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 2024-2025

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. Alojamientos turísticos en Granada

En esta práctica se trabajará con datos de alojamientos turísticos en la ciudad de Granada obtenidos a través de la plataforma de reservas online Booking.com. Las búsquedas se realizaron entre los días 25 de Octubre de 2024 y 10 de Noviembre de 2024. En cada búsqueda, se extraen los 400 primeros resultados obtenidos. y se consultan intervalos de estancias de 2 y 3 días para 2, 4, 6 y 8 huéspedes. En total se dispone de un total de 194.497 resultados.

Cada resultado de la búsqueda contiene la siguiente información:

- *Ranking position*: posición del anuncio en la búsqueda
- *Hotel name*: nombre del alojamiento
- *Price*: precio del alojamiento en la búsqueda
- *Deal*: oferta del alojamiento en la búsqueda
- *Location*: barrio donde se ubica

- *Distance*: distancia en metros al centro de la ciudad
- *Type*: tipo de alojamiento
- *Quality*: calidad del alojamiento en estrellas para hoteles (de 1 a 5) o cuadrados para otros alojamientos (de 1 a 5)
- *Rating*: puntuación del alojamiento
- *Special*: mención especial del alojamiento
- *Review*: número de reseñas recibidas para calcular el Rating
- *Description*: descripción del alojamiento
- *Bedrooms*: número de dormitorios
- *Living Rooms*: número de salones
- *Bathrooms*: número de cuartos de baño
- *Kitchens*: número de cocinas
- *Surface Area (m2)*: superficie en m2
- *Total Beds*: número total de camas
- *Individual Beds*: número de camas individuales (entre 90 y 130 cm)
- *Double Beds*: número de camas dobles (entre 131 y 150 cm)
- *Double Large Beds*: número de camas dobles grandes (entre 151 y 180 cm)
- *Double Extralarge Beds*: número de camas dobles extragrandes (entre 181 y 210 cm)
- *Sofa Beds*: número de sofás cama
- *Bunk Beds*: número de literas
- *Search in advance*: días de antelación respecto a la fecha de check in cuando se realizó la búsqueda
- *Guests*: número de personas adultas alojadas en la búsqueda
- *Average selected price*: precio medio de un conjunto de alojamientos seleccionados. Se escogen los alojamientos que aparecen entre los 50 primeros puestos de la búsqueda (ranking), con una distancia al centro de la ciudad inferior o igual a 700 metros, puntuación (rating) igual o superior a 9 y un número de reseñas (review) igual o superior a 10. Si hay menos de 7 anuncios con estas condiciones, se relaja la condición a una puntuación de 8.7 y un número de reseñas mínimo de 5.

- *Total of apartments*: número total de apartamentos disponibles para las fechas y número de personas de la búsqueda
- *Total of hotels*: número total de hoteles disponibles para las fechas y número de personas de la búsqueda
- *Query date*: fecha en la que se realizó las búsqueda
- *Check in*: fecha de entrada al alojamiento en la búsqueda
- *Check out*: fecha de salida del alojamiento en la búsqueda

Además, se dispone de un segundo conjunto de resultados en el que se han agrupado los resultados por alojamiento. Contiene alojamientos únicos basado en nombre del alojamiento y superficie (pues hay casos de diferentes alojamientos bajo un mismo nombre). De existir más de uno con igual nombre y superficie, se consideran como uno solo escogiéndose el resultado más reciente entre las búsquedas. Este conjunto contiene, además de los campos descritos anteriormente, la siguiente información:

- *Price avg*: precio medio del alojamiento en todas las búsquedas en las que apareció
- *Ranking position avg*: posición del anuncio en todas las búsquedas en las que apareció
- *Number of views*: número de veces que aparece el alojamiento en las búsquedas realizadas. Si no aparece el anuncio, puede ser que bien alojamiento no aparece entre los primeros 400 puestos o que no está disponible para la búsqueda correspondiente.
- *Price difference*: diferencia promedio de precio del alojamiento respecto al precio medio de alojamientos seleccionados en cada búsqueda. Se compara el precio del alojamiento frente a los alojamientos que aparecen entre los 50 primeros puestos de la búsqueda (ranking), con una distancia al centro de la ciudad inferior o igual a 700 metros, puntuación (rating) igual o superior a 9 y un número de reseñas (review) igual o superior a 10. Si hay menos de 7 anuncios con estas condiciones, se relaja la condición a una puntuación de 8.7 y un número de reseñas mínimo de 5. Por tanto, no representa el precio medio de la oferta de alojamiento total, sino el precio medio para anuncios de calidad y bien posicionados en la búsqueda devuelta por Booking.com.

