International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(16s) 2026 pp. 1014-1023
https://doi.org/10.70917/ijcisim-2026-4627
Article

Underlying Structure in the Clustering of Peruvian Departments According to Demographic Indicators, 2022

Leticia de los Ángeles Corella Prieto[1], Raúl Alfredo Sánchez Ancajima[2], Gaspar Chávez Dioses[3], Kelvin Howard Pizarro Romero[4], José Luis González Márquez[5]

[1] Graduate School, Universidad Nacional de Tumbes, Tumbes, Perú.
Email: lcorella@untumbes.edu.pe
ORCID: 0009-0006-5147-7185
[2] Universidad Nacional de Tumbes, Tumbes, Perú.
[3] Advisor, Universidad Nacional de Tumbes, Tumbes, Perú.
ORCID: 0000-0002-0096-3654
[4] Universidad Técnica de Machala, Ecuador.
Email: khpizarro2016@gmail.com
[5] Universidad Estatal Península de Santa Elena, Ecuador.
Corresponding Author: Leticia de los Ángeles Corella Prieto

Abstract: Territorial gaps in demographic indicators hinder the effective targeting of public policies in Peru. This study aimed to segment the 24 departments and the Constitutional Province of Callao (2022) to identify differentiated demographic profiles that support territorial planning. Following a quantitative, non-experimental, cross-sectional design, six standardized indicators were used: crude birth rate, total fertility rate, crude death rate, life expectancy at birth, infant mortality rate, and crude nuptiality rate. k-means (Lloyd, MacQueen, and Hartigan–Wong variants) and k-medoids (PAM) were applied, and cluster quality was assessed through average silhouette width, the elbow method, the gap statistic, and inter-algorithm stability (Jaccard index). Principal component analysis (PCA) supported interpretation (PC1 = 59.5 %; PC2 = 24.2 %; 83.8 % cumulative variance), and the resulting structure was contrasted with a Ward hierarchical dendrogram. The results converged on k = 2 as the main solution, given its higher silhouette (≈ 0.33), the absence of negative widths, and high replicability across algorithms; an advanced cluster (low birth and fertility rates, high life expectancy, very low infant mortality) and a lagging cluster with the opposite pattern emerged. The slightly higher crude mortality in the advanced cluster is consistent with population aging. As an operational disaggregation, k = 4 distinguished a metropolitan-southern pole, an Andean-Amazonian bloc facing greater challenges, and two intermediate transitional groups. ANOVA confirmed significant between-group differences in most indicators. The contribution is a reproducible, visualizable typology that strengthens territorial planning and the prioritization of resources for maternal-child health and basic services.
Keywords: cluster; dendrogram; demographic indicators; k-means; silhouette; Peru

1. INTRODUCTION

Demographic indicators are essential tools for quantitatively characterizing the social and health status of populations and for guiding government decision-making, since many health indicators are computed from demographic data (Pan American Health Organization, 2018). Phenomena such as fertility, birth, mortality, nuptiality, and migration are summarized in indicators that describe the intensity, dynamics, and evolution of populations (National Statistics Institute of Spain, 2021). However, a country's health situation and mortality rates are not homogeneous across its regional subpopulations, as documented in various Latin American cities and regions (Bilal et al., 2019; Queiroz et al., 2020).

Since 2015, the Latin American region has experienced rising extreme poverty and widening inequalities in income and access to health services, with negative effects on quality of life (Abramo et al., 2020). The demographic transition advances at uneven rates: the regional total fertility rate stood at around 1.85 births per woman in 2022, below replacement level (ECLAC, 2019), and the COVID-19 pandemic caused the largest worldwide loss of life expectancy at birth recorded in the region (United Nations, 2022). These dynamics coexist with marked subnational asymmetries.

