International Journal of Population Data Science

Journal Website: www.ijpds.org

Drift in Data Linkage Models

Leah Quinn[1]

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

Data linkage refers to identifying the same entity across different records or datasets. Drift refers to a change in the characteristics of data, and the subsequent impact on model performance. A literature review was conducted to investigate drift's impact on data linkage methods. The scope included papers covering how to determine drift's existence and severity, its impact on linkage quality, and how to mitigate against it within linkage contexts. Linkage was found to be a severely lacking area in drift research. Drift literature focuses instead on machine learning models, where it is consistently shown to significantly degrade model performance and quality of outputs. Key themes from the review suggest that: There are numerous ways to identify the occurrence of drift. Drift has a notable negative impact on performance in a range of model types and contexts. The conditions which cause drift are present within many linkages. There is little to no research into drift within different methods for record linkage. We emphasise the need for research in this area to establish how best to measure drift within linkages and what severity of drift can be tolerated before the linkage quality is affected. Study into the most effective ways to mitigate drift when conducting linkage tasks may also help to avoid quality implications for later analysis.

https://doi.org/10.23889/ijpds.v11i5.3726