International Journal of Population Data Science

Journal Website: www.ijpds.org

Improving name comparison similarity scores to reduce number of records for clerical review

Palfy, M¹, Vasquez, SA¹, and Garcia, AF¹

¹SA NT DataLink

Introduction

Many well established string comparators are currently used in data linkage. Jaro-Winkler distance is SA NT DataLink's metric of choice for comparing personal names. However, due to Jaro-Winkler's lower specificity we investigated if output scores could be transformed to produce scores more closely matching those assigned manually.

age variables were affected, the resultant outcome was encouraging enough to consider exploring other possibilities for replicating clerical review knowledge in other comparators and metrics to reduce the demands for clerical review.

Objectives and Approach

Our objective was to reduce the need for clerical review by modifying the Jaro-Winkler distance metric output scores. Clerical reviewers assigned similarity scores to pairs of first or last names from a database of approximately 2,000 random cases. By plotting the Jaro-Winkler scores against those assigned by the reviewers, a distinct radical function shape was observed. We then transformed the Jaro-Winkler scores by applying a power function where we gradually changed the exponent until we obtained the best fit with our clerically assigned scores. From the next linkage, two separate outputs were created (original and modified) and the results compared.

Results

To assess the best fit we calculated the sum of squared errors for each of tested exponent values ranging from 1.1 to 6.0 (with 0.1 steps). The minimum sum of squared errors was achieved with exponent value of 4.6. We performed a probabilistic linkage for one decade of the Birth Registry records looking for familial links. Two separate linkage runs were conducted and clerically reviewed. In the second run, names were compared using the modified Jaro-Winkler comparator. This resulted in a reduced number of false positives. Though the lower-end threshold of the clerically reviewed "grey area" had to be lowered, the overall range was narrower resulting in less record pairs for clerical review.

Conclusion/Implications

By transforming the Jaro-Winkler scores, we reduced the number of records requiring clerical review. While only three link-