

International Journal of Population Data Science

Journal Website: www.ijpds.org



Large-scale geocoding via iterative deduplication: Building a unified address database in Brazil with CIDACS-RL

Robespierre Pita¹, Maria Ichihara², Gabriel Rocha³, Carlos Santos¹, Leandro Ribeiro², Pablo Pita², Samila Sena², Fernanda Eustáquio², José Gomes², and Rafael Silva²

¹Federal University of Bahia, Salvador, Brazil; CIDACS, Salvador, Brazil

²CIDACS, Salvador, Brazil

³CIDACS, Salvador, Brazil; Federal University of Bahia, Salvador, Brazil

This work aims to construct a dataset of unique addresses from Brazilian administrative, socioeconomic, and health databases. We assess the potential of iterative deduplication to generate a parent-duplicate data structure that supports address management and downstream information retrieval tasks. Our approach followed three steps: local deduplication within each database to establish parent-duplicate relationships; cross-database deduplication to identify global parent records and propagate their identifiers; and iterative deduplication of global parent records, updating assignments across databases. Graph-based modeling was used to identify shared connected components, enhancing geocoding coverage. Deduplication accuracy was evaluated through manual review of state-stratified samples ($N = 2,000$), with independent assessment by at least two reviewers and disagreements resolved by a third. Five data sources from Northeastern Brazil were used:

the National Register of Addresses from 2010 (CNEFE 2010; $N = 17,839,562$), CNEFE 2022 ($N = 30,545,117$), the 100 Million Brazilian Cohort baseline (CadÚnico; $N = 54,985,455$), Mortality (SIM; $N = 6,431,199$), and Live Birth (SINASC; $N = 17,199,287$). Local deduplication substantially reduced data volume, identifying 8.25M, 531 thousand, and 583 thousand parent records for the CadÚnico, SIM, and SINASC, corresponding to reductions of 77.84%, 91.74%, and 96.61%. After iterative cross-database deduplication. The resulting database comprised 28.6M records, including 24.8M unique parent addresses and 5.9M duplicates. Average deduplication precision across all steps exceeded 0.9. Iterative deduplication is an effective strategy for enhancing data management and quality across heterogeneous datasets, facilitating large-scale spatial analysis for public health and climate change research.