Peru, with 24 departments and the Constitutional Province of Callao, eight natural regions, and remarkable cultural and climatic diversity, exhibits wide territorial gaps in social, economic, and health dimensions. Infant mortality rates in departments such as Ayacucho or Puno far exceed those of Lima or Arequipa, and life expectancy in the Amazon is lower than in the main cities (National Institute of Statistics and Informatics, 2020). Against this backdrop, grouping homogeneous departments into macro-regions—based on a compact set of key demographic indicators—would facilitate the design of differentiated policies and a more efficient allocation of resources, a strategy that in the European Union has enabled investment to be targeted at territories with similar socioeconomic characteristics (European Commission, 2020; Iammarino et al., 2019).

Cluster analysis (*clustering*) is an unsupervised learning technique that groups observations while maximizing internal homogeneity and between-group heterogeneity. Among partitional methods, the k-means algorithm and its Lloyd, MacQueen, and Hartigan–Wong variants are the most widely used (Morissette & Chartier, 2013), whereas k-medoids (PAM) offers robustness to outliers by relying on real medoids (Schubert & Rousseeuw, 2021). The internal quality of partitions is commonly evaluated with the silhouette coefficient (Rousseeuw, 1987), recognized—together with the Davies–Bouldin index—as one of the most informative measures for convex clusters (Chicco et al., 2025). Several recent studies have successfully applied the combination of PCA and clustering to classify territories according to demographic and socioeconomic profiles and to inform place-based policies (Soares et al., 2003; Pavone et al., 2020; Panzabekova et al., 2025; Ayberkin et al., 2026; Polyanskaya, 2025).

Despite this evidence, there is no reproducible typology of Peruvian departments built from basic demographic indicators and validated through multiple internal criteria. The present study addresses this gap, with the general objective of *uncovering the underlying structure in the clustering of Peruvian departments according to demographic indicators in 2022*. The specific objectives were: (a) to describe the indicators using measures of central tendency and dispersion; (b) to determine the optimal number of clusters with the elbow, average silhouette, and gap statistic methods; (c) to interpret the resulting clusters using the k-means variants; and (d) to assess internal validity through silhouette width. The general hypothesis holds that the obtained structure allows homogeneous patterns among departments to be identified according to their demographic indicators.

2. MATERIALS AND METHODS

2.1 Design and study units

The research followed a quantitative approach with a descriptive-explanatory scope and multivariate treatment, a non-experimental design, and a cross-sectional temporal structure, since the variables were observed without manipulation at a single point in time (Hair et al., 2019; Hernández-Sampieri & Mendoza, 2018). The units of analysis were the 24 departments and the Constitutional Province of Callao (n = 25). Sampling was non-probabilistic by convenience, using the complete census-administrative record available for 2022.

2.2 Variables and instrument

Six continuous demographic indicators for 2022 were analyzed, obtained from the official publications of the National Institute of Statistics and Informatics (INEI): crude birth rate (CBR, per thousand), total fertility rate (TFR, children per woman), crude death rate (CDR, per thousand), life expectancy at birth (LEB, years), infant mortality rate (IMR, per thousand live births), and crude nuptiality rate (CNR, per thousand). Data collection was performed by electronic record-keeping, requiring no psychometric validation as these are secondary records.

2.3 Preprocessing

Data were checked to rule out missing values and, prior to clustering, were standardized (mean 0, standard deviation 1) to homogenize scales and prevent higher-variance indicators from dominating the distances. Euclidean distance was used over the standardized variables. Principal component analysis (PCA) served as an exploratory and low-dimensional visualization tool.

2.4 Determining the number of clusters

The optimal number of groups was assessed with three complementary criteria: the elbow method on the within-cluster sum of squares (WSS), the average silhouette width (Rousseeuw, 1987), and the gap statistic (Tibshirani et al., 2001). Additionally, the consensus of multiple indices was contrasted using the NbClust package in R.