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 (por ejemplo, resultados en dos barrios distintos de Granada, o en hoteles frente a apartamentos). 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. 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 la oferta de alojamientos turísticos en Granada. 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 viernes **15 de diciembre** de 2024 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`.

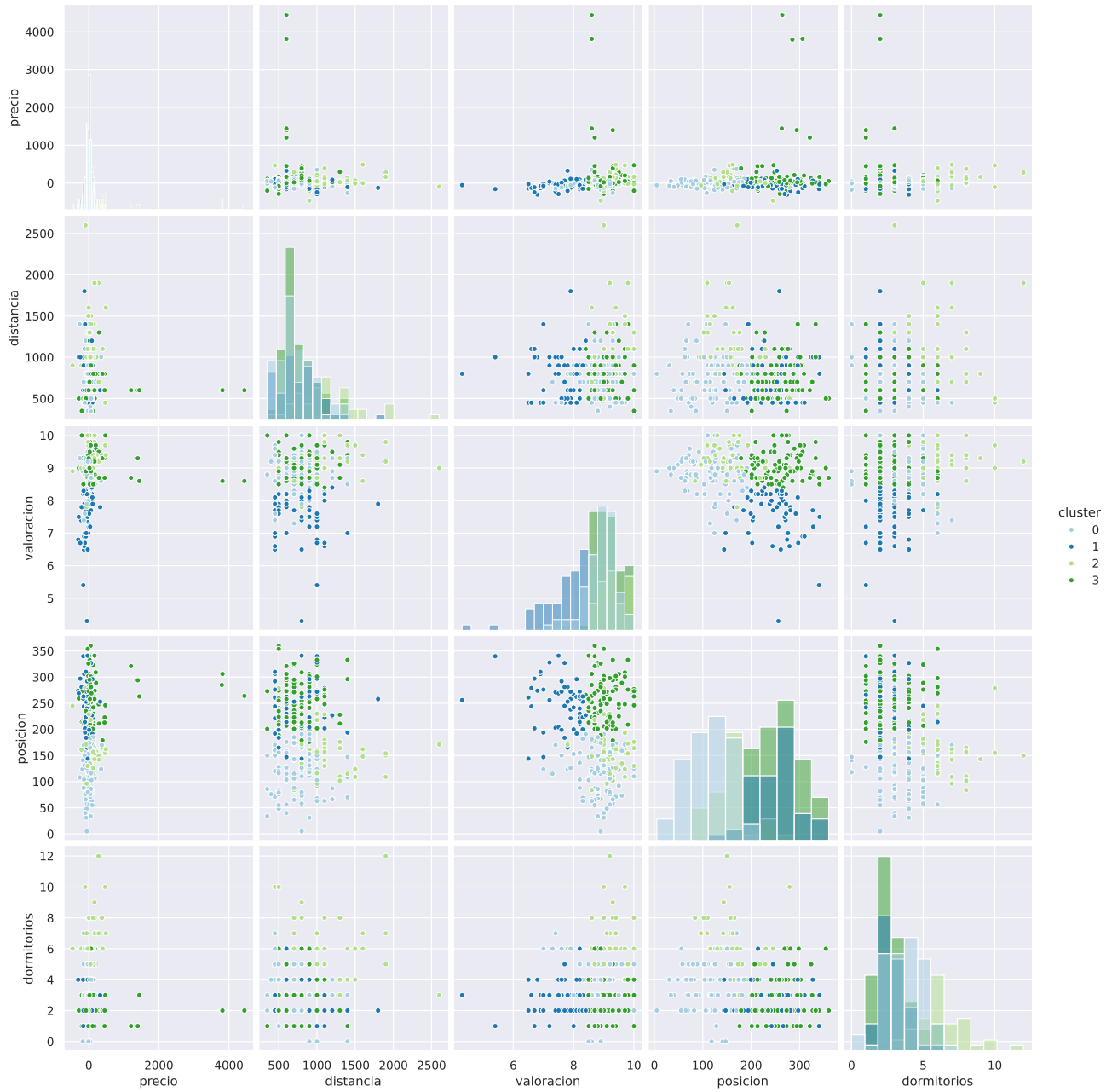

Figura 2.1: Ejemplo de resultado de K-means con $k = 4$ relacionando diferencia de precio, distancia al centro, valoración, posición dentro la búsqueda y número de dormitorios

