

Cluster Analysis: Theoretical Development and Application to Socioeconomic Data of Ecuador

Jennifer Avilés ; Ysaí Ronquillo ; Jorge Chamba 

Facultad de Ciencias Naturales y Matemáticas, Escuela Superior Politécnica del Litoral (ESPOL), Ecuador,

jaaviles@espol.edu.ec, yronquil@espol.edu.ec, jechamba@espol.edu.ec

Abstract– *This study proposes a socioeconomic vulnerability typology for Ecuadorian parishes based on official data from the 2022 Population and Housing Census. A total of 308 parishes are analyzed using 17 socioeconomic and basic services indicators, previously standardized. Hierarchical cluster analysis is applied using Ward's method with Manhattan distance, evaluating multiple linkage criteria and distance metrics. The optimal number of clusters is determined through the agglomerative coefficient, silhouette width, and dendrogram inspection, resulting in a robust five-cluster solution. The validity of the classification is further assessed through a comparison with the K-means algorithm, achieving a 93% agreement between both clustering approaches. The resulting clusters reveal distinct territorial patterns in socioeconomic conditions and access to basic infrastructure, enabling the identification of priority areas for targeted public policy interventions. These findings demonstrate the usefulness of cluster analysis as a data-driven tool for territorial planning and decision-making, providing a reproducible and adaptable methodological framework applicable to other geographic and socioeconomic contexts.*

Keywords– *Cluster Analysis, Ward's Method, Socioeconomic Vulnerability, Ecuador Census 2022, Public Policy.*

Análisis Cluster: Desarrollo Teórico y Aplicación a Datos Socioeconómicos del Ecuador

Jennifer Avilés ; Ysaí Ronquillo ; Jorge Chamba 

Facultad de Ciencias Naturales y Matemáticas, Escuela Superior Politécnica del Litoral (ESPOL), Ecuador,
jaaviles@espol.edu.ec, yronquil@espol.edu.ec, jechamba@espol.edu.ec

Resumen– El presente trabajo expone un estudio detallado de análisis cluster aplicado a datos socioeconómicos del Censo Nacional de Ecuador 2022. Se analizan 308 parroquias mediante 17 indicadores socioeconómicos y de acceso a servicios básicos, previamente estandarizados. Se emplea análisis de conglomerados jerárquico evaluando múltiples configuraciones de enlace y métricas de distancia, siendo el método de Ward con la distancia de Manhattan la configuración con la clasificación más óptima. La selección del número consistente de conglomerados se fundamenta en el coeficiente aglomerativo, el ancho de silueta y la inspección del dendrograma, identificándose una partición robusta de cinco clusters. La validez de los resultados se refuerza mediante una comparación con el algoritmo K-means, obteniéndose una concordancia promedio del 93% entre ambas soluciones. Los clusters resultantes revelan patrones territoriales diferenciados en términos de condiciones socioeconómicas y acceso a infraestructura básica, lo que permite identificar zonas prioritarias para la formulación de políticas públicas focalizadas. Los resultados evidencian el potencial del análisis de conglomerados como herramienta para el apoyo a la planificación territorial y la toma de decisiones basadas en datos, ofreciendo un marco reproducible y adaptable a otros contextos geográficos.

Palabras clave– Análisis Cluster, Método de Ward, Vulnerabilidad Socioeconómica, Censo Ecuador 2022, Política Pública.

I. INTRODUCCIÓN

La identificación de desigualdades socioeconómicas y territoriales constituye un desafío central para la planificación y formulación de políticas públicas en América Latina. En países con alta heterogeneidad territorial, como Ecuador, las condiciones de vida, el acceso a servicios básicos y las oportunidades socioeconómicas presentan marcadas diferencias entre regiones y unidades administrativas locales [16]. Estas disparidades dificultan la asignación eficiente de recursos y la correcta implementación de intervenciones focalizadas, especialmente cuando se carece de herramientas analíticas que sintetizan información multidimensional de manera objetiva y reproducible.

