International Journal of Population Data Science

Journal Website: www.ijpds.org

Publishing Linked Data with Correction Weights

An-Chiao Liu[1], Peter Lugtig[1], Sander Scholtus[2], and Ton de Waal[3]

[1]Utrecht University, Utrecht, Netherlands

[2]Statistics Netherlands, Den Haag, Netherlands

[3]Statistics Netherlands, Den Haag, Netherlands; Tilburg University, Tilburg, Netherlands

Data linkage is often used to make inferences from distinct data sources. The data may be linked by a set of variables with high discrimination, for example, a social security number or some personal information. A good linking variable often implies a high disclosure risk in identifying the corresponding person. To prevent the disclosure risk, the data linker may choose to remove the linking variables before publishing the linked datasets. The design of the original data sources, the response pattern, and the unlinked part of the original data are often also unknown to the secondary user. However, the quality of the linked data is constrained by the original data sources and the linking process. Without considering the potential error or bias in linked datasets, naïvely treating them as error-free in secondary analysis may result in biased inference. To indicate the quality of the linked dataset without sacrificing privacy, we propose publishing correction weights alongside the linked datasets. The weights are generated given the information in the original data sources and the quality of the linkage. Both selection issues of the sample and measurement issues of the linking variables are addressed in the constructed weights, and we allow the possibility of having multiple potential links for a record. Secondary users may apply design-based estimators for subsequent analyses based on the correction weights, or apply sensitivity analysis given different sample inclusion criteria. An example is presented, and the option of secondary analysis given the constructed weights is discussed.

https://doi.org/10.23889/ijpds.v11i5.3531