**Grant Number:** N/A

**Sponsor:** N/A

**Project title:** SIPHER Synthetic Population for Individuals in Great Britain 2019-2021 – Supplementary Resources

The following files have been archived:

| File name | File description (Short description of content, sample size, format, any linking between different types of data, i.e. survey and interviews/focus groups) |
| --- | --- |
| Folder: “01\_constraints/” | This folder contains all aggregate-level population statistics data for Lower layer Super Output Areas / Data Zones in Great Britain which were used to create the SIPHER Synthetic Population. In total 8 constraint files are stored in the overarching folder **“01\_constraints/”.**  This folder contains the file **“readme\_licenses\_sources.txt”** which provides a detailed overview on all sources for constraint files and the applicable license agreements.  In addition, we have provided all underlying R-code pipelines in a sub-folder **“…/code\_for\_constraints/**” to allow for full reproducibility of these constraint files. Please note that constraint tables were obtained via web-queries means that an internet connection is required when reproducing these. |
| Folder: “02\_us\_modified/” | This folder contains a reproducible pipeline of R-code which details how we have modified the Understanding Society survey data files k\_hhresp.tab and k\_indresp.tab to align with the aggregate-level population statistics data (provided in the “01\_constraints/” folder). The resulting modified survey dataset was utilised when creating the SIPHER Synthetic Population. Please note that no Understanding Society survey datasets shared with this deposit. To fully re-run the pipelines, these files (SN6614) need to be downloaded manually and provided in the “**…/RData/”** subfolder. We have indicated with placeholders which exact files and where these files need to be stored. All code files are stored in the **“…/RCode/”** subfolder. Once the pipeline has been run successfully, a recoded survey dataset will appear in subfolder **“…/ROutput/”**. |
| Folder: “03\_validation/” | This folder contains a reproducible pipeline of R-code which provides examples of our quality control and validation process. In addition, the R-code in this folder demonstrates how to load, link and analyse the SIPHER Synthetic Population Dataset in combination with Understanding Society Survey data (SN6614).  Please note that neither the SIPHER Synthetic Population nor any survey datasets are provided. These files need to be downloaded separately (SN6614). We have indicated with placeholders where these data need to be stored to re-run the pipeline. All code files are stored in the **“…/RCode/”** subfolder. Once the pipeline is run successfully, the output of the validation pipeline will appear in subfolder **“…/ROutput/”**. |

**Publications**:

This deposit supplements the data resources “SIPHER Synthetic Population for Individuals in Great Britain 2019-2021” and its respective user guide. The dataset is available via the UK Data Service under the following DOI: <https://doi.org/10.5255/UKDA-SN-9277-1>

A previous version of the SIPHER Synthetic Population has been described in full detail here:

* Wu, G., Heppenstall, A., Meier, P. et al. A synthetic population dataset for estimating small area health and socio-economic outcomes in Great Britain. Sci Data 9, 19 (2022). <https://doi.org/10.1038/s41597-022-01124-9>