

International Journal of Population Data Science

Journal Website: www.ijpds.org



Swansea University
Prifysgol Abertawe

Creating a Data Cleaning and Pre-Processing Module for Generalisable Data Linkage

Josie Plachta¹, Mary Cleaton¹, Leah Quinn¹, Alex Mackay¹, and Zoe White¹

¹Office for National Statistics

Objective and Approach

The Office for National Statistics (ONS) are developing a generalisable tool to facilitate the linkage of various datasets to its population-spine. However, a generalisable process requires that a variety of input datasets can be adaptively pre-processed – which is a problem for bespoke methodologies. Key requirements of the cleaning pipeline include minimal input from the user, and scalability and efficiency to work on Big Data.

Approach

The pipeline must recognise and adjust the pre-processing steps applied based on the variables present and user requirements. It must accept, preprocess, and derive consistent standardised variables from a variety of input variables and formats, including complex data characteristics.

Results

The MVP pipeline successfully met the requirements. It is based on a three-level hierarchy of functions, allowing flexibility and complexity in data preparation. With minimal user input, a variety of important linkage variables are cleaned, and additional variables derived consistently.

Conclusions & Implications

The module has shown promising results at scale, successfully pre-processing datasets of over 91 million records. It will be a valuable tool for increasing the ease and efficiency of record linkage to the ONS' population-spine. This will make linked data more accessible and increase the consistency of linked datasets, improving compatibility for onward linkages and the comparability of results.

Future work will involve applying this cleaning method to a wider range of different datasets to further test the generalisability of the method, and increasing the adaptability of the module to allow for even greater variation in the input datasets.

