

# International Journal of Population Data Science

Journal Website: [www.ijpds.org](http://www.ijpds.org)



## The impact of drift on data linkage models

Rad Siddiqui<sup>1</sup> and Leah Quinn<sup>1</sup>

<sup>1</sup>Office for National Statistics, Manchester, United Kingdom

Data drift refers to the change in the statistical properties of input data; this data is collected for the training of a mathematical model.

We investigated the impact of drift on data linkage models (with further interest in machine-learning models), by researching how to determine drift existence and severity, impact on linkage, and how to mitigate against it. A literature review was conducted of scientific papers and journals focusing on these three key areas.

This review was done to assess the true potential impact of drift on a linkage model, and about what steps should be taken if it can be properly mitigated against.

The key findings are as follows:

- There are different types of drift – such as data, concept and performance. Research suggests it is not only necessary to determine the existence of drift, but also its type.
- Drift can critically damage model performance, and conditions that generate drift can be found in both linkage and non-linkage contexts.
- In particular, due to the potentially high level of mess, complexity and error that can come with it, moving towards big data could prove challenging and can result in poor linkage results; its non-stationary nature is understood to damage model performance.
- We can combat drift by utilising high quality model training data, avoiding over-training on highly specific contexts, and ensuring that subpopulations are not under-represented.

Unfortunately, study around the conjunction of linkage, drift and machine learning is developmentally lacking and limited; drift is an under-researched linkage concern that we should be aware of and demands extensive study.

We emphasize the need for more extensive studies into drift within linkage due to its potential to alter model outcomes. We see that the conditions that foster drift are potentially universal, and can therefore exist in any linkage task.

