Aplicación de Clustering Jerárquico a Estadísticas de Jugadores de la NBA (temporada 2024–2025)

Autores:
Fabricio-Javier Rivadeneira-Zambrano
Universidad Laica Eloy Alfaro de Manabí, **ULEAM**
fabricio.rivadeneira@uleam.edu.ec
Manta, Ecuador

Silvia-Mercedes Carvajal-Rivadeneira
Unidad Educativa Julio Pierregrosse, **UEJP**
scarvajal@juliopierregrosse.edu.ec
Manta, Ecuador

Rodolfo-Andrés Rivadeneira-Zambrano
Universidad Técnica de Manabí, **UTM**
rodolfo.rivadeneira@utm.edu.ec
Portoviejo, Ecuador

Isaac-Fabricio Rivadeneira-Carvajal
Universidad Laica Eloy Alfaro de Manabí, **ULEAM**
e1316793783@live.uleam.edu.ec
Manta, Ecuador

Joshua-Javier Rivadeneira-Carvajal
Universidad Laica Eloy Alfaro de Manabí, **ULEAM**
e1316793775@live.uleam.edu.ec
Manta, Ecuador

DOI: https://doi.org/10.56124/encriptar.v9i17.010

Resumen

Se aplicó el método de clustering jerárquico a un conjunto de datos con estadísticas agregadas de 731 jugadores de la NBA correspondientes a la temporada 2024–2025. El objetivo fue identificar perfiles de rendimiento sin imponer etiquetas previas, utilizando una estrategia no supervisada basada en distancias. Para ello, se seleccionaron variables cuantitativas de producción (puntos, rebotes, asistencias), volumen de uso (minutos jugados, intentos de tiro y tiros libres) y contribuciones defensivas (robos, tapones, pérdidas y faltas). Tras un preprocesamiento orientado a hacer comparables las escalas (normalización) se construyó una matriz de disimilaridad y se generó el Dendrograma, evaluando el corte del diagrama mediante criterios de separación

y coherencia interna. El análisis condujo a una partición de tres grupos bien diferenciados: un clúster de jugadores de alta carga ofensiva y gran volumen, un clúster de rol/soporte con menor uso y aportes más equilibrados, y un clúster intermedio que combina eficiencia con participación moderada. Estos resultados facilitan comparaciones entre jugadores, resumen patrones dominantes del torneo y aportan una base para análisis posteriores (segmentación y toma de decisiones).

Palabras clave: agrupamiento jerárquico; NBA; medidas de disimilaridad; segmentación; análisis no supervisado.

Application of Hierarchical Clustering to NBA Player Statistics (2024–2025 Season)

ABSTRACT

Hierarchical clustering was applied to a dataset containing aggregated statistics for 731 NBA players from the 2024–2025 season. The goal was to uncover performance profiles without predefined labels, using an unsupervised, distance-based approach. We considered quantitative variables capturing offensive output (points, rebounds, assists), usage/volume (minutes played, field-goal and free-throw attempts), and defensive contributions (steals, blocks), together with indicators such as turnovers and personal fouls. After preprocessing to make variables comparable across scales (normalization), a dissimilarity matrix was computed and a dendrogram was constructed; the final partition was selected by cutting the hierarchy using separation and internal consistency criteria. The analysis yielded three clearly differentiated groups: a high-usage, high-production cluster; a role/support cluster with lower usage and more balanced contributions; and an intermediate cluster characterized by moderate involvement paired with relatively efficient production. Overall, the resulting segmentation supports player comparison, summarizes dominant season patterns, and provides a practical baseline for downstream tasks such as roster construction, and decision support.

Keywords: clustering; NBA; dissimilarity measures; segmentation; unsupervised learning.

1. Introducción

El presente trabajo aplica clustering jerárquico a estadísticas agregadas de 731 jugadores de la NBA correspondientes a la temporada 2024–2025, con el objetivo de identificar perfiles de desempeño sin recurrir a etiquetas previas. En analítica deportiva, la segmentación de jugadores permite sintetizar grandes volúmenes de información (por ejemplo, volumen de uso, eficiencia y contribuciones defensivas) en un conjunto reducido de "tipos" o perfiles comparables. Esto resulta útil para apoyar procesos de exploración, comparaciones de rol, evaluación de compatibilidad entre jugadores y generación de hipótesis para análisis posteriores.

