



**UNIVERSIDAD
DE GRANADA**

**Departamento de Ciencias de la
Computación e Inteligencia Artificial**

E.T.S. de Ingenierías Informática y de Telecomunicación

Inteligencia de Negocio

Guion de Prácticas

**Práctica 2:
Análisis Relacional mediante Segmentación**

Curso 2025-2026

Grado en Ingeniería Informática

Grado en Ingeniería Informática y Matemáticas

Grado en Ingeniería Informática y Administración y Dirección de Empresas

Práctica 2

Segmentación mediante *Clustering*

1. Objetivos y Evaluación

En esta segunda práctica de la asignatura Inteligencia de Negocio veremos el uso de técnicas de aprendizaje no supervisado para análisis relacional mediante segmentación. Se trabajará con un conjunto de datos sobre el que se aplicarán distintos algoritmos de agrupamiento (*clustering*). A la luz de los resultados obtenidos se deberán crear informes y análisis lo suficientemente profundos.

La práctica se calificará hasta un **máximo de 2 puntos**. Se valorará el acierto en los recursos de análisis gráficos empleados, la complejidad de los experimentos realizados, la interpretación de los resultados, la organización y redacción del informe, etc.

2. Condiciones de Vida 2024

En esta práctica se trabajará con datos de Encuesta de Condiciones de Vida (ECV), basada en criterios armonizados para todos los países de la Unión Europea. El conjunto de datos fue obtenido a través del sitio web oficial del Instituto Nacional de Estadística y contiene información detallada sobre distintos aspectos de las condiciones de vida de los hogares. En total, se registraron 29.781 resultados, abarcando variables relacionadas con renta, vivienda, identificación y características del hogar, composición familiar, ingresos y finanzas, gastos y capacidad de afrontarlos, carencias materiales, situación de los menores y cuidado de personas dependientes.

El conjunto de datos se organiza en dos ficheros:

- *diccionario.xlsx*: El fichero proporciona la información detallada de cada variable, incluyendo su tipo y formato.
- *ECV.csv*: El fichero contiene los datos de la encuesta, con 29.781 registros y 184 columnas que reflejan las respuestas de los hogares.

Estos dos ficheros permiten conocer tanto la estructura como el contenido completo del conjunto de datos.

El objetivo de la práctica es definir algunos casos de estudio de interés (fijando condiciones en algunas variables), aplicar distintos algoritmos de *clustering*, analizar la calidad de las soluciones obtenidas y, finalmente, interpretar los resultados para explicar los distintos perfiles o grupos encontrados.

3. Tareas a Realizar

La práctica consiste en aplicar y analizar técnicas de agrupamiento para descubrir grupos en el conjunto de datos bajo estudio. El trabajo se realizará empleando bibliotecas y paquetes de Python, principalmente `numpy`, `pandas`, `scikit-learn`, `matplotlib` y `seaborn`. Se recomienda consultar las siguientes páginas web:

- <http://scikit-learn.org/stable/modules/clustering.html>
- <http://www.learndatasci.com/k-means-clustering-algorithms-python-intro/>
- http://hdbSCAN.readthedocs.io/en/latest/comparing_clustering_algorithms.html
- <https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/>
- <http://seaborn.pydata.org/generated/seaborn.clustermap.html>

Nos interesaremos en segmentar la muestra seleccionando previamente grupos de interés según las variables categóricas y/u ordinales. Queda a elección libre del alumno/a escoger varios casos (al menos tres) y realizar el estudio sobre ellos. En cada caso de estudio, si se quiere, se puede realizar un análisis comparado entre dos subconjuntos de datos según la división realizada por una variable categórica y/o numérica. Será necesario también aplicar una normalización para que las métricas de distancia y la visualización funcionen correctamente.

En cada caso de estudio se analizarán 5 algoritmos distintos de agrupamiento (siendo al menos uno de ellos K-means y uno jerárquico) obteniéndose el tiempo de ejecución y métricas de rendimiento tales como el coeficiente *silhouette* y el índice Calinski-Harabasz. Además, se analizará el efecto de algunos parámetros determinantes (por ejemplo, el valor de k si el algoritmo necesita fijarlo *a priori*) en al menos 2 algoritmos distintos para cada caso de estudio.

