capa `flatten` de WDA ($n = 1{,}100$), para los valores de k evaluados. Columna izquierda: proyección PCA. Columna derecha: proyección t-SNE.

La Figura 5.12 recoge las matrices de migración entre estas configuraciones, que permiten seguir el destino de cada grupo al variar el número de clusters.



(a) $k = 2 \rightarrow k = 3$

(b) $k = 3 \rightarrow k = 4$

(c) $k = 2 \rightarrow k = 4$

Figura 5.12: Matrices de migración entre las particiones de la capa `flatten` de WDA. Cada celda indica cuántas muestras del grupo de la fila pasan al grupo de la columna.

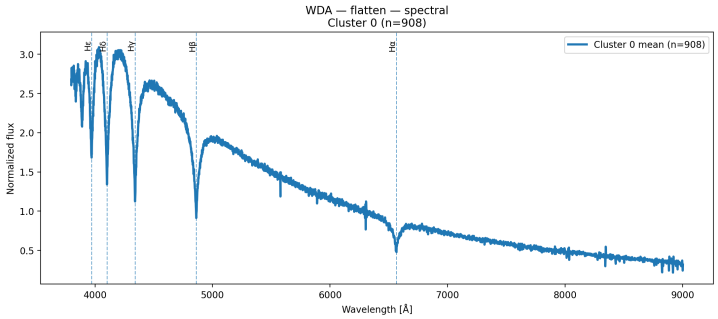
La lectura conjunta de las tres matrices revela un comportamiento asimétrico. El grupo minoritario que aparece ya con $k = 2$, de 72 muestras, se conserva a lo largo de todas las configuraciones: 50 de sus 72 elementos constituyen el grupo C2 de la partición con $k = 3$ (52 muestras en total), y 51 de esos 52 se mantienen como C2 al pasar a $k = 4$. Se trata, por tanto, de un núcleo de alrededor de medio centenar de espectros que ninguna de las particiones fragmenta.

El resto de la muestra, en cambio, no exhibe una estructura estable. Al pasar de $k = 2$

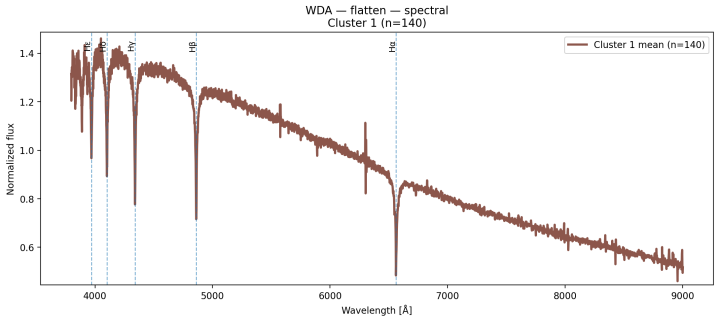
a $k = 3$, el grupo mayoritario de 1,028 muestras cede 140 elementos a un grupo nuevo, lo que constituye una subdivisión legítima. Pero en la transición de $k = 3$ a $k = 4$ el patrón se rompe: el grupo C1 de la partición con $k = 4$, de 249 muestras, se forma con 204 elementos procedentes de C0 y 44 procedentes de C1, es decir, reúne fragmentos de dos grupos distintos de la partición anterior. Una subdivisión jerárquica exigiría que cada grupo nuevo proviniera íntegramente de uno solo de los precedentes.

En consecuencia, la capa `flatten` sostiene un único agrupamiento robusto, el núcleo minoritario de unas 50 muestras, mientras que la mayoría de los espectros se redistribuye de forma distinta con cada valor de k, sin que ninguna frontera interna llegue a conservarse.

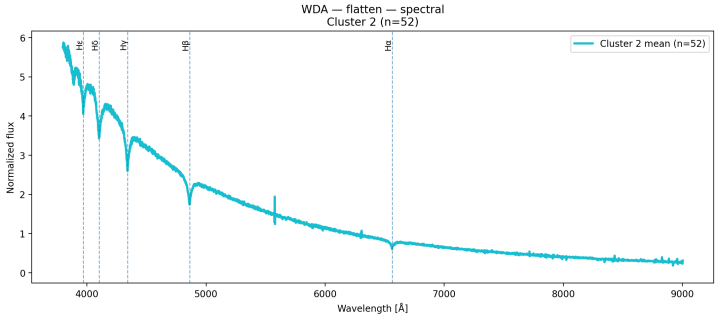
Finalmente se sintetizan los espectros representativos de la partición seleccionada por el eigengap para esta capa, $k = 3$. La Figura 5.13 presenta por separado el espectro promedio de cada uno de los 3 grupos, la Figura 5.14 los superpone junto al espectro promedio de la clase, lo que permite apreciar las diferencias sistemáticas entre ellos. Las Figuras 5.15 y 5.16 presentan el resultado equivalente para la partición obtenida con DBSCAN sobre esta misma capa, donde el conjunto de puntos etiquetados como ruido se trata como un grupo más, dado que su espectro promedio también resulta informativo.



(a) Cluster 0



(b) Cluster 1



(c) Cluster 2

Figura 5.13: Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 3$) sobre la capa `flatten` de WDA.

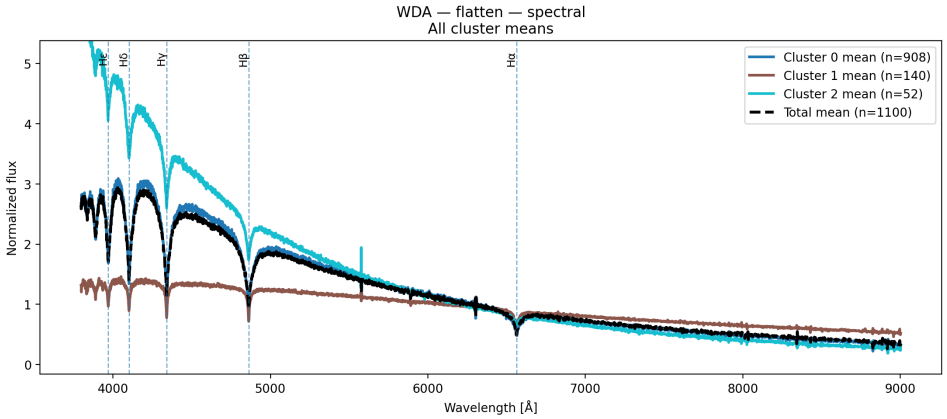
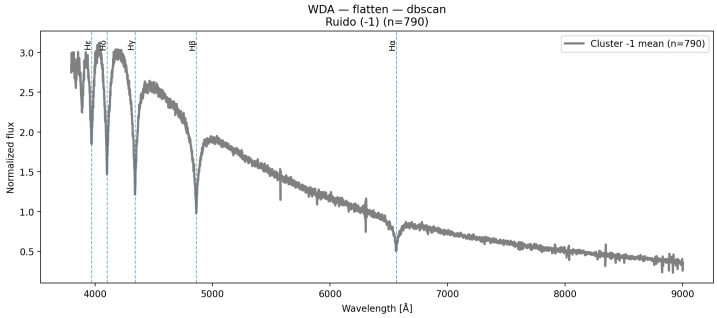
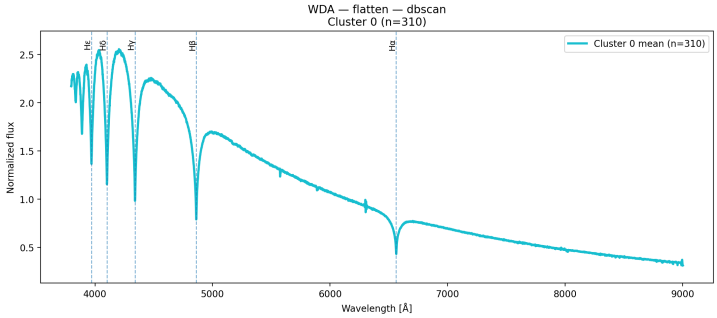


Figura 5.14: Comparación de los 3 espectros representativos obtenidos con Spectral Clustering ($k = 3$) sobre la capa `flatten`, superpuestos al espectro promedio de la clase WDA.



(a) Ruido (-1) ($n = 790$)



(b) Cluster 0 ($n = 310$)

Figura 5.15: Espectro promedio de cada grupo obtenido con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$) sobre la capa `flatten` de WDA, incluido el conjunto de puntos etiquetados como ruido.

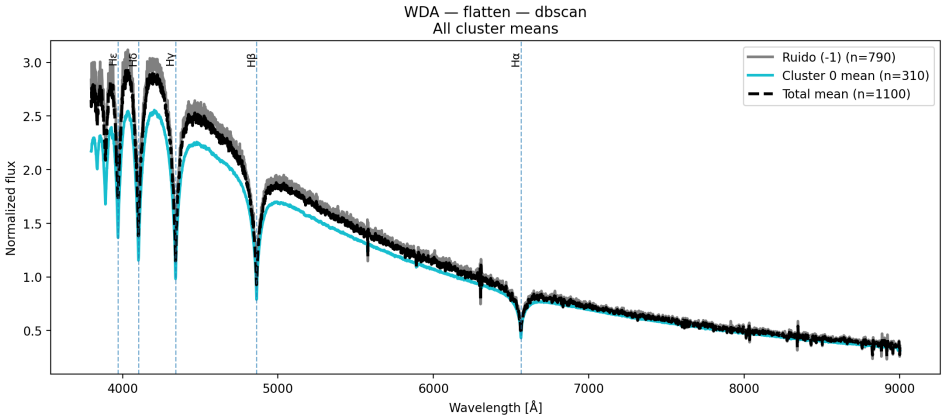
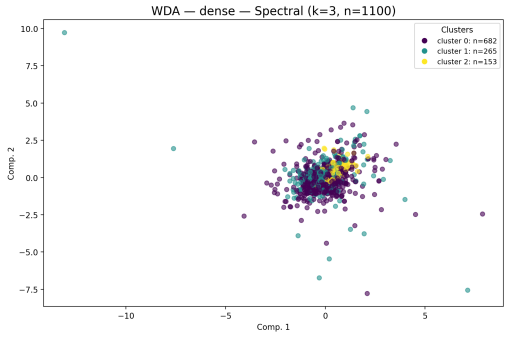


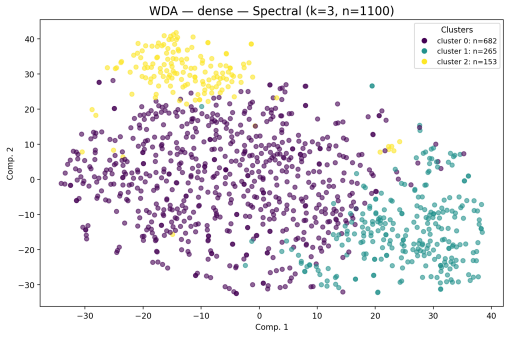
Figura 5.16: Comparación de los espectros representativos de los grupos de DBSCAN sobre la capa `flatten`, incluido el ruido, superpuestos al espectro promedio de la clase WDA.

5.11.4. Capa `dense`

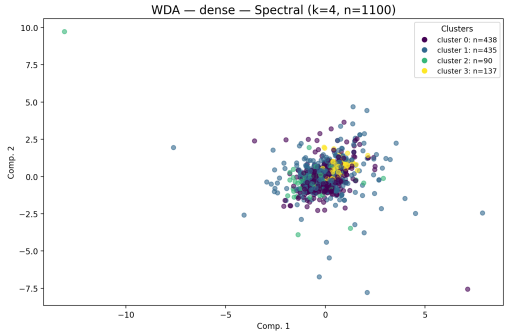
La Figura 5.17 presenta las particiones obtenidas sobre esta capa para $k \in \{3, 4, 5\}$, en ambas proyecciones.



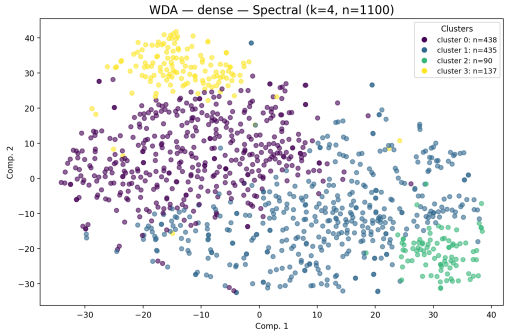
(a) $k = 3$, PCA



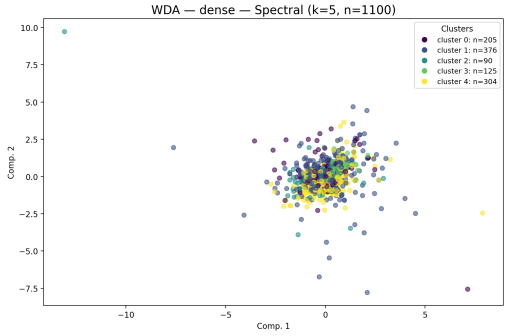
(b) $k = 3$, t-SNE



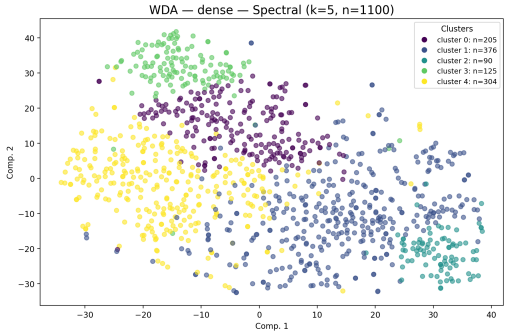
(c) $k = 4$, PCA



(d) $k = 4$, t-SNE



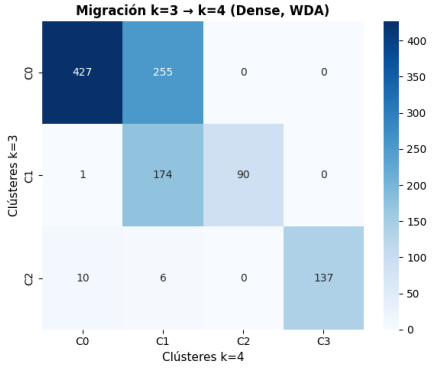
(e) $k = 5$, PCA



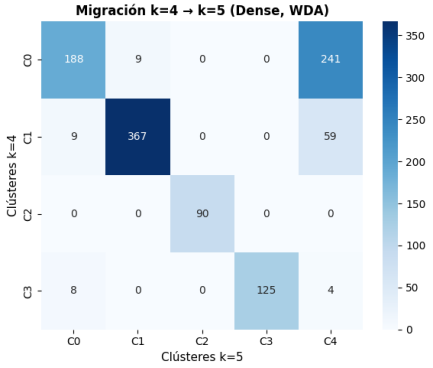
(f) $k = 5$, t-SNE

Figura 5.17: Particiones de Spectral Clustering sobre la capa `dense` de WDA ($n = 1{,}100$), para los valores de k evaluados. Columna izquierda: proyección PCA. Columna derecha: proyección t-SNE.

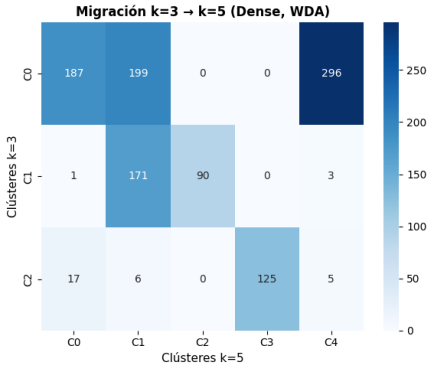
La Figura 5.18 recoge las matrices de migración entre estas configuraciones, que permiten seguir el destino de cada grupo al variar el número de clusters.



(a) $k = 3 \rightarrow k = 4$



(b) $k = 4 \rightarrow k = 5$



(c) $k = 3 \rightarrow k = 5$

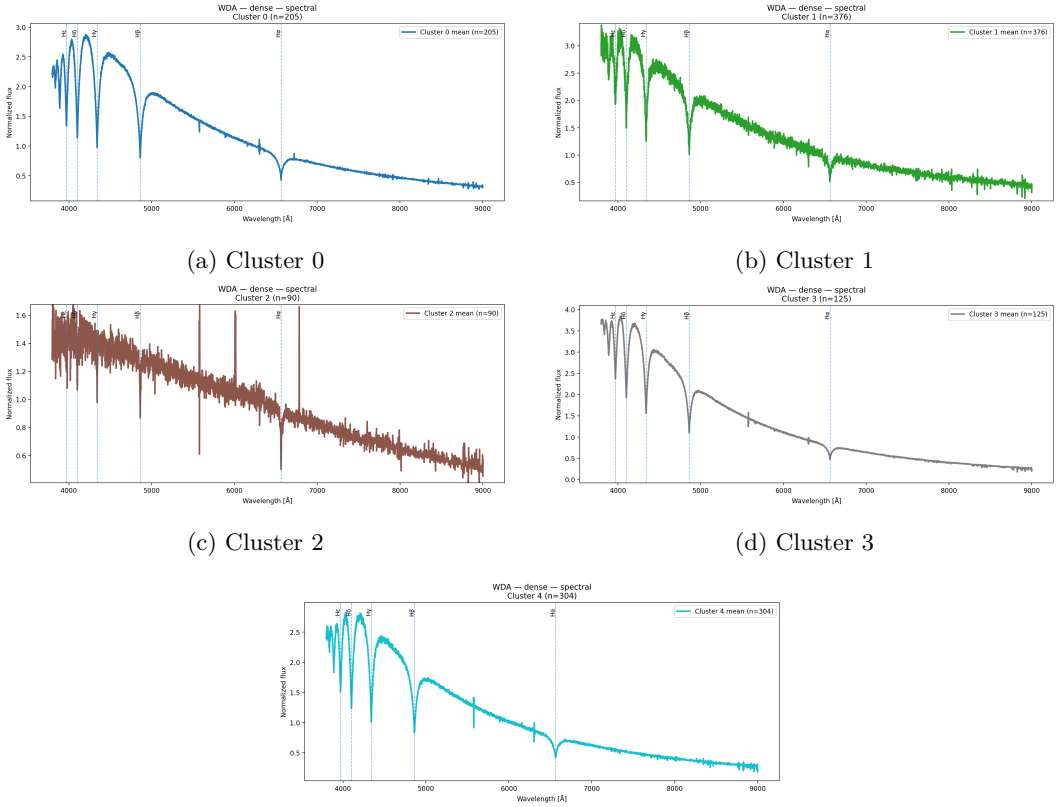
Figura 5.18: Matrices de migración entre las particiones de la capa `dense` de WDA. Cada celda indica cuántas muestras del grupo de la fila pasan al grupo de la columna.

Las matrices de `dense` describen el comportamiento más estable de las capas analizadas. Al pasar de $k = 3$ a $k = 4$ emerge un grupo de 90 muestras (C2), procedente en su totalidad del grupo C1 de la partición anterior, y ese grupo permanece intacto al pasar a $k = 5$: las 90 muestras se conservan sin pérdida ni incorporación alguna. Un segundo grupo presenta estabilidad comparable, aunque no perfecta: las 153 muestras de C2 en $k = 3$ dan lugar a 137 en C3 de $k = 4$ y a 125 en C3 de $k = 5$.

El resto de la muestra reproduce el patrón de reorganización ya observado en `flatten`. El grupo C1 de $k = 4$, de 435 muestras, combina 255 elementos de C0 con 174 de C1 de la partición con $k = 3$, y el grupo C4 de $k = 5$, de 304 muestras, combina 241 de C0 con 59 de C1 de la partición con $k = 4$. En ambos casos los grupos nuevos se construyen atravesando las fronteras previas en lugar de respetarlas.

Se concluye que dense contiene al menos dos subestructuras identificables y persistentes: una de 90 espectros y otra de alrededor de 130, inmersas en una población mayoritaria que carece de divisiones internas estables. El **grupo de** 90 **muestras**, por su invariancia completa ante el cambio de k, constituye el **candidato más sólido a subgrupo real de toda la exploración**.