En este contexto, el análisis de clusters se ha consolidado como una técnica ampliamente utilizada para la clasificación de territorios a partir de múltiples indicadores socioeconómicos, permitiendo identificar patrones latentes y tipologías territoriales sin imponer estructuras a priori [1,4]. Diversos estudios han demostrado su utilidad en el análisis de pobreza multidimensional, desarrollo regional y planificación territorial; sin embargo, en el caso ecuatoriano, la mayoría de

las aplicaciones existentes se han concentrado en escalas agregadas, como provincias o cantones, o han considerado conjuntos limitados de variables, lo que restringe la capacidad de capturar heterogeneidades locales a escalas administrativas más finas como las parroquias, una división político-territorial de tercer nivel que representan el escalafón administrativo de menor rango dentro de la organización del estado.

Adicionalmente, una limitación recurrente en la literatura aplicada es la escasa validación de las soluciones de agrupamiento y la falta de transparencia en los criterios utilizados para la selección de variables, métricas de distancia y número óptimo de conglomerados. Estas debilidades metodológicas reducen la reproducibilidad de los resultados y limitan su utilidad como insumo técnico para la toma de decisiones basada en evidencia.

Motivado por estas brechas, el presente estudio tiene como objetivo principal identificar y caracterizar tipologías de vulnerabilidad socioeconómica a nivel parroquial en Ecuador, utilizando información oficial del VIII Censo de Población y VII de Vivienda 2022 del Instituto Nacional de Estadística y Censos (INEC) [5]. Para ello, se analizan 308 parroquias mediante un conjunto de 17 indicadores relacionados con condiciones socioeconómicas y acceso a servicios básicos. Se emplea un enfoque de análisis de conglomerados jerárquico, evaluando distintas configuraciones metodológicas y aplicando criterios cuantitativos para la selección del número óptimo de clusters, complementado con una validación mediante el algoritmo K-means.

La principal contribución de este trabajo radica en la construcción de una clasificación territorial robusta y reproducible a escala parroquial, que permite identificar patrones espaciales diferenciados de vulnerabilidad socioeconómica. Los resultados obtenidos constituyen un insumo técnico relevante para la planificación territorial y la priorización de intervenciones públicas, y ofrecen un marco metodológico adaptable a otros contextos geográficos y conjuntos de datos censales.

II. FUNDAMENTOS TEÓRICOS

El análisis cluster es una técnica multivariante de aprendizaje no supervisado cuyo objetivo es agrupar observaciones de manera que los elementos dentro de un mismo grupo sean lo más similares posible entre sí y, al mismo tiempo, lo más diferentes posible de los elementos

pertenecientes a otros grupos [1,4]. A diferencia de los métodos de clasificación supervisada, el análisis de conglomerados no requiere etiquetas previas, lo que lo convierte en una herramienta especialmente adecuada para la exploración de estructuras latentes en datos socioeconómicos multidimensionales.

Dentro de las técnicas de agrupamiento, los métodos jerárquicos permiten construir una estructura anidada de conglomerados que puede representarse mediante un dendrograma, facilitando la interpretación del proceso de agrupamiento y la selección del número de clusters [2,4]. Estos métodos se caracterizan por no requerir la especificación previa del número de grupos y por ofrecer una visión global de las relaciones de similitud entre las observaciones. En particular, el algoritmo de agrupamiento jerárquico aglomerativo de Ward emplea un criterio de optimización basado en la minimización de la varianza intra-grupo en cada iteración de fusión. Al priorizar la reducción de la varianza intragrupo (heterogeneidad interna), el método maximiza la homogeneidad local, lo que resulta en una partición del espacio de datos caracterizada por conglomerados compactos [6].