El análisis deberá apoyarse en visualizaciones tales como nubes de puntos (*scatter matrix*), dendrogramas (en agrupamiento jerárquico), mapas de temperatura (*heatmap*), gráfico de burbujas con la distancia relativa entre los centros de los clústers mediante *multidimensional scaling*, etc. Todas las figuras que aparecen a continuación son ejemplos basados en un caso de estudio diferente relacionado con apartamentos turísticos en Granada. Por ejemplo, en la figura 2.1 se incluye un *scatter matrix* de un conjunto de variables numéricas obtenido por K-means ($k = 4$) y en la figura 2.2 la distribución del coeficiente de *silhouette* en cada agrupamiento. Se recomienda que sobre estas visualizaciones se construyan tablas que caractericen aproximadamente cada grupo observando las agrupaciones realizadas. Para esa interpretación,

puede ayudar el uso de gráficas de los centroides como la de la figura 2.3 o gráficos de distribución como el de la figura 2.4. En la web de la asignatura se incluye un *script* de ejemplo que puede servir como punto de partida para realizar la práctica.

A partir de los resultados obtenidos se deberán extraer conclusiones sobre las condiciones de vida. Se valorará el acierto en la selección de casos de estudio que mejor reflejen los grupos encontrados en los datos.

4. Esquema de la Documentación

La documentación entregada deberá ajustarse al siguiente esquema (debe respetarse la numeración y nombre de las secciones):

1. **Introducción:** se hablará sobre el problema abordado y todas las consideraciones generales que se deseen indicar.
2. **Caso de estudio X:** se incluirá una sección por cada caso de estudio analizado (el epígrafe describirá el subconjunto de datos bajo estudio). En ella se explicará en detalle en un primer apartado qué caso se analiza y por qué (deberá indicarse el número de datos que representa el caso de estudio). Se incluirá una tabla comparativa con los resultados de los algoritmos de *clustering* (que incluirá, al menos, el número de *clusters* obtenidos, el valor de las métricas y el tiempo de ejecución en cada algoritmo) y tantas otras tablas para el análisis de los parámetros (una tabla por algoritmo). Cada sección contendrá las visualizaciones necesarias para analizar el problema y junto a cada visualización se incluirá una tabla que caracterice cada *cluster*. Se añadirá un apartado final titulado “Interpretación de la segmentación” que incluirá las conclusiones generales a las que haya llegado el alumno a la luz de los resultados en el correspondiente caso de estudio. En cada sección deberán incluirse extractos de los *scripts* que el alumno considere relevantes para destacar el trabajo realizado.
3. **Contenido adicional:** opcionalmente, cualquier tarea adicional a las descritas en este guion puede presentarse en esta sección.
4. **Bibliografía:** referencias y material consultado para la realización de la práctica.

Las tablas de resultados no deberán ser capturas de pantalla, sino tablas creadas en el procesador de texto empleado. No se aceptarán otras secciones distintas de estas. Además, la primera página de la documentación incluirá una portada con el nombre completo del alumno, grupo de prácticas y dirección email. También se incluirá una segunda página con el índice del documento donde las diferentes secciones y páginas estarán enlazadas en el pdf.

5. Entrega

La fecha límite de entrega será el domingo **14 de diciembre** de 2025 hasta las **23:59**. La entrega se realizará a través de la web de la asignatura en <https://prado.ugr.es/>. En ningún caso se aceptan entregas a través de enlaces como Dropbox, Google Drive, WeTransfer o similares. En un único fichero zip se incluirá la documentación, los *scripts* de Python empleados y cualquier otro archivo que el alumno considere relevante. El nombre del archivo zip será el siguiente (sin espacios): P2-apellido1-apellido2-nombre.zip. La documentación tendrá el mismo nombre pero con extensión pdf. Es decir, la alumna “María Teresa del Castillo Gómez” subirá el archivo P2-delCastillo-Gómez-MaríaTeresa.zip que contendrá, entre otros, el archivo P2-delCastillo-Gómez-MaríaTeresa.pdf.

