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

A descriptive clustering of municipal water profiles

Table of contents

Summary 1
Question 1
Data 1
Method 1
What sets the clusters apart 2
Outcomes by cluster 2
Correlation is not causation 3
Limits 3
Sources 3

Summary

Using 11,235 municipality-zone-year rows (1,090 municipalities, 2016-2023), k-means finds 4 groups of municipal water profiles. The one with the highest utility-supplied IRCA (median np.float64(35.2)) is "Rural, low-coverage, high-risk water". Infant mortality, Saber 11 scores and dropout differ across clusters, but these are **descriptive associations, not causal effects of water quality**.

Question

How do Colombian municipalities group by the safety of their drinking water and the context around it, and how do infant mortality, school results and dropout look in each group?

Data

One row is a municipality x zone (urban/rural) x year, from the public dataset in `data/` (INS, DNP TerriData, DANE, CHIRPS, ICFES, MEN; see Sources). **IRCA is the utility-supplied ("prestadores") IRCA from INS open data**, not INS's national total; it runs from 0 (no risk) to 100 (unfit) and is much more often missing for rural zones. Only rows of municipalities with an IRCA value in 2016-2023 are used. Department capitals have no urban/rural coverage split, so their urban coverage is filled from the municipal total (214 rows, flagged in the data).

Method

Nine profile features (IRCA, water and sewer coverage, multidimensional poverty, rural share, altitude, rainfall, municipal performance, log water transfers per person) were standardised within each year and clustered with k-means (`random_state` = 42, 20 starts). k = 4 was chosen from the elbow, with silhouette 0.19 and stability across five seeds (ARI >= 0.999). Outcomes (infant mortality, Saber 11, dropout) were **not** used to cluster; they only describe the clusters.

What sets the clusters apart

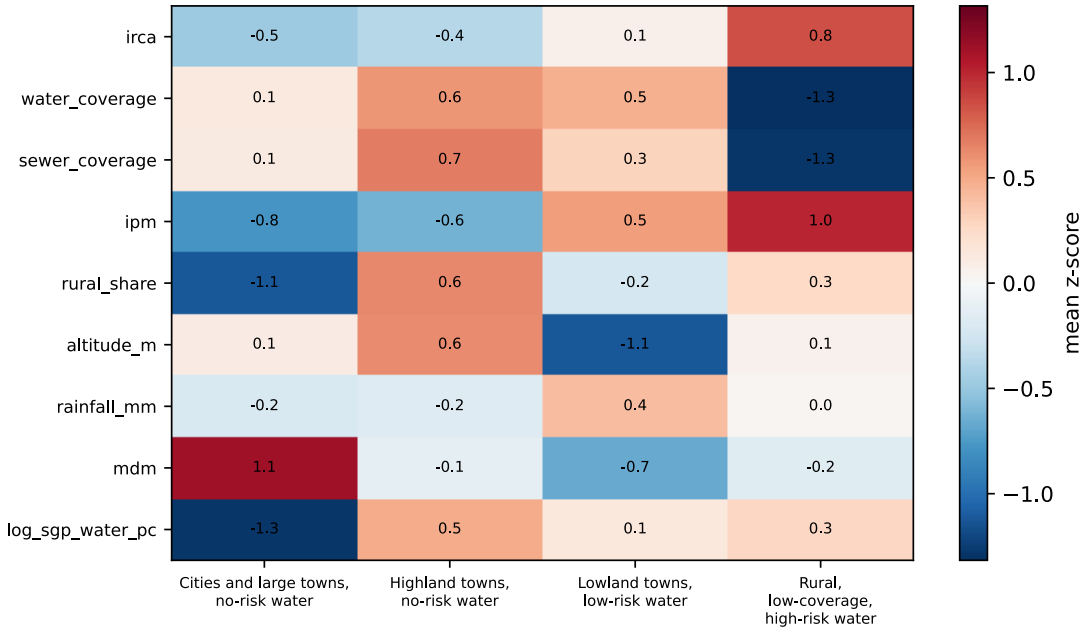


Figure 1: Cluster mean z-scores (within year) of the nine profile features; blue is below the yearly average and red above.