2.5 Clustering algorithms

Partitions were estimated with the three k-means variants—Lloyd, MacQueen, and Hartigan–Wong—and with k-medoids (PAM), together with agglomerative hierarchical clustering using Ward's method for structural contrast. The three k-means variants differ in the centroid update rule: Lloyd recomputes in batch, MacQueen incrementally, and Hartigan–Wong reassigns an observation only if doing so reduces the total sum of squares, which tends to confer slightly better performance (Morissette & Chartier, 2013; Vouros et al., 2021).

2.6 Validation and analysis

Internal validity was evaluated through silhouette width (individual and global), the Dunn index, and explained variance (ratio of between-cluster to total sum of squares). The stability of assignments was examined with the Jaccard index via resampling. To characterize the profiles, standardized centroids were compared and a one-way analysis of variance (ANOVA) was applied per indicator, with Tukey multiple comparisons. All processing was carried out in R (*cluster*, *factoextra*, and *NbClust* packages); the figures in this article were reproduced independently from the same dataset to ensure the traceability of results.

3. RESULTS

3.1 Descriptive characterization

Table 1 summarizes the six indicators. The crude birth rate averaged 18.6 per thousand (13.6–24.9) and fertility 2.16 children per woman, close to replacement level. Mean life expectancy was 76.8 years, with moderate dispersion. Infant mortality showed the greatest relative heterogeneity (mean 14.4; SD 3.70; maximum 23.0 in Puno), reflecting territorial inequalities in maternal-child health.

Table 1. *Descriptive summary of the six demographic indicators, Peru 2022 (n = 25).*

Indicator	Mean	Median	SD	Minimum	Maximum
Crude birth rate (‰)	18.64	18.20	2.79	13.60	24.90
Total fertility rate	2.16	2.10	0.35	1.60	3.00
Crude death rate (‰)	5.85	5.80	0.93	4.50	8.90

Life expectancy at birth (yrs)	76.79	76.20	1.98	74.30	80.20
Infant mortality rate (‰)	14.35	14.30	3.70	7.60	23.00
Crude nuptiality rate (‰)	2.21	2.40	0.71	0.73	3.52

SD: standard deviation. Compiled from INEI (2022).

3.2 Exploratory structure via PCA

The first two principal components explained 83.8 % of total variability (PC1 = 59.5 %; PC2 = 24.2 %), so the factorial plane adequately summarizes the structure. PC1 contrasted departments with high birth, fertility, and infant mortality rates (Loreto, Ucayali, Amazonas, Huancavelica) with those of higher life expectancy (Lima, Callao, Arequipa, Moquegua, Tacna), while PC2 mainly captured variation in crude mortality and nuptiality (Figure 1). Junín and Puno appeared as atypical observations owing to their high mortality and infant mortality, respectively.

