International Journal of Population Data Science

Journal Website: www.ijpds.org

Synthetic data generator for testing record linkage routines in Brazil.

Trentin, V¹, Bastos, V¹, Costa, M^{2,1}, Camargo, K³, Sobrino, R¹, Guillen, LC¹, and Coeli, C¹

¹Federal University of Rio de Janeiro

²COPPE

³Universidade do Estado do Rio de Janeiro

Introduction

Record linkage has been increasingly used in Brazil. However, only a few studies report the quality of the linkage process. Synthetic test data can be used to evaluate the quality of data linkage.

Objectives and Approach

To develop a synthetic data generator that creates test datasets with similar attributes and error characteristics found in the Brazilian databases.

We analyzed the 2013 mortality database from Rio de Janeiro State to know the characteristics and frequency distribution of the database attributes (name, mother's name, sex, date of birth and address).

We used Python and C++ to customize and add routines to GeCo (<http://dlrep.org/dataset/GeCo>), a personal data generation tool developed by Tran et al. (DOI:[10.1145/2505515.2508207](https://doi.org/10.1145/2505515.2508207)).

Results

Brazilian names have specific characteristics that distinguish them from other countries' patterns: multiple family names are usual, as are composite first names, and, despite that, homonyms are frequent. Family names may include the full extension or only parts of either the father and mother's respective family names, or both, so there is a wide variation in progeny family names and not necessarily a common family name for all family members.

Conclusion/Implications

Due to the specific national characteristics of name building in Brazil, modeling synthetic data is particularly challenging and needs to have more flexible rules in order to generate databases that will actually allow assessing the quality of data linkage processes.