SIPHER's Synthetic Population for Individuals in Great Britain: Creation, Validation, and Examples of Application

Authors: Andreas Höhn[1], Nik Lomax[2], Kashif Zia[1], Fraser Bell[3], Emma Comrie[4], Gillian Fergie[1], Alison Heppenstall[1], Robin Purshouse[5], Jo Winterbottom[1], and Petra Meier[1]

Affiliations: [1] University of Glasgow, [2] University of Leeds, [3] Greater Manchester Combined Authority, [4] Public Health Scotland, , [5] University of Sheffield

Funding: This work by the SIPHER Consortium was supported by the UK Prevention Research Partnership (MR/S037578/2), which is funded by the British Heart Foundation, Cancer Research UK, Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Health and Social Care Research and Development Division (Welsh Government), Medical Research Council, National Institute for Health Research, Natural Environment Research Council, Public Health Agency (Northern Ireland), The Health Foundation and Wellcome.

Acknowledgement: This research was conducted as part of the Systems Science in Public Health and Health Economics Research - SIPHER Consortium and we thank the whole team for valuable input and discussions that have informed this work. We are very grateful for the opportunity to work with the UK Household Longitudinal Study (Understanding Society) and would like to thank the UK Data Service for hosting the access.

Ethical Approval: The University of Essex Ethics Committee has approved all data collection for the Understanding Society main survey. No additional ethical approval was necessary for other data sources used in our analyses. This data was consistently open and publicly available, subject to acceptance of an End User Licence Agreement.

1. Background

Unlike other European countries, such as the Nordic countries, Great Britain (GB) does not have a comprehensive population register-based system. This limits the availability of data which allows researchers, analysts, and policymakers to explore the interaction of health with other life domains such as employment or housing at the individual level. In some cases, survey data can be considered a suitable alternative as it provides a representative sample of the population at individual level. However, survey data does typically not allow for a small-area perspective. While special license linkages of survey data may enable a more detailed geographical resolution, sampling strategies are often not representative for the geographical resolution of interest (Benzeval 2020). In addition, the small number of survey respondents across small areas often presents a further limitation as it reduces statistical power.

Creating full-scale synthetic populations via spatial microsimulation can provide an opportunity to overcome limitations surrounding the availability of attribute-rich individual-level data and statistical power (Harland et al 2012). In this study, we describe the creation and validation of a full-scale synthetic population dataset, representative of the adult population in GB at a small-area level. Our work represents an update of the methodology and data sources described in Wu et al. (2022), and will be accompanied by a shared dataset that will be made available via the UK Data Service (*DOI pending*).

The dataset was created and validated by the System Science in Public Health and Health Economics Research (SIPHER) consortium (Meier et al. 2019), which is a multi-disciplinary group of scientists and government partners at local, regional, and national levels. Highlighting the utility of this dataset, we will illustrate how the data set is currently used as part of spatial assessments and dynamic microsimulation applications across the SIPHER consortium (Lomax et al. 2023, Clay et al. 2023).

2. Data and Methods

SIPHER's Synthetic Population for Individuals was created by combining microdata from the general license version of the Understanding Society, also known as the UK Household Longitudinal Survey (UKHLS), wave 11 / "k" (representing a collection period spanning 2019 and 2020), with aggregate-level data from the 2011 UK Census for England, Wales and Scotland and population statistics data reflecting the year 2020 (referred to as constraints). Data for Northern Ireland were not used because of a lack of alignment with 2011 Census outputs. Understanding Society is the largest and longest running UK panel study, designed to be representative of the UK population with a sample size of 39,802 households at Wave 1. All data sources (survey, census and population estimates) are generally free and publicly available. **Table 1** provides an overview of the utilised variables from the Understanding Society main survey.

Table 1: Overview of all utilised Understanding Society main survey variables.

Constraint	Variables in Understanding Society	Source
Age/Sex	age_dv sex	k_indresp.tab
Highest qualification	hiqual_dv	k_indresp.tab
Ethnicity	racel_dv	k_indresp.tab
Marital status	marstat	k_indresp.tab
Economic activity	jbstat	k_indresp.tab
General health	scsf1	k_indresp.tab
Household tenure	tenure_dv	k_hhresp.tab
Household type (Household Composition)	hhtype_dv	k_hhresp.tab

