

International Journal of Population Data Science

Journal Website: www.ijpds.org



Generating realistic synthetic data

Zoe White¹

¹ONS, Fareham, United Kingdom

Synthetic data has many different uses such as testing pipelines and training people for using new methods and technologies, which could be used within data linkage. The synthetic data generators project objectives are to create high fidelity synthetic data. To have high fidelity data, the synthetic data generator would need to include all variables, e.g. forename, surname and date of birth, and to include realistic errors, for example the frequency and types of errors. This ensures the synthetic data behaves similarly to the real underlying data. An initial workshop with users gained the requirements for; variables, data size, realistic errors and bias. For example, swapped forename and middle names for error, or missing children in census sized data. We also carried out

research into different synthetic data generators in Python and Pyspark, as well as meeting our user requirements. Our final approach was to create the functions and documentation for running the generators with errors and guidance on how to run the generator. Results, Conclusions and Implementation The synthetic data generator will be run end to end with user specifying where errors and biases are for a particular dataframe and compared to the genuine sensitive data. The usability of the code is highly important for ability to create synthetic data for developing code or in the training of new team members. Therefore, user testing of the synthetic data generator will be carried out to gather feedback, which will inform version 2 development or implemented within the code.