El análisis se implementó en IBM SPSS Statistics versión 27, utilizando el procedimiento de Hierarchical Cluster (IBM Corp., 2021). Esta elección ofrece dos ventajas prácticas. Primero, permite definir explícitamente los componentes esenciales del clústering basado en distancias: (i) la estandarización de variables, (ii) la medida de disimilaridad y (iii) el criterio de vinculación (o método de agregación). Segundo, entrega salidas orientadas a la interpretación (matriz de proximidades, historial de aglomeración y dendrograma), facilitando justificar el corte final del diagrama y describir con claridad cómo se formaron los grupos (IBM Corp., s. f.).

2. Materiales y métodos

2.1. Datos, variables y criterio de análisis

El conjunto de datos se basa en estadísticas de temporada y considera variables de volumen y participación (por ejemplo, partidos jugados, minutos jugados e intentos de tiro), de producción (puntos, rebotes, asistencias) y de acciones defensivas y disciplina (robos, tapones, pérdidas y faltas).

Al tratarse de estadísticas agregadas, la lectura de un perfil debe entenderse como un indicador general de rendimiento a lo largo de la

temporada. En ese sentido, el objetivo no es predecir un resultado, sino organizar los casos (jugadores) en grupos relativamente homogéneos, de modo que los miembros de un mismo clúster presenten patrones estadísticos similares y difieran, en promedio, de los jugadores en otros clústeres.

En clústering jerárquico aglomerativo, cada jugador inicia como un clúster propio y el algoritmo va fusionando pares de clústeres sucesivamente hasta llegar a un solo grupo. La estructura resultante es una jerarquía que puede “cortarse” en diferentes niveles para obtener distintas particiones. En la práctica, esto permite comparar alternativas: por ejemplo, una solución con 2 clústeres podría separar “alto uso” vs. “bajo uso”, mientras que 3 o 4 clústeres pueden capturar matices adicionales (Miyamoto, 2022).

3. Metodología y Obtención de datos

Los datos corresponden a estadísticas de temporada de la NBA 2024–2025 y fueron extraídos de la tabla de totales de jugadores publicada en Basketball-Reference, en la página NBA (Basketball-Reference, 2025).

El análisis se desarrolló en IBM SPSS Statistics versión 27 mediante el procedimiento Hierarchical Cluster (IBM Corp., 2021). Para garantizar comparabilidad entre variables, se aplicó estandarización por variable (puntuaciones z). La disimilaridad entre jugadores se calculó con la Distancia Euclidiana Cuadrática y el criterio de vinculación utilizado fue Complete Linkage. La solución final se definió inspeccionando el dendrograma y el historial de aglomeración, seleccionando un corte que produjo tres clústeres interpretables.

3.1. Estandarización por variable

Dado que las variables de entrada se expresan en escalas distintas (conteos, minutos y porcentajes), se realizó una estandarización por variable (puntuaciones z). Sin este paso, variables con rango numérico grande (como

minutos jugados o puntos) tienden a dominar la medida de distancia, reduciendo el aporte de indicadores de eficiencia o defensa. La estandarización hace que cada variable tenga, aproximadamente, media cero y desviación estándar unitaria, permitiendo que el clústering refleje patrones multivariados y no sólo diferencias en magnitud. Asimismo, esta decisión facilita interpretar los perfiles como combinaciones relativas de fortalezas y debilidades, en lugar de niveles absolutos.

Adicionalmente, la preparación de datos implicó verificar la coherencia de los tipos de variable, revisar rangos plausibles y considerar el tratamiento de valores faltantes. En análisis basados en distancia, la presencia de datos faltantes puede distorsionar la matriz de proximidades si no se controla adecuadamente; por ello, se trabajó con una configuración consistente con el cálculo de distancias en SPSS y con el objetivo de preservar la comparabilidad entre casos (IBM Corp., 2021).

Inicialmente los datos cargados conforman una tabla compuesta de 569 filas por 31 variables, de las cuales se eliminaron variables categóricas y algunas no relevantes para el estudio, por lo que se procede a seleccionar a las variables numéricas para aplicar agrupamientos (ver Tabla1).

Tabla 1. Descripción y análisis de variables seleccionadas.