La medida de similitud o distancia desempeña un rol fundamental en el desempeño de los algoritmos de clustering. En aplicaciones socioeconómicas, donde los datos suelen presentar escalas heterogéneas y distribuciones asimétricas, el uso de métricas robustas como la distancia Manhattan ha mostrado un buen desempeño al reducir la influencia de valores extremos [2,10]. Para garantizar la comparabilidad entre variables, es práctica común aplicar procesos de estandarización previa, como la normalización de *z-scores*.

La evaluación de la calidad de una solución de agrupamiento constituye un aspecto clave del análisis. Entre los criterios internos más utilizados se encuentran el coeficiente aglomerativo y el ancho de silueta, este último propuesto por Rousseeuw [7], que permite evaluar la coherencia interna de los clusters y su separación relativa. Adicionalmente, la comparación entre diferentes algoritmos de agrupamiento jerárquico y el algoritmo K-means [3], contribuye a validar la estabilidad y robustez de la clasificación obtenida.

III. DATOS Y VARIABLES

El presente estudio utiliza información oficial proveniente del VIII Censo de Población y VII de Vivienda 2022, elaborado por el Instituto Nacional de Estadística y Censos del Ecuador (INEC) [5]. El censo constituye la fuente más reciente y completa de información demográfica, socioeconómica y habitacional a nivel nacional, lo que garantiza la representatividad y actualidad de los datos empleados en el análisis.

La unidad de análisis corresponde a las parroquias, considerada como el nivel político administrativo más desagregado del país para el cual se dispone de información censal homogénea y sistemática. En total, se analizaron 308 parroquias pertenecientes a provincias claves del territorio ecuatoriano, excluyéndose aquellas unidades administrativas con información incompleta o inconsistencias en los registros, a fin de asegurar la calidad del conjunto de datos utilizado.

A partir de la base censal, se seleccionó un conjunto de 17 indicadores socioeconómicos y de acceso a servicios básicos, diseñados para capturar de manera integral la calidad de vida y vulnerabilidad territorial. Estos indicadores incluyen variables relacionadas con condiciones de vivienda, disponibilidad de infraestructura básica, características demográficas y aspectos socioeconómicos de la población. La selección de variables se fundamentó en criterios de relevancia teórica, disponibilidad de información y uso recurrente en estudios de caracterización territorial y análisis de desigualdad socioeconómica.

Previo a la aplicación de las técnicas de agrupamiento, los indicadores fueron sometidos a un proceso de preprocesamiento que incluyó la revisión de valores atípicos y la estandarización mediante el popular método de normalización de la puntuación Z (*z-scores*), de manera que las variables tengan una media de 0 y una desviación estándar de 1. Este procedimiento permitió garantizar que todos los datos tengan la misma escala y evita que aquellas variables con mayor varianza dominen el proceso de clustering.

La Tabla 1 presenta el conjunto de indicadores considerados en el estudio, junto con su definición conceptual y estadísticos descriptivos básicos. Esta estructura de datos proporciona la base para la identificación de patrones latentes y la posterior construcción de tipologías territoriales de vulnerabilidad socioeconómica.

TABLA I
INDICADORES SELECCIONADOS PARA EL ANÁLISIS

Dimensión	Indicadores Representativos
Educación	Tasa de analfabetismo, Población sin educación básica, Población con educación superior
Servicios Básicos	Viviendas sin agua por red pública, Sin alcantarillado, Sin recolección de basura
Vivienda	Déficit cualitativo, Déficit cuantitativo, Hacinamiento, Materiales precarios
Conectividad	Hogares sin internet, Hogares sin computadora
Economía	Tasa de dependencia demográfica, Pobreza por NBI (Necesidades Básicas Insatisfechas)

IV. METODOLOGÍA

Con el objetivo de identificar tipologías territoriales de vulnerabilidad socioeconómica, se aplicó un enfoque de análisis de clusters basado en métodos jerárquicos y no

jerárquicos. El procedimiento metodológico se estructuró en tres etapas principales: (i) definición de la medida de disimilitud y el método de agrupamiento, (ii) determinación del número óptimo de clusters y (iii) validación de la solución obtenida.

