

International Journal of Population Data Science

Journal Website: www.ijpds.org



Handling missing data when using Goldstein et al.'s Scalelink method of data linkage

Mary Cleaton¹, Gavin Thomson¹, Josie Plachta¹, and Rachel Shipsey¹

¹Office for National Statistics, Newport, United Kingdom

Scalelink is an innovative probabilistic data linkage method based on multiple correspondence analysis. Unlike the current gold-standard probabilistic algorithm, Fellegi-Sunter, it does not assume linkage variable independence. All real-world linkage contains dependent linkage variables. Thus, avoiding this assumption has potential to improve data linkage quality. However, to date there are no recommendations regarding handling missing data values when using Scalelink. As all real-world data contains missingness, this means Scalelink cannot currently be used for real-world linkage. Seven candidate methods were identified from the literature as being used in multiple correspondence analysis. They were iteratively tested, first using tiny artificial datasets and subsequently using random samples of real-world census data. These tests were performed using a novel Python/PySpark implementation of

Scalelink. Initial testing resulted in a three-method short-list: the Missing Single method (which treats missingness as an agreement state) and two variants of the Missing Insertion method (which removes missingness by imputing donor values). Following testing, results were analysed by comparing precision, recall and the F1 metric. From this, the Missing Single method was clearly unsuitable for use with Scalelink. However, both of the other methods worked well, although they should be used for different types of missingness. Despite imputation being highly counter-intuitive for linkage, this work demonstrates that it is the best way to handle missingness when using Scalelink. This finding supports further testing of Scalelink in more realistic scenarios, increasing its potential to improve probabilistic data linkage and facilitate more reliable decision-making based on linked datasets.