Nombre de Variable	Descripción
G	Partidos jugados.
GS	Partidos como titular.
MP	Minutos jugados.
FG	Canastas de campo.
FGA	Intentos de tiro de campo.
3P	Triples anotados.
3PA	Intentos de triple.
2P	Dobles anotados.
2PA	Intentos de doble.
eFG%	Porcentaje efectivo de tiro. Ajusta por el hecho de que un tiro de 3 puntos vale un punto más que uno de 2 puntos.
FT	Tiros libres anotados.
FTA	Intentos de tiro libre.
ORB	Rebotes ofensivos.
DRB	Rebotes defensivos.
TRB	Rebotes totales.
AST	Asistencias.
STL	Robos.
BLK	Tapones.
TOV	Pérdidas.
PF	Faltas personales.
PTS	Puntos.

Fuente: (Basketball-Reference, 2025).

Se observa en la Tabla 1, que es necesario la estandarización de las variables a fin de evitar sesgos en el análisis final debido a sus diferentes unidades.

3.2 Medida de disimilaridad: Distancia Euclidiana Cuadrática

Como medida de distancia se utilizó la Distancia Euclidiana Cuadrática. Conceptualmente, esta medida suma, a través de las variables estandarizadas, las diferencias al cuadrado entre dos jugadores.

Esto tiene dos implicaciones: primero, resalta discrepancias grandes (las penaliza más fuertemente) y segundo, mantiene una interpretación geométrica directa en el espacio de características. En un contexto deportivo, esta propiedad puede ser útil para separar con mayor nitidez a jugadores con roles muy distintos (por ejemplo, especialistas defensivos vs. principales generadores ofensivos). En SPSS, la elección de esta distancia define la matriz de proximidades sobre la cual opera el procedimiento jerárquico (IBM Corp., 2021)

Además, como una estrategia de validación adicional, se replicó el cálculo de disimilaridad con otra medida de distancia, la distancia euclidiana no cuadrática, manteniendo la estandarización por variable y el mismo método de vinculación que se detalla en el punto siguiente, y al comparar la partición en clústeres no se observó diferencias grandes respecto a la solución original, indicando posiblemente que la estructura de los datos está dominada por patrones robustos de magnitud/volumen (como: minutos jugados y conteos acumulados) que permanecen coherentes bajo transformaciones razonables de disimilaridad o debido a variables altamente correlacionadas (MP con PTS, FGA o FG), por lo que distintas distancias tienden a preservar el orden relativo de similitud entre jugadores, generando dendrogramas y cortes comparables.

3.3 Método de agrupamiento: Complete Linkage

El método jerárquico seleccionado fue Complete Linkage (vecino más lejano). Este criterio define la distancia entre dos clústeres como la máxima distancia entre cualquier par de jugadores pertenecientes a clústeres distintos. En comparación con otros esquemas de vinculación, complete linkage tiende

a producir clústeres más compactos, ya que evita fusionar grupos si existe algún par de casos excesivamente distante. En aplicaciones con estadísticas de jugadores, esta característica favorece la interpretación: los grupos resultantes suelen tener menor dispersión interna y, por tanto, describirse con mayor claridad (IBM Corp., s. f.; Miyamoto, 2022).

Desde una perspectiva práctica, este método ayuda a mitigar el riesgo de formar clústeres “alargados” que mezclen perfiles muy diferentes. Por ejemplo, si un clúster agrupa a jugadores de rol con baja participación, complete linkage tiende a impedir que el grupo absorba a un anotador de alto uso cuya distancia máxima con algunos miembros del clúster sea grande. Esta propiedad puede
resultar especialmente valiosa cuando se busca construir tipologías de rol

3.4 Corte del dendrograma y solución de tres clústeres

El algoritmo aglomerativo produce una jerarquía completa; por tanto, el número final de clústeres se decide mediante un corte. En este trabajo se inspeccionaron el dendrograma y el historial de aglomeración para identificar saltos notables en el nivel de fusión (incrementos pronunciados en la disimilaridad), los cuales suelen indicar que, a partir de cierto punto, las fusiones comienzan a combinar grupos ya bastante diferentes. Con base en esta evidencia se seleccionó una solución de tres clústeres de casos/individuos. Este resultado ofrece un compromiso entre simplicidad e interpretabilidad: permite distinguir perfiles generales (por ejemplo, alto volumen ofensivo, rol/soporte y un grupo intermedio) sin perder completamente la heterogeneidad natural de la liga.