A. Análisis de clusters jerárquico

En una primera etapa, se empleó el análisis de clusters jerárquico aglomerativo, el cual permite construir una estructura progresiva de agrupamiento sin requerir la especificación previa del número de clusters. Entre los distintos métodos de enlace evaluados, se seleccionó el método de Ward, dado que minimiza el incremento de la varianza intra-grupo en cada etapa de fusión, favoreciendo la obtención de conglomerados compactos y homogéneos [6].

Como medida de disimilitud se utilizó la distancia Manhattan, ampliamente recomendada en contextos socioeconómicos debido a su robustez frente a valores extremos y su adecuada interpretación cuando se trabaja con variables estandarizadas [2,10]. La combinación del método de Ward con esta métrica fue evaluada frente a otras alternativas, seleccionando aquella que presentó un mayor coeficiente aglomerativo, indicador que mide el grado de cohesión interna de la estructura jerárquica.

El proceso de agrupamiento se representó mediante un dendrograma, el cual permitió analizar visualmente las posibles particiones del conjunto de datos y orientar la selección del número de conglomerados.

B. Selección del número óptimo de clusters

La determinación del número óptimo de clusters se realizó a partir de un enfoque combinado que integra criterios cuantitativos y análisis gráfico. En particular, se consideraron el coeficiente aglomerativo, el ancho promedio de la silueta y la inspección del dendrograma. El índice de silueta, propuesto por Rousseeuw [7], evalúa simultáneamente la cohesión interna y la separación entre clusters, proporcionando una medida global de la calidad de la partición.

A partir de estos criterios, se identificó una solución de cinco clusters como la más adecuada para representar la estructura subyacente de los datos, al ofrecer un balance favorable entre interpretabilidad, cohesión interna y diferenciación entre grupos.

C. Validación de la solución de agrupamiento

Con el fin de evaluar la robustez de la clasificación obtenida mediante el método jerárquico, se realizó una validación cruzada utilizando el algoritmo K-means [3]. Para este propósito, se fijó el número de clusters en cinco y se compararon las asignaciones resultantes con las obtenidas a partir del método de Ward.

La concordancia entre ambas soluciones fue cuantificada mediante el porcentaje de coincidencia en la asignación de las

parroquias a los clusters, alcanzándose un nivel de acuerdo del 93%. Este resultado sugiere una alta estabilidad de la estructura de agrupamiento identificada y refuerza la consistencia de las tipologías territoriales obtenidas.

Todos los análisis fueron realizados utilizando el software estadístico R. R (v4.3.2; R Core Team 2023) [15], empleando paquetes especializados para análisis multivariado y visualización de resultados. Este enfoque metodológico garantiza la reproducibilidad del estudio y facilita su adaptación a otros contextos territoriales o conjuntos de datos censales.

V. RESULTADOS Y DISCUSIÓN

A. Selección del Mejor Método

La comparación de los coeficientes aglomerativos reveló diferencias significativas entre los métodos. El método de Ward con distancia Manhattan obtuvo el coeficiente más alto (0.98), lo que indica una estructura de agrupamiento muy fuerte. En contraste, el enlace promedio con distancia euclidiana mostró un desempeño inferior (0.80), tendiendo a generar el efecto de 'encadenamiento' donde las observaciones se añaden una a una a un gran grupo, sin revelar estructura real.

TABLA II
COMPARACIÓN DE COEFICIENTES AGLOMERATIVOS

Método de Enlace	Distancia	Coef. Aglomerativo
Ward	Manhattan	0.98 (Seleccionado)
Promedio	Euclidiana	0.80

Basado en estos resultados, se procedió con la solución Ward-Manhattan. El análisis del dendrograma (Fig. 1) muestra saltos grandes en la distancia de fusión, sugiriendo un corte natural.

