International Journal of Population Data Science

Journal Website: www.ijpds.org

Record Linkage Methodology for the Social Data Linkage Environment at Statistics Canada

Babyak, Colin^{1*} and Saidi, Abdelnasser¹

¹Statistics Canada

Objectives

The objectives of this talk are to introduce Statistics Canada's Social Data Linkage Environment (SDLE) and to explain the methodology behind the creation of the central depository and how both deterministic and probabilistic record linkage techniques are used to maintain and expand the environment.

Approach

We will start with a brief overview of the SDLE and then continue with a discussion of how both deterministic linkages and probabilistic linkages (using Statistic Canada's generalized record linkage software, G-Link) have been combined to create and maintain a very large central depository, which can in turn be linked to virtually any social data source for the ultimate end goal of analysis.

Results

Although Canada has a population of about 36 million people, the central depository contains some 300 million records to represent them, due to multiple addresses, names, etc. Although this allows for a significant reduction in missing links, it raises the spectre of additional false positive matches and has added computational complexity which we have had to overcome.

Conclusion

The combination of deterministic and probabilistic record linkage strategies has been effective in creating the central depository for the SDLE. As more and more data are linked to the environment and we continue to refine our methodology, we can now move on to the ultimate goal of the SDLE, which is to analyze this vast wealth of linked data.

*Corresponding Author:

Email Address: colin.babyak@canada.ca (C. Babyak)