

International Journal of Population Data Science

Journal Website: www.ijpds.org



Enhancing geocoding in Brazil through language model-based preprocessing

Robespierre Pita¹, Gabriel Rocha², Maria Ichihara³, Katie Harron⁴, Paloma Brito⁵, Roberto Carreiro³, Bethania Almeida³, Pablo Ramos³, Marcos Barreto⁶, and Mauricio Barreto³

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

²CIDACS, Salvador, Brazil; Federal University of Bahia, Salvador, Brazil

³CIDACS, Salvador, Brazil

⁴University College London, London, United Kingdom

⁵Federal University of Bahia, Salvador, Brazil

⁶London School of Economics and Political Science, London, United Kingdom; CIDACS, Salvador, Brazil

Similarity-based geocoding pipelines are highly sensitive to lexical noise in address data. This work evaluates the potential of language models to enhance and standardize address inputs during preprocessing, thereby improving geocoding accuracy in CIDACS-RL, which relies on Jaro-Winkler similarity and is sensitive to string length and prefix agreement. We extended CIDACS-RL with a preprocessing step that removes street types (e.g., street, avenue, lane) and stop words and expands numeric and abbreviated address components. Two pre-trained language models were included in the experiments. Records from nine Northeastern Brazilian states in the cohort baseline ($N = 54,985,455$) were geocoded by linking addresses to census tracts or coordinates from the Brazilian National Register of Addresses (CNEFE; $N = 30,545,117$). After linkage, we manually assessed the accuracy of our extension

against previous CIDACS-RL version using stratified samples ($N = 2,000$) from each run to determine the cutoff points. The geocoding was evaluated using accuracy and precision. Overall, the proposed extension achieved competitive performance, with higher mean precision (0.86, $SD = 0.01$) and accuracy (0.94, $SD = 0.024$) compared to the original CIDACS-RL, which achieved a mean precision of 0.84 ($SD = 0.02$) and a mean accuracy of 0.93 ($SD = 0.024$). The largest gains were observed in states with shorter average street name lengths, such as Alagoas and Pernambuco, where precision increased by approximately 3% to 10%. Our findings support that geocoding tasks may benefit from pre-trained language models for preprocessing steps, addressing the limitation of well-posed similarity measures.