To combine survey and population statistics data, we used the JavaScript-based Flexible Modelling Framework (FMF) software (Harland 2013). Using a combinatorial optimisation algorithm (simulated annealing), the FMF creates a full-scale synthetic population by repeatedly assigning individuals from the Understanding Society main survey to small areas in GB in such a way that the overall result resembles the information provided in the constraint files. We utilised data on the following constraints to inform the FMF algorithm: 2020 population estimates (age/sex); 2011 Census data for highest qualification, ethnicity, marital status, economic activity, general health, household type ("household composition"), and household tenure. This means that SIPHER's Synthetic Population is representative with respect to the above characteristics at the small-area level. In detail, SIPHER's Synthetic Population for Individuals is based on the spatial resolution of Lower Super Output Areas (LSOAs, ~ 2000 individuals per area) for England and Wales and Data Zones (DZs, ~ 700 individuals per area) for Scotland. An overview of all utilised population statistics data used as constraints and informing the FMF algorithm is provided in **Table 2.**

The process of aligning survey and population statistics data, as well as the underlying spatial microsimulation process, have previously been described in full detail elsewhere (Wu et al. 2022). The dataset, alongside a suite of examples and supplementary material documenting creation and validation has been submitted to

the UK Data Service and will soon become widely available for all registered UK Data Service users.

Table 2: Overview of all utilised aggregate-level population statistics data (*constraints*) used to inform the spatial microsimulation algorithm of the Flexible Modelling Framework software.

Constraint	**Year**	**Source**
Age/Sex	2020	NOMIS API
Highest qualification	2011	census 2011 tables QS501EW/SC - Highest level of qualification; NOMIs API ID 554 (for E&W only)
Ethnicity	2011	census 2011 tables LC6201EW/SC - Economic activity by ethnic group by age; NOMIs API ID 818 (for E&W only)
Marital status	2011	census 2011 tables LC6201EW/SC - Economic activity by ethnic group by age; NOMIS API ID 818 (for E&W only)
Economic activity	2011	census 2011 tables LC6201EW/SC - Economic activity by ethnic group by age; NOMIS API ID 818 (E&W only)
General health	2011	census 2011 tables QS302EW/SC - General health; NOMIS API ID 531 (for E&W only). All raw data was reflective of all ages and did not provide the opportunity to distinguish age groups. We have implemented an adjustment to only reflect the age range 16+.
Household tenure	2011	for England and Wales census 2011 table LC3408EW - Long-term health problem or disability by tenure by age; NOMIS API ID 1403 (for E&W only). For Scotland this table does not exist with individuals as unit of observation, and we therefore diverted to table QS403SC. The raw data for Scotland was reflective of all ages and did not provide the opportunity to distinguish age groups. We have therefore implemented an adjustment to only reflect the age range 16+.
Household type (“composition”)	2011	census 2011 tables LC1109EW/SC - Household composition by age by sex; NOMIS API ID 849 (for E&W only)

3. Results of Validation

As all synthetic data sources, SIPHER's Synthetic Population for Individuals requires a thorough validation process. This is to ensure that the data aligns with the provided constraints and accurately reflects the distribution of attributes it has been designed to capture (Edwards and Tanton 2012). To validate our data set, we compared aggregate-level statistics for small areas obtained from SIPHER's Synthetic Population with (1) aggregate-level information from the utilised constraint tables (*internal validation*), and (2) aggregate-level information of related external data sources (*external validation*).

Results of the internal validation, comparing the number of simulated individuals within each LSOA and DZ in SIPHER's Synthetic Population that represent a range of attributes with the expected number of individuals as provided by the constraint file are summarised in **Figure 1.**