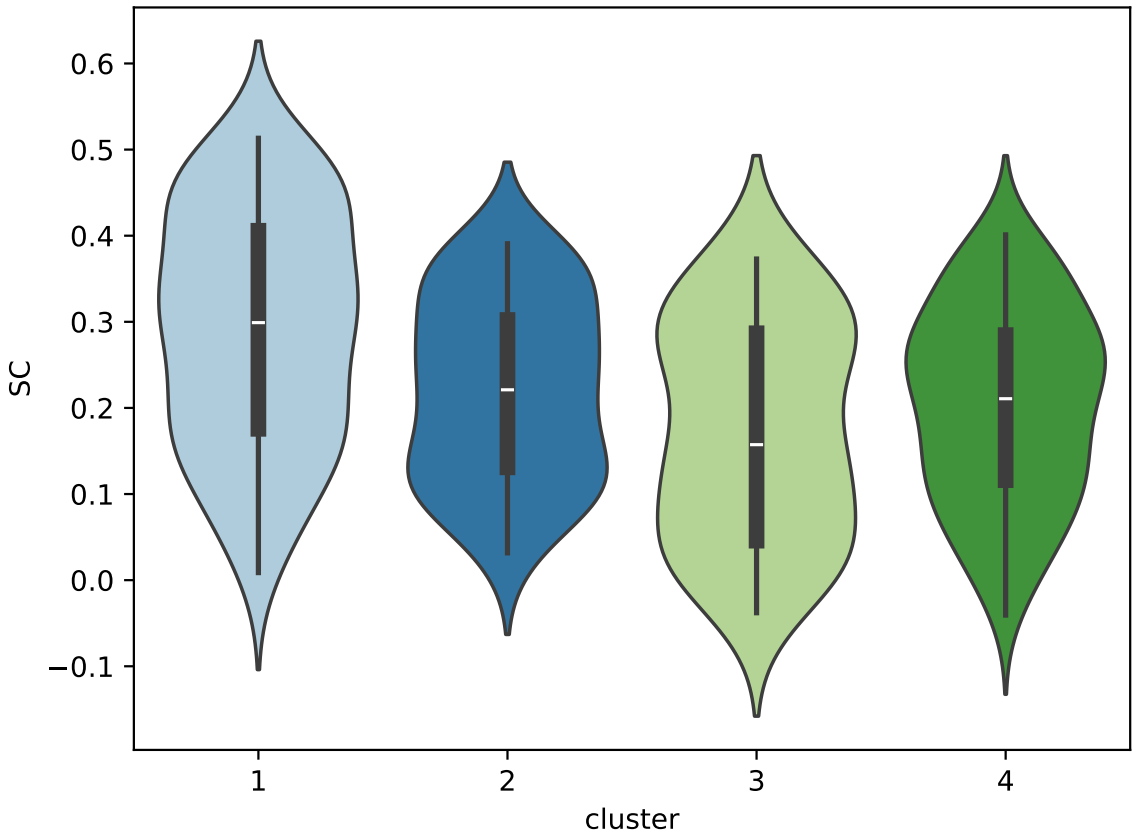

Figura 2.2: Distribución del coeficiente de *silhouette* en cada agrupamiento

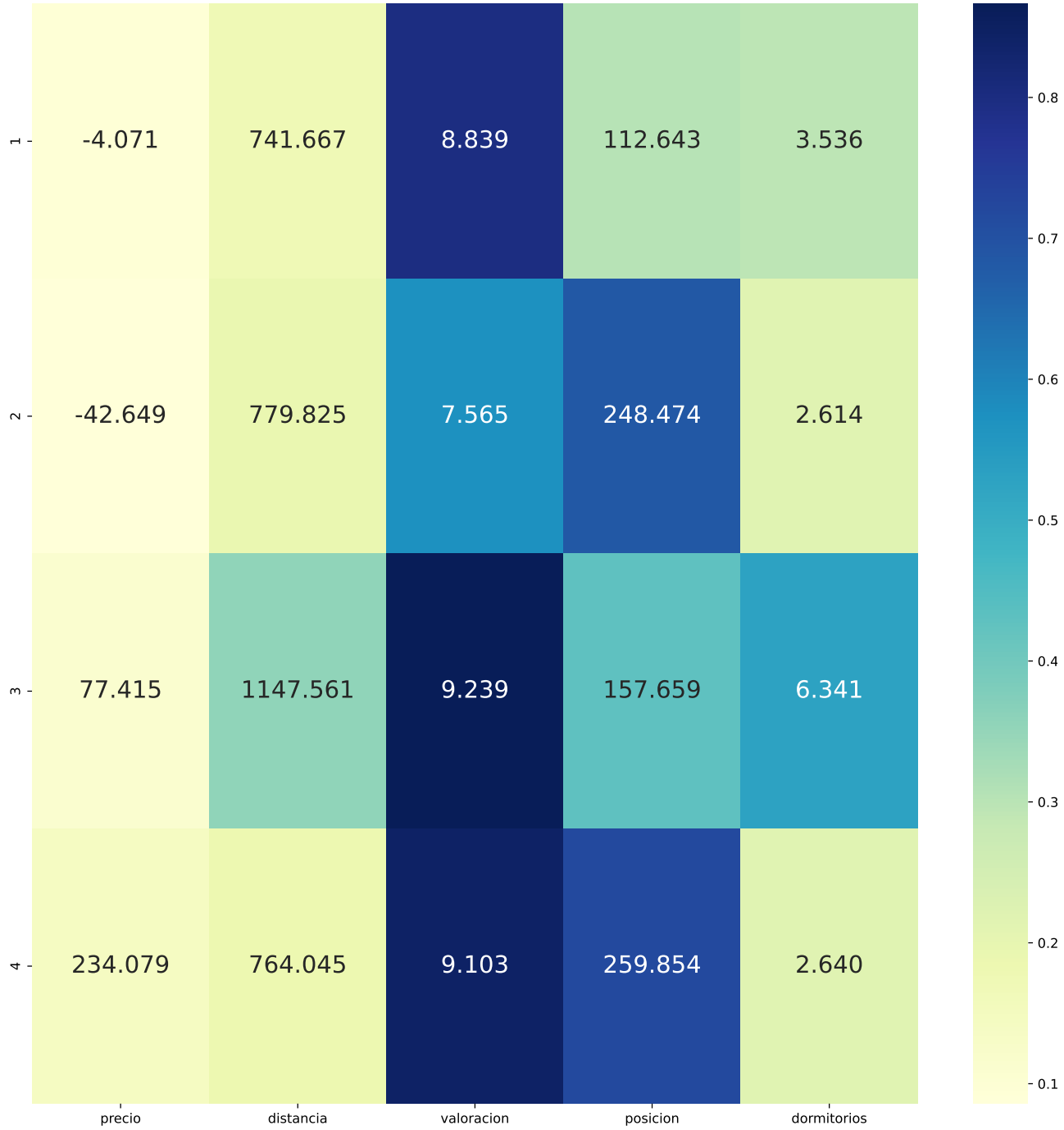