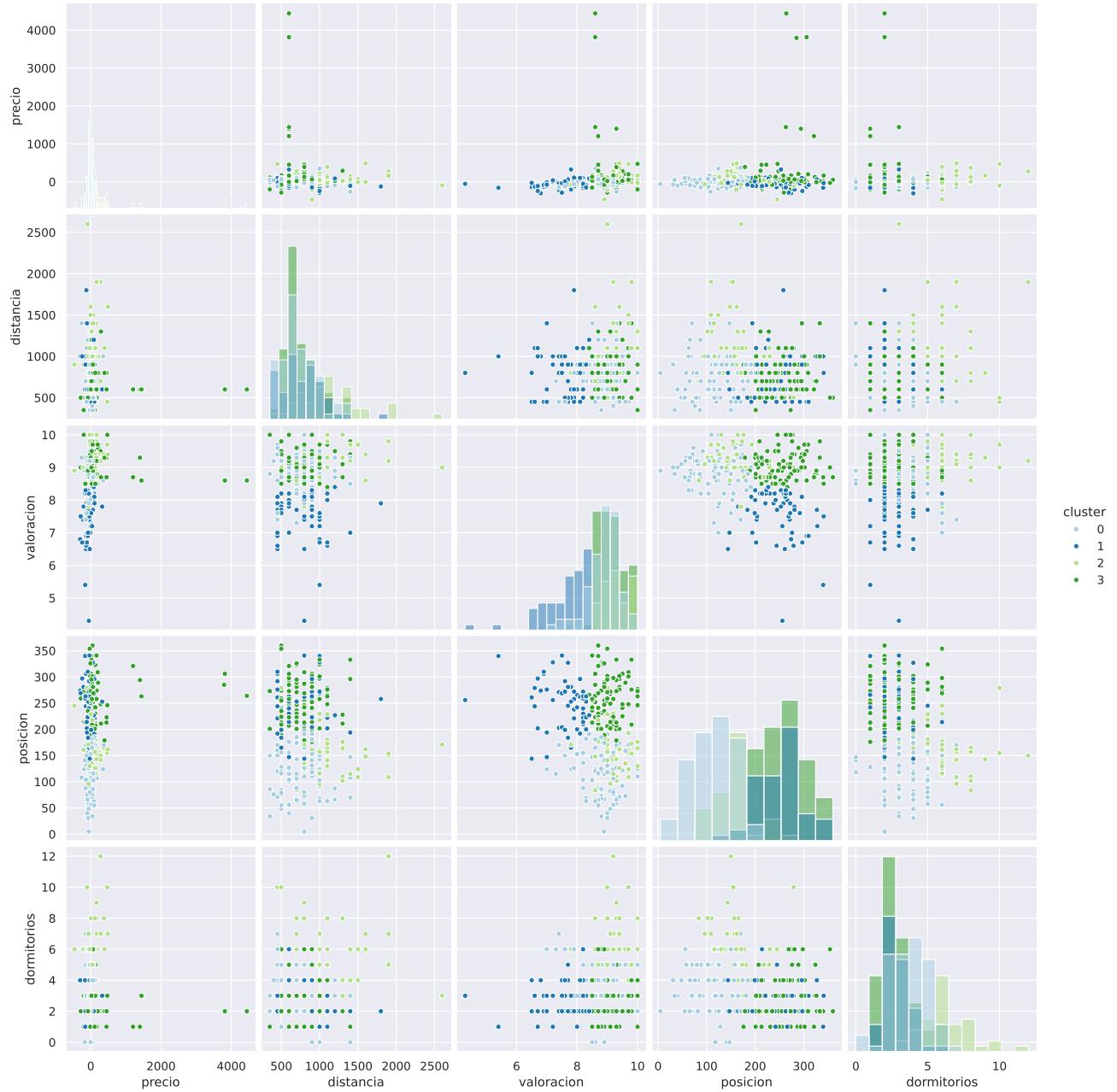


Figura 2.1: Ejemplo de resultado de K-means con $k = 4$ relacionando diferencia de precio, distancia al centro, valoración, posición dentro la búsqueda y número de dormitorios