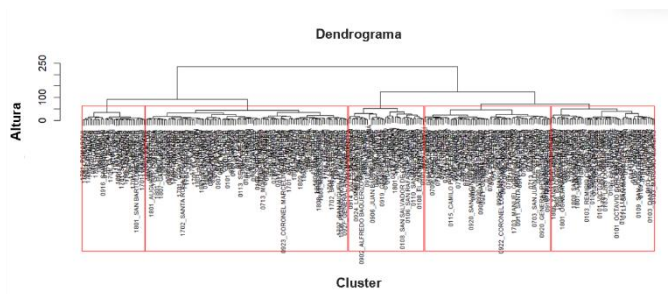


Fig. 1. Dendrograma del agrupamiento jerárquico (Método Ward, Distancia Manhattan).

El método del codo (Fig. 2) corrobora esta decisión, mostrando que la varianza intra-cluster deja de disminuir significativamente después de $k = 5$.

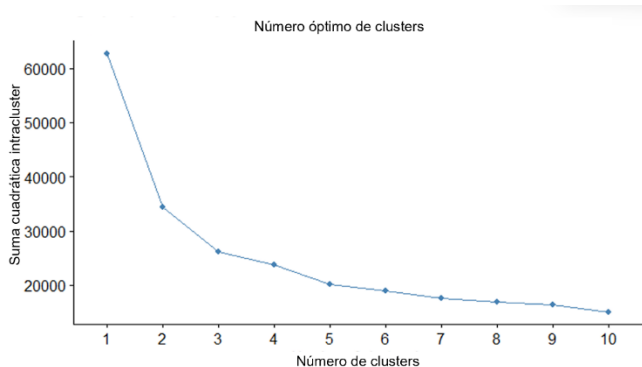


Fig. 2 Gráfico del Codo (Elbow Method) para la determinación de k.

B. Caracterización de los Clusters

La segmentación en 5 clusters revela una estratificación socioeconómica clara:

Cluster 1: Nivel Medio-Alto (La clase media emergente).

Este es el grupo más grande (109 parroquias). Se caracteriza por tener indicadores socioeconómicos positivos, aunque no excelentes. En general, las parroquias de este grupo tienen una cobertura de servicios básicos alta, así como ventajas en condición de vivienda y empleo respecto, pero existen brechas en la educación de nivel superior y acceso a tecnologías avanzadas. Representa a las zonas periurbanas consolidadas y cabeceras cantonales secundarias.

Cluster 2: Alta Precariedad (El cinturón de pobreza).

Conformado por 37 parroquias, este es el grupo más crítico. Las parroquias de este grupo presentan los peores indicadores en casi todas las dimensiones, analfabetismo superior al 15%, un mayor porcentaje de viviendas con piso de tierra o materiales precarios, y falta crítica de alcantarillado. Geográficamente, incluye zonas marginales y rurales aisladas de la sierra y costa.

Cluster 3: Nivel Alto (La élite urbana).

Compuesto por 38 parroquias. Aquí se concentran parroquias con los mejores indicadores del país: cobertura completa (o cercana al 100%) de servicios eléctricos y de agua, mejores condiciones de vivienda, mayor acceso a tecnología y empleo, analfabetismo cercano a cero y alta tasa de profesionales universitarios. Corresponde a las parroquias urbanas centrales de Quito (ej. Iñaquito) y zonas exclusivas de Guayaquil (ej. Samborondón satélite).

Cluster 4: Vulnerabilidad Rural (Déficit de Infraestructura y Educación).

Agrupar a 56 parroquias. Su característica distintiva no es la pobreza monetaria extrema, sino la falta de acceso a la educación. Tienen viviendas aceptables, pero carecen de

alcantarillado y recolección de basura, dependiendo de pozos sépticos y quema de residuos, la mayor parte de los hogares no disponen de equipos de cómputo. La característica de este grupo es típica de la ruralidad productiva agrícola.

