Water safety profiles and infant mortality in Colombian municipalities, 2016-2023

Water-quality workshop

Abstract We describe how Colombian municipalities group by drinking-water safety and context, using real public data. K-means on nine standardised profile features, including the utility-supplied IRCA, finds four clusters. Infant mortality, Saber 11 scores and school dropout are used only to describe the clusters, never to form them. Within clusters the association between IRCA and infant mortality is weak. All comparisons are descriptive and do not establish a causal effect of water safety.

1. Introduction

Safe drinking water is a precondition for child health, yet water safety varies widely across Colombian municipalities and between urban and rural zones. The national surveillance index IRCA (*Índice de Riesgo de la Calidad del Agua*) runs from 0 (no risk) to 100 (unfit) and is classified into bands by Resolución 2115 de 2007 [1]. The open INS table used here reports the **utility-supplied ("prestadores") IRCA** [2], not INS's national total. We ask how municipalities group by water safety and context, and how outcomes look in each group.

2. Methods

The unit of analysis is a municipality x zone (urban or rural) x year row, 2016-2023, built from public sources [2], [3], [4], [5], [6], [7], [8], [9]. We keep rows of municipalities with an IRCA value (11,235 rows, 1,090 municipalities). For department capitals, which have no urban/rural coverage split, urban coverage is filled from the municipal total (214 flagged rows).

Each of nine profile features x_{ij} is standardised within year,

$$z_{ij} = \frac{x_{ij} - \bar{x}_{j,t(i)}}{s_{j,t(i)}}, \tag{1}$$

where $\bar{x}_{j,t(i)}$ and $s_{j,t(i)}$ are the mean and standard deviation of feature j in the year $t(i)$ of row i. K-means [10] minimises the within-cluster sum of squares

$$W = \sum_{c=1}^{k} \sum_{i \in C_c} \lVert z_i - m_c \rVert^2, \tag{2}$$

with m_c the centroid of cluster C_c. We choose k by the elbow of Equation 2 and the silhouette [11]

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{3}$$

where $a(i)$ is the mean distance to the other points of its cluster and $b(i)$ the smallest mean distance to another cluster. Stability across seeds is measured with the adjusted Rand index [12]. Outcomes are never used for clustering.

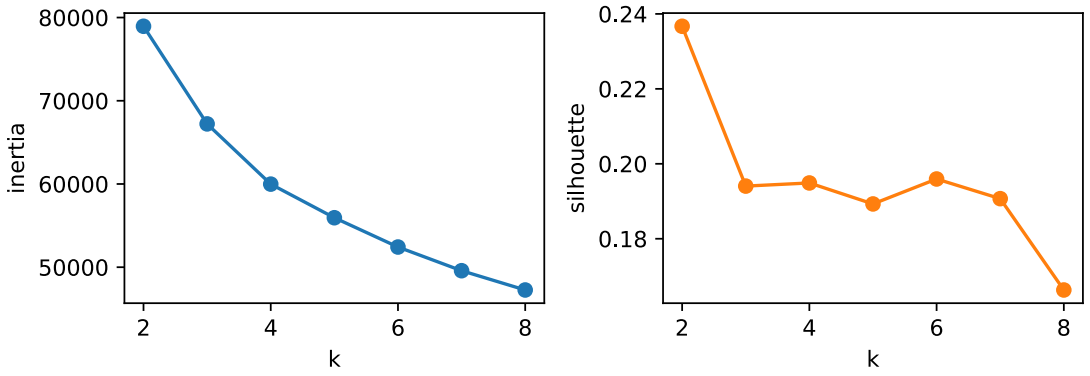


Figure 1: Inertia (left) and mean silhouette (right) by number of clusters k.

3. Results

We selected $k = 4$: the next cluster lowers inertia by only about 7%, the silhouette is 0.19 and the adjusted Rand index across five seeds is at least 0.999. The clusters are "Cities and large towns, no-risk water", "Highland towns, no-risk water", "Lowland towns, low-risk water", "Rural, low-coverage, high-risk water". Figure 2 shows what separates them.