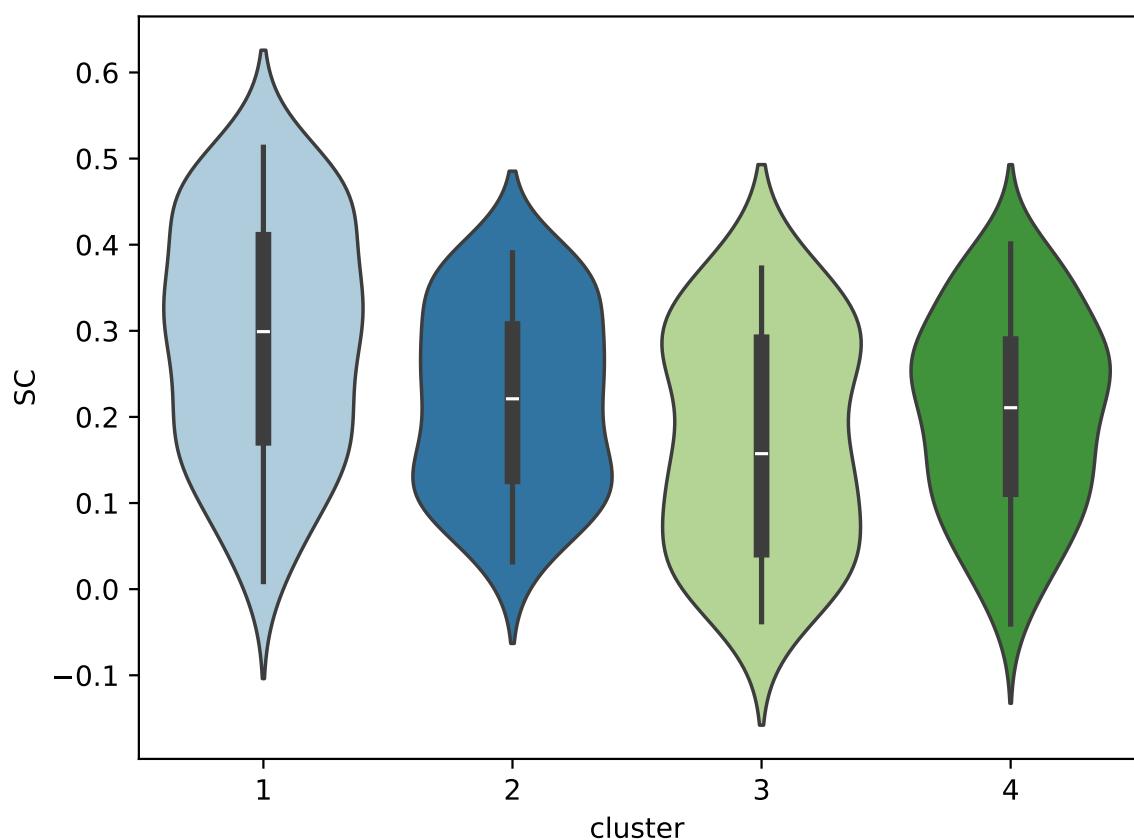


Figura 2.2: Distribución del coeficiente de *silhouette* en cada agrupamiento

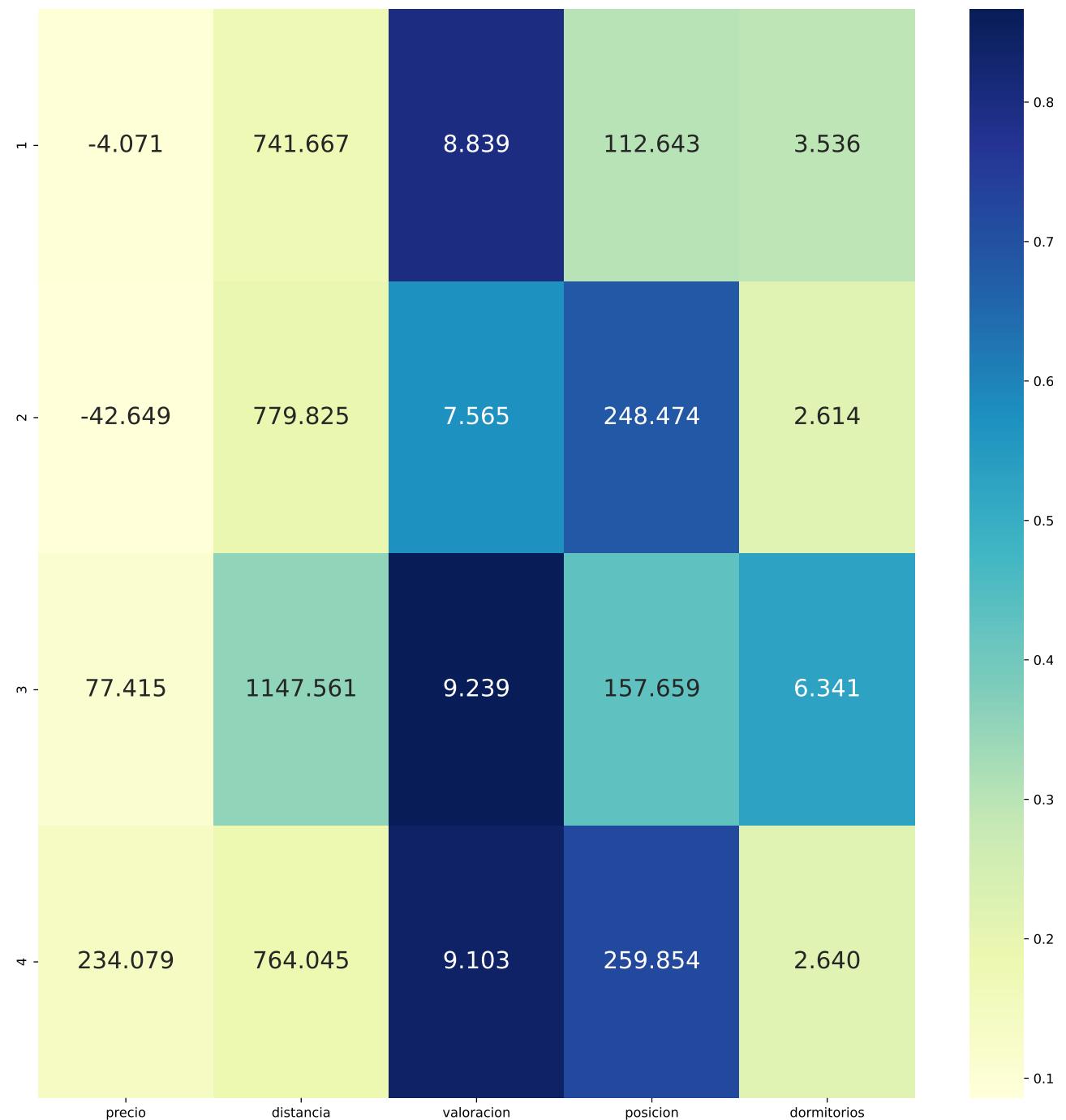