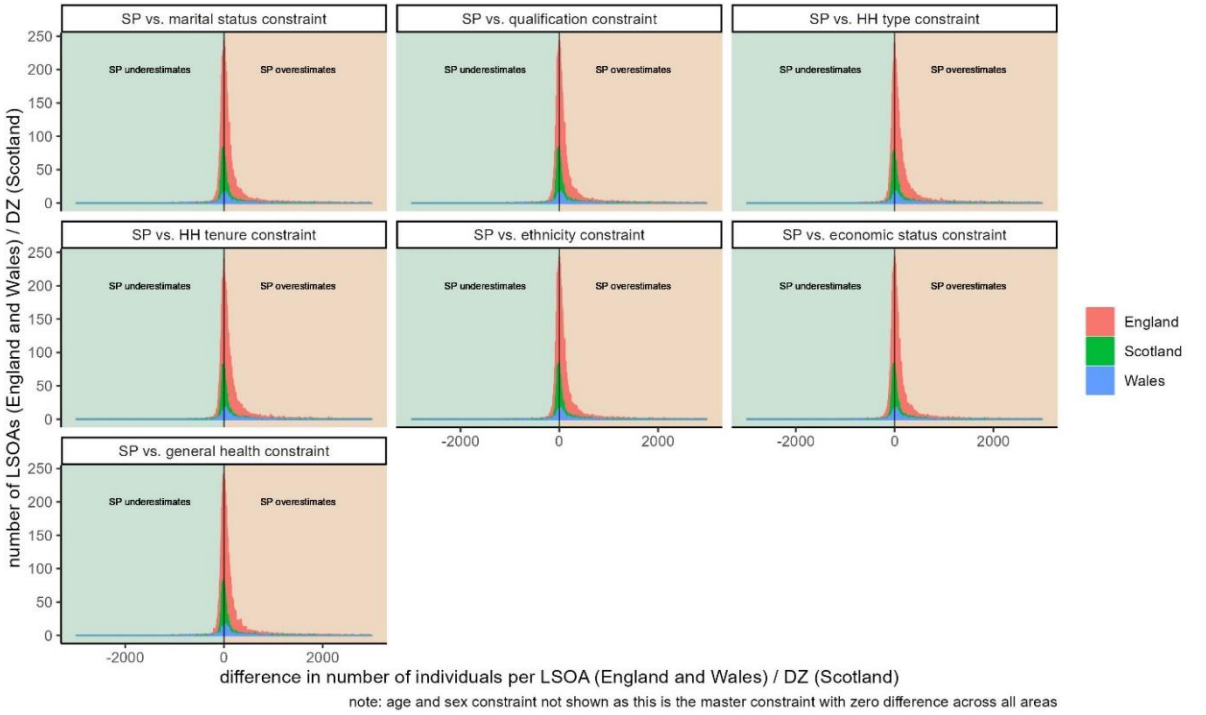

Figure 1: Overview of differences in the total number of individuals across areas; comparing SIPHER's Synthetic Population for Individuals vs. utilised constraint tables representing aggregate-level population statistics for all Lower Super Output Areas in England and Wales as well as Data Zones in Scotland. Note: "SP" refers to our synthetic dataset.

Deviations in the total number of simulated individuals per area by attribute in the synthetic data set in comparison with the expected number of individuals can occur either due to a misalignment of survey and population statistics data, or emerge from a poor performance of the spatial microsimulation algorithm when populating areas. Ideally, differences should be as small as possible, with deviations following a normal distribution around a median and mean of zero. As shown in **Figure 1**,

SIPHER's Synthetic Population achieved an excellent fit in terms of number of individuals for each of the utilised constraint tables.

SIPHER's Synthetic Population is based on 8 constraints reflecting a wide range of sociodemographic characteristics as well as individuals' general health status. While information on occupational class has not been used as a constraint, it can be assumed that the concept should be well represented in SIPHER's Synthetic Population due to its strong association with the 8 included constraints. To explore this, we compared area-level patterns in NS-Sec classification (a measure of occupational structure) across Scottish DZs obtained directly from SIPHER's Synthetic Population with data on the NS-Sec classification reported in the 2011 Census.

As shown in **Figure 2**, SIPHER's Synthetic Population showed a strong similarity with a 5-category based NS-SeC classification. Despite a good reflection of area-level patterns in occupational class, care is required when interpreting findings for residual categories of variables which were not used as constraints as well as all variables for which the association which the constraint variables have not been explored. For example, as shown in **Figure 2**, our synthetic dataset might not be well suited to capture individuals forming the "other" category with respect to occupational class – although a good fit is achieved for all other categories.