Outcomes by cluster

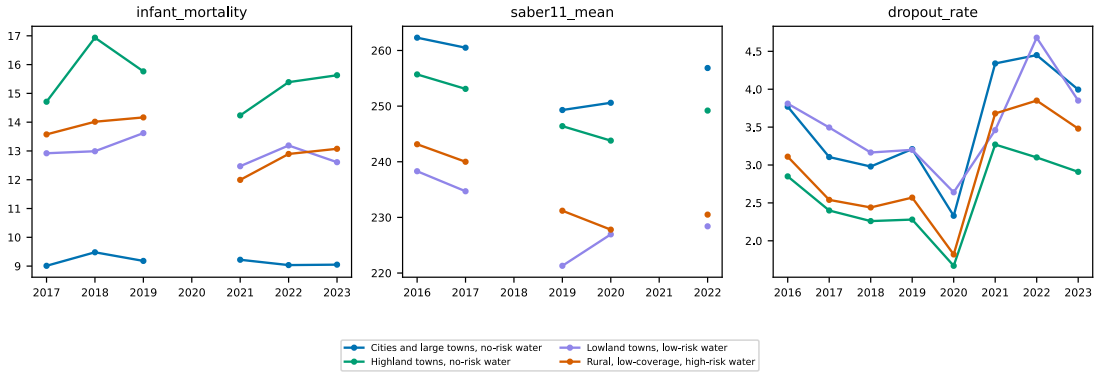


Figure 2: Median infant mortality (per 1,000 live births), Saber 11 mean score and school dropout (%) by cluster and year. Lines break where a source has no data.

Infant mortality changes method in 2017 and Saber 11 has no data for 2018 and 2021, so year-to-year changes should not be read as real trends.

Correlation is not causation

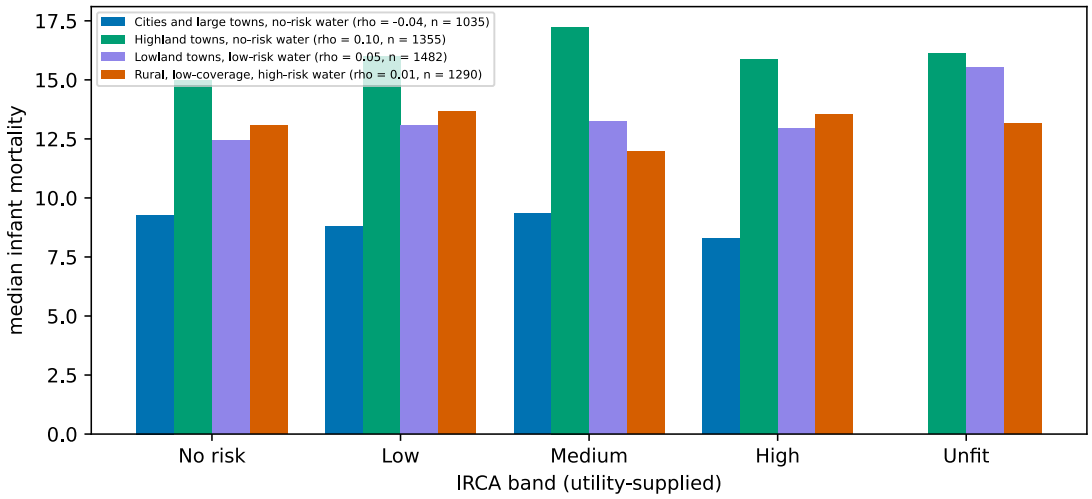


Figure 3: Median infant mortality (per 1,000 live births) by IRCA band (utility-supplied) within each cluster. Bars are medians; the Spearman rho between IRCA and infant mortality is in the legend.

Within each cluster the association between IRCA and infant mortality is weak (rho from −0.04 to 0.10). This is descriptive and does not establish a causal effect of water safety. Poverty, health services, altitude and rurality may confound it.

Limits

- Utility-supplied IRCA only; rural zones are often unsampled.
- Infant mortality changes method in 2017 and is volatile in small municipalities.
- Saber 11 covers 2014-2017, 2019, 2020 and 2022; school location is not residence.
- Missing coverage was imputed (year x zone median) in 653 rows.
- Clusters are descriptive groupings, not causal categories.

Sources

- INS, “Calidad del Agua para Consumo Humano en Colombia” (IRCA), datos.gov.co `nxt2-39c3`.
- DNP, TerriData (coverage, rurality, IPM, MDM, IDF, SGP, infant mortality), https://terridata.dnp.gov.co.
- DANE, DIVIPOLA `gdxc-w37w` and population projections, https://www.dane.gov.co.
- CHIRPS v3.0 rainfall, https://doi.org/10.15780/G2JQ0P.
- NASA SRTM via OpenTopoData, https://doi.org/10.5067/MEaSUREs/SRTM/SRTMGL3.003.
- ICFES Saber 11, datos.gov.co `kgxf-xxbe`, `rnvb-vnyh`; MEN dropout, `nudc-7mev`.
- Details: `data/README.md`. Resolución 2115 de 2007, art. 15 (IRCA bands).