Cluster 5: Transición (Vulnerabilidad Habitacional).

Incluye 68 parroquias. Las parroquias de este grupo tienen un elevado porcentaje de viviendas consideradas como no aceptables, así como de viviendas sin acceso a red pública de agua, además de ser uno de los grupos con mayores casos de hacinamiento. Son zonas densamente pobladas en proceso de urbanización desordena.

TABLA III
ESTADÍSTICAS DESCRIPTIVAS PROMEDIO POR CLUSTER (Z-SCORES)

Variable	Cluster 1	Cluster 2	Cluster 3
Analfabetismo	-0.5	2.1	-1.2
Sin Alcantarillado	-0.2	1.8	-1.5
Hacinamiento	0.1	1.5	-0.9
Sin Internet	-0.3	1.2	-1.8
Dependencia	0.0	1.1	-1.0

C. Validación y Comparación

Para verificar la robustez de estos resultados, se ejecutó el algoritmo K-means fijando $k = 5$. La matriz de confusión (Tabla IV) muestra una coincidencia del 93%.

TABLA IV
MATRIZ DE CONFUSIÓN (WARD VS K-MEANS)

Coincidencia	Cantidad
Parroquias coincidentes	285 (93%)
Parroquias reasignadas	23 (7%)

Las discrepancias se encontraron principalmente en el Cluster 5 (Transición), que K-means tendió a dividir entre los clusters 1 y 4. Esto sugiere que el Cluster 5 es el más variable y representa una zona gris en la estratificación social.

Se presenta (Fig. 3) el agrupamiento de parroquias por cluster de las provincias representativas del Ecuador (Guayas, Pichincha, Azuay, El Oro y Tungurahua) con la aplicación del enlace de Ward.

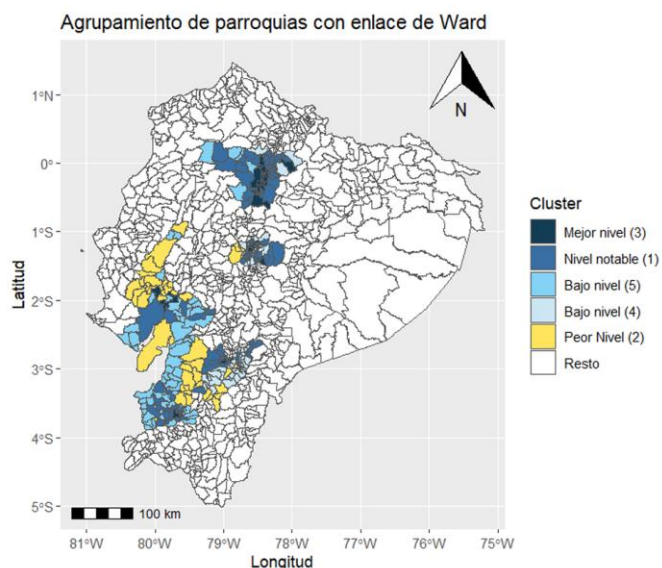


Fig. 3. Mapa de distribución espacial de los clusters en Ecuador.

VI. CONCLUSIONES Y TRABAJO FUTURO

Este estudio ha demostrado la utilidad del análisis de clusters como una herramienta eficaz para desglosar la complejidad socioeconómica del Ecuador a partir de la información del Censo de Población y Vivienda 2022. Los resultados confirman que el uso del método de Ward combinado con la distancia Manhattan permite obtener tipologías territoriales coherentes y estables en contextos censales caracterizados por alta variabilidad y la presencia de valores atípicos, al minimizar la varianza interna de los conglomerados y garantizar una mayor robustez del proceso de agrupamiento.