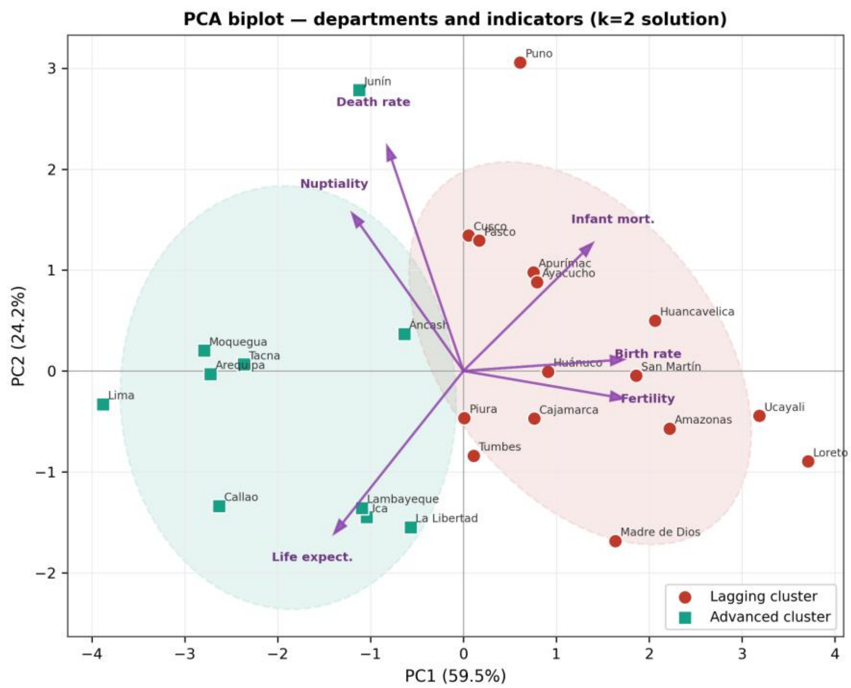

Figure 1. PCA biplot: projection of the 25 territories and the six indicators onto PC1–PC2, colored by the k = 2 partition. Ellipses represent the dispersion of each cluster.

3.3 Optimal number of clusters

The average silhouette width peaked at k = 2 (≈ 0.33), while the elbow method placed the inflection point around k = 3–4, and the gap statistic, together with several NbClust indices, supported solutions of 2 and 4 groups (Figure 2). Given this convergence, both configurations were examined in detail: k = 2 as the main structure, for parsimony and greater cohesion, and k = 4 as an operational disaggregation.

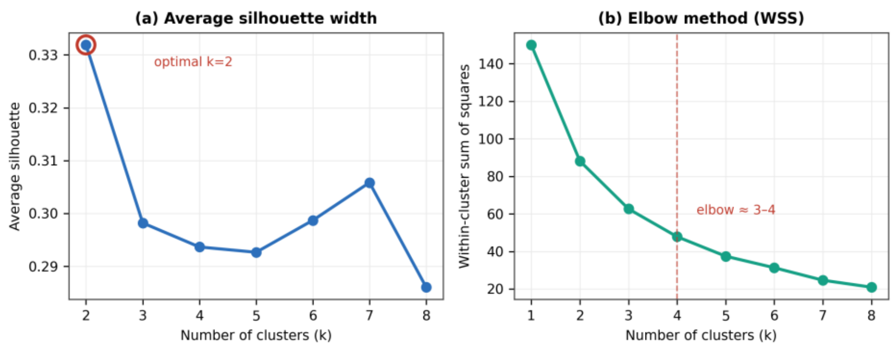

Figure 2. Selection of the number of clusters. (a) Average silhouette width by k (maximum at k = 2). (b) Within-cluster sum of squares (elbow at k ≈ 3–4).

3.4 Main solution (k = 2)

The two-group partition clearly separated the departments along PC1 (Figure 3a). The advanced cluster (n = 10: Lima, Callao, Arequipa, Moquegua, Tacna, Ica, La Libertad, Lambayeque, Áncash, and Junín) showed standardized centroids of low birth rate (−0.82) and fertility (−0.78), high life expectancy (+0.98), and very low infant mortality (−1.00). The lagging cluster (n = 15) exhibited the inverse pattern, with greater reproductive pressure and poorer infant survival outcomes (Table 2, Figure 4). Crude mortality was slightly higher in the advanced cluster (+0.44), consistent with more aged population structures.

Table 2. *Standardized centroids per indicator for the k = 2 partition (national mean = 0).*

Indicator	Advanced cluster	Lagging cluster
Crude birth rate	−0.82	+0.55
Total fertility rate	−0.78	+0.52
Crude death rate	+0.44	−0.29
Life expectancy at birth	+0.98	−0.65
Infant mortality rate	−1.00	+0.66
Crude nuptiality rate	+0.53	−0.35

Negative values indicate levels below the national mean; positive values, above it.