Figura 2.3: Centros de los grupos de la figura 2.1

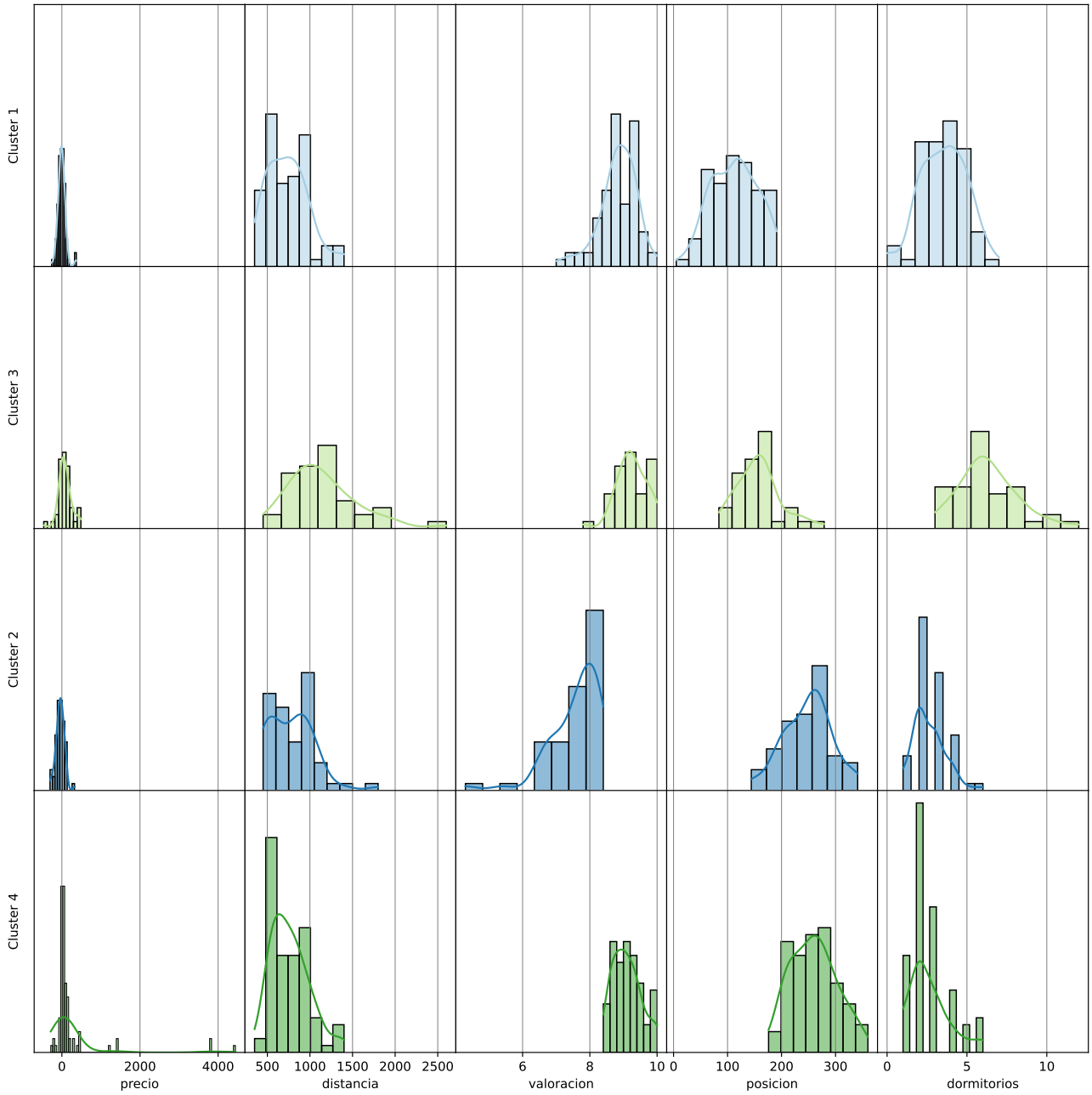

Figura 2.4: Distribución de los grupos de la figura 2.1 ordenados de acuerdo a su posición en los resultados de búsqueda (aquí se muestran histogramas y curvas de distribución de probabilidad, pero también se puede simplificar empleando boxplots, por ejemplo, cuando la distribución tiende a la normalidad)