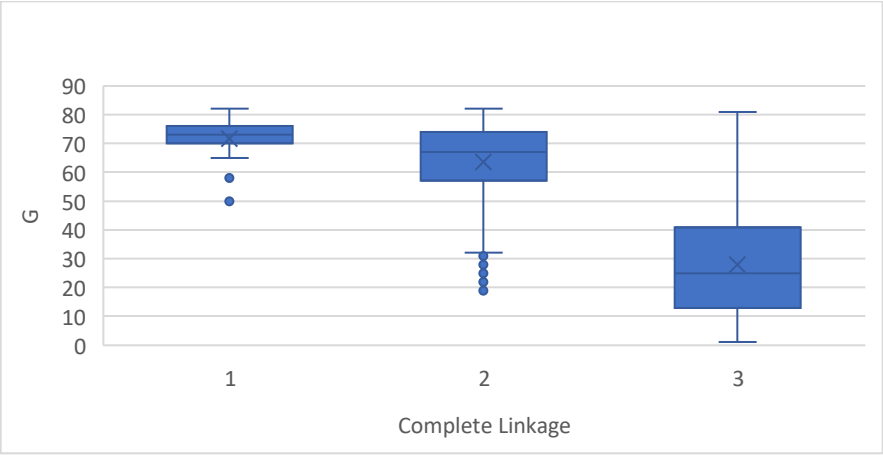
Finalmente, es importante destacar que los clústeres no deben interpretarse como “categorías verdaderas” o inmutables, sino como una segmentación dependiente de las variables incluidas, la estandarización, la distancia y el criterio de vinculación. Por ello, el clústering jerárquico debe en-

tenderse como una herramienta exploratoria: ayuda a organizar los datos y generar interpretaciones, pero sus conclusiones deben contrastarse con conocimiento del dominio y, cuando sea pertinente, con análisis complementarios. En términos metodológicos, la posibilidad de evaluar alternativas (distintas distancias o vinculaciones) y comparar la estabilidad de los grupos forma parte de las buenas prácticas en clústering jerárquico (Dhulipala et al., 2021).

4. Resultados

Se presentan los siguientes resultados en bloxplots, para comparar de manera gráfica y práctica los diferentes grupos de jugadores acorde a las variables analizadas. Los gráficos no son de todas las variables para no alargar la extensión de este trabajo.

Figura 1. Bloxplot de los conglomerados, según variable G.

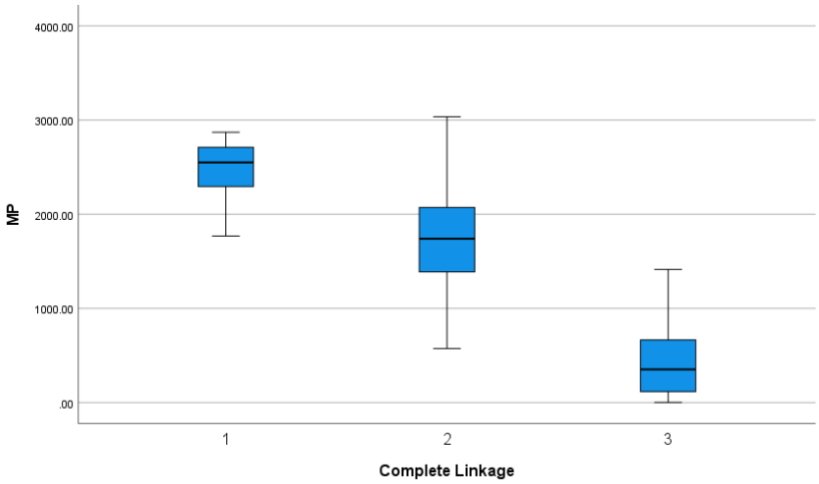


Fuente: Autor (2026).

De la figura 1, G (Partidos jugados). El clúster 1 presenta el mayor promedio de partidos jugados (71.65), seguido del clúster 2 (63.54), mientras que el clúster 3 registra una participación significativamente menor (28.03).