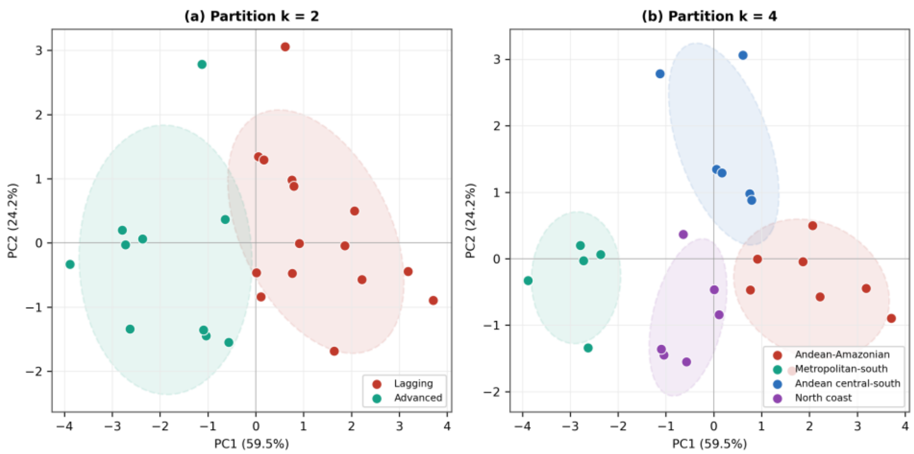

Figure 3. PC1–PC2 factorial map. (a) k = 2 partition (advanced vs. lagging). (b) k = 4 partition, with dispersion ellipses per group.

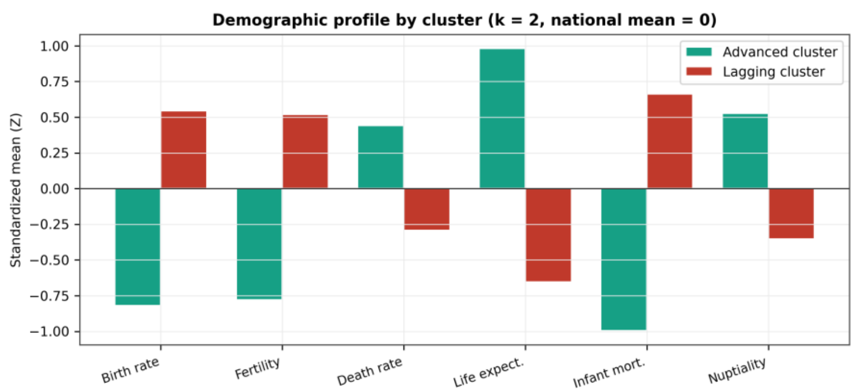

Figure 4. Mean demographic profile per cluster (k = 2) in standardized units. The advanced cluster concentrates features of an advanced demographic transition.

The silhouette plot (Figure 5) confirmed an acceptable overall quality (average ≈ 0.33), with no relevant negative widths and only two borderline cases near zero, supporting the robustness of the two-group solution.

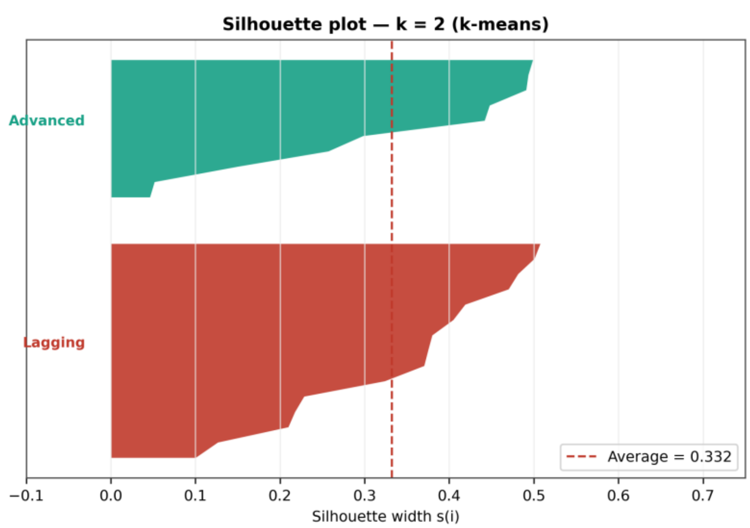

Figure 5. Silhouette plot of the k = 2 partition (k-means). The dashed line marks the average silhouette.