Finalmente se sintetizan los espectros representativos de la partición seleccionada por el eigengap para esta capa, $k = 5$. La Figura 5.19 presenta por separado el espectro promedio de cada uno de los 5 grupos, la Figura 5.20 los superpone junto al espectro promedio de la clase, lo que permite apreciar las diferencias sistemáticas entre ellos. Las Figuras 5.21 y 5.22 presentan el resultado equivalente para la partición obtenida con DBSCAN sobre esta misma capa, donde el conjunto de puntos etiquetados como ruido se trata como un grupo más, dado que su espectro promedio también resulta informativo.



(a) Cluster 0

(b) Cluster 1

(c) Cluster 2

(d) Cluster 3

(e) Cluster 4

Figura 5.19: Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 5$) sobre la capa dense de WDA.

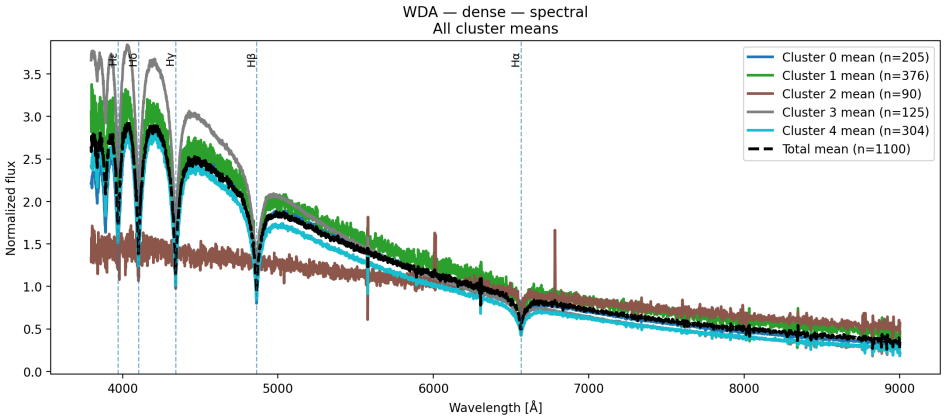
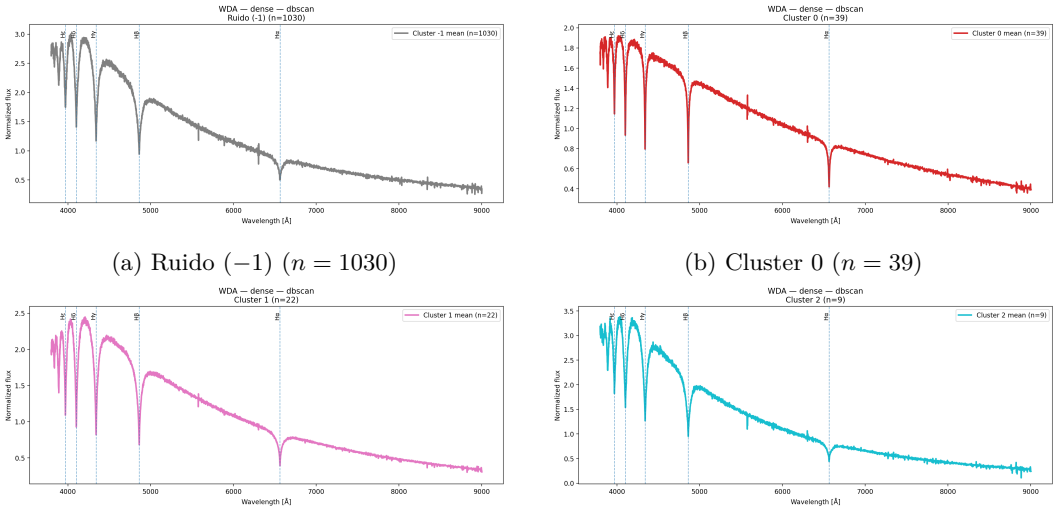


Figura 5.20: Comparación de los 5 espectros representativos obtenidos con Spectral Clustering ($k = 5$) sobre la capa `dense`, superpuestos al espectro promedio de la clase WDA.



(a) Ruido (-1) ($n = 1030$)

(b) Cluster 0 ($n = 39$)

(c) Cluster 1 ($n = 22$)

(d) Cluster 2 ($n = 9$)

Figura 5.21: Espectro promedio de cada grupo obtenido con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$) sobre la capa `dense` de WDA, incluido el conjunto de puntos etiquetados como ruido.

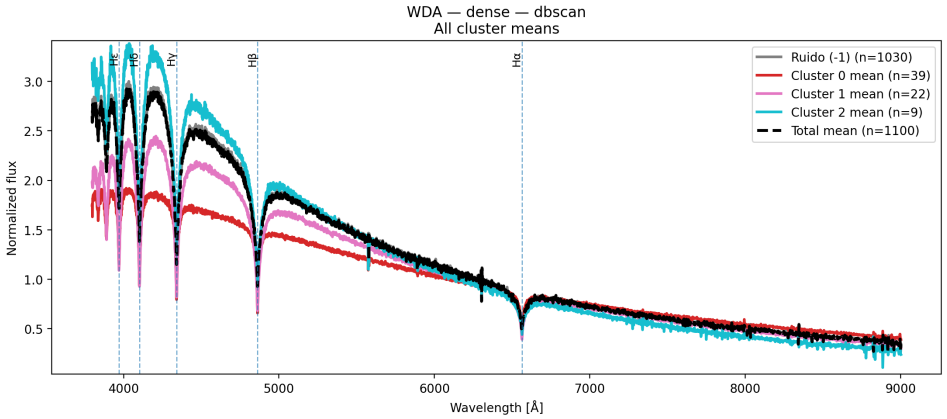
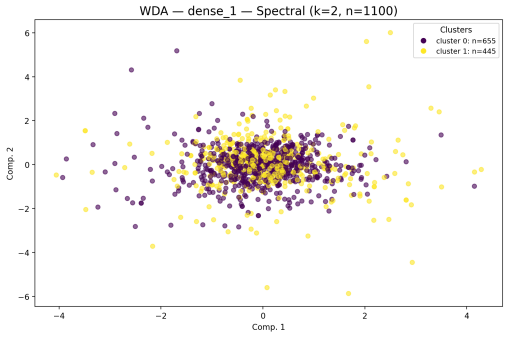


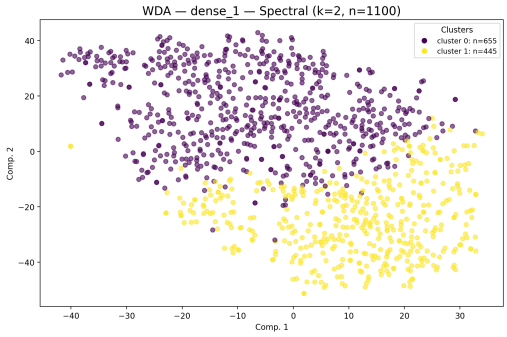
Figura 5.22: Comparación de los espectros representativos de los grupos de DBSCAN sobre la capa `dense`, incluido el ruido, superpuestos al espectro promedio de la clase WDA.

5.11.5. Capa `dense_1`

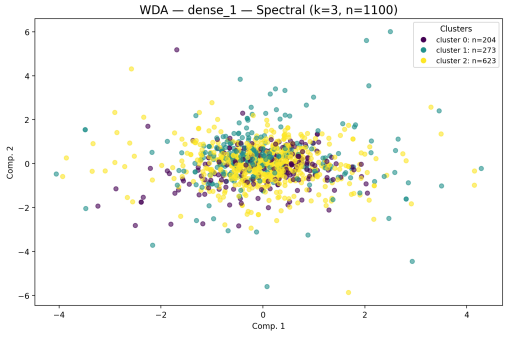
La Figura 5.23 presenta las particiones obtenidas sobre esta capa para $k \in \{2, 3, 4\}$, en ambas proyecciones.



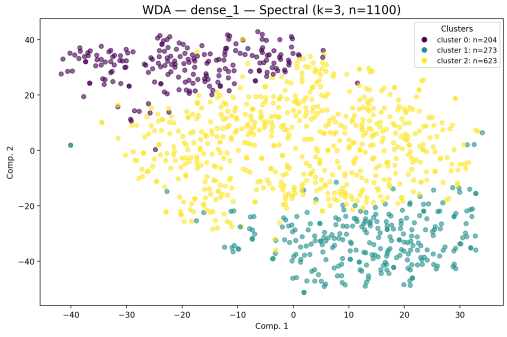
(a) $k = 2$, PCA



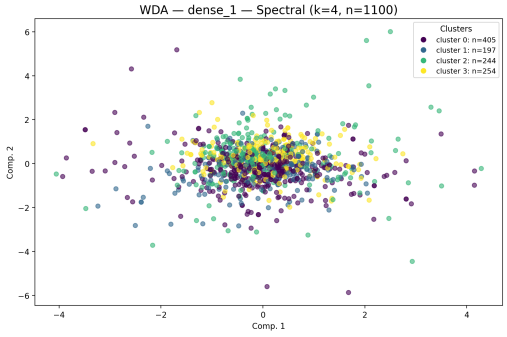
(b) $k = 2$, t-SNE



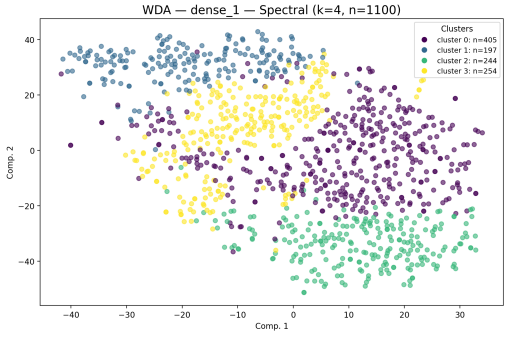
(c) $k = 3$, PCA



(d) $k = 3$, t-SNE



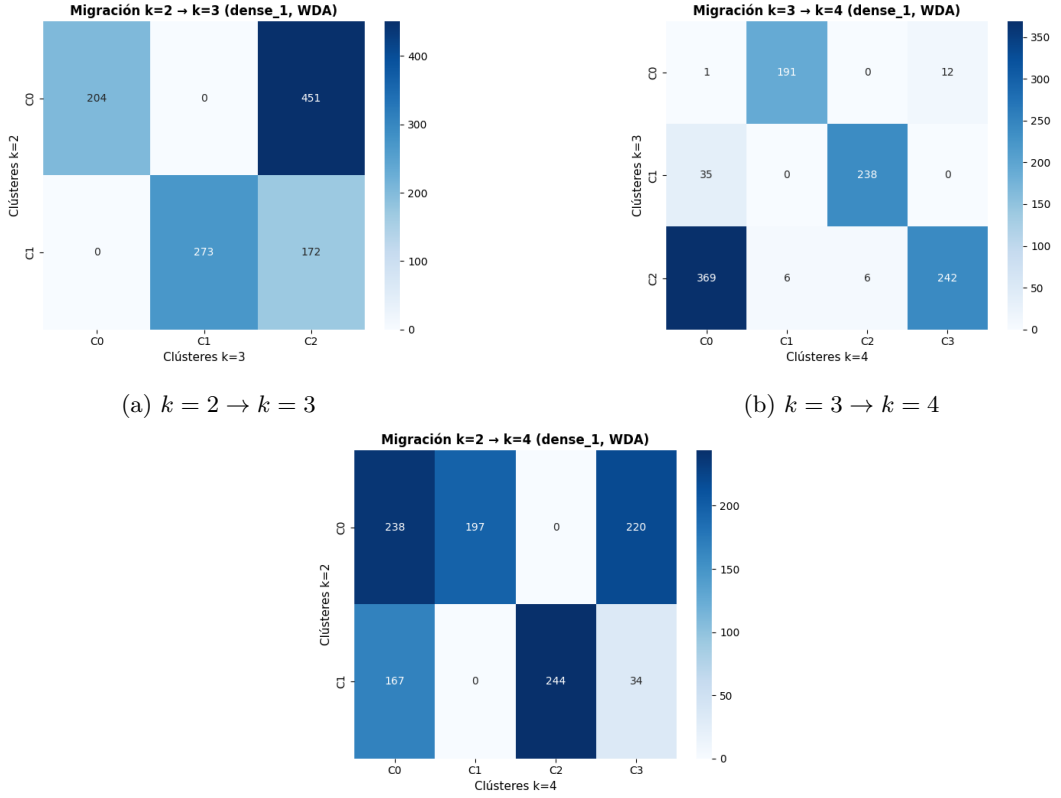
(e) $k = 4$, PCA



(f) $k = 4$, t-SNE

Figura 5.23: Particiones de Spectral Clustering sobre la capa `dense_1` de WDA ($n = 1{,}100$), para los valores de k evaluados. Columna izquierda: proyección PCA. Columna derecha: proyección t-SNE.

La Figura 5.24 recoge las matrices de migración entre estas configuraciones, que permiten seguir el destino de cada grupo al variar el número de clusters.



(a) $k = 2 \to k = 3$

(b) $k = 3 \to k = 4$

(c) $k = 2 \to k = 4$

Figura 5.24: Matrices de migración entre las particiones de la capa `dense_1` de WDA. Cada celda indica cuántas muestras del grupo de la fila pasan al grupo de la columna.

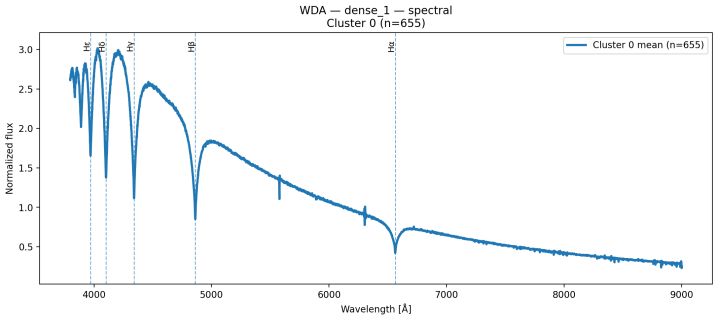
El caso de `dense_1` es el que se aparta con mayor claridad de una organización jerárquica. La matriz de $k = 2 \to k = 3$ muestra que la partición en tres grupos no constituye un refinamiento de la anterior: el grupo C2 de $k = 3$, de 623 muestras, se forma reuniendo 451 elementos de C0 con 172 de C1, esto es, fusionando fragmentos situados a ambos lados de la frontera trazada con $k = 2$. Los otros dos grupos sí la respetan, ya que C0 procede íntegramente de C0 y C1 íntegramente de C1, pero el grupo mayoritario la ignora por completo.

La transición de $k = 3$ a $k = 4$ resulta más ordenada: el grupo de 623 muestras se divide en dos, de 369 y 242 elementos, mientras que C0 y C1 se trasladan casi sin cambios a C1 (191 de 204) y C2 (238 de 273) respectivamente. La matriz de $k = 2 \to k = 4$ confirma, no obstante, que el cruce inicial persiste: el grupo C0 de $k = 4$ reúne 238 muestras de un lado

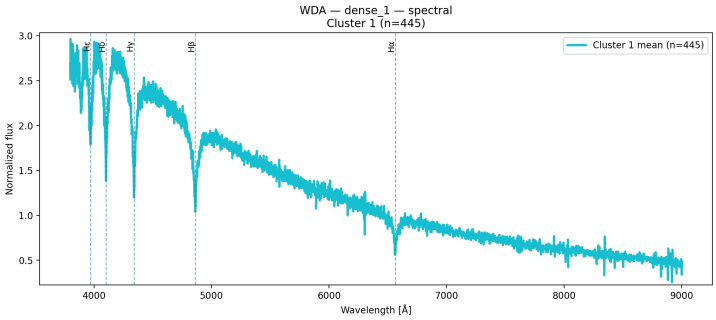
de la frontera original y 167 del otro.

Que la primera transición redistribuya la muestra en lugar de subdividirla indica que la frontera hallada con $k = 2$ no corresponde a una separación real del espacio latente, sino a un corte impuesto sobre una distribución sin discontinuidades. Esta lectura es coherente con los valores de Silhouette obtenidos en esta capa, próximos a cero para todos los k evaluados (0,083, 0,024 y −0,006).

Finalmente se sintetizan los espectros representativos de la partición seleccionada por el eigengap para esta capa, $k = 2$. La Figura 5.25 presenta por separado el espectro promedio de cada uno de los 2 grupos, la Figura 5.26 los superpone junto al espectro promedio de la clase, lo que permite apreciar las diferencias sistemáticas entre ellos. Las Figuras 5.27 y 5.28 presentan el resultado equivalente para la partición obtenida con DBSCAN sobre esta misma capa, donde el conjunto de puntos etiquetados como ruido se trata como un grupo más, dado que su espectro promedio también resulta informativo.



(a) Cluster 0



(b) Cluster 1

Figura 5.25: Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 2$) sobre la capa `dense_1` de WDA.

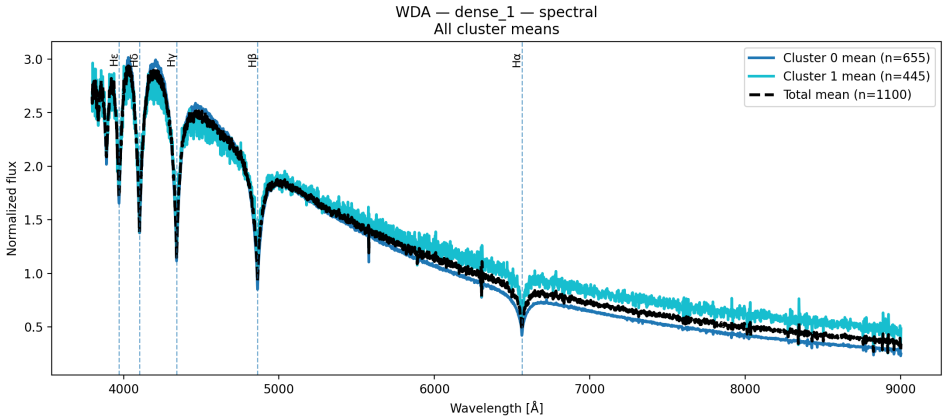
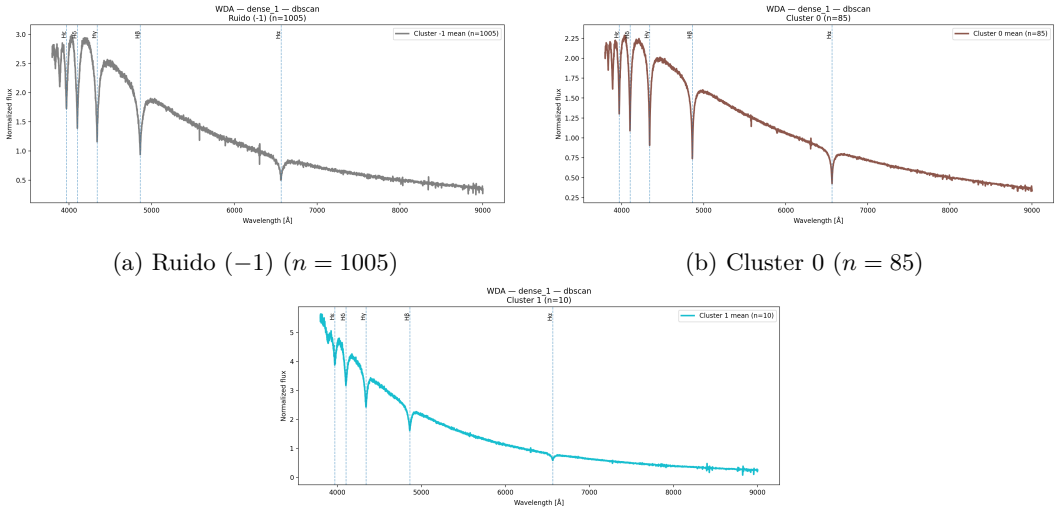


Figura 5.26: Comparación de los 2 espectros representativos obtenidos con Spectral Clustering ($k = 2$) sobre la capa `dense_1`, superpuestos al espectro promedio de la clase WDA.