Esto sugiere que el primer grupo agrupa jugadores con alta disponibilidad y rol estable a lo largo de la temporada, el segundo a jugadores de rotación con presencia frecuente pero más variable, y el tercero a jugadores con participación esporádica (por ejemplo, contratos de corto plazo, lesiones o rol situacional)

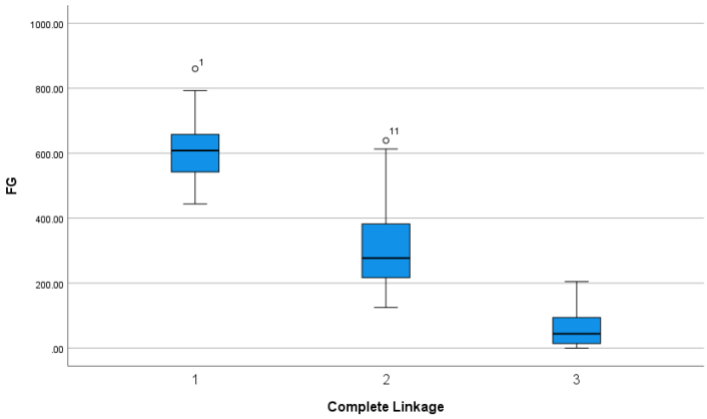
Figura 2. Bloxplot de los conglomerados, según variable MP.



Fuente: Autor (2026).

En la figura 2, MP (Minutos jugados). La diferencia entre clústeres se amplifica en minutos: clúster 1 promedia 2474.48 MP, clúster 2 1724.49 MP y clúster 3 428.67 MP. En conjunto con G, esto indica que la separación principal de la segmentación está asociada al volumen de participación (tiempo en cancha), lo que a su vez impacta los totales acumulados en el resto de las variables.

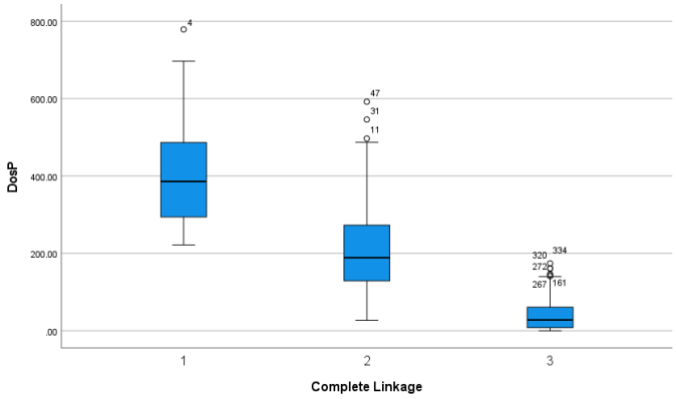
Figura 3. Bloxplot de los conglomerados, según variable FG.



Fuente: Autor (2026).

De figura 3, FG (Canastas de campo). El clúster 1 alcanza 610.91 FG en promedio, aproximadamente el doble del clúster 2 (306.77) y muy por encima del clúster 3 (58.95). Este patrón es consistente con roles ofensivos más centrales en el clúster 1, una contribución ofensiva intermedia en el clúster 2 y una producción limitada en el clúster 3, explicada principalmente por el bajo tiempo de juego.

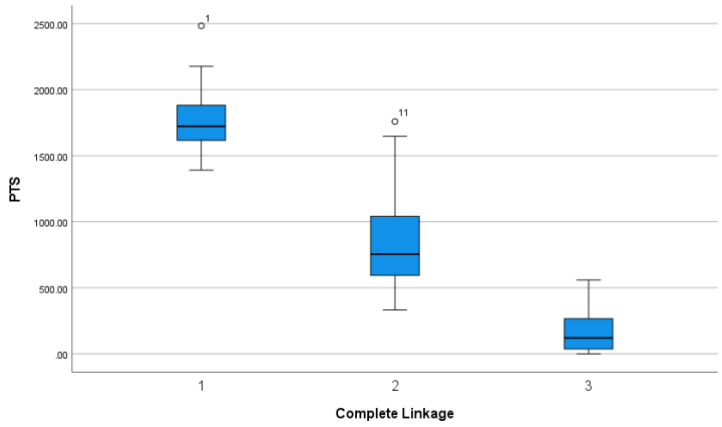
Figura 4. Bloxplot de los conglomerados, según variable DosP.



Fuente: Autor (2026).

De Figura 4, DosP (Dobles anotados). En dobles, el clúster 1 (412.65) supera al clúster 2 (210.00) y al clúster 3 (38.82), lo cual sugiere que la anotación cerca del aro o en tiros de dos puntos está fuertemente asociada al volumen general del jugador. La brecha entre clústeres respalda la interpretación de que el clúster 1 concentra perfiles con mayor carga ofensiva total.

Figura 5. Bloxplot de los conglomerados, según variable PTS.



Fuente: Autor (2026).