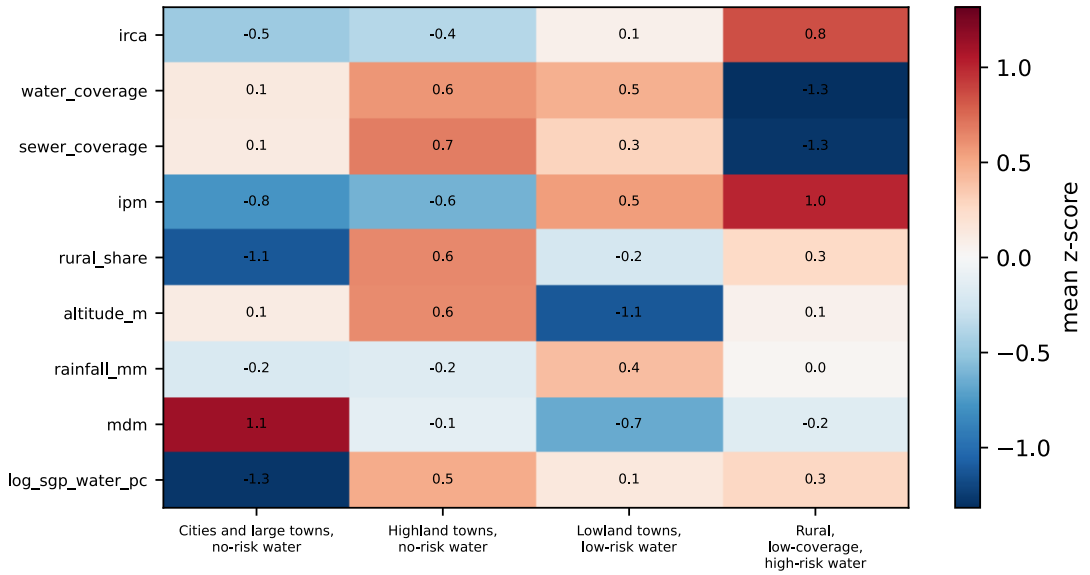


Figure 2: Cluster mean z-scores (within year) of the nine profile features.

Figure 3 shows the outcomes by cluster; lines break in years without data.

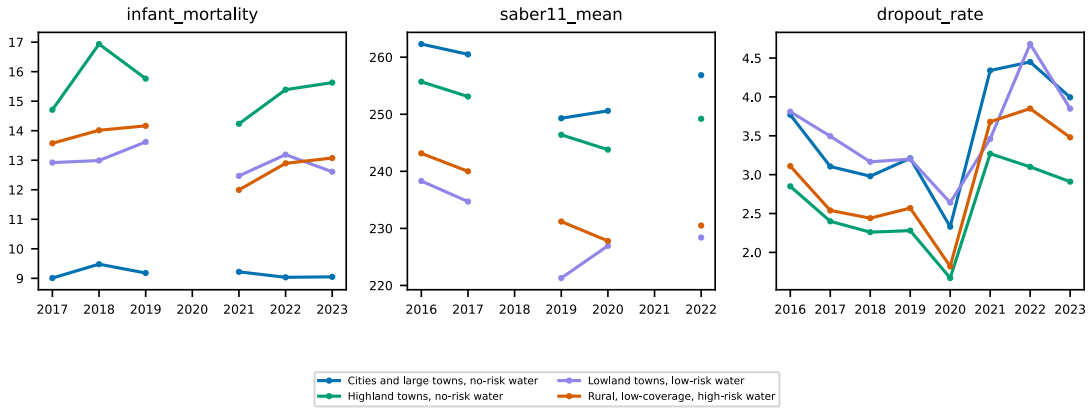


Figure 3: Median infant mortality per 1,000 live births, Saber 11 score and school dropout (%) by cluster and year.

Within clusters, Spearman's ρ between IRCA and infant mortality ranges from -0.04 to 0.10, a weak association.

4. Conclusions

Municipal water-safety profiles form four descriptive groups. Their outcome differences mirror poverty, rurality and altitude as much as IRCA, and within clusters the IRCA-infant mortality association is weak. This is not evidence that water quality does or does not cause any outcome. Limits: utility-supplied IRCA only, rural zones often unsampled, imputed coverage, a change of infant-mortality method in 2017 and no Saber 11 data for 2018 and 2021.

References

[1] Ministerio de la Protección Social and Ministerio de Ambiente, "Resolución 2115 de 2007, art. 15: IRCA classification." 2007.

[2] Instituto Nacional de Salud, "Calidad del Agua para Consumo Humano en Colombia (utility-supplied IRCA)." [Online]. Available: https://www.datos.gov.co/resource/nxt2-39c3

[3] Departamento Nacional de Planeación, "TerriData: coverage, rurality, poverty, performance, fiscal and health indicators." [Online]. Available: https://terridata.dnp.gov.co/

[4] Departamento Administrativo Nacional de Estadística, "DIVIPOLA municipal codes (gdxc-w37w), population projections and MGN 2023 boundaries." [Online]. Available: https://www.dane.gov.co/

[5] ICFES, "Resultados únicos Saber 11."

[6] Ministerio de Educación Nacional, "Estadísticas en educación por municipio (school dropout)."

[7] C. Funk and others, "The Climate Hazards Center Infrared Precipitation with Stations, Version 3," *Scientific Data*, vol. 13, p. 718, 2026, doi: 10.1038/s41597-026-07096-4.

[8] Climate Hazards Center, "CHIRPS version 3.0 data repository." doi: 10.15780/G2JQ0P.

[9] NASA JPL, "NASA Shuttle Radar Topography Mission Global 3 arc second." NASA LP DAAC, 2013. doi: 10.5067/MEaSUREs/SRTM/SRTMGL3.003.

[10] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982, doi: 10.1109/TIT.1982.1056489.

[11] P. J. 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, doi: 10.1016/0377-0427(87)90125-7.

[12] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985, doi: 10.1007/BF01908075.