DSSV 2023 Data Science, Statistics & Visualisation
ECDA 2023 European Conference on Data Analysis

Hausdorff Distance: A Powerful Tool for Matching Households and Individuals in Different Databases

Thais Pacheco Menezes[1], Michael Fop[1], and Thomas Brendan Murphy[1,2]

[1]School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland
[2]Insight Centre for Data Analytics, University College Dublin, Belfield, Dublin 4, Ireland.

Matching households and individuals from different databases can be difficult due to the absence of unique identifiers, typographical errors, and changes in attributes over time. The tools of record linkage are of great assistance in this task [1]. This work defines a general multi-step record linkage procedure that allows the incorporation of household information to improve the process of matching entities across different databases. We propose using the Hausdorff distance to estimate the probability of a match between households in multiple files. Subsequently, the probabilities of a match between individuals within matched households are computed using a logistic regression model based on attribute level distances. The estimated probabilities are then employed in a linear programming optimization framework to infer one-to-one matches between the individuals. Furthermore, the methodology is developed for the application of linking the Italian Survey of Household Income and Wealth of 2014 and 2016. The approach yields around 70% of correct matches found when the household information is considered in the process of matching individuals. A comparison with when the matching is done directly shows that it is highly beneficial to include such information.

Keywords: household information, Hausdorff distance, matching databases, record linkage

[1] T. N. Herzog, F. J. Scheuren, and W. E. Winkler, *Data Quality and Record Linkage Techniques*, vol. 1. Springer, 2007.