(a) Ruido (-1) ($n = 1005$)

(b) Cluster 0 ($n = 85$)

(c) Cluster 1 ($n = 10$)

Figura 5.27: Espectro promedio de cada grupo obtenido con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$) sobre la capa `dense_1` de WDA, incluido el conjunto de puntos etiquetados como ruido.

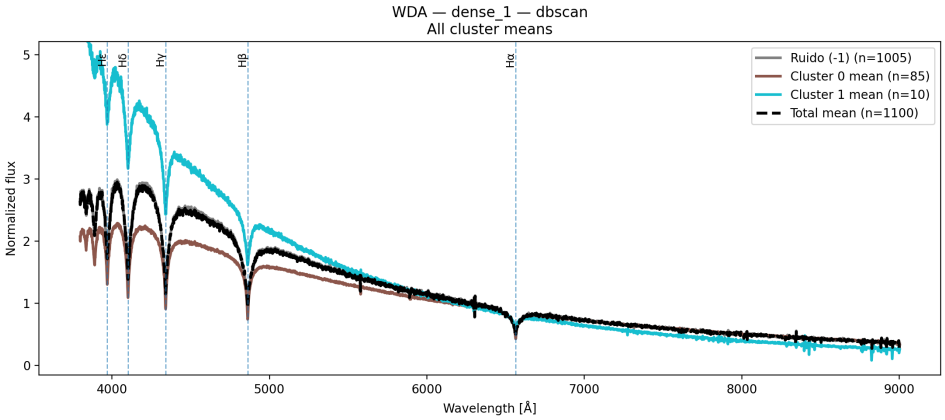
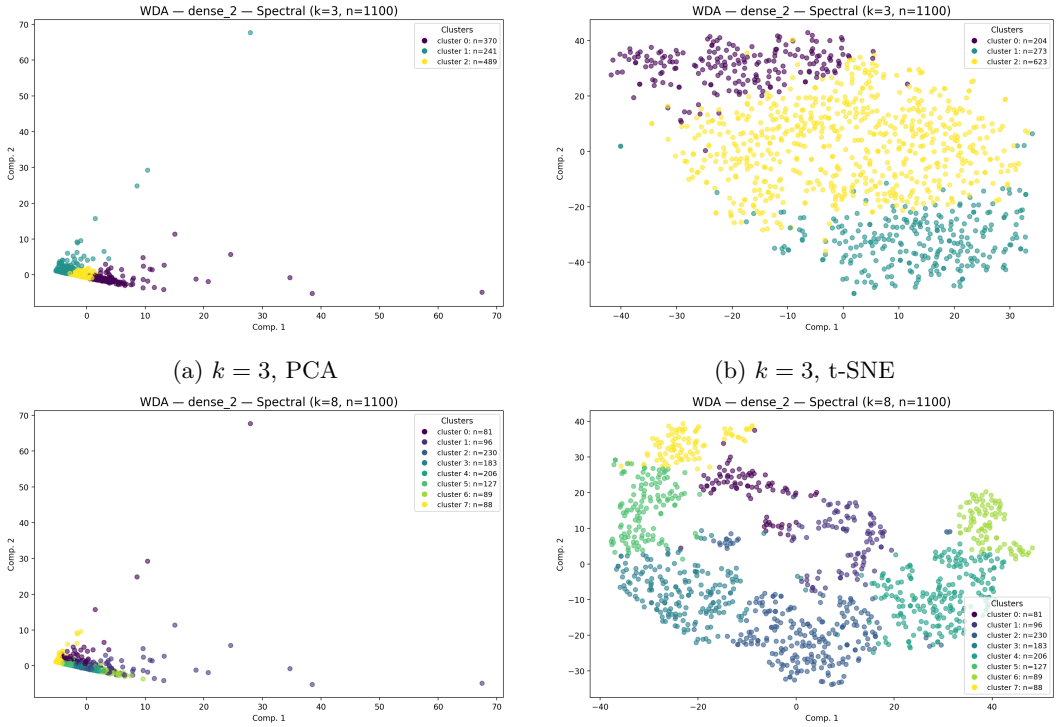


Figura 5.28: Comparación de los espectros representativos de los grupos de DBSCAN sobre la capa `dense_1`, incluido el ruido, superpuestos al espectro promedio de la clase WDA.

5.11.6. Capa `dense_2`

La Figura 5.29 presenta las particiones obtenidas sobre esta capa para $k \in \{3, 8\}$, en ambas proyecciones.



(a) $k = 3$, PCA

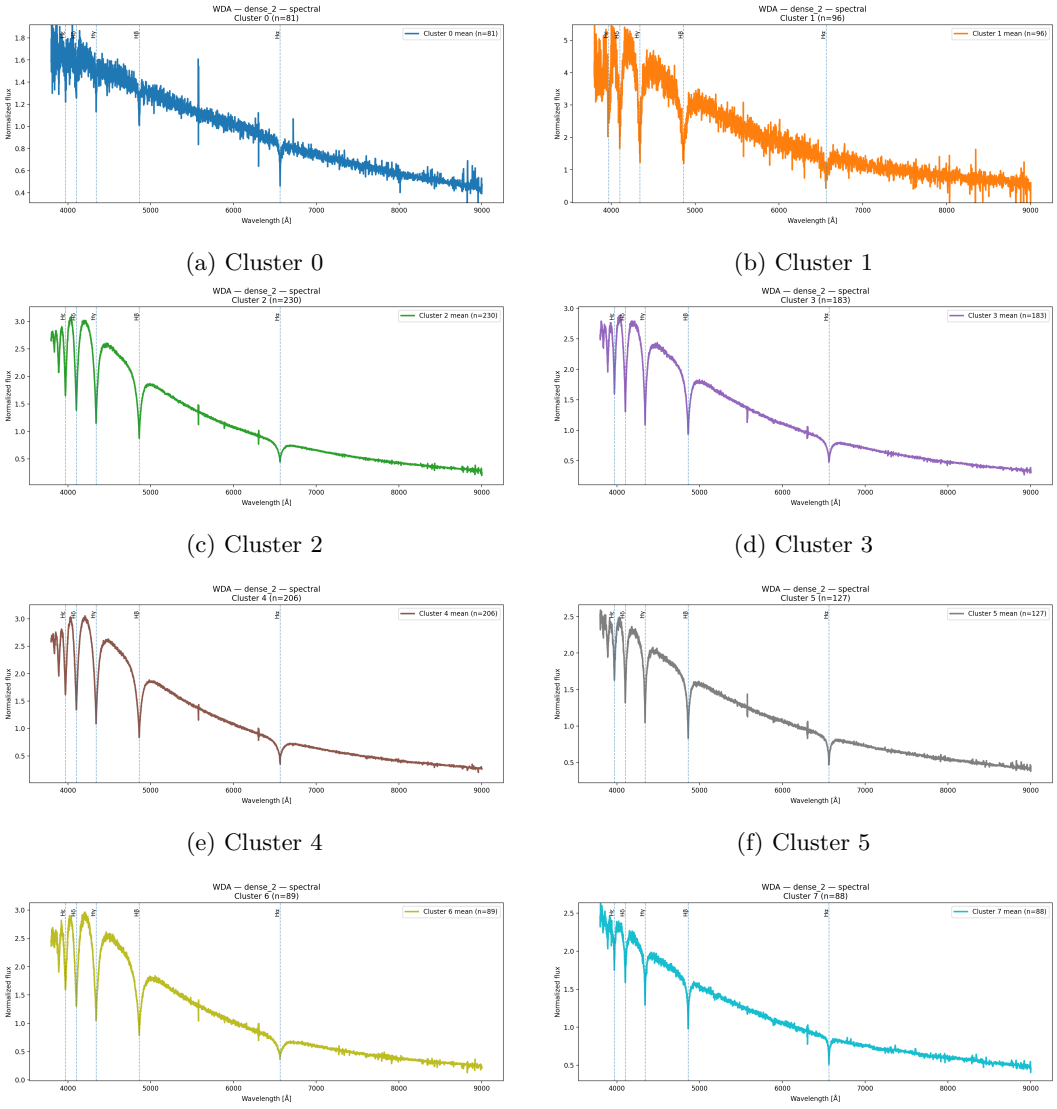
(b) $k = 3$, t-SNE

(c) $k = 8$, PCA

(d) $k = 8$, t-SNE

Figura 5.29: Particiones de Spectral Clustering sobre la capa `dense_2` de WDA ($n = 1{,}100$), para los valores de k evaluados. Columna izquierda: proyección PCA. Columna derecha: proyección t-SNE.

Finalmente se sintetizan los espectros representativos de la partición seleccionada por el eigengap para esta capa, $k = 8$. La Figura 5.30 presenta por separado el espectro promedio de cada uno de los 8 grupos, la Figura 5.31 los superpone junto al espectro promedio de la clase, lo que permite apreciar las diferencias sistemáticas entre ellos. Las Figuras 5.32 y 5.33 presentan el resultado equivalente para la partición obtenida con DBSCAN sobre esta misma capa, donde el conjunto de puntos etiquetados como ruido se trata como un grupo más, dado que su espectro promedio también resulta informativo.



(a) Cluster 0

(b) Cluster 1

(c) Cluster 2

(d) Cluster 3

(e) Cluster 4

(f) Cluster 5

(g) Cluster 6

(h) Cluster 7

Figura 5.30: Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 8$) sobre la capa `dense_2` de WDA.

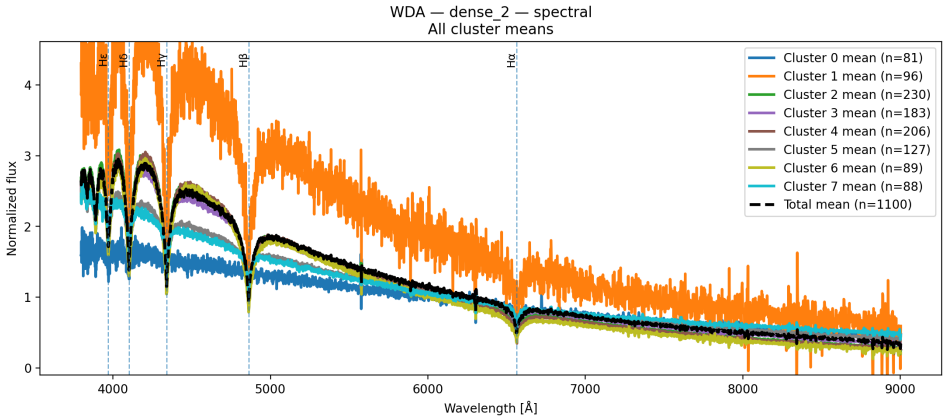
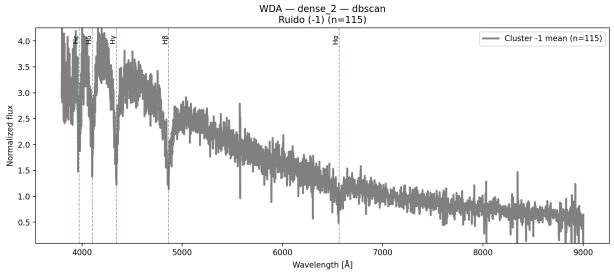
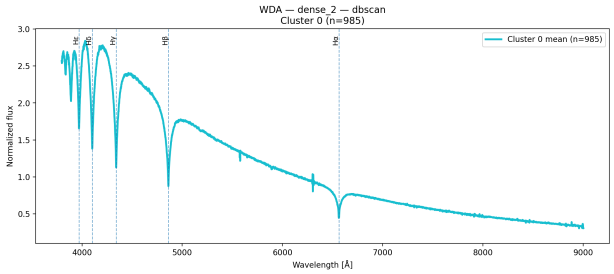


Figura 5.31: Comparación de los 8 espectros representativos obtenidos con Spectral Clustering ($k = 8$) sobre la capa `dense_2`, superpuestos al espectro promedio de la clase WDA.



(a) Ruido (-1) ($n = 115$)



(b) Cluster 0 ($n = 985$)

Figura 5.32: Espectro promedio de cada grupo obtenido con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$) sobre la capa `dense_2` de WDA, incluido el conjunto de puntos etiquetados como ruido.

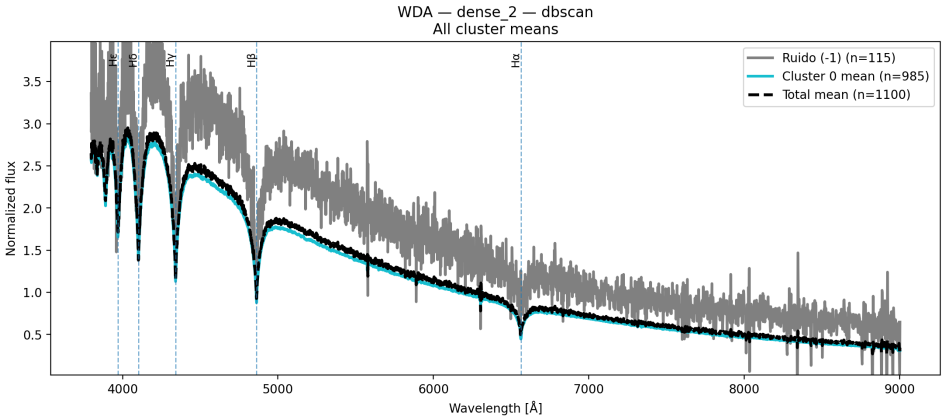


Figura 5.33: Comparación de los espectros representativos de los grupos de DBSCAN sobre la capa `dense_2`, incluido el ruido, superpuestos al espectro promedio de la clase WDA.

5.11.7. Métricas internas de las particiones

La Tabla 5.8 reúne las métricas internas de todas las configuraciones de Spectral Clustering evaluadas sobre WDA. Su lectura ordena las capas de forma inequívoca: la partición de `dense_2` con $k = 8$ obtiene simultáneamente el mejor valor en las tres métricas, con un Silhouette de 0,1382 frente al máximo de 0,0831 del resto de la tabla, el único Davies–Bouldin por debajo de 2 y un Calinski–Harabasz de 84,41 que triplica con holgura el de cualquier otra configuración. Las capas restantes entregan Silhouette próximo a cero o negativo en cuanto k supera el valor 2.

Tabla 5.8: Métricas internas de las particiones de Spectral Clustering evaluadas sobre la clase WDA (n = 1,100). Silhouette y Calinski–Harabasz: valores más altos indican mejor partición. Davies–Bouldin: valores más bajos.

Capa	k	**Silhouette**	**Davies–Bouldin**	**Calinski–Harabasz**
`flatten`	2	0.0668	2.288	21.49
`flatten`	3	−0,0472	2.662	20.24
`flatten`	4	−0,0258	3.301	18.87
`dense`	3	−0,0692	4.257	22.01
`dense`	4	−0,0642	4.053	21.47
`dense`	5	−0,0577	3.820	20.76
`dense_1`	2	0.0831	5.122	25.13
`dense_1`	3	0.0238	4.401	24.49
`dense_1`	4	−0,0056	4.106	22.98
`dense_2`	8	**0.1382**	**1.5860**	**84.41**

Ese orden, sin embargo, admite tres reparos que conviene formular antes de darle valor interpretativo.

El primero concierne a la magnitud absoluta de los valores. El Silhouette Score está acotado en el intervalo $[-1, 1]$, y suele considerarse que solo por encima de 0,25 indica una estructura de grupos apreciable. La mejor partición de toda la exploración alcanza 0,1382, de modo que *la mejor* no equivale aquí a *buena*: incluso sobre `dense_2` los grupos resultan poco compactos en relación con la distancia que los separa. Este resultado es coherente con la geometría descrita en la Sección 5.8, en la que los espacios latentes aparecen como una única concentración de densidad sin separaciones marcadas, y explica por qué la métrica no llega a valores altos en ninguna configuración.

El segundo es que las métricas no coinciden entre sí. Entre las capas distintas de `dense_2`, el Silhouette sitúa en primer lugar a `dense_1` con $k = 2$ (0,0831), que es sin embargo la configuración con el peor Davies–Bouldin de la tabla completa (5,122). El índice de Davies–Bouldin, por su parte, prefiere `flatten` con $k = 2$ (2,288), que el Silhouette relega al segundo lugar. El Calinski–Harabasz, en tanto, se mantiene entre 18,87 y 25,13 en las nueve configuraciones ajenas a `dense_2`, es decir, no discrimina entre ellas. Tres índices que ordenan de tres maneras distintas no constituyen un criterio de decisión.

El tercero afecta a la comparación entre capas. Como se advirtió en la Sección 5.6, `dense_2` es la única capa que no atraviesa la reducción por PCA, por poseer ya de forma nativa las 32 dimensiones del espacio común. Parte de su ventaja métrica podría proceder de ese tratamiento diferenciado y no de la representación misma.

Un aspecto sí resulta informativo, y es la evolución de las métricas al aumentar k dentro

de cada capa. En `flatten` y en `dense_1` el mejor valor se obtiene en $k = 2$ y se degrada al subdividir, lo que indica que las divisiones adicionales no corresponden a estructura presente en el espacio. En `dense` ocurre lo contrario: en el rango evaluado el Silhouette mejora de forma sostenida al pasar de $k = 3$ a $k = 5$ ($-0{,}0692$, $-0{,}0642$ y $-0{,}0577$) y el Davies–Bouldin lo acompaña (4,257, 4,053 y 3,820). Es la única capa en que subdividir mejora la partición, y coincide exactamente con aquella en que las matrices de migración de la Figura 5.18 muestran la emergencia de un grupo estable de 90 muestras al pasar de $k = 3$ a $k = 4$. Las dos evidencias son independientes y apuntan en el mismo sentido.

El contraste decisivo, no obstante, se produce al confrontar esta tabla con la lectura física de los espectros representativos que desarrolla el Capítulo 6. Los tres hallazgos con contenido astrofísico de la clase WDA proceden de configuraciones que la tabla puntúa mal. La separación por temperatura efectiva sobre `flatten` se obtiene con $k = 3$, cuyo Silhouette es $-0{,}0472$, el grupo frío de 90 espectros, el candidato a subgrupo más sólido de toda la exploración, procede de `dense` con $k = 5$, cuyo Silhouette es $-0{,}0577$, y la separación por anchura de las líneas de Balmer procede de DBSCAN sobre esa misma capa. La configuración mejor puntuada, `dense_2` con $k = 8$, no dio lugar a ninguna interpretación física reportada.

Esta divergencia no invalida las métricas, sino que delimita lo que miden. El Silhouette y el Davies–Bouldin evalúan compacidad y separación en términos de distancias euclidianas respecto de centroides, mientras que Spectral Clustering no construye sus grupos sobre esa noción sino sobre la conectividad del grafo de vecindad. Una partición que separa correctamente dos regiones alargadas o curvas del espacio latente recibe por construcción un Silhouette bajo, sin que ello signifique que la separación sea espuria. El caso de DBSCAN sobre `dense` lo ilustra en la dirección opuesta: su Silhouette de 0,289 es el más alto de todos los reportados para WDA, pero se calcula únicamente sobre las 70 muestras que el algoritmo no descartó como ruido, es decir, sobre un 6 % de la clase previamente depurado de sus elementos ambiguos. Comparar ese valor con los de la tabla carecería de sentido.

Se desprende de lo anterior el criterio con que las métricas se emplean en el resto del capítulo. Sirven para descartar configuraciones degeneradas, para verificar que una partición no colapsa y para comparar capas bajo una regla homogénea, y en ese uso son necesarias. No sirven, en cambio, para decidir qué partición merece ser examinada, decisión que en este trabajo recae en la estabilidad ante el cambio de k que revelan las matrices de migración y en el contenido de los espectros representativos.

5.12. Comportamiento de DBSCAN sobre la clase WDA

Las métricas recogidas en la Tabla 5.8 corresponden únicamente a las particiones de Spectral Clustering. Resta describir el comportamiento de DBSCAN sobre esta misma clase, cuyas métricas no figuran en aquella tabla por corresponder a un algoritmo de naturaleza distinta, que no requiere fijar el número de grupos y que etiqueta parte de la muestra como

ruido.