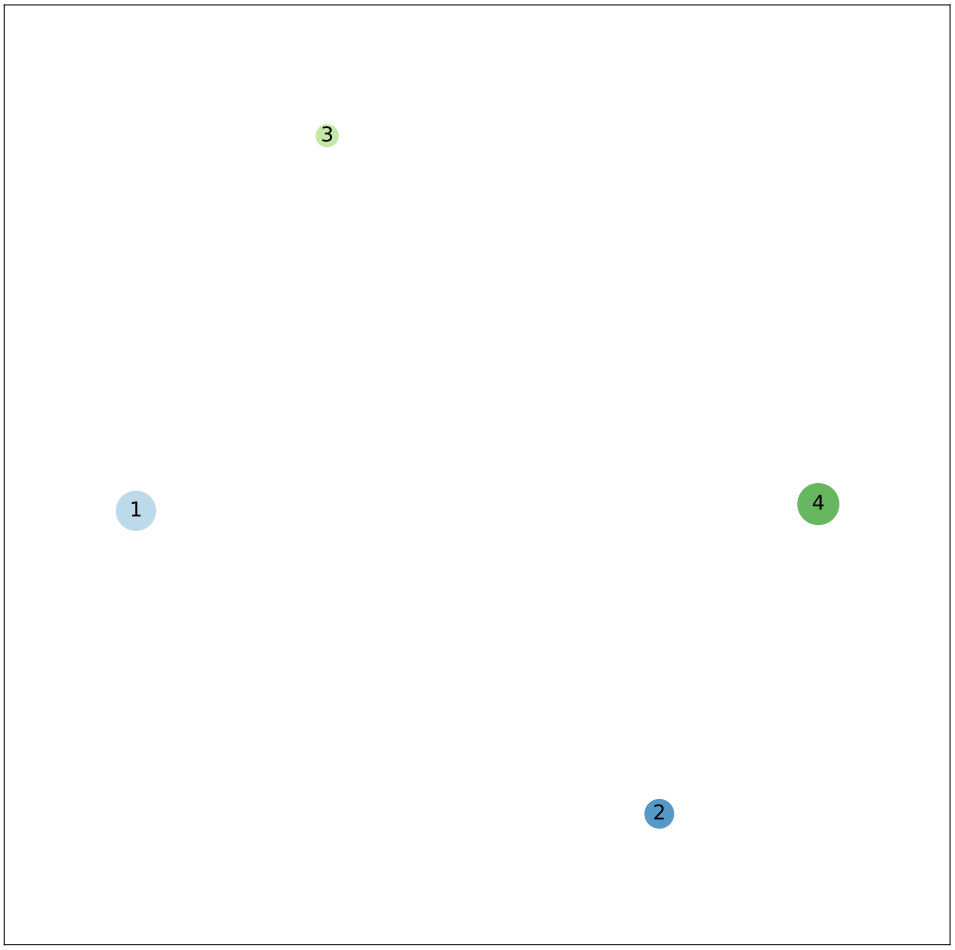

Figura 2.5: Distancia relativa entre centroides de clusters (radio del círculo proporcional al número de objetos en cada clúster)

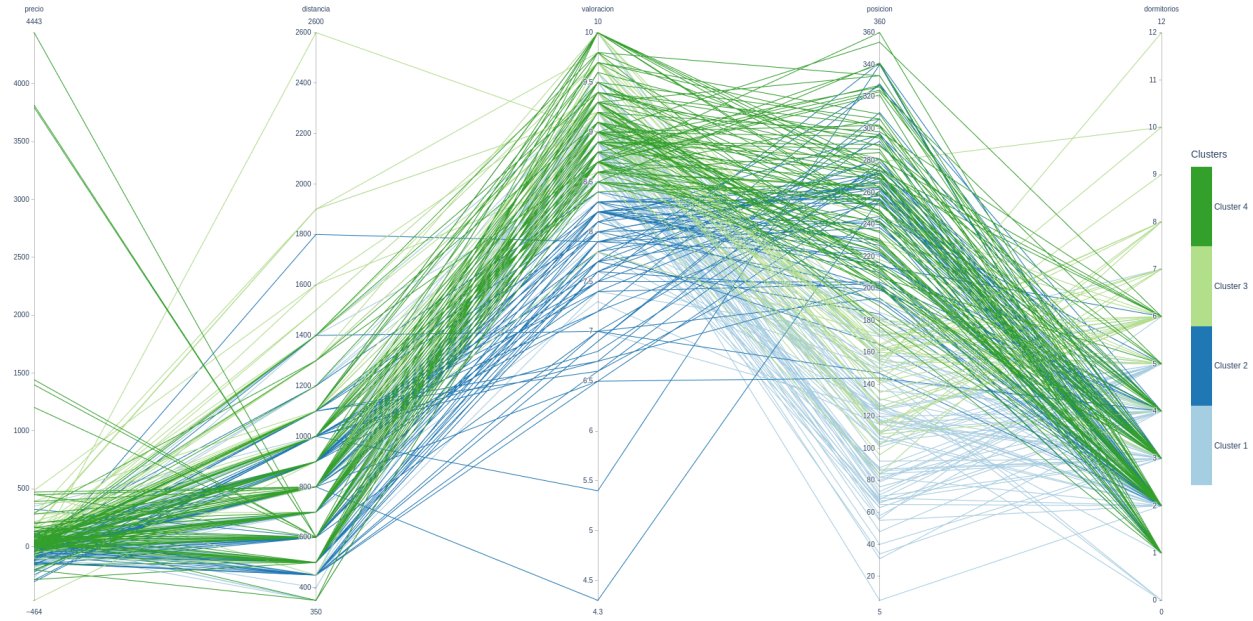

Figura 2.6: Coordenadas paralelas para interpretar en detalle; es mejor usar la versión interactiva HTML