3.5 Operational disaggregation (k = 4)

The four-group partition raised explained variance to 68.1 % and revealed coherent internal substructures (Figure 3b): a metropolitan-southern pole (Lima, Callao, Arequipa, Moquegua, Tacna) with the best indicators; an Andean-Amazonian bloc (Amazonas, Cajamarca, Huancavelica, Huánuco, Loreto, Madre de Dios, San Martín, Ucayali) facing the greatest challenges in birth and infant mortality rates; an Andean central-south group (Apurímac, Ayacucho, Cusco, Junín, Pasco, Puno); and a north-coast group (Áncash, Ica, La Libertad, Lambayeque, Piura, Tumbes) with intermediate conditions. This disaggregation increases structural separation (higher Dunn index) at the cost of lower overall cohesion, so it is interpreted as a complement for the fine-grained targeting of interventions.

3.6 Stability and significance

The Lloyd, MacQueen, and Hartigan–Wong variants converged on virtually identical assignments, with the latter slightly higher in silhouette, and the Jaccard index revealed high group stability across resamples. The per-indicator ANOVA (Table 3) rejected the equality of means between the two clusters for most variables—very markedly for infant mortality (F = 45.0) and life expectancy (F = 41.3)—whereas only crude mortality did not reach significance at k = 2. In the k = 4 solution, all indicators were significant, confirming its greater discriminant capacity.

Table 3. *Analysis of variance (ANOVA) between clusters (k = 2, k-means).*

Indicator	**F**	**p-value**	**Significance**
Infant mortality rate	44.98	7.7×10^{-7}	***
Life expectancy at birth	41.27	1.5×10^{-6}	***
Crude birth rate	18.63	2.6×10^{-4}	***
Total fertility rate	15.74	6.1×10^{-4}	***
Crude nuptiality rate	5.28	3.1×10^{-2}	*
Crude death rate	3.44	7.7×10^{-2}	n.s.

**** $p < 0.001$; * $p < 0.05$; n.s.: not significant.*

3.7 Hierarchical contrast

The Ward dendrogram (Figure 6) reproduced the two-macro-cluster structure when cut at a height of ≈ 7–10: a compact metropolitan-southern cluster and a broad, heterogeneous Andean-Amazonian bloc, with internal merges at intermediate heights consistent with the k = 4 solution. The coherence among the partitional clustering, the hierarchical clustering, and the factorial map reinforces the validity of the obtained typology.

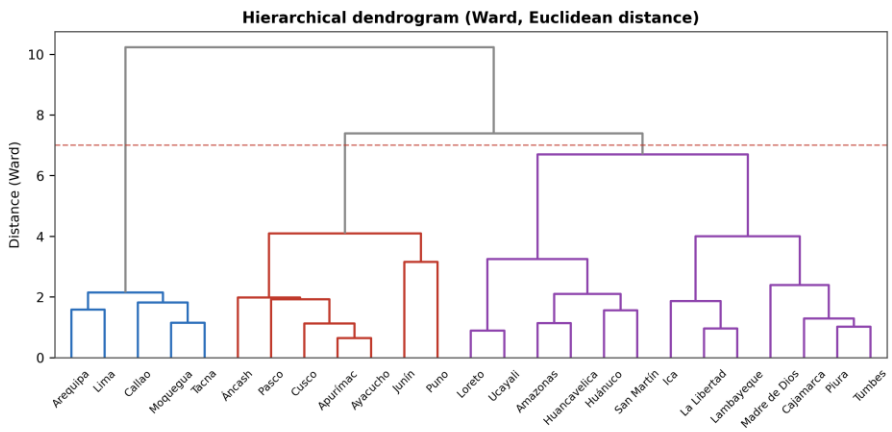

Figure 6. Hierarchical dendrogram (Ward's method, Euclidean distance). The cut at height ≈ 7 delimits the groups consistent with the k = 4 partition.