Figura 2.3: Centros de los grupos de la figura 2.1

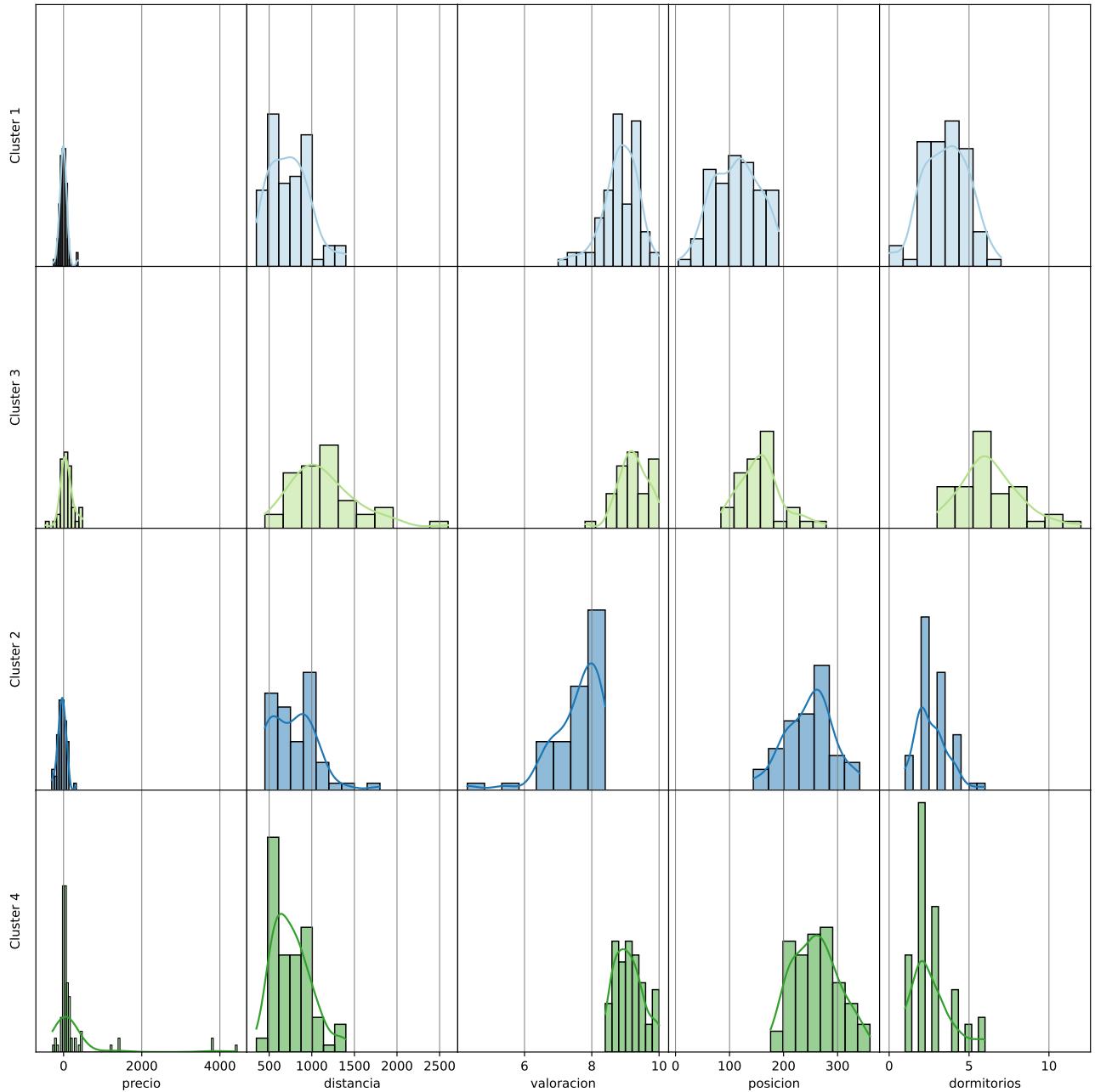


Figura 2.4: Distribución de los grupos de la figura 2.1 ordenados de acuerdo a su posición en los resultados de búsqueda (aquí se muestran histogramas y curvas de distribución de probabilidad, pero también se puede simplificar empleando boxplots, por ejemplo, cuando la distribución tiende a la normalidad)