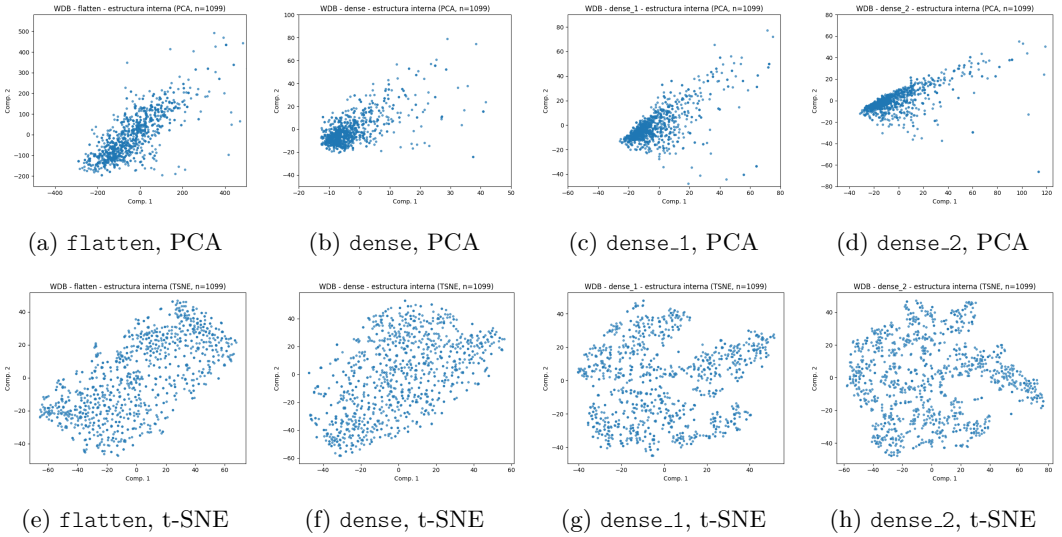
En cuanto a DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$), su comportamiento varía de forma notable entre capas. Sobre `flatten` y `dense_2` no llega a formar más de un grupo, con 790 y 115 puntos etiquetados como ruido respectivamente, de modo que sus métricas internas no quedan definidas. Sobre `dense` identifica tres grupos con 1,030 puntos de ruido (sil = 0,289, DB = 1,24, CH = 25,6) y sobre `dense_1`, dos grupos con 1,005 puntos de ruido (sil = 0,146, DB = 1,55, CH = 10,4). En los dos casos en que forma grupos, estos reúnen entre el 3 % y el 6 % de la muestra, lo que confirma la fuerte sensibilidad del algoritmo a la densidad del espacio latente ya discutida en la Sección 5.9.

5.13. Aplicación del pipeline a la clase WDB

Todo el desarrollo descrito hasta este punto se construyó utilizando la clase WDA: la exploración visual de los espacios latentes, la selección del número de clusters, los primeros experimentos con los tres algoritmos y la consolidación del pipeline en funciones reutilizables. Con el procedimiento ya definido, esta sección lo aplica a la clase WDB sin modificar la metodología, lo que permite a la vez obtener resultados para esta clase y verificar que el pipeline es reutilizable más allá de aquella con la que fue desarrollado. Por tratarse de una aplicación y no de una construcción, la exposición omite las justificaciones metodológicas ya expuestas y se limita a los resultados.

5.13.1. Exploración de los espacios latentes

Al igual que en WDA, el primer paso consiste en verificar si los espacios latentes de WDB exhiben agrupamientos discernibles antes de aplicar ningún algoritmo. La Figura 5.34 presenta las proyecciones PCA y t-SNE de las cuatro capas.



(a) `flatten`, PCA (b) `dense`, PCA (c) `dense_1`, PCA (d) `dense_2`, PCA

(e) `flatten`, t-SNE (f) `dense`, t-SNE (g) `dense_1`, t-SNE (h) `dense_2`, t-SNE

Figura 5.34: Proyecciones 2D de los espacios latentes de las cuatro capas para la clase WDB ($n = 1{,}100$). Fila superior: PCA. Fila inferior: t-SNE. Al igual que en WDA, no se aprecian agrupamientos evidentes en ninguna capa ni proyección.

El resultado reproduce el de WDA: en ninguna capa ni bajo ninguna de las dos proyecciones se aprecian subgrupos separados a simple vista. La muestra forma en todos los casos una nube única, sin discontinuidades que permitan delimitar poblaciones por inspección visual. Esto confirma que la subestructura, de existir, no es accesible desde la proyección bidimensional y justifica aplicar sobre WDB el mismo procedimiento de clustering sobre el espacio completo de 32 dimensiones.

Cabe señalar que esta observación tiene además una consecuencia práctica sobre la elección de algoritmos. Como se estableció en la Sección 5.9, K-Means produce particiones degeneradas sobre nubes unimodales sin separaciones claras, al limitarse a aislar los puntos más alejados del centro. Dado que la geometría de los espacios latentes de WDB es equivalente a la de WDA, ese algoritmo se descarta también aquí sin repetir el experimento, y el análisis prosigue directamente con Spectral Clustering.

5.13.2. Selección de k mediante el eigengap

El número de clusters se determina con el mismo criterio empleado en WDA: la heurística del eigengap del Laplaciano, aplicada en la Sección 5.9. Las Figuras 5.35 a 5.38 muestran el espectro de autovalores de cada capa, y la Tabla 5.9 resume los valores obtenidos junto con la magnitud del salto que los sustenta.

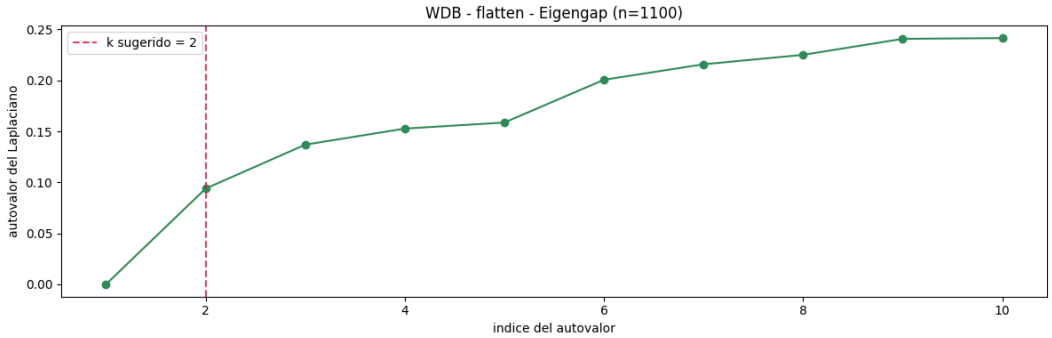


Figura 5.35: Autovalores del Laplaciano normalizado para la capa `flatten` de WDB ($n = 1{,}100$). El mayor salto entre autovalores consecutivos indica $k = 2$.

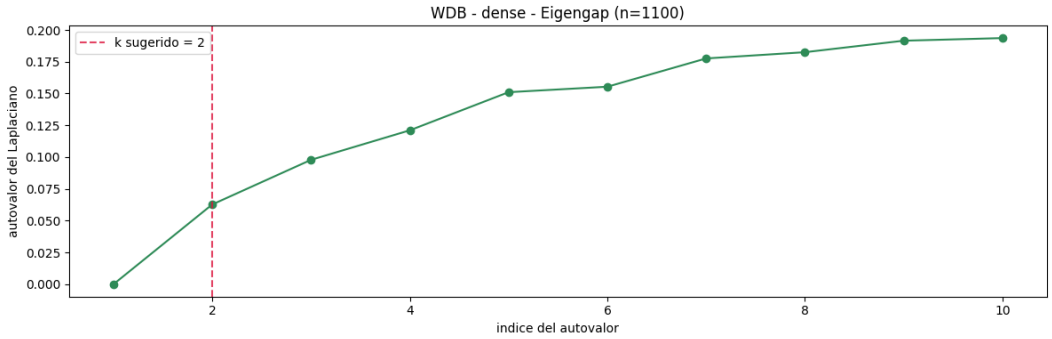


Figura 5.36: Autovalores del Laplaciano normalizado para la capa `dense` de WDB ($n = 1{,}100$). El mayor salto entre autovalores consecutivos indica $k = 2$.

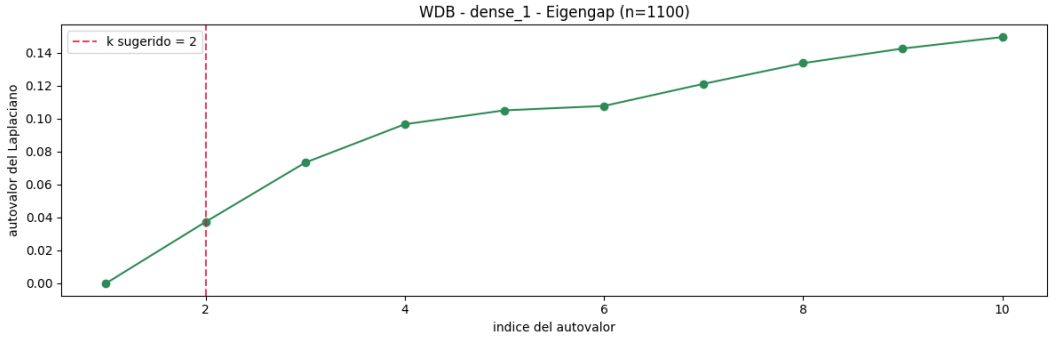


Figura 5.37: Autovalores del Laplaciano normalizado para la capa `dense_1` de WDB ($n = 1{,}100$). El mayor salto entre autovalores consecutivos indica $k = 2$.

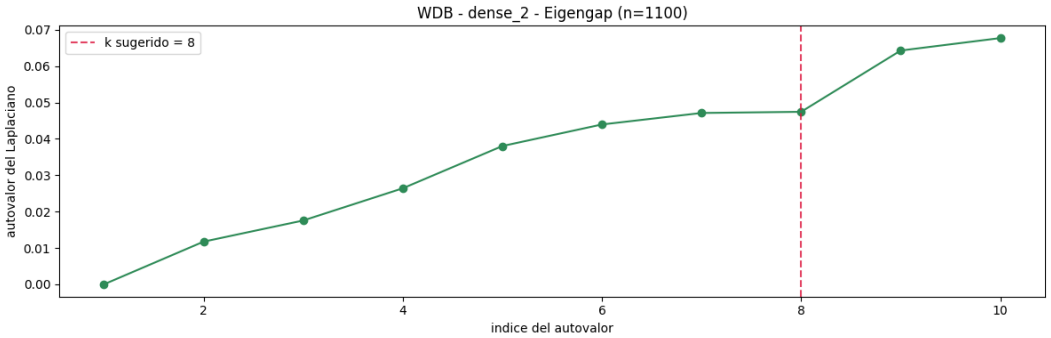


Figura 5.38: Autovalores del Laplaciano normalizado para la capa `dense_2` de WDB ($n = 1{,}100$). El mayor salto entre autovalores consecutivos indica $k = 8$.

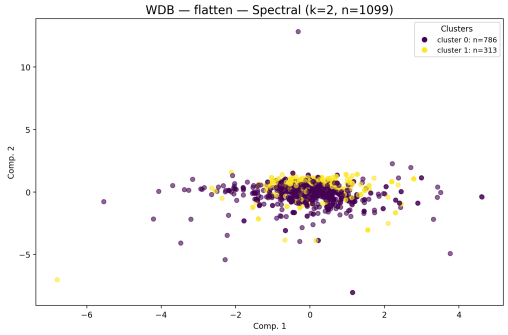
Tabla 5.9: Número de clusters indicado por el eigengap del Laplaciano en cada capa de la clase WDB ($n = 1{,}100$), con la magnitud del salto entre autovalores consecutivos en que se sustenta.

Capa	***k* sugerido**	**Salto**
`flatten`	2	0,0431
`dense`	2	0,0351
`dense_1`	2	0,0360
`dense_2`	8	0,0169

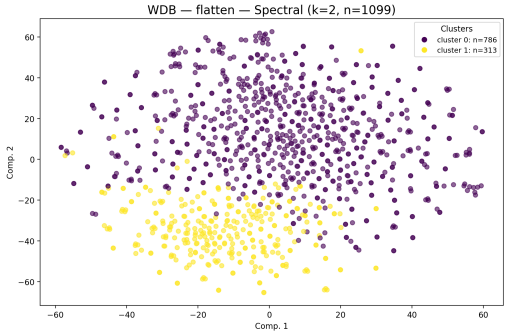
Las tres capas más superficiales coinciden en $k = 2$, mientras que `dense_2` indica $k = 8$. Conviene notar que la magnitud del salto decrece de forma sostenida al avanzar hacia la salida, desde 0,0431 en `flatten` hasta 0,0169 en `dense_2`: el criterio señala un valor en las cuatro capas, pero lo hace con evidencia progresivamente más débil. El $k = 8$ de `dense_2` es, en particular, el peor sustentado de los cuatro, lo que aconseja tratarlo con reserva y evaluar también valores contiguos, tal como se procedió en WDA.

5.13.3. Capa `flatten`

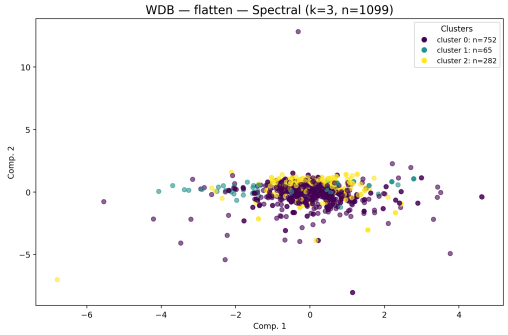
Se evalúan $k = 2$, valor indicado por el eigengap, y $k = 3$ como valor contiguo superior. No se examina $k = 1$ por carecer de sentido como partición. La Figura 5.39 presenta ambas configuraciones en las dos proyecciones.



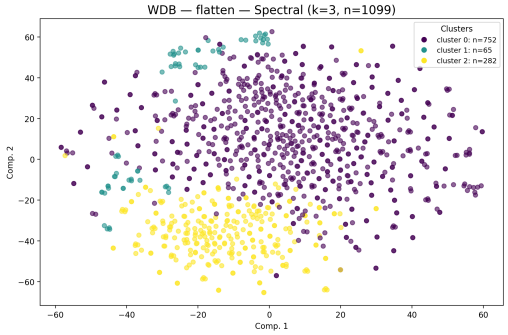
(a) $k = 2$, PCA



(b) $k = 2$, t-SNE



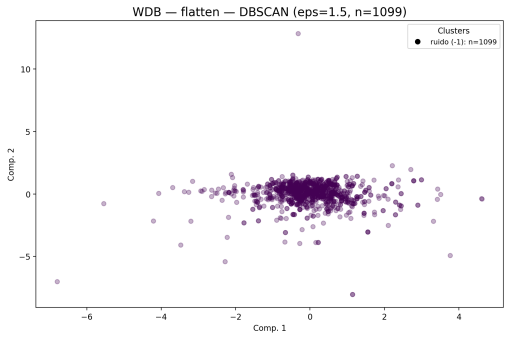
(c) $k = 3$, PCA



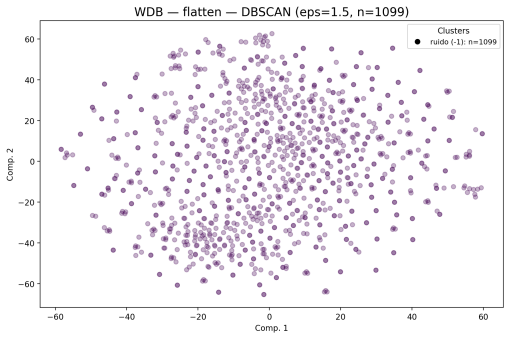
(d) $k = 3$, t-SNE

Figura 5.39: Particiones de Spectral Clustering sobre la capa `flatten` de WDB ($n = 1{,}100$). Columna izquierda: proyección PCA. Columna derecha: t-SNE.

La Figura 5.40 muestra el resultado de DBSCAN sobre esta capa.



(a) Proyección PCA

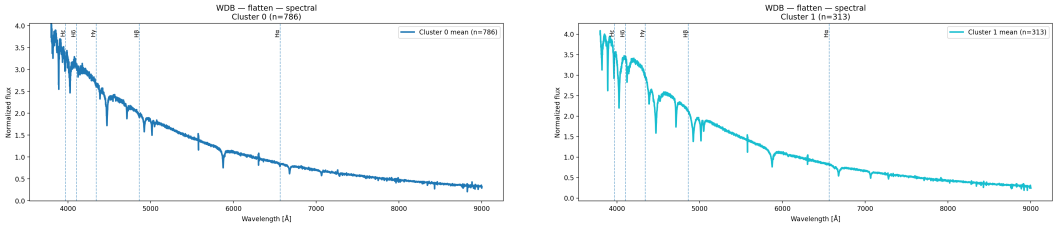


(b) Proyección t-SNE

Figura 5.40: Partición obtenida con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$) sobre la capa `flatten` de WDB, en ambas proyecciones.

No se sintetizan espectros representativos para esta partición. Al no formarse ningún cluster, el único grupo existente es el del ruido, que reúne la totalidad de la muestra: su espectro promedio coincide por construcción con el promedio de la clase y no aporta información alguna sobre subestructura.

Los espectros representativos de la partición con $k = 2$ se presentan en las Figuras 5.41 y 5.42.



(a) Cluster 0 (b) Cluster 1

Figura 5.41: Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 2$) sobre la capa `flatten` de WDB.

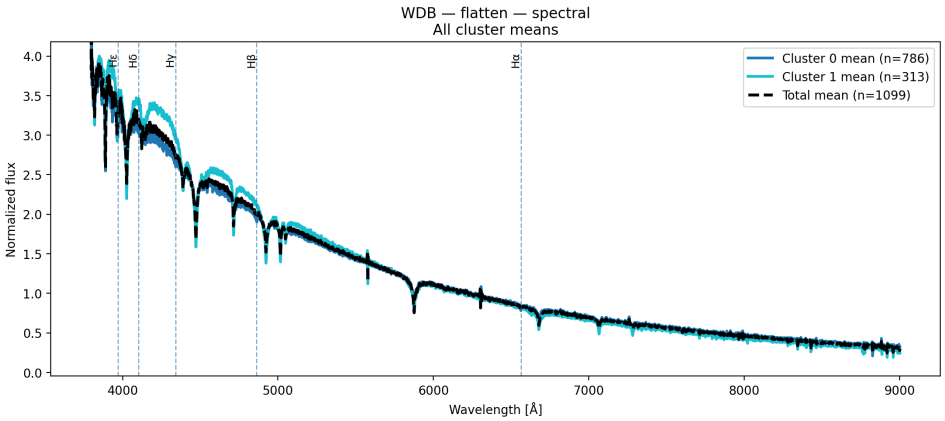
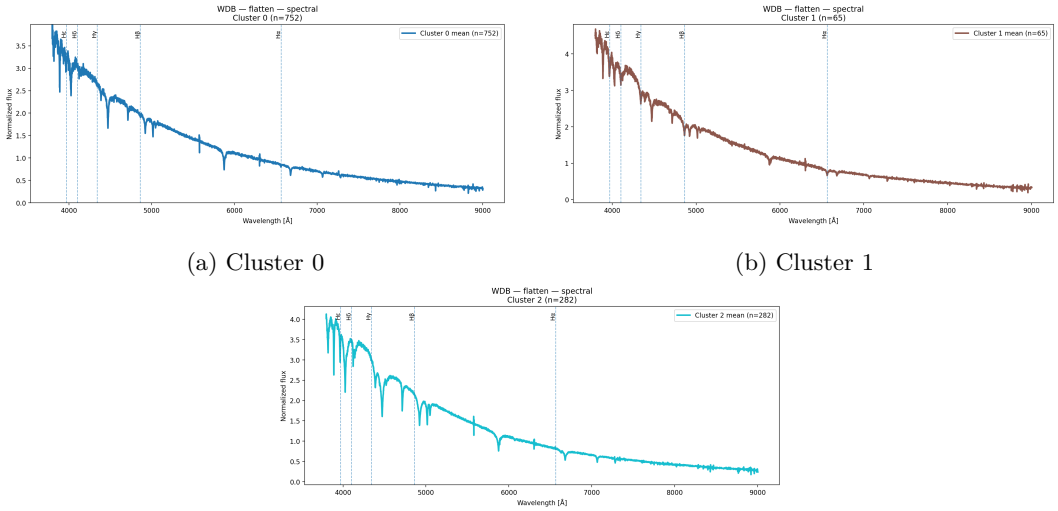


Figura 5.42: Comparación de los 2 espectros representativos ($k = 2$) sobre la capa `flatten`, superpuestos al espectro promedio de la clase WDB.