4. DISCUSSION

The findings confirm the hypothesis that the selected demographic indicators capture a homogeneous, interpretable territorial structure. The convergence of the silhouette, the gap statistic, and the Jaccard index on the k = 2 solution—with positive widths and no incorrect assignments—is consistent with the premise that positive silhouette values reflect adequate separation and correct assignment of observations (Rousseeuw, 1987; Jin & Han, 2011), and with evidence that the silhouette coefficient ranks among the most informative internal indices for convex clusters (Chicco et al., 2025).

The stability across the k-means variants supports the robustness of the clustering: the near-coincidence of Lloyd, MacQueen, and Hartigan–Wong is congruent with the partitional nature of the method, which minimizes within-group variability, and with the finding that Hartigan–Wong tends to perform slightly better (Morissette & Chartier, 2013; Vouros et al., 2021). The complementary use of k-medoids (PAM), robust to outliers (Schubert & Rousseeuw, 2021), and of the Ward dendrogram provided methodological triangulation.

Substantively, the dichotomy between an advanced urban-coastal pole and a lagging Andean-Amazonian bloc is consistent with the heterogeneous behavior of Latin American demographic indicators, attributable to differences in socioeconomic conditions and access to health services (ECLAC, 2019; Bilal et al., 2019). This result replicates, for the Peruvian case, the coast–interior pattern described in multivariate territorial classifications of other countries (Soares et al., 2003) and the usefulness of the PCA–clustering pairing to inform place-based policies (Pavone et al., 2020; Ayberkin et al., 2026; Panzabekova et al., 2025; Polyanskaya, 2025). The slightly higher—non-significant—crude mortality of the advanced cluster is consistent with the population aging typical of advanced stages of the demographic transition.

Among the limitations, the cross-sectional nature precludes causal or trend inferences, and the small size (n = 25) restricts the complexity of the recoverable structure, favoring parsimonious solutions. Non-probabilistic sampling and the use of secondary records condition temporal generalization: the 2022 typology constitutes a diagnostic baseline whose relationship with later years would require new evidence. Future research could incorporate additional socioeconomic and health indicators, a longitudinal approach, and density-based methods to test the robustness of the boundaries between groups.

5. CONCLUSIONS

The descriptive analysis revealed marked territorial disparities, especially in infant mortality and life expectancy, which justify the use of these indicators as the basis for clustering. The elbow, average silhouette, and gap statistic criteria converged on k = 2 as the highest-quality solution (silhouette ≈ 0.33; 41.2 % explained variance), with two clear profiles: an advanced cluster—low birth and fertility rates, high life expectancy, and very low infant mortality—and a lagging cluster with the opposite pattern. The k-means variants yielded consistent results, with Hartigan–Wong slightly superior, and ANOVA confirmed significant between-group differences for most indicators.

As an operational disaggregation, k = 4 distinguished a metropolitan-southern pole, an Andean-Amazonian bloc facing greater challenges, and two intermediate transitional groups, useful for the fine-grained design of interventions. The typology—reproducible and visualizable on the PC1–PC2 plane and coherent with the hierarchical dendrogram—strengthens territorial planning: it suggests prioritizing maternal-child health, reproductive education, and basic services in the lagging departments (particularly Huancavelica, Ayacucho, Apurímac, Puno, Cajamarca, Amazonas, Loreto, Ucayali, and Madre de Dios), and healthy-aging and chronic-disease management strategies in the advanced cluster, enabling a more efficient and equitable allocation of resources.

Declarations

Funding: the authors declare that they received no external funding. **Conflict of interest:** the authors declare no conflicts of interest. **Data availability:** the indicators come from official INEI publications and are publicly available.

