International Journal of Population Data Science

Journal Website: www.ijpds.org

Automatically finding threshold in probabilistic linkage

Zoe White[1]

[1]Office for National Statistics, Fareham, United Kingdom

Objective

Identifying thresholds for the acceptance of record pairs as matches in probabilistic linkage can be clerically expensive. Our objective is to create a pipeline for probabilistic linkage, which scores blocked pairs using Ministry of Justice's Splink package. The pipeline also finds two thresholds in an automated and generalisable approach.

Methods

As we would like a generalisable approach, setting a constant threshold is not possible, due to changing score ranges from using different administrative data. To identify thresholds and potential approaches for automating the thresholds finding, two datasets of links, were used to create graphs. The two datasets of links were made using deterministic and probabilistic linkage methods. The two methods of automating threshold finding were found by analysing graphs and were coded within the pipeline.

We will present the two different automatic and generalised methods used to create the respective thresholds.

Results

Precision and recall values were calculated at different thresholds by using both clerical matching and a gold standard dataset. The automatic and generalisable methods of finding the probabilistic threshold produced reasonable results that satisfy requirements of the probabilistic pipeline.

The two automated threshold methods were integrated into the pipeline and tested end to end. The results were consistent with graph-based threshold calculations.

Conclusion

Our research and methods work well on high quality data, but these datasets are not representative of all administrative data. Low quality datasets have not been tested within our methods; therefore, we cannot confirm their performance on them. Evaluation of all methods within the probabilistic pipeline is also required. Including evaluating the quality (precision, recall, bias, etc), efficiency and ease of use.

https://doi.org/10.23889/ijpds.v10i3.3163