Los espectros representativos de la partición con $k = 3$ se presentan en las Figuras 5.43 y 5.44.



(a) Cluster 0

(b) Cluster 1

(c) Cluster 2

Figura 5.43: Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 3$) sobre la capa `flatten` de WDB.

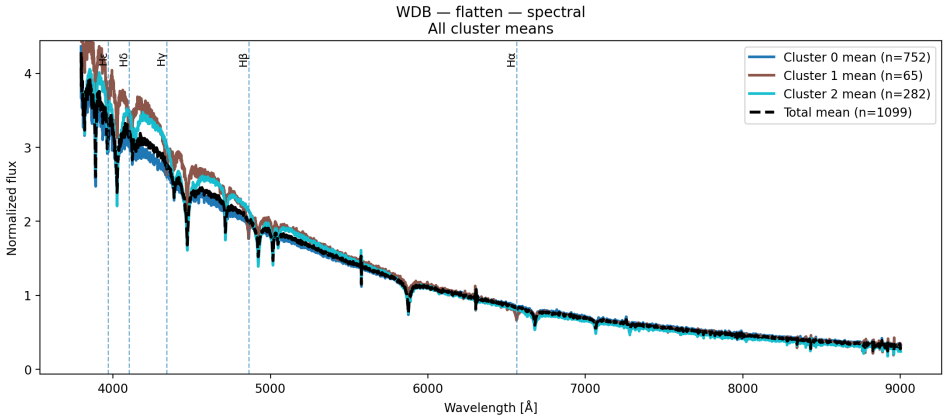
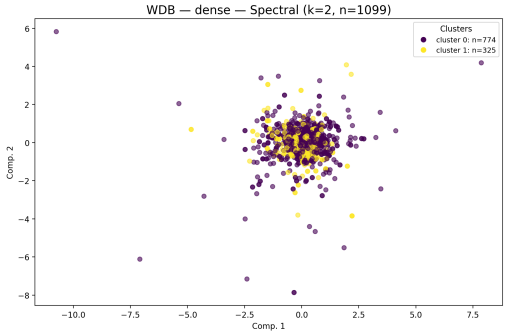


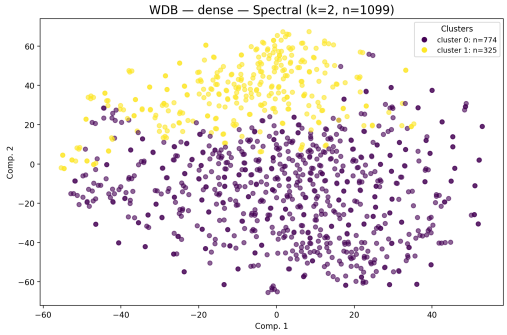
Figura 5.44: Comparación de los 3 espectros representativos ($k = 3$) sobre la capa `flatten`, superpuestos al espectro promedio de la clase WDB.

5.13.4. Capa `dense`

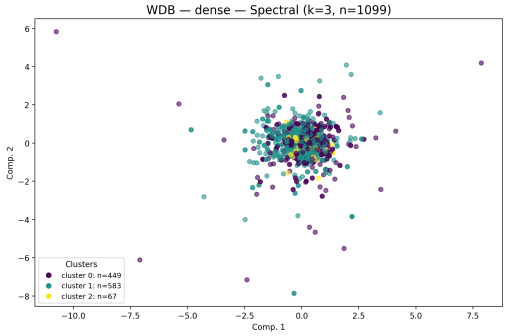
Se evalúan $k = 2$, valor indicado por el eigengap, y $k = 3$ como valor contiguo superior. No se examina $k = 1$ por carecer de sentido como partición. La Figura 5.45 presenta ambas configuraciones en las dos proyecciones.



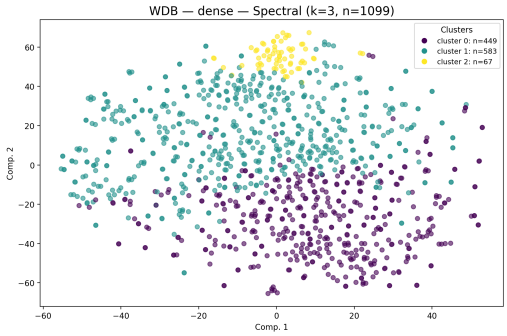
(a) $k = 2$, PCA



(b) $k = 2$, t-SNE



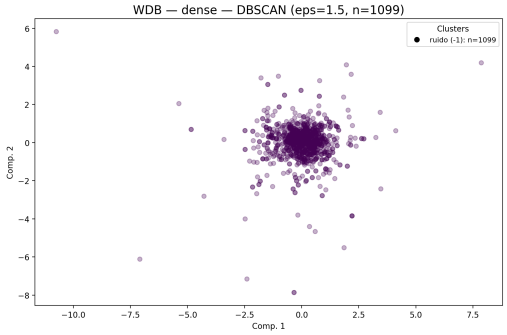
(c) $k = 3$, PCA



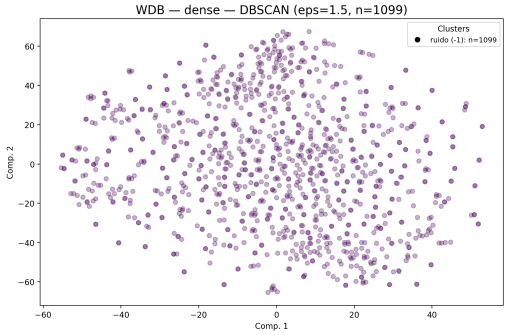
(d) $k = 3$, t-SNE

Figura 5.45: Particiones de Spectral Clustering sobre la capa `dense` de WDB ($n = 1{,}100$). Columna izquierda: proyección PCA. Columna derecha: t-SNE.

La Figura 5.46 muestra el resultado de DBSCAN sobre esta capa.



(a) Proyección PCA

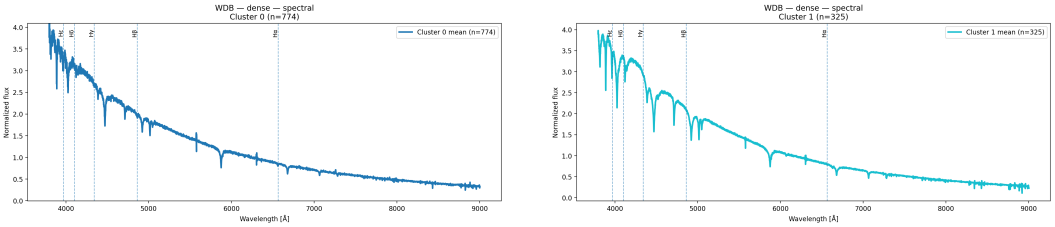


(b) Proyección t-SNE

Figura 5.46: Partición obtenida con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$) sobre la capa `dense` de WDB, en ambas proyecciones.

No se sintetizan espectros representativos para esta partición. Al no formarse ningún cluster, el único grupo existente es el del ruido, que reúne la totalidad de la muestra: su espectro promedio coincide por construcción con el promedio de la clase y no aporta información alguna sobre subestructura.

Los espectros representativos de la partición con $k = 2$ se presentan en las Figuras 5.47 y 5.48.



(a) Cluster 0 (b) Cluster 1

Figura 5.47: Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 2$) sobre la capa `dense` de WDB.

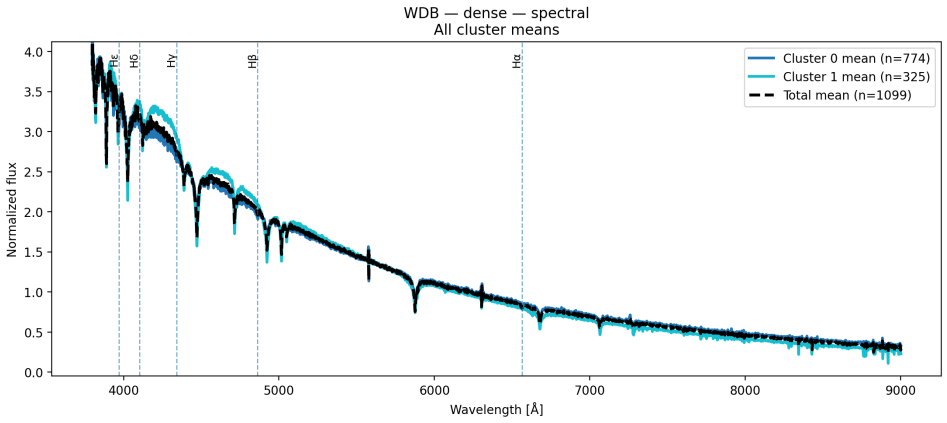
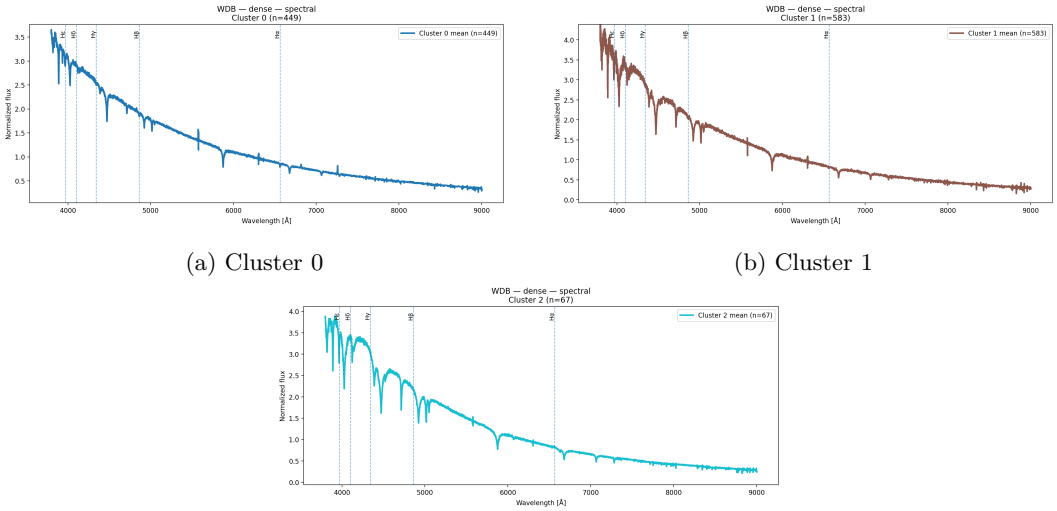


Figura 5.48: Comparación de los 2 espectros representativos ($k = 2$) sobre la capa `dense`, superpuestos al espectro promedio de la clase WDB.

Los espectros representativos de la partición con $k = 3$ se presentan en las Figuras 5.49 y 5.50.



(a) Cluster 0

(b) Cluster 1

(c) Cluster 2

Figura 5.49: Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 3$) sobre la capa `dense` de WDB.

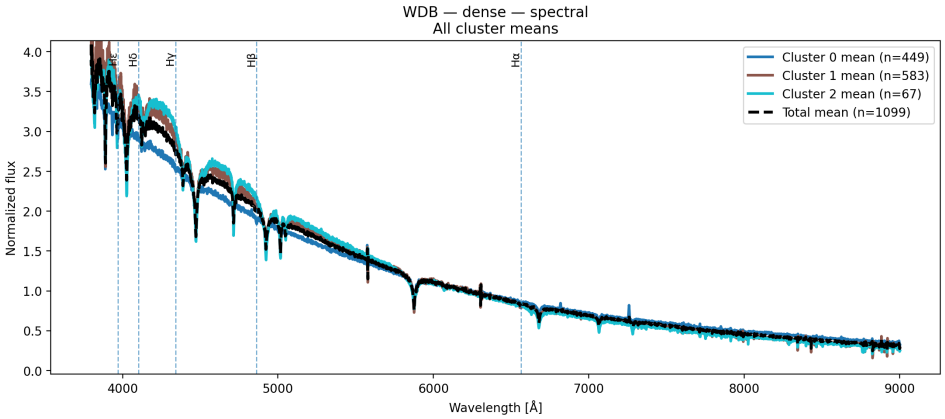
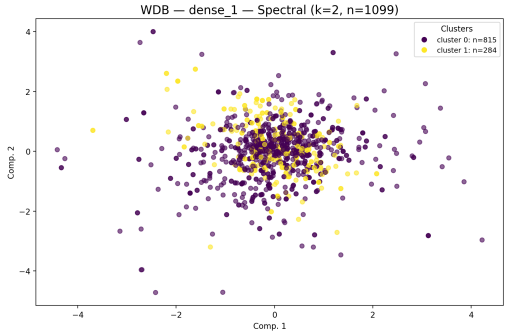


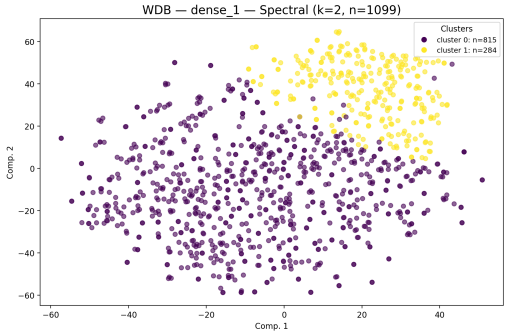
Figura 5.50: Comparación de los 3 espectros representativos ($k = 3$) sobre la capa `dense`, superpuestos al espectro promedio de la clase WDB.

5.13.5. Capa `dense_1`

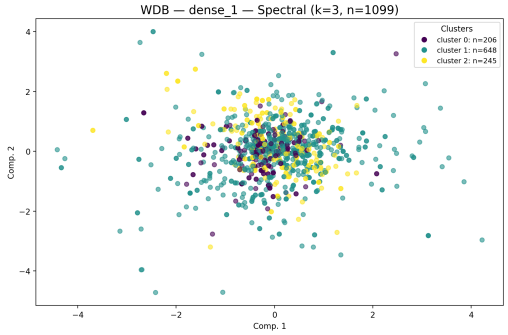
Se evalúan $k = 2$, valor indicado por el eigengap, y $k = 3$ como valor contiguo superior. No se examina $k = 1$ por carecer de sentido como partición. La Figura 5.51 presenta ambas configuraciones en las dos proyecciones.



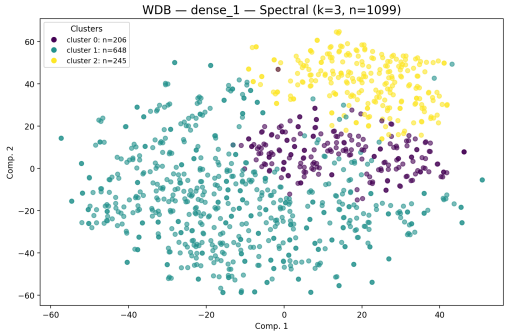
(a) $k = 2$, PCA



(b) $k = 2$, t-SNE



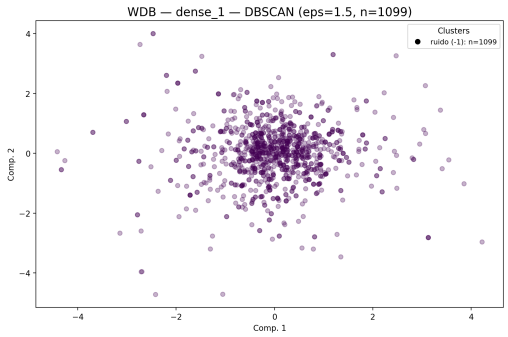
(c) $k = 3$, PCA



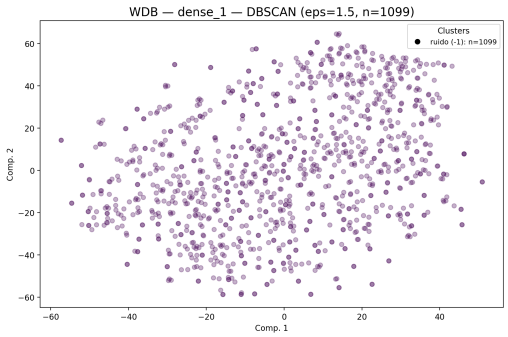
(d) $k = 3$, t-SNE

Figura 5.51: Particiones de Spectral Clustering sobre la capa `dense_1` de WDB ($n = 1{,}100$). Columna izquierda: proyección PCA. Columna derecha: t-SNE.

La Figura 5.52 muestra el resultado de DBSCAN sobre esta capa.



(a) Proyección PCA

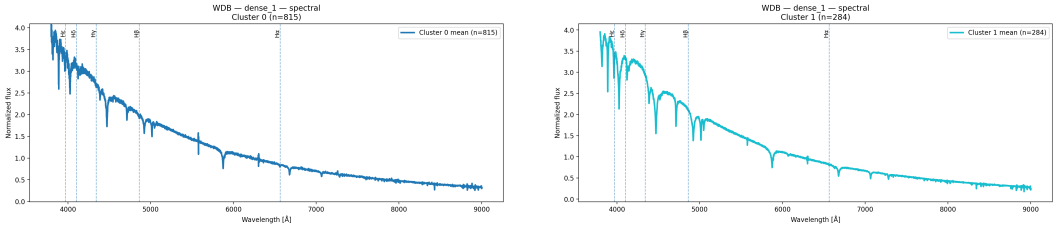


(b) Proyección t-SNE

Figura 5.52: Partición obtenida con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$) sobre la capa `dense_1` de WDB, en ambas proyecciones.

No se sintetizan espectros representativos para esta partición. Al no formarse ningún cluster, el único grupo existente es el del ruido, que reúne la totalidad de la muestra: su espectro promedio coincide por construcción con el promedio de la clase y no aporta información alguna sobre subestructura.

Los espectros representativos de la partición con $k = 2$ se presentan en las Figuras 5.53 y 5.54.



(a) Cluster 0 (b) Cluster 1

Figura 5.53: Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 2$) sobre la capa `dense_1` de WDB.