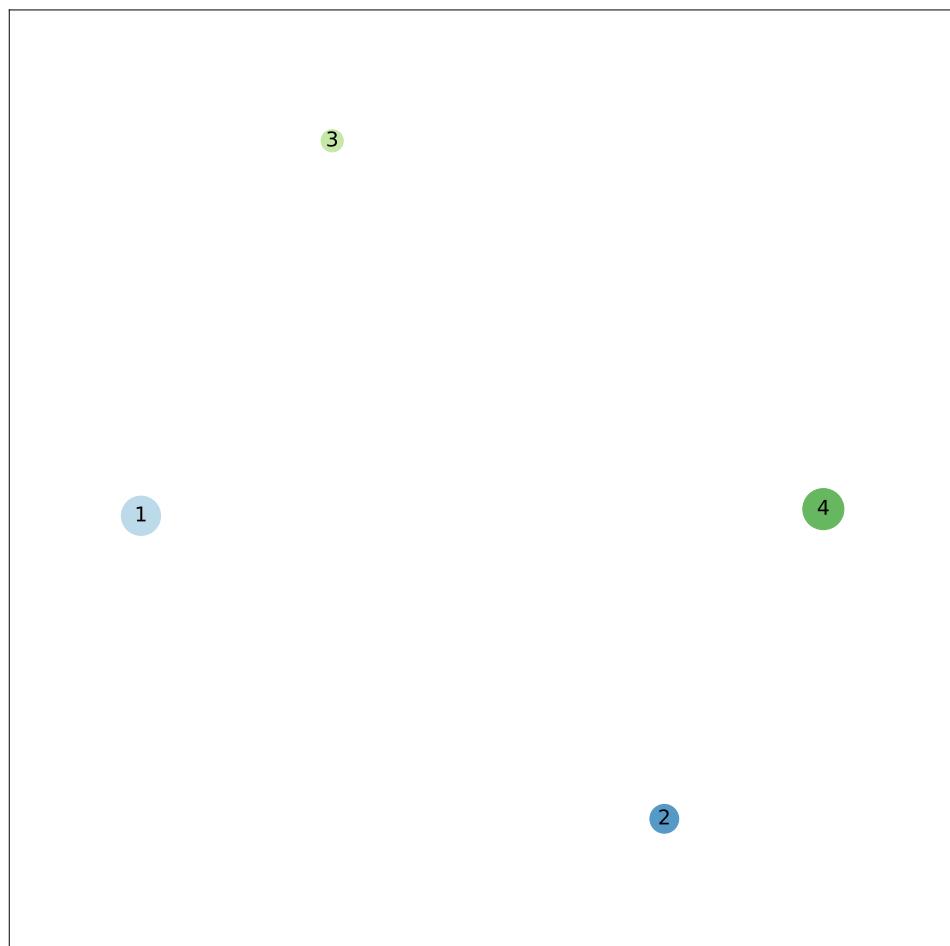


Figura 2.5: Distancia relativa entre centroides de clusters (radio del círculo proporcional al número de objetos en cada clúster)

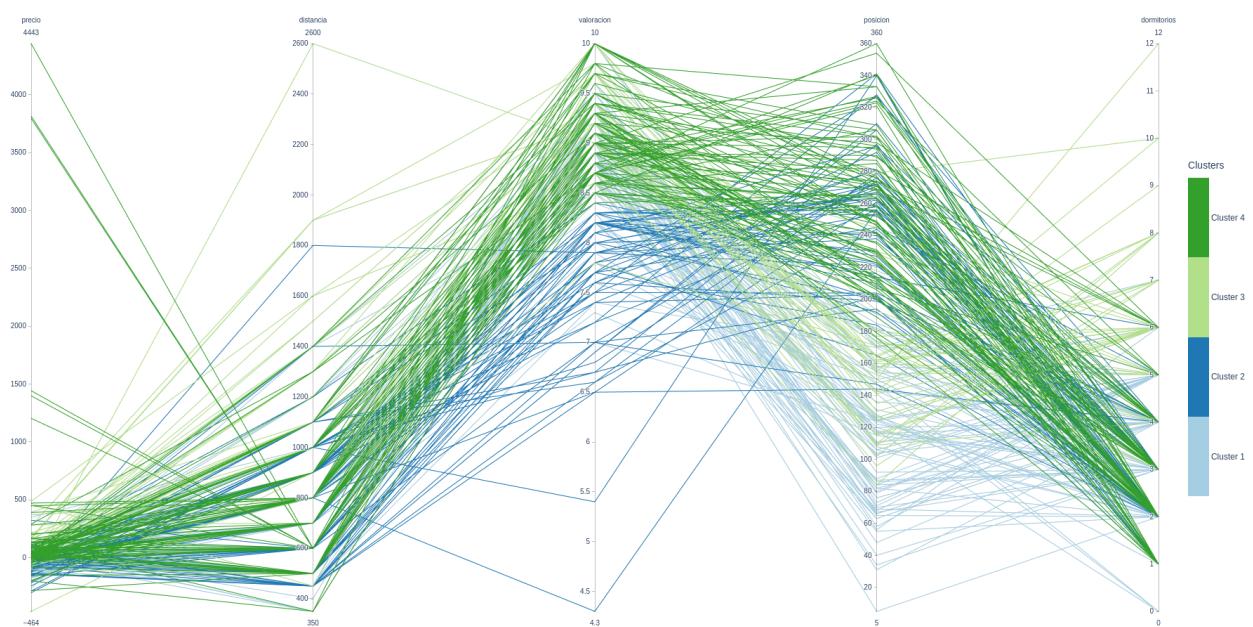


Figura 2.6: Coordenadas paralelas para interpretar en detalle; es mejor usar la versión interactiva HTML