De la figura 5, PTS (Puntos). La variable PTS muestra la separación más evidente: clúster 1 promedia 1768.96 puntos, clúster 2 831.54 y clúster 3 159.30. Este resultado confirma que los clústeres capturan diferencias sustantivas en producción ofensiva acumulada, y que la variable de volumen (MP) actúa como un determinante estructural de los totales

5. Discusión: normalización por minutos jugados

Dado que varias de las variables analizadas son totales acumulados o variables basadas en el volumen (PTS, FG, FGA, TRB, AST), una parte importante de las diferencias entre clústeres puede explicarse simplemente por el tiempo en cancha (MP).

Para complementar la interpretación basada en volumen, se podría incorporar indicadores normalizados por minutos, típicamente expresados por 36 minutos según la NBA (ej., PTS/36, FG/36, AST /36). Esta normalización permite distinguir con mayor claridad entre (i) jugadores que producen mucho porque juegan mucho y (ii) jugadores que, aun con menor participación, muestran alta productividad relativa.

En el contexto de esta segmentación, la normalización por minutos podría ajustar la lectura del clúster 3: aunque sus totales son bajos, algunos casos podrían presentar tasas por minuto comparables a las de los clústeres 1 y 2, lo que sugeriría eficiencia en muestras de tiempo reducidas. De manera similar, en el clúster 2 podrían identificarse perfiles con alta producción relativa que no se reflejan plenamente en los totales acumulados. Ergo, analizar simultáneamente variables basadas en volumen y variables normalizadas por tiempo ayuda a separar rol de productividad o eficiencia, y puede mejorar la formación de los conglomerados cuando existen grandes diferencias en minutos jugados.

6. Conclusiones

Los resultados del clustering jerárquico (complete linkage) evidencian tres perfiles bien diferenciados de jugadores en la temporada 2024–2025 a partir de variables de volumen, producción y contribución defensiva.

El Clúster 1 concentra a jugadores de alta disponibilidad y carga de minutos (G=71.65; MP=2474), con altos volúmenes ofensivos (FGA=1294; PTS=1769) y elevada participación en la creación de juego (AST=465). Este

grupo puede interpretarse como el de principales piezas de rotación y alto impacto, con producción sostenida y una eficiencia (eFG%=0.549) similar a la del clúster intermedio.

El Clúster 2 refleja un perfil de rotación con participación moderada (G=63.54; MP=1724) y una producción ofensiva intermedia (PTS=832; FGA=650). A pesar de menor volumen que el clúster 1, presenta valores competitivos en rebote (TRB=322) y una eficiencia comparable (eFG%=0.548), sugiriendo jugadores de soporte que contribuyen de forma equilibrada.

El Clúster 3 agrupa a jugadores con baja participación o uso esporádico (G=28.03; MP=429; GS=4.73), con producción y volumen reducidos (PTS=159; FGA=130) y menores contribuciones acumuladas en asistencias, robos y tapones. La mayor variabilidad relativa (DE altas respecto a la media) sugiere heterogeneidad interna asociada a jugadores con pocos minutos, roles muy específicos o estancias parciales en la temporada.

En conjunto, la segmentación obtenida respalda que las variables analizadas capturan principalmente diferencias de volumen de participación y carga ofensiva, mientras que la eficiencia (eFG%) muestra menor separación entre los dos clústeres con mayor uso. Estos clústeres facilitan la comparación de perfiles, la exploración de roles y la formulación de análisis posteriores (por ejemplo, caracterización detallada por posición o normalización por minuto) para refinar la interpretación del rendimiento.

6. Referencias

Basketball-Reference. (2025). NBA player totals — 2024–25 season. Recuperado el 3 de febrero de 2026, de https://www.basketball-reference.com/leagues/NBA_2025_totals.html

Dhulipala, L., Eisenstat, D., Lącki, J., Mirrokni, V., & Shi, J. (2021). Hierarchical agglomerative graph clustering in nearly-linear time (arXiv:2106.05610).

IBM Corp. (2021). IBM SPSS Statistics Algorithms (Version 27). IBM. https://public.dhe.ibm.
com/software/analytics/spss/documentation/statistics/27.0/en/client/Manuals/IBM_
SPSS_Statistics_Algorithms.pdf

IBM Corp. (s. f.). Hierarchical cluster analysis method. IBM Documentation. Recuperado el 3 de febrero de 2026, de https://www.ibm.com/docs/en/spss-statistics/cd?topic=analysis-hierarchical

Miyamoto, S. (2022). Theory of agglomerative hierarchical clustering. Springer.