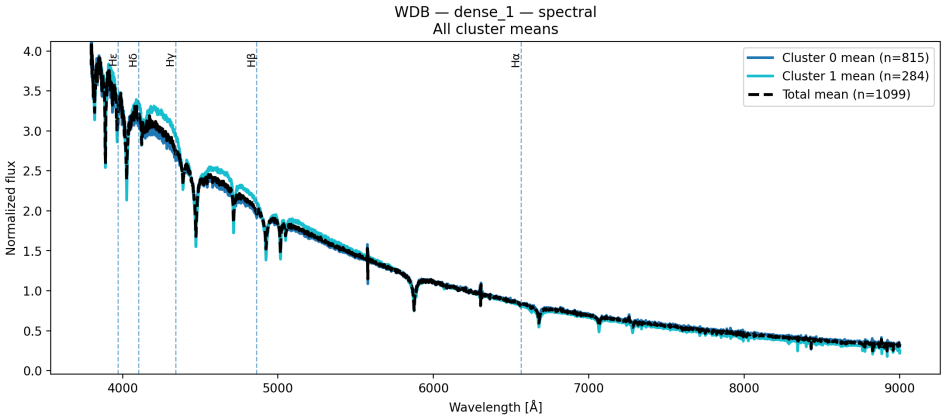
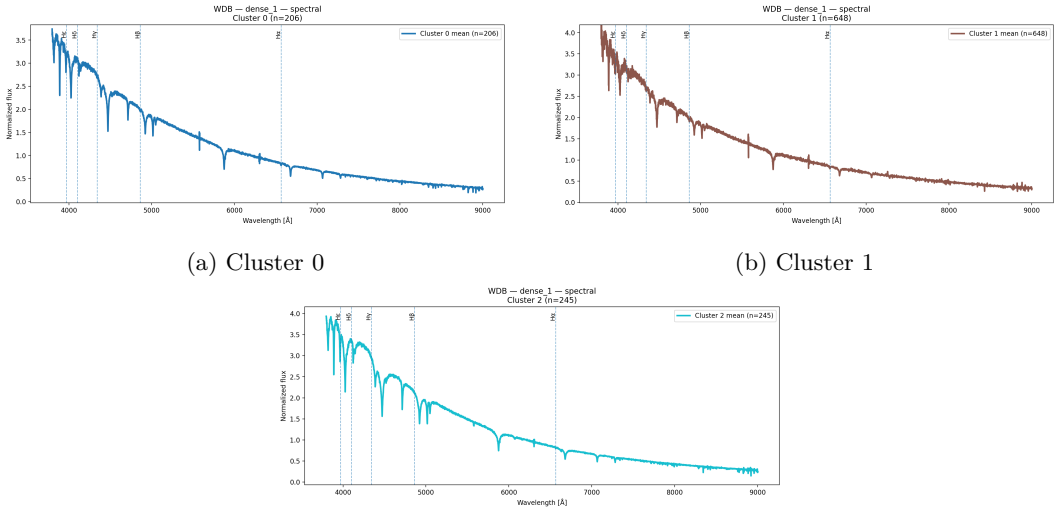


Figura 5.54: Comparación de los 2 espectros representativos ($k = 2$) sobre la capa `dense_1`, superpuestos al espectro promedio de la clase WDB.

Los espectros representativos de la partición con $k = 3$ se presentan en las Figuras 5.55 y 5.56.



(a) Cluster 0

(b) Cluster 1

(c) Cluster 2

Figura 5.55: Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 3$) sobre la capa `dense_1` de WDB.

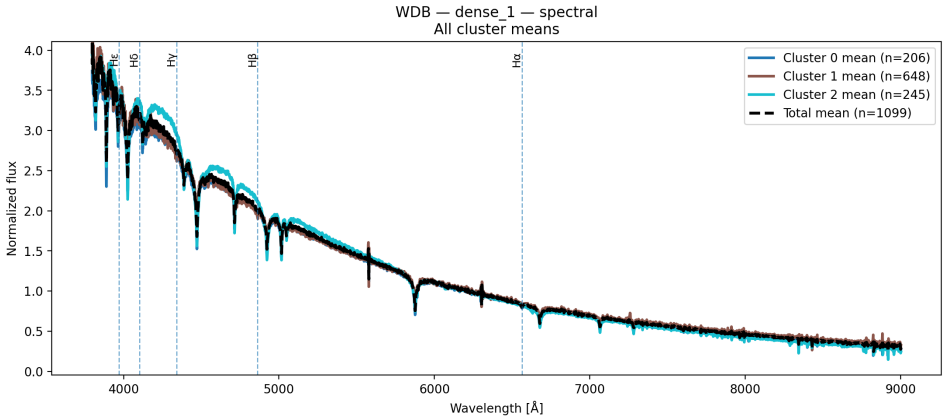
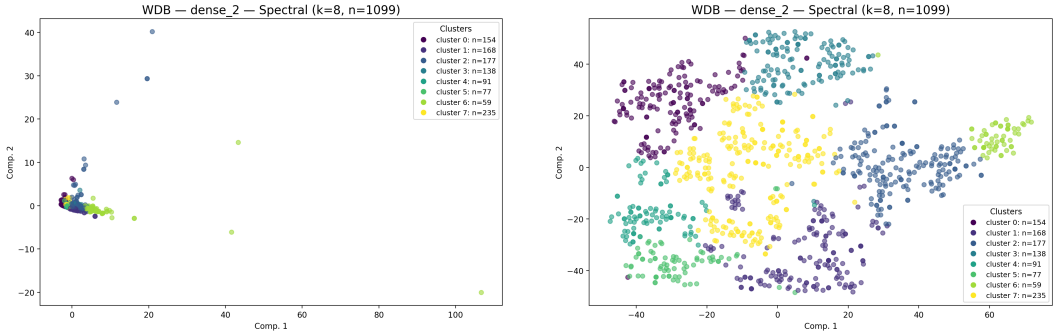


Figura 5.56: Comparación de los 3 espectros representativos ($k = 3$) sobre la capa `dense_1`, superpuestos al espectro promedio de la clase WDB.

5.13.6. Capa `dense_2`

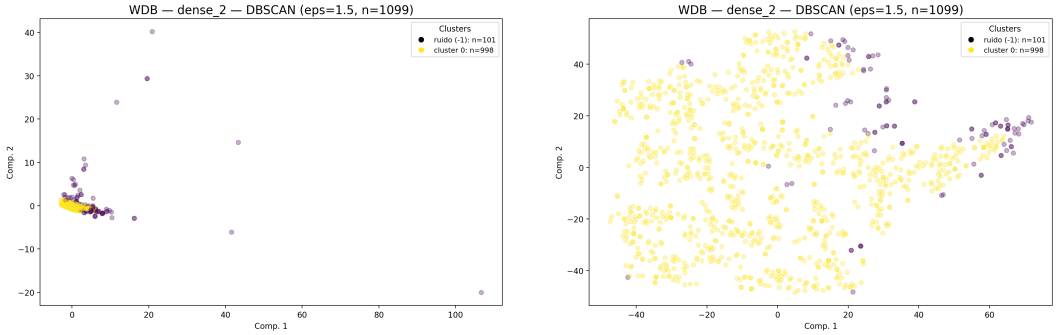
Se evalúa $k = 8$, valor indicado por el eigengap para esta capa. La Figura 5.57 presenta la partición en ambas proyecciones.



(a) $k = 8$, PCA (b) $k = 8$, t-SNE

Figura 5.57: Particiones de Spectral Clustering sobre la capa `dense_2` de WDB ($n = 1{,}100$). Columna izquierda: proyección PCA. Columna derecha: t-SNE.

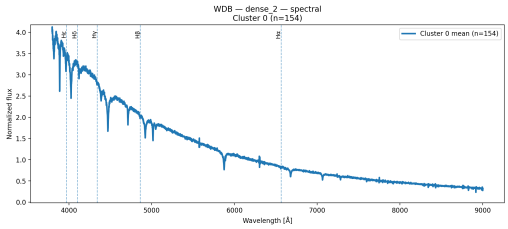
La Figura 5.58 muestra el resultado de DBSCAN sobre esta capa.



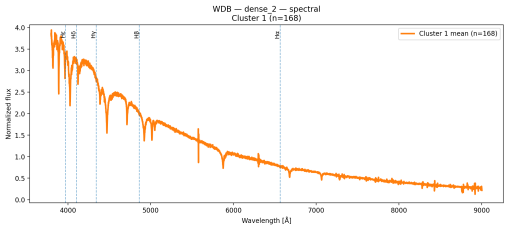
(a) Proyección PCA (b) Proyección t-SNE

Figura 5.58: Partición obtenida con DBSCAN ($\varepsilon = 1{,}5$ y `min_samples` $= 10$) sobre la capa `dense_2` de WDB, en ambas proyecciones.

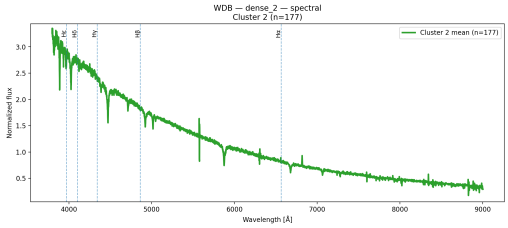
Los espectros representativos de la partición con $k = 8$ se presentan en las Figuras 5.59 y 5.60.



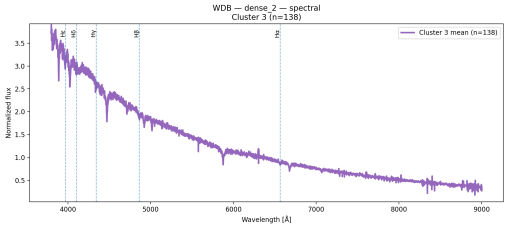
(a) Cluster 0



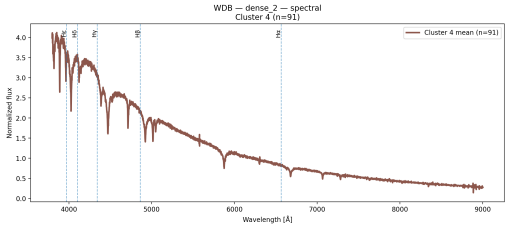
(b) Cluster 1



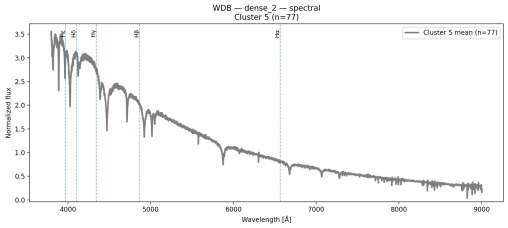
(c) Cluster 2



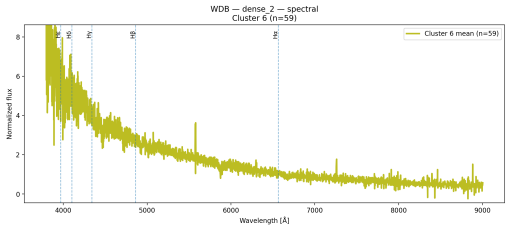
(d) Cluster 3



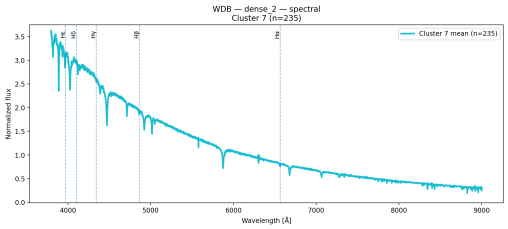
(e) Cluster 4



(f) Cluster 5



(g) Cluster 6



(h) Cluster 7

Figura 5.59: Espectro promedio de cada grupo obtenido con Spectral Clustering ($k = 8$) sobre la capa `dense_2` de WDB.

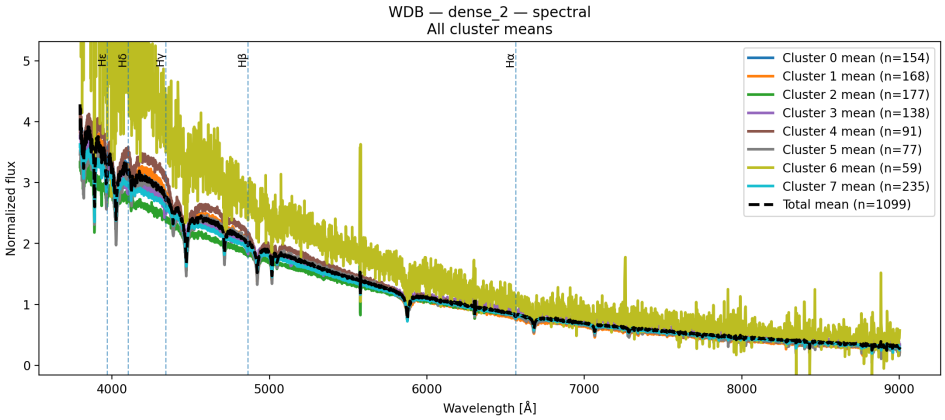
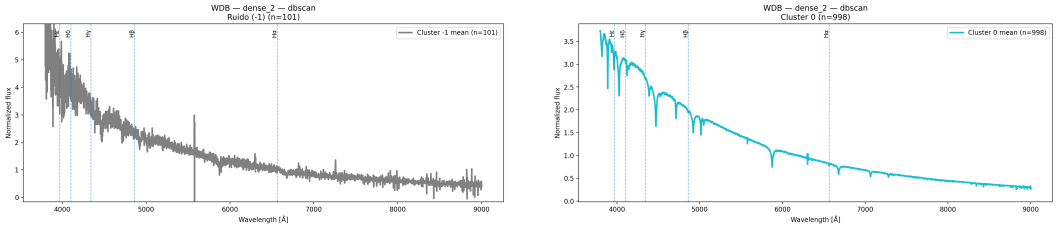


Figura 5.60: Comparación de los 8 espectros representativos ($k = 8$) sobre la capa `dense_2`, superpuestos al espectro promedio de la clase WDB.

A diferencia de las capas anteriores, aquí sí procede sintetizar los espectros de la partición de DBSCAN, dado que el algoritmo separa la muestra en dos grupos: el cluster principal y el conjunto de puntos relegados al ruido. Las Figuras 5.61 y 5.62 los recogen.



(a) Ruido (-1) (b) Cluster 0

Figura 5.61: Espectro promedio de cada grupo obtenido con DBSCAN sobre la capa `dense_2` de WDB.

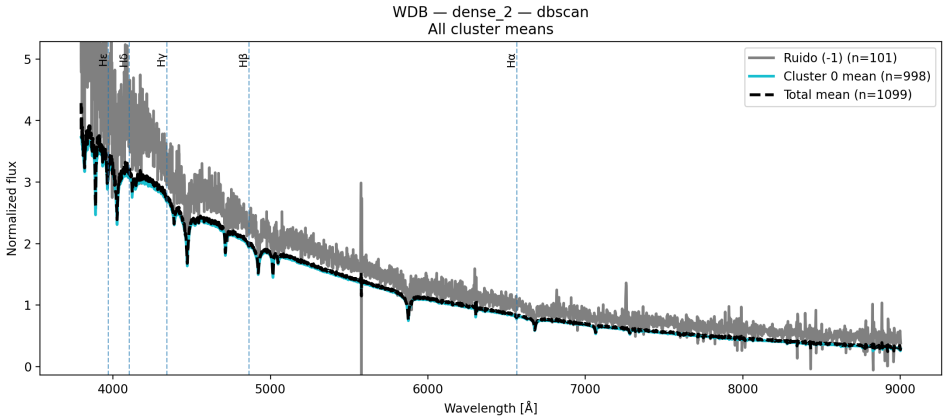


Figura 5.62: Comparación de los espectros de la partición de DBSCAN sobre la capa `dense_2`, superpuestos al espectro promedio de la clase WDB.

5.13.7. Comportamiento de DBSCAN sobre la clase WDB

Las figuras anteriores muestran que DBSCAN no produce una partición utilizable en ninguna de las cuatro capas, si bien el modo en que falla difiere entre ellas. Sobre `flatten`, `dense` y `dense_1` el algoritmo no forma ningún cluster: la totalidad de las 1,099 muestras válidas queda etiquetada como ruido, de modo que no existe partición alguna y las métricas internas carecen de definición. Sobre `dense_2` el comportamiento se invierte: el algoritmo reúne 998 muestras en un único cluster y relega solo 101 a la categoría de ruido, lo que tampoco constituye una partición, puesto que un grupo que contiene el 91 % de la muestra no separa nada.

Ambos desenlaces son manifestaciones del mismo problema. Con el radio fijado en $\varepsilon = 1{,}5$, ninguna vecindad de las capas superficiales alcanza los diez puntos que exige `min_samples` para constituir una región densa, mientras que en `dense_2`, donde la representación es más compacta y las distancias entre puntos sistemáticamente menores, ese mismo radio abarca prácticamente toda la nube. El algoritmo transita así de un extremo al otro sin pasar por un régimen intermedio en que llegue a delimitar más de un grupo.

Este resultado reproduce, de forma más notable, lo observado en WDA. Allí DBSCAN sí formaba grupos en dos de las cuatro capas, pero siempre reuniendo una fracción mínima de la muestra y descartando el resto como ruido, y el análisis de los espectros mostró que esos grupos se distinguían por temperatura efectiva antes que por composición. En WDB ni siquiera se alcanza ese régimen. La explicación es la misma que se desarrolló en la Sección 6.1.1: los espacios latentes presentan una densidad que varía de forma continua, sin los valles de baja densidad que esta familia de algoritmos necesita para delimitar regiones. Que el fenómeno se repita en una clase distinta, con un número de muestras

y una composición atmosférica diferentes, sugiere que se trata de una propiedad de las representaciones que produce la red y no de una particularidad de WDA.

En consecuencia, para WDB el análisis se apoya exclusivamente en Spectral Clustering, y los resultados de DBSCAN se reportan como constatación de que, con un radio fijado de antemano y común a todas las capas, el algoritmo no entrega información utilizable sobre esta clase. Ello no descarta a los métodos basados en densidad, sino que condiciona su empleo a una selección de ε adaptada a cada espacio latente, o bien a variantes que prescindan de un umbral global, según se detalla en el Capítulo 7.

5.13.8. Métricas internas de las particiones

La Tabla 5.10 reúne las métricas internas de las configuraciones evaluadas sobre WDB. El orden que produce coincide con el observado en WDA: la partición de `dense_2` con $k = 8$ obtiene el mejor valor en las tres métricas, mientras que las tres capas más superficiales quedan comprendidas en una franja estrecha y sin jerarquía clara entre ellas, con Silhouette próximo a cero o negativo.

Tabla 5.10: Métricas internas de las particiones de Spectral Clustering evaluadas sobre la clase WDB (n = 1,100). Silhouette y Calinski–Harabasz: valores más altos indican mejor partición. Davies–Bouldin: valores más bajos.

Capa	k	**Silhouette**	**Davies–Bouldin**	**Calinski–Harabasz**
`flatten`	2	0.0241	4.602	24.81
`flatten`	3	0.0081	4.016	20.00
`dense`	2	0.0080	4.674	25.27
`dense`	3	−0,0482	4.746	19.04
`dense_1`	2	−0,0278	4.086	27.35
`dense_1`	3	−0,0564	4.083	24.10
`dense_2`	8	**0.0505**	**1.3963**	**57.23**

La ventaja de `dense_2` no se reparte por igual entre las tres métricas. En Silhouette alcanza 0,0505 frente al 0,0241 de la mejor configuración restante, poco más del doble, y en Calinski–Harabasz 57,23 frente a 27,35, en una proporción semejante. Es el índice de Davies–Bouldin el que expresa la diferencia con nitidez: 1,3963 frente a un mínimo de 4,016 en el resto de la tabla, es decir, casi tres veces mejor y el único valor de la clase por debajo de 4. La distancia entre `dense_2` y las capas superficiales es, por tanto, real, pero solo una de las tres métricas la refleja de forma inequívoca.

Conviene situar además estas cifras en su escala. Ninguna configuración de WDB alcanza un Silhouette de 0,06, y el mejor valor de la clase queda por debajo del que obtenían en WDA capas tan poco favorables como `flatten` o `dense_1` con $k = 2$ (0,0668 y 0,0831

respectivamente). Juzgada únicamente por sus métricas internas, la clase WDB carecería de estructura de grupos apreciable en cualquiera de las cuatro capas.

Ese juicio, sin embargo, sería equivocado, y la clase WDB ofrece la demostración más clara de ello en todo el trabajo. El hallazgo principal, la separación de los espectros según la presencia e intensidad de las líneas de hidrógeno que desarrolla la Sección 6.2.1, procede de `flatten` con $k = 3$: una configuración cuyo Silhouette es 0,0081, indistinguible de cero, y cuyo Calinski–Harabasz de 20,00 es el segundo más bajo de la tabla completa. El segundo hallazgo, el rasgo en emisión próximo a los 7,300 Å que describe la Sección 6.2.2, procede de `dense`, la capa peor puntuada del conjunto. La configuración mejor puntuada, `dense_2` con $k = 8$, no dio lugar a ninguna interpretación física.