REFERENCES

1. Abramo, L., Cecchini, S., & Ullmann, H. (2020). Enfrentar las desigualdades en salud en América Latina: el rol de la protección social. Ciência & Saúde Coletiva, 25(5), 1587–1598. https://doi.org/10.1590/1413-81232020255.32802019
2. Ayberkin, D., et al. (2026). Analysis of Turkey's demographic and social structure using PCA and K-means clustering. MANAS Sosyal Araştırmalar Dergisi.
3. Bilal, U., Alazraqui, M., Caiaffa, W., et al. (2019). Inequalities in life expectancy in six large Latin American cities from the SALURBAL study: an ecological analysis. The Lancet Planetary Health, 3(12), e503–e510. https://doi.org/10.1016/S2542-5196(19)30235-9
4. Chicco, D., et al. (2025). The Silhouette coefficient and the Davies–Bouldin index are more informative than Dunn, Calinski–Harabasz, Shannon entropy and Gap statistic for unsupervised clustering internal evaluation. PeerJ Computer Science.
5. Economic Commission for Latin America and the Caribbean (ECLAC). (2019). Latin America and the Caribbean will reach its peak population around 2058. ECLAC.
6. European Commission. (2020). Regional policy and macro-regional strategies. European Union.
7. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage.
8. Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C, 28(1), 100–108. https://doi.org/10.2307/2346830
9. Hernández-Sampieri, R., & Mendoza, C. (2018). Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta. McGraw-Hill.
10. Iammarino, S., Rodríguez-Pose, A., & Storper, M. (2019). Regional inequality in Europe: evidence, theory and policy implications. Journal of Economic Geography, 19(2), 273–298. https://doi.org/10.1093/jeg/lby021
11. Jin, X., & Han, J. (2011). Partitional clustering. In Encyclopedia of Machine Learning. Springer. https://doi.org/10.1007/978-0-387-30164-8_631
12. Morissette, L., & Chartier, S. (2013). The k-means clustering technique: general considerations and implementation in Mathematica. Tutorials in Quantitative Methods for Psychology, 9(1), 15–24. https://doi.org/10.20982/tqmp.09.1.p015
13. National Institute of Statistics and Informatics (INEI). (2020). Peru: mortality and life expectancy indicators by department. INEI.
14. National Institute of Statistics and Informatics (INEI). (2022). Peru: demographic indicators by department, 2022. INEI. https://www.inei.gob.pe
15. National Statistics Institute of Spain. (2021). Basic demographic indicators: methodology. INE.
16. Pan American Health Organization. (2018). Health indicators: conceptual and operational aspects. PAHO.
17. Panzabekova, A., et al. (2025). Cluster analysis of the demographic characteristics of Kazakhstan's regions and their relationship with income distribution. Bulletin of the L. N. Gumilyov ENU, Economic Series.
18. Pavone, P., et al. (2020). Multidimensional clustering of EU regions: a contribution to orient public policies in reducing regional disparities. Social Indicators Research, 149, 863–887.

19. Polyanskaya, E. (2025). Clustering of regions of the Far East by the level of health determinants. DEMIS. Demographic Research.
20. Queiroz, B., Lima, E., Freire, F., & Gonzaga, M. (2020). Temporal and spatial trends of adult mortality in small areas of Brazil, 1980–2010. Genus, 76(36), 1–22. https://doi.org/10.1186/s41118-020-00105-3
21. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
22. Schubert, E., & Rousseeuw, P. J. (2021). Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA and CLARANS algorithms. Information Systems, 101, 101804. https://doi.org/10.1016/j.is.2021.101804
23. Soares, J. O., Marques, M. M. L., & Monteiro, C. M. F. (2003). A multivariate methodology to uncover regional disparities: a contribution to improve European Union and governmental decisions. European Journal of Operational Research, 145(1), 121–135.
24. Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B, 63(2), 411–423.
25. United Nations. (2022). Latin America and the Caribbean lost nearly 3 years of life expectancy at birth between 2019 and 2021 due to the COVID-19 pandemic. ECLAC–CELADE.
26. Vouros, A., Langdell, S., & Croucher, M. (2021). An empirical comparison between stochastic and deterministic centroid initialisation for K-means variations. Machine Learning, 110(8), 1975–2003. https://doi.org/10.1007/s10994-021-06021-7