La identificación de cinco clusters socioeconómicos diferenciados evidencia que las condiciones de vulnerabilidad no son homogéneas en el territorio y que las políticas públicas de carácter universal resultan, en muchos casos, ineficientes. Desde una perspectiva aplicada, los resultados permiten avanzar hacia un enfoque de políticas focalizadas, ajustadas a las necesidades específicas de cada perfil territorial. En particular, el Cluster 2 presenta un nivel de rezago multidimensional que sugiere la necesidad de intervenciones de emergencia en ámbitos como transferencias monetarias, condiciones de vivienda y alfabetización. El Cluster 4 requiere prioritariamente inversión en infraestructura sanitaria, especialmente en recolección de basura y alcantarillado, más que programas de asistencia económica directa, también es necesario mejorar el acceso a la educación superior. Por su parte, el Cluster 5 demanda estrategias orientadas al mejoramiento de vivienda y a la planificación urbana, con el objetivo de reducir los niveles de hacinamiento identificados, incrementar el acceso a la red pública de agua y mejorar la habitabilidad de las viviendas.

La validación realizada mediante el algoritmo K-means, que mostró una alta concordancia con la solución jerárquica, refuerza la estabilidad y consistencia de las tipologías obtenidas, consolidando su potencial como insumo técnico para la planificación territorial y la toma de decisiones basada en evidencia.

La coincidencia del 93% entre Ward y K-means confirma que las estructuras de vulnerabilidad en Ecuador están fuertemente polarizadas y son detectables independientemente del algoritmo utilizado, aunque los métodos jerárquicos demostraron mayor sensibilidad para capturar los clusters de transición (Cluster 5).

No obstante, el estudio presenta algunas limitaciones. El análisis se basa en datos transversales, lo que impide capturar la evolución temporal de la vulnerabilidad socioeconómica, y no incorpora de manera explícita restricciones espaciales en el proceso de agrupamiento. Como línea de trabajo futuro, se propone la inclusión de variables espaciales y métodos de clustering con restricciones de contigüidad, con el fin de garantizar la formación de regiones geográficamente continuas que faciliten la implementación y administración logística de los programas sociales.

REFERENCIAS

- [1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.
- [2] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, 5th ed. Chichester, UK: John Wiley & Sons, 2011.
- [3] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, 1967, pp. 281–297.
- [4] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley & Sons, 1990.
- [5] Instituto Nacional de Estadística y Censos (INEC), "VIII Censo de Población y VII de Vivienda 2022," [En línea]. Disponible: <https://www.ecuadorencifras.gob.ec>.
- [6] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.
- [7] P. 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.
- [8] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York: Springer, 2013.
- [9] A. C. Rencher and W. F. Christensen, *Methods of Multivariate Analysis*, 3rd ed. Wiley, 2012.
- [10] S. Saraçlı, N. Doğan, and I. Doğan, "Comparison of hierarchical cluster analysis methods by cophenetic correlation," *Journal of Inequalities and Applications*, vol. 2013, art. 203, 2013.
- [11] ECLAC, *Social Panorama of Latin America and the Caribbean 2024: Challenges of non-contributory social protection*, Santiago: United Nations, 2024.
- [12] J. Martínez and A. Aguilar, "Socioeconomic Urban Environment in Latin America: Towards a Typology of Cities," *Sustainability*, vol. 15, no. 8, p. 6380, 2023.

- [13]P. Ruiz-García et al., "Spatial Inequalities and the Sensitivity of Social Vulnerability in Ecuador," *Land*, vol. 14, no. 11, p. 2110, 2025].
- [14]S. Kumar et al., "Multidimensional Poverty Clustering using K-Means Algorithm with Dimensionality Reduction," ResearchGate Preprints, Jan. 2026.
- [15]R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2023. <https://www.R-project.org/>.
- [16]K. Esparza, K. Campoverde and R. Correa, "Índice de Desarrollo de las Regiones de Ecuador". *Revista Ecuatoriana y Política*, No. 37, pp. 21-26, 2023.