Un detalle de esa comparación tiene consecuencias metodológicas directas. Sobre `flatten`, tanto el eigengap como el Silhouette y el Calinski–Harabasz prefieren $k = 2$ antes que $k = 3$, y solo el índice de Davies–Bouldin invierte esa preferencia (4,016 frente a 4,602). De haberse seguido cualquiera de los dos primeros criterios de forma estricta, la partición con $k = 3$ no se habría examinado y el resultado principal del trabajo no se habría encontrado. Ello respalda la práctica adoptada en la Sección 5.9, consistente en no limitar la síntesis de espectros al valor de k seleccionado sino extenderla a sus valores contiguos.

Entre las capas superficiales las métricas tampoco concuerdan entre sí. El Calinski–Harabasz sitúa en primer lugar a `dense_1` con $k = 2$ (27,35), configuración cuyo Silhouette es negativo ($-0{,}0278$) y que este último índice relega al sexto lugar de siete. Un desacuerdo de esa magnitud entre dos medidas calculadas sobre la misma partición confirma que ninguna de ellas puede emplearse aisladamente como criterio de selección.

Cabe por último comparar ambas clases sobre la configuración que comparten, `dense_2` con $k = 8$. WDA obtiene mejor Silhouette (0,1382 frente a 0,0505) y mejor Calinski–Harabasz (84,41 frente a 57,23), pero WDB obtiene el mejor Davies–Bouldin de las dos (1,3963 frente a 1,5860). La diferencia entre clases no es por tanto uniforme: en WDB los grupos aparecen tan bien separados como en WDA en relación con su dispersión interna, y lo que decae es la cohesión de cada muestra con el grupo al que se le asigna. Esa diferencia no puede atribuirse al tamaño muestral, dado que ambos experimentos se ejecutaron sobre $n = 1{,}100$ espectros. Admite al menos dos lecturas que el presente análisis no permite separar: que la subestructura interna de WDB sea efectivamente menos marcada, o que la representación de la red esté mejor adaptada a WDA, posibilidad verosímil si el conjunto con que fue entrenada refleja la distribución de la Tabla 5.3, en la que WDA supera en un orden de magnitud a WDB.

5.14. Cierre del capítulo

Este capítulo construyó el pipeline sobre la clase WDA y lo aplicó a la clase WDB. Quedaron fijados el criterio de selección de k para cada algoritmo, la configuración de

DBSCAN y el conjunto de métricas con que se evalúa cada partición, y se obtuvieron para ambas clases las particiones de las cuatro capas latentes junto con sus espectros representativos.

De la evaluación realizada se desprenden dos conclusiones de orden metodológico. La primera es que las métricas internas ordenan las cuatro capas de forma inequívoca y coincidente en ambas clases, situando a `dense_2` por delante del resto, pero que ese orden no basta para decidir qué partición merece ser examinada: los valores obtenidos son bajos en términos absolutos, los tres índices no concuerdan entre sí y ninguno informa sobre la naturaleza de lo que separan. La segunda es que el análisis de migración entre particiones con distinto k resultó el criterio más discriminante de los empleados, por permitir distinguir los grupos que persisten de los que se reorganizan por completo.

Establecer si las particiones obtenidas corresponden a propiedades físicas de las estrellas excede lo que estos criterios permiten decidir, y requiere el examen de los espectros representativos sintetizados para cada grupo. A esa tarea se dedica el Capítulo 6.

Capítulo 6

RESULTADOS E INTERPRETACIÓN ASTROFÍSICA

El Capítulo 5 construyó el pipeline, fijó sus hiperparámetros y evaluó las particiones que produce sobre las clases WDA y WDB mediante métricas internas y matrices de migración. Ese análisis estableció qué configuraciones resultan estables y cuáles se reorganizan ante cualquier cambio, pero dejó abierta la pregunta que motiva el trabajo: si los grupos obtenidos corresponden o no a propiedades físicas de las estrellas que los componen.

Este capítulo aborda esa pregunta a partir de los espectros representativos sintetizados para cada grupo, contrastando las particiones con la fenomenología conocida de las enanas blancas. La exposición sigue el orden de las dos clases y culmina en una discusión conjunta de los resultados metodológicos y astrofísicos obtenidos.

6.1. Interpretación física de las particiones de la clase WDA

6.1.1. Particiones de DBSCAN: separación por temperatura efectiva

Las particiones obtenidas con DBSCAN se sometieron a revisión desde la física estelar, a partir de los espectros representativos de cada grupo. La conclusión es uniforme para las cuatro capas en WDA: los grupos que el algoritmo de densidad separa no se distinguen por su composición atmosférica ni por la presencia de líneas particulares, sino por la **temperatura efectiva** de las estrellas que los componen.

La evidencia está en el nivel del continuo. Como se describió en la Sección 5.4, cada

espectro se normaliza dividiéndolo por su flujo en una longitud de onda fija próxima a los 6,158 Å. Tras esa operación, el valor que toma el espectro en el extremo azul deja de ser una medida de brillo absoluto y pasa a expresar el cociente entre el flujo azul y el rojo, es decir, la pendiente del continuo. La Tabla 6.1 recoge ese valor para el espectro promedio de cada grupo.

Tabla 6.1: Flujo normalizado en el extremo azul del espectro promedio de cada grupo obtenido con DBSCAN sobre la clase WDA ($n = 1{,}100$). El valor equivale a la razón entre el flujo azul y el flujo en la longitud de onda de normalización, y constituye por tanto una medida de la pendiente del continuo.

Capa	**Ruido** (-1)	**Cluster 0**	**Cluster 1**	**Cluster 2**
`flatten`	2.84	2.17	—	—
`dense`	2.69	1.75	1.98	3.19
`dense_1`	2.67	2.07	5.51	—
`dense_2`	3.36	2.57	—	—

Las diferencias son sistemáticas y de magnitud considerable: en `dense` los valores recorren un rango que va de 1,75 a 3,19, y en `dense_1` el contraste entre los dos grupos alcanza un factor cercano a 2,7. Un espectro que, normalizado en el mismo punto, presenta más flujo en el azul posee un continuo más inclinado hacia longitudes de onda cortas, lo que en una atmósfera estelar corresponde a una temperatura efectiva mayor. Ordenar los grupos por este valor equivale, en consecuencia, a ordenarlos por temperatura.

En cuanto a las líneas de la serie de Balmer, no se apreciaron diferencias relevantes entre los grupos. Esto resulta esperable, dado que la totalidad de la muestra pertenece a la clase WDA, de modo que todos los espectros presentan atmósferas dominadas por hidrógeno y comparten el mismo conjunto de líneas.

Este hallazgo tiene una consecuencia doble para los objetivos del trabajo. Por una parte constituye un resultado negativo respecto de la búsqueda de subclases: la temperatura efectiva es el eje de variación mejor conocido y más estudiado dentro de las enanas blancas de tipo DA, de modo que recuperarlo no supone identificar una subpoblación nueva sino redescubrir un parámetro ya establecido. Por otra, constituye una validación del pipeline, puesto que demuestra que las particiones obtenidas sobre el espacio latente responden a una propiedad física real y no a artefactos numéricos.

Conviene, no obstante, no agotar el resultado en esa doble lectura, porque el modo en que se obtiene tiene interés propio. Ordenar un conjunto de espectros de menor a mayor temperatura a partir de la inclinación de su continuo es una operación que el astrónomo realiza de forma rutinaria por inspección visual, y que solo requiere del ajuste de atmósferas modelo cuando se busca un valor numérico y no un ordenamiento. El procedimiento aquí

descrito llega a ese mismo ordenamiento sin disponer de ninguno de los elementos en que se apoya el ojo experto: no recibe valores de temperatura, no se le indica que la pendiente del continuo constituya una magnitud relevante y ni siquiera se le señala qué región del espectro debe observar. La red fue entrenada para discriminar entre doce categorías espectrales, y el clustering opera sobre sus representaciones internas sin acceso a etiqueta alguna. Que de ese procedimiento emerja una partición ordenada por temperatura implica que la propiedad quedó codificada en el espacio latente de una forma que un algoritmo de densidad puede recuperar por sí solo, pese a no habérsele pedido en ningún momento.

Cabe delimitar el alcance de la afirmación. El procedimiento no mide la temperatura ni la asigna a cada objeto: agrupa espectros que, examinados a posteriori, resultan estar ordenados por ella, y esa lectura la aporta el análisis físico posterior y no el algoritmo. La correspondencia entre el flujo del extremo azul y la temperatura efectiva es, además, cualitativa mientras no se contraste con los valores publicados para estos objetos. Con esas reservas, el hallazgo anticipa en WDA lo que la Sección 6.2.1 muestra de forma más nítida sobre la clase WDB, donde la partición reproduce sin supervisión una distinción taxonómica que también se establece por inspección visual del espectro.

6.1.2. Capa `flatten`: la temperatura como único eje de separación

La partición de Spectral Clustering con $k = 3$ sobre `flatten` reparte la muestra en grupos de 908, 140 y 52 espectros que se ordenan de forma nítida por la pendiente del continuo. El grupo 0, que reúne el 83 % de la muestra, presenta un espectro promedio prácticamente indistinguible del promedio de la clase, el grupo 2 exhibe una pendiente marcadamente mayor, correspondiente a las estrellas más calientes, y el grupo 1 muestra la pendiente más baja, propia de objetos de menor temperatura efectiva. La partición equivale, en la práctica, a segmentar la clase en tres rangos de temperatura: uno central que domina la muestra y dos laterales.

El contraste con DBSCAN sobre esta misma capa resulta instructivo. Ambos algoritmos separan por la misma propiedad física, pero lo hacen con distinta eficacia: DBSCAN se limita a aislar un núcleo denso de 310 espectros y relega 790 a la categoría de ruido, mientras que Spectral Clustering particiona la totalidad de la muestra en tres rangos bien delimitados. La lectura conjunta indica que, en esta capa, la representación conserva esencialmente la forma global del continuo, cuya variación está dominada por la temperatura efectiva. Dicho de otro modo: `flatten` no aporta un eje de variación adicional al ya conocido.

6.1.3. Capa `dense`: anchura de las líneas y gravedad superficial

Los resultados sobre `dense` son los más informativos de la exploración, porque en esta capa los grupos dejan de distinguirse únicamente por el continuo. El examen de los espectros representativos de la partición de DBSCAN, que forma tres grupos de 39, 22 y 9 espectros,

revela una segunda diferencia sistemática: la **anchura de las líneas de absorción**, y en particular la de Hα. El grupo 0 presenta una línea sensiblemente más estrecha y profunda que la de los grupos 1 y 2, y entre estos dos últimos se aprecia a su vez una diferencia del mismo tipo.

Esta observación es relevante porque la anchura de las líneas de Balmer en una enana blanca no es un rasgo arbitrario, sino un indicador de la gravedad superficial. El ensanchamiento de estas líneas está dominado por el *efecto Stark*, esto es, por la perturbación que los campos eléctricos de las partículas cargadas circundantes ejercen sobre los niveles energéticos del hidrógeno. Una gravedad superficial mayor se traduce en líneas más anchas, de modo que la anchura de Hα constituye una medida indirecta de $\log g$, parámetro que los catálogos publican de forma rutinaria para las enanas blancas del SDSS.

En consecuencia, la partición de DBSCAN sobre `dense` estaría separando los espectros a lo largo de un eje físico distinto del observado en `flatten`: no solo por temperatura efectiva, sino también por gravedad superficial. Ello sugiere que, a esta profundidad, la red ha combinado la información del continuo con la del perfil de las líneas, lo que constituye una representación más rica que la de las capas superficiales.

6.1.4. El grupo frío de `dense` con $k = 5$

La partición de Spectral Clustering con $k = 5$ sobre `dense` aísla un grupo, el cluster 2, de 90 espectros cuyo espectro promedio arranca en un flujo normalizado de 1,34, frente al valor de 2,65 del promedio de la clase. Se trata del nivel más bajo de los cinco grupos y, con diferencia, del conjunto más frío de la partición. Es además el grupo que las matrices de migración de la Figura 5.18 señalan como el más estable de toda la exploración, dado que se conserva sin pérdida alguna al variar el número de clusters, lo que refuerza el interés de examinarlo en detalle.

Su espectro promedio presenta tres rasgos que lo distinguen del resto.

- **Asimetría en Hα.** La línea de absorción no es simétrica respecto de su centro, a diferencia de lo que se observa en los demás grupos.
- **Alta dispersión interna.** El promedio exhibe un aspecto ruidoso que no procede del ruido instrumental de un espectro individual, sino de la heterogeneidad del grupo.
- **Rasgos no atribuibles al hidrógeno.** Se aprecian dos características intensas que no corresponden a ninguna transición de la serie de Balmer: una en torno a los 6,000 Å en emisión y otra algo por encima de los 5,000 Å en absorción. Que ambas sobrevivan al promediado de 90 espectros implica que están presentes en una fracción sustantiva de ellos, y no en unos pocos objetos aislados.

Ambos rasgos se verificaron sobre los espectros individuales del grupo y no únicamente sobre su promedio, dado que una característica visible en el espectro medio podría proceder

de unos pocos objetos de gran amplitud. La inspección confirma que varios espectros presentan de forma individual la característica en emisión próxima a los 6,000 Å, de modo que no se trata de un efecto del promediado.

Queda pendiente, sin embargo, determinar si el origen de esa característica es astrofísico o instrumental.

La característica en absorción situada algo por encima de los 5,000 Å se observa asimismo en espectros individuales del grupo, y es la que concentra el mayor interés. Su identificación excede el alcance de este trabajo, pero una línea de absorción ajena a la serie de Balmer, presente en un conjunto de espectros que resulta además ser el más estable de toda la exploración, constituye precisamente el tipo de indicio que motiva la búsqueda de subclases. Su caracterización, junto con la medición precisa de la longitud de onda de ambos rasgos y con la explicación de la asimetría observada en Hα, se plantea como línea de trabajo futuro y requiere un análisis espectroscópico especializado.

6.2. Interpretación física de las particiones de la clase WDB

6.2.1. Capa `flatten`: separación por presencia de hidrógeno

La partición de Spectral Clustering con $k = 3$ sobre `flatten` reparte la muestra en grupos de 752, 65 y 282 espectros. A diferencia de lo observado en WDA, aquí los tres grupos presentan pendientes de continuo semejantes, de modo que la temperatura efectiva no es el criterio que los separa. Lo que los distingue es la **presencia e intensidad de las líneas de absorción del hidrógeno**:

- el grupo 1, el más reducido con 65 espectros, exhibe líneas de Balmer marcadas en todas las transiciones de la serie.
- el grupo 0, que reúne la mayor parte de la muestra, muestra las mismas líneas pero de forma muy sutil, apenas perceptibles aunque presentes.
- el grupo 2 no presenta absorción alguna en Hα, Hβ ni Hγ.

Este resultado requiere una precisión previa. La clase WDB agrupa enanas blancas cuya atmósfera está dominada por helio, de modo que cabría esperar la ausencia total de rasgos de hidrógeno. Sin embargo, existe consenso en la comunidad astronómica en que todas las enanas blancas contienen hidrógeno en alguna proporción, por reducida que sea, y aquellas de tipo DB en que ese hidrógeno resulta detectable reciben la designación DBA.

La relevancia del hallazgo reside en que la partición reproduce esa distinción sin haber recibido información alguna sobre ella. Todos los espectros de la muestra comparten la misma etiqueta, WDB, y el pipeline no dispone en ningún punto de datos sobre el contenido de hidrógeno: opera únicamente sobre las representaciones latentes del flujo. Que

el agrupamiento resultante ordene los espectros según la intensidad de sus líneas de Balmer significa que esa propiedad está codificada en el espacio latente y que el algoritmo la ha recuperado de forma no supervisada.

Su interés práctico es igualmente destacable. La identificación de objetos DBA se realiza habitualmente mediante inspección visual: un especialista examina cada espectro y determina si las líneas de Balmer están presentes. Un procedimiento capaz de segmentar una muestra según ese criterio, sin supervisión y sobre miles de espectros, constituye una vía de automatización de esa tarea. Conviene subrayar que este resultado no se buscó de forma deliberada, sino que emergió del análisis de los espectros representativos, lo que ilustra el valor de la etapa de síntesis espectral dentro del pipeline propuesto.

A diferencia de la separación por temperatura observada en WDA, que recupera un eje de variación ya conocido y continuo, esta separación corresponde a una distinción taxonómica establecida.

6.2.2. Capa `dense`: un rasgo en emisión próximo a los 7.300 Å

La partición con $k = 3$ sobre `dense` divide la muestra en grupos de 449, 583 y 67 espectros. El espectro representativo del grupo 0 presenta una característica que no aparece en ningún otro grupo de esta partición ni en las de las restantes capas: un **rasgo en emisión en torno a los** 7,300 **Å**.

Tres circunstancias hacen que esta observación merezca atención. La primera es que el rasgo no presenta el aspecto irregular propio del ruido, sino un perfil definido. La segunda es que el grupo reúne 449 espectros, de modo que una característica que sobrevive al promediado de ese número de objetos debe estar presente en una fracción considerable de ellos. La tercera es su especificidad: no se observa en ninguna otra combinación de capa y partición de las examinadas.

Para descartar que la característica fuese un producto del promediado, se examinaron individualmente los espectros del grupo. La revisión confirma que el rasgo aparece por separado en varios de ellos, y no únicamente en la media del conjunto, de modo que no procede de unos pocos objetos de amplitud desmedida. Que un grupo que reúne el 41 % de la muestra comparta una característica ausente en todas las demás particiones constituye, por sí solo, un resultado que merece destacarse.

La identificación del rasgo excede el alcance de este trabajo, que no contempla el ajuste de modelos de atmósfera ni el análisis de abundancias químicas. Queda pendiente, en particular, establecer si su origen es intrínseco a los objetos o si procede de la reducción de los datos, dado que la región del espectro en que aparece es susceptible de residuos instrumentales y atmosféricos.

6.2.3. Profundidad de la representación y naturaleza de los rasgos detectados

Los dos resultados anteriores admiten una lectura conjunta relativa al papel que desempeña la profundidad de la capa. La separación observada en `flatten` descansa sobre un rasgo directamente perceptible: la presencia o ausencia de líneas de absorción es lo primero que un observador humano advierte al examinar un espectro, y `flatten` no es sino la salida aplanada de los bloques convolucionales, esto es, un conjunto de descriptores morfológicos locales próximos a esa lectura visual.

La separación observada en `dense` es de otra naturaleza. Aislar un grupo definido por un rasgo estrecho situado en una región concreta del espectro requiere combinar los descriptores locales de la capa anterior mediante las transformaciones afines y no lineales que introduce la capa densa. No basta con detectar la presencia de una estructura: hay que construir una combinación de características que resulte sensible a ella y no a las demás. En este sentido, la capa `dense` opera sobre lo perceptible de forma inmediata, sino sobre relaciones derivadas de ello.

Esta observación proporciona un argumento adicional a favor de evaluar múltiples capas en lugar de escoger una sola a priori, tal como se planteó en el Capítulo 3 al discutir el sesgo funcional. Capas de distinta profundidad no ofrecen versiones mejores o peores de la misma información, sino tipos distintos de información: las superficiales conservan la morfología directa del espectro, mientras que las profundas construyen descriptores derivados capaces de aislar rasgos que ninguna característica local expresa por sí sola.

6.3. Discusión

El resultado metodológico central del trabajo concierne al papel que desempeña la profundidad de la capa. Juzgadas por sus métricas internas, las cuatro capas se ordenan de forma inequívoca y coincidente en ambas clases: `dense_2`, la última representación latente antes de la salida de clasificación, produce las particiones más cohesivas y mejor separadas, con el mejor valor de Silhouette, Davies–Bouldin y Calinski–Harabasz tanto en WDA como en WDB. Ese ordenamiento es coherente con el sesgo funcional descrito en el Capítulo 3: al estar orientada a reforzar las fronteras entre las doce clases con que la red fue entrenada, `dense_2` organiza el espacio en regiones compactas y separables.

El ordenamiento se invierte, sin embargo, en cuanto el criterio deja de ser métrico y pasa a ser el contenido físico de las particiones. Los hallazgos interpretables del trabajo proceden todos de capas peor puntuadas: la separación por temperatura efectiva, de `flatten`, la separación por gravedad superficial y el grupo frío de 90 espectros de `dense`, y la separación de los objetos según la presencia de líneas de hidrógeno, de `flatten` sobre WDB, en una configuración cuyo Silhouette es indistinguible de cero. La capa mejor puntuada en ambas

clases no dio lugar a ninguna lectura física.

De ambas observaciones se sigue que la profundidad de la capa no determina la calidad de la representación sino el tipo de estructura que puede recuperarse de ella. Se sugiere que, las capas superficiales conservan la morfología directa del espectro y separan por rasgos que un observador humano advierte de inmediato, las intermedias construyen descriptores derivados capaces de aislar características que ninguna propiedad local expresa por sí sola, y la más profunda entrega una partición fina y bien delimitada, pero organizada en torno a las fronteras de la tarea con que la red fue entrenada antes que a la fenomenología de los espectros. Evaluar las cuatro capas, en lugar de escoger una a priori, constituye por tanto una decisión metodológica y no un análisis redundante.

En cuanto a los algoritmos, Spectral Clustering resultó más consistente que DBSCAN sobre estos espacios latentes. DBSCAN mostró una fuerte sensibilidad a su parámetro de densidad: con la misma configuración generó desde la totalidad de los puntos como ruido, en tres de las cuatro capas de WDB, hasta agrupar el 91 % de la muestra en un solo cluster en la capa restante. El barrido de ε presentado en la Sección 5.9 (Tabla 5.7) permite atribuir ese comportamiento a la geometría de los embeddings: su densidad varía de forma continua, sin valles pronunciados de baja densidad, de modo que un umbral global resulta poco discriminante y el resultado depende críticamente del valor escogido. Con un radio fijo, en consecuencia, los resultados de DBSCAN son poco informativos.

Ello no descarta a la familia de métodos basados en densidad, sino que precisa la condición bajo la cual podría resultar productiva: una selección adaptativa de ε, ajustada a la escala de distancias propia de cada espacio latente, o bien métodos que no dependan de un umbral global único, como HDBSCAN. Conviene añadir que la preferencia por Spectral Clustering queda sustentada frente a K-Means y DBSCAN, que son los algoritmos efectivamente evaluados sobre estos espacios latentes, y no frente al conjunto completo de alternativas identificadas en el Capítulo 4. Establecer si alguna de las familias restantes ofrece un mejor desempeño requiere ejecutarlas bajo las mismas condiciones, lo que se plantea como trabajo futuro.

El análisis de migración entre particiones constituye, en este contexto, el criterio más informativo de los empleados. A diferencia de las métricas internas, que resumen la calidad de una partición en un único número, permite seguir el destino de cada grupo al variar el número de clusters y distinguir así los agrupamientos que persisten de los que se reorganizan por completo. A él se debe la identificación del grupo de 90 espectros de la capa `dense`, que se conserva sin pérdida alguna ante el cambio de k y que resultó ser, además, el que ofreció la lectura espectral más rica de la clase WDA.

En conjunto, el marco desarrollado sobre la clase WDA se aplicó sin modificaciones a WDB, lo que verifica su reutilización, y produjo en esta segunda clase el resultado más próximo al objetivo declarado del trabajo. Determinar si los subgrupos restantes

corresponden a subclases no descritas exige el contraste con parámetros astrofísicos externos y el análisis espectroscópico de los rasgos señalados, líneas que detalla el Capítulo 7.

Capítulo 7

CONCLUSIONES

7.1. Resumen

Este trabajo construyó un marco metodológico de aprendizaje no supervisado para la búsqueda de subclases de enanas blancas a partir de sus espectros. El pipeline implementado abarca el preprocesamiento de espectros del SDSS, la extracción de representaciones latentes desde cuatro capas de una red neuronal convolucional preentrenada, la reducción de dimensionalidad mediante PCA a un espacio común de 32 dimensiones, la aplicación comparativa de K-Means, Spectral Clustering y DBSCAN, la evaluación de las particiones mediante métricas internas (Silhouette, Davies–Bouldin, Calinski–Harabasz y, para los métodos de densidad, DBCV) y el análisis de migración entre configuraciones con distinto número de clusters. La etapa final del procedimiento es la síntesis de espectros representativos por grupo, que permite contrastar las particiones obtenidas con la fenomenología astrofísica mediante las líneas de las series de Balmer y de helio. El marco se desarrolló sobre la clase WDA y se aplicó sin modificaciones a la clase WDB.

7.2. Conclusiones

Los resultados obtenidos permiten extraer cinco conclusiones principales.

Primero, la profundidad de la capa no determina la calidad de la representación sino el tipo de estructura que puede recuperarse de ella. Esta es la conclusión metodológica central del trabajo, y matiza la expectativa inicial de que existiera una capa inequívocamente mejor. Se sugiere que, las capas superficiales conservan la morfología directa del espectro: la partición obtenida sobre `flatten` separa por rasgos que un observador humano advierte de inmediato, como la pendiente del continuo o la presencia de líneas de absorción. Las capas intermedias, en cambio, construyen descriptores derivados

mediante combinaciones afines y no lineales de los anteriores, y son las que aíslan rasgos que ninguna característica local expresa por sí sola. Capas de distinta profundidad ofrecen, por tanto, tipos distintos de información, lo que justifica evaluarlas todas en lugar de escoger una a priori.

Segundo, se identifica una fuerte sensibilidad de DBSCAN a su parámetro de densidad, que vuelve poco informativos los resultados obtenidos con un radio fijo. El barrido del radio de vecindad sobre las capas de WDA muestra que el algoritmo transita abruptamente entre descartar la práctica totalidad de la muestra y colapsarla en un único grupo, y que solo en una ventana estrecha de valores llega a formar más de uno. Con el radio fijado en un mismo valor para todas las capas, en WDB no llega a producirse partición: en tres de las cuatro capas la totalidad de las muestras queda etiquetada como ruido, y en la restante se reúne el 91 % en un solo cluster. El comportamiento admite una explicación común: la densidad de estos embeddings varía de forma continua, sin valles pronunciados de baja densidad, de modo que un umbral global de densidad resulta poco discriminante y el resultado depende críticamente del valor escogido.

Esta observación no descarta a la familia de métodos basados en densidad, sino que señala la condición bajo la cual podría resultar productiva. Progresar por esta vía exige abandonar el radio fijo en favor de una selección adaptativa de ε, ajustada a la escala de distancias propia de cada espacio latente, o bien recurrir a métodos que no dependan de un umbral global único, como HDBSCAN, capaz de extraer grupos de densidad variable a partir de una jerarquía. Que el comportamiento se repita en dos clases de composición atmosférica distinta sugiere, en todo caso, que la propiedad pertenece a las representaciones que produce la red y no a una clase en particular.

Tercero, en la clase WDA las particiones recuperan ejes de variación físicos ya conocidos. El análisis de los espectros representativos muestra que la separación responde principalmente a la temperatura efectiva, legible en la pendiente del continuo tras la normalización. Sobre la capa `dense` aparece además un segundo eje: la anchura de las líneas de Balmer, dominada por el efecto Stark y por tanto ligada a la gravedad superficial de la estrella. Ambos son parámetros establecidos y de variación continua, de modo que recuperarlos no constituye el hallazgo de una subclase nueva. Sí constituye, en cambio, evidencia de que las particiones responden a propiedades físicas reales y no a artefactos numéricos.

Cuarto, en la clase WDB la partición reproduce una distinción taxonómica sin haber recibido información sobre ella. Sobre la capa `flatten`, los grupos obtenidos se ordenan según la presencia e intensidad de las líneas de absorción del hidrógeno, separando los objetos con líneas marcadas, los que las presentan de forma sutil y aquellos en que están ausentes. Esa distinción se sugiere que corresponde a la clasificación DBA, que en la práctica astronómica se establece mediante inspección visual de cada espectro. Dado que todas las

muestras comparten la misma etiqueta y que el pipeline no dispone en ningún punto de información sobre el contenido de hidrógeno, el resultado indica que dicha propiedad está codificada en el espacio latente y que el algoritmo la recupera de forma no supervisada. Es el resultado que más se aproxima al objetivo declarado del trabajo, y sugiere una vía de automatización de una tarea que hoy se realiza manualmente.

Quinto, las métricas internas no bastan para interpretar una partición. A lo largo del trabajo, los hallazgos con contenido físico no provinieron de los valores de Silhouette, Davies–Bouldin o Calinski–Harabasz, sino del examen de los espectros representativos y del seguimiento de los grupos mediante las matrices de migración. Las métricas cumplen la función de descartar configuraciones degeneradas y de comparar capas bajo un criterio homogéneo, pero no informan sobre la naturaleza de lo que separan. El contraste entre ambas clases lo ilustra: la capa `dense_2` obtiene sistemáticamente las mejores métricas en WDA y en WDB, y sin embargo los dos hallazgos con contenido físico, la separación por anchura de línea y la separación por presencia de hidrógeno, se produjeron sobre capas distintas y peor puntuadas. Guiarse únicamente por las métricas habría llevado a descartar precisamente las particiones más informativas. El caso más claro se produce en WDB: sobre `flatten`, tanto el criterio de selección de k como dos de las tres métricas prefieren $k = 2$ frente al $k = 3$ del que procede el hallazgo principal del trabajo.

En conjunto, el marco propuesto demuestra que las representaciones latentes de una red preentrenada, combinadas con clustering sobre espacios reducidos y con síntesis espectral por grupo, permiten identificar de manera reproducible subestructura dentro de las clases espectrales de enanas blancas, y que al menos una de las estructuras recuperadas corresponde a una distinción reconocida por la comunidad astronómica. Determinar si las restantes constituyen subclases no descritas exige el contraste con parámetros astrofísicos externos, que constituye la continuación natural del trabajo.

7.3. Trabajo Futuro

El pipeline desarrollado fue diseñado de forma modular, de modo que admite extensiones sin cambios estructurales. Se identifican las siguientes líneas de trabajo futuro, ordenadas según su proximidad a los resultados obtenidos.

Identificación del rasgo en emisión de la clase WDB. El grupo hallado sobre la capa `dense` presenta una característica en emisión próxima a los 7,300 Å que no aparece en ninguna otra partición, y que se verificó también sobre los espectros individuales del grupo. Su identificación exige un análisis espectroscópico dedicado que este trabajo no aborda. Es la verificación de mayor prioridad, por ser la que podría convertir una observación en un hallazgo.

Caracterización del grupo frío de la clase WDA. La partición sobre `dense` con $k = 5$ aísla un grupo de 90 espectros que se conserva sin pérdida alguna ante variaciones

del número de clusters, y cuyo espectro promedio presenta asimetría en Hα y dos rasgos no atribuibles al hidrógeno, verificados también sobre los espectros individuales del grupo. La línea de absorción situada algo por encima de los 5,000 Å es la que concentra el mayor interés, y su identificación exige un análisis espectroscópico dedicado. Para la característica en emisión próxima a los 6,000 Å corresponde establecer antes si su origen es instrumental, comparando su posición en longitud de onda de reposo con la que ocupa en longitud de onda observada.

Control del tratamiento diferenciado de la capa `dense_2`. Las cuatro capas se reducen a 32 dimensiones mediante PCA salvo `dense_2`, que ya posee esa dimensión y se conserva sin transformar. Dado que esa capa es la que obtiene las mejores métricas en ambas clases, corresponde repetir la comparación aplicando PCA también sobre ella, o bien omitiéndola en todas, para determinar qué parte de su ventaja procede de la representación y qué parte del tratamiento recibido.

Selección adaptativa de los hiperparámetros de DBSCAN. El radio de vecindad ε se mantuvo fijo en 1,5 a lo largo de los experimentos, priorizando la comparabilidad entre capas y clases por sobre el ajuste individual de cada espacio. Dado que con ese radio fijo los resultados del algoritmo resultaron poco informativos, progresar por la vía de los métodos basados en densidad exige sustituirlo por un valor ajustado a la escala de distancias propia de cada espacio latente. El procedimiento habitual para ello es el gráfico de k-distance, propuesto por Ester et al. [15]. Una alternativa complementaria consiste en recurrir a HDBSCAN [16], que sustituye el umbral global de densidad por una jerarquía y elimina con ello la dependencia de ε.

Ampliación de las familias de clustering. El trabajo se consolidó sobre tres algoritmos representativos de las familias particional, de grafos y de densidad. Las restantes alternativas identificadas en el Capítulo 4 constituyen una extensión directa: GMM, Agglomerative Clustering (cuyo dendrograma permite inspeccionar la jerarquía de subgrupos) y HDBSCAN, que elimina la dependencia del parámetro ε. Estas alternativas pueden además combinarse con reductores de dimensionalidad no lineales como UMAP.

Extensión a la clase genérica WD. El trabajo se centró en las clases WDA y WDB, las dos con mayor número de muestras. La etiqueta genérica WD, que reúne las enanas blancas sin subclase determinada, resulta de interés por su carácter heterogéneo, que la convierte en el caso donde una subestructura no reconocida es más plausible.

Escalamiento a la muestra completa. Los experimentos se realizaron sobre submuestras de 1,100 espectros por clase. Extender el análisis a la totalidad de los espectros disponibles permitiría establecer si las particiones obtenidas se mantienen al aumentar la muestra o si dependen del subconjunto analizado, y resultaría especialmente informativo en el caso de WDA, cuyas 13,723 muestras disponibles superan en un orden de magnitud a las empleadas.

BIBLIOGRAFÍA

[1] A. Bédard, "The spectral evolution of white dwarfs: where do we stand?" *Astrophysics and Space Science*, vol. 369, p. 43, 2024, doi:10.1007/s10509-024-04307-5.

[2] D. Koester, "White Dwarf Spectra and Atmosphere Models," *Memorie della Società Astronomica Italiana*, vol. 75, pp. 282–291, 2008, arXiv:0812.0482.

[3] P.-E. Tremblay and P. Bergeron, "Spectroscopic Analysis of DA White Dwarfs: Stark Broadening of Hydrogen Lines Including Nonideal Effects," *The Astrophysical Journal*, vol. 696, no. 2, pp. 1755–1770, 2009, doi:10.1088/0004-637X/696/2/1755.

[4] D. Saumon, S. Blouin, and P.-E. Tremblay, "Current challenges in the physics of white dwarf stars," *Physics Reports*, vol. 988, pp. 1–63, 2022, doi:10.1016/j.physrep.2022.09.001.

[5] D. Koester, B. T. Gänsicke, and J. Farihi, "The frequency of planetary debris around young white dwarfs," *Astronomy & Astrophysics*, vol. 566, p. A34, 2014, doi:10.1051/0004-6361/201423691.

[6] L. Ferrario, D. de Martino, and B. T. Gänsicke, "Magnetic White Dwarfs," *Space Science Reviews*, vol. 191, pp. 111–169, 2015, doi:10.1007/s11214-015-0152-0.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[8] D. Baron, "Machine Learning in Astronomy: a practical overview," 2019, arXiv:1904.07248. doi:10.48550/arXiv.1904.07248.

[9] G. Traven, G. Matijevič, T. Zwitter *et al.*, "The GALAH Survey: Classification and Diagnostics with t-SNE Reduction of Spectral Information," *The Astrophysical Journal Supplement Series*, vol. 228, p. 24, 2017, doi:10.48550/arXiv.1612.02242.

[10] O. Vincent, P. Bergeron, and P. Dufour, "Data-driven selection and spectral classification of white dwarf stars," *Monthly Notices of the Royal Astronomical Society*, vol. 521, no. 1, pp. 760–781, 2023.

[11] O. Vincent, M. A. Barstow, S. Jordan, C. Mander, P. Bergeron, and P. Dufour, "Classification and parameterisation of a large Gaia sample of white dwarfs using XP spectra," *Astronomy & Astrophysics*, vol. 682, p. A5, 2024, doi:10.1051/0004-6361/202347694.

[12] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014, pp. 3320–3328.

[13] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[14] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1.* University of California Press, 1967, pp. 281–297.

[15] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).* AAAI Press, 1996, pp. 226–231.

[16] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," in *Advances in Knowledge Discovery and Data Mining (PAKDD 2013), Lecture Notes in Computer Science, vol. 7819.* Springer, 2013, pp. 160–172.

[17] U. von Luxburg, "A Tutorial on Spectral Clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007, doi:10.1007/s11222-007-9033-z.

[18] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 14 (NIPS 2001).* MIT Press, 2002, pp. 849–856.

[19] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[20] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[21] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[22] D. Moulavi, P. A. Jaskowiak, R. J. G. B. Campello, A. Zimek, and J. Sander, "Density-Based Clustering Validation," in *Proceedings of the 2014 SIAM International Conference on Data Mining (SDM)*, 2014, pp. 839–847.

[23] J. C. Dunn, "Well-Separated Clusters and Optimal Fuzzy Partitions," *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, 1974, doi:10.1080/01969727408546059.

[24] M. Halkidi and M. Vazirgiannis, "Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set," in *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM 2001)*, 2001, pp. 187–194.

[25] X. L. Xie and G. Beni, "A Validity Measure for Fuzzy Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991.

[26] J. S. Almeida and C. Allende Prieto, "Automated unsupervised classification of the Sloan Digital Sky Survey stellar spectra using k-means clustering," *The Astrophysical Journal*, vol. 763, p. 50, 2013, arXiv:1211.5321.

[27] R. Garcia-Dias, C. Allende Prieto, J. S. Almeida, and I. Ordovás-Pascual, "Machine learning in APOGEE: Unsupervised spectral classification with K-means," *Astronomy & Astrophysics*, vol. 612, p. A98, 2018, doi:10.1051/0004-6361/201732134.

[28] D. Fraix-Burnet, C. Bouveyron, and J. Moultaka, "Unsupervised classification of SDSS galaxy spectra," *Astronomy & Astrophysics*, vol. 649, p. A53, 2021, doi:10.1051/0004-6361/202040046.

[29] A. B. A. Queiroz, C. Chiappini, A. Pérez-Villegas *et al.*, "The Milky Way bar and bulge revealed by APOGEE and Gaia EDR3," *Astronomy & Astrophysics*, vol. 656, p. A156, 2021, doi:10.1051/0004-6361/202039030.

[30] A. Castro-Ginard, C. Jordi, X. Luri *et al.*, "Hunting for open clusters in Gaia EDR3: 628 new open clusters found with OCfinder," *Astronomy & Astrophysics*, vol. 661, p. A118, 2022, doi:10.1051/0004-6361/202142568.

[31] L. A. Gutiérrez-Soto, R. Lopes de Oliveira, S. Akras *et al.*, "Mapping Hα excess candidate point sources in the southern hemisphere using S-PLUS data," *Astronomy & Astrophysics*, vol. 695, p. A104, 2025, doi:10.1051/0004-6361/202453167.

[32] L. Spina, L. Magrini, G. G. Sacco *et al.*, "The Gaia-ESO Survey: Chemical tagging in the thin disk," *Astronomy & Astrophysics*, vol. 668, p. A16, 2022, doi:10.1051/0004-6361/202243316.

[33] F. Bufano, C. Bordiu, T. Cecconello *et al.*, "Sifting the debris: Patterns in the SNR population with unsupervised ML methods," *Astronomy & Astrophysics*, vol. 690, p. A203, 2024, doi:10.1051/0004-6361/202451096.

[34] I. A. Guez and M. Claire, "Reading Between the Rainbows: Comparative Exoplanet Characterisation through Molecule Agnostic Spectral Clustering," 2024, arXiv:2410.16986. doi:10.48550/arXiv.2410.16986.

[35] M. Sasdelli, E. E. O. Ishida, R. Vilalta, M. Aguena, V. C. Busti, H. Camacho, A. M. M. Trindade, F. Gieseke, R. S. de Souza, Y. T. Fantaye, and P. A. Mazzali, "Exploring the spectroscopic diversity of type Ia supernovae with DRACULA: a machine learning approach," *Monthly Notices of the Royal Astronomical Society*, vol. 461, no. 2, pp. 2044–2059, 2016, doi:10.1093/mnras/stw1478.