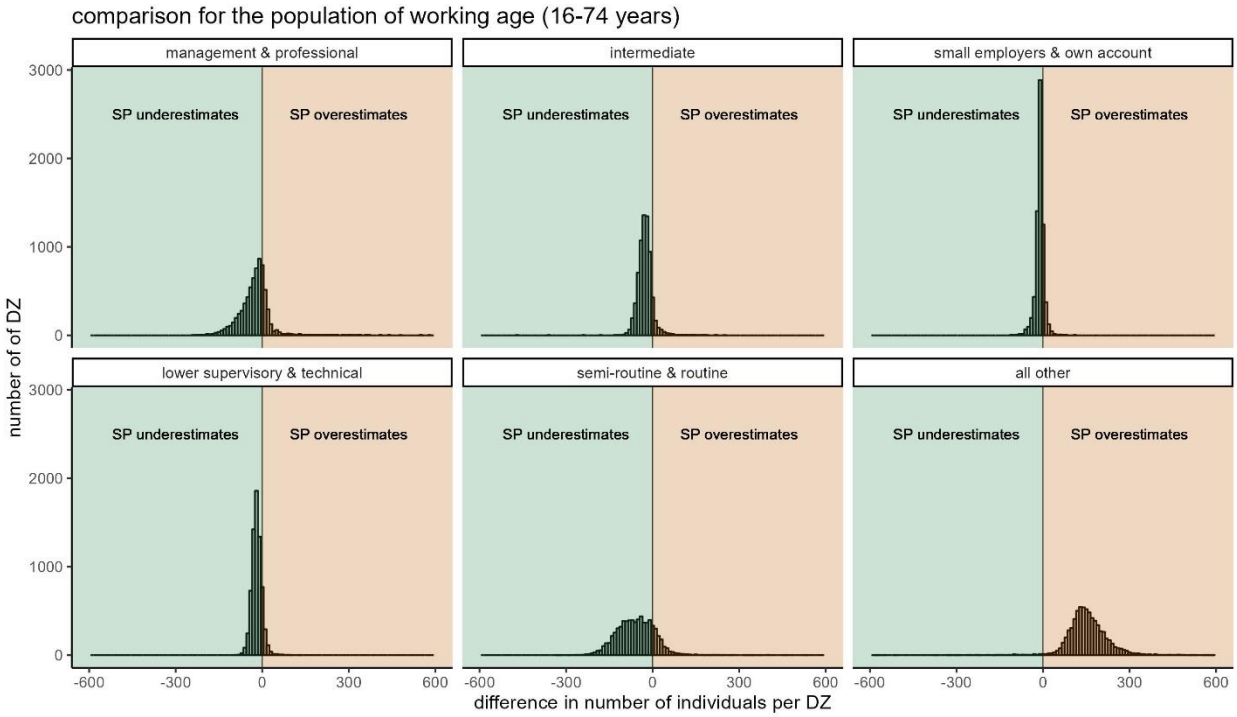

Figure 2: Comparison of area-level patterns in NS-Sec classifications across Scottish Data Zones: SIPHER's Synthetic Population vs. Census 2011 data. Note: information on occupational class has not been used when creating SIPHER's Synthetic Population, but its distribution within Scottish Data Zones is used as a point reference in this comparison.

4. Examples of Application

SIPHER's Synthetic Population for Individuals in Great Britain enables a wide range of applications. For example, the data can be used to derive descriptive statistics for needs assessments and prioritisation decisions. Furthermore, the data can be used as an input population in pre-intervention modelling of potential policy interventions via a dynamic microsimulation model (e.g., Clay et al 2023), or inform model parameters such as behavioural rules of agent-based models (ABMs). Here, one unique advantage of SIPHER's Synthetic Population for Individuals is that it enables a dedicated small-area perspective.

Figure 3 illustrates this advantage by presenting raw average Shortform-12 (SF-12) summary scores for physical (PCS) and mental (MCS) health, for all Lower Super Output Areas (LSOAs) of the Greater Manchester Combined Authority (GMCA), derived from more than 2 million individual-level data points.

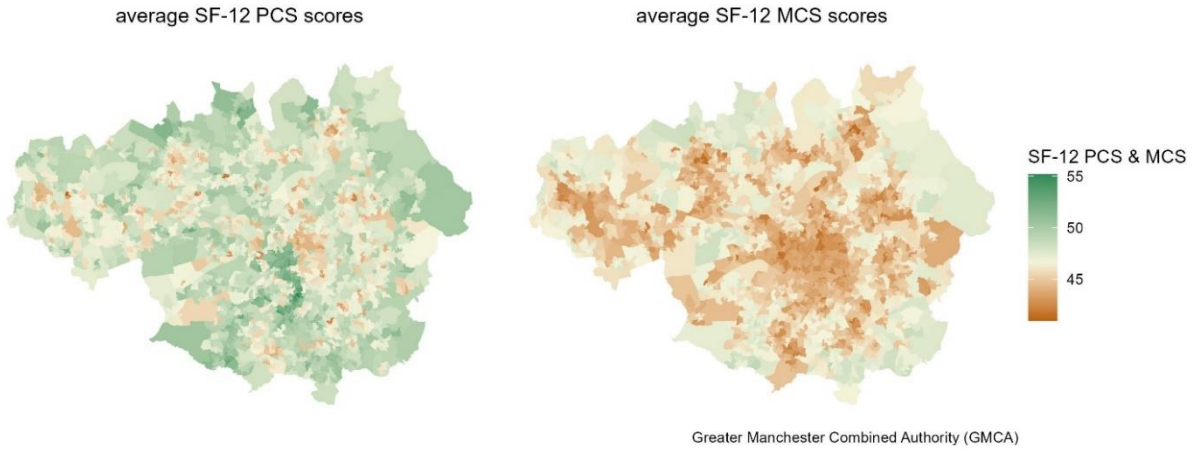

Figure 3: Spatial plotting of Shortform-12 physical (PCS) and mental (MCS) health summary scores for all Lower Super Output Areas of the Greater Manchester Combined Authority, reflecting census 2011 boundary definitions.

Across the SIPHER consortium, our synthetic population data set is currently used primarily in applications assessing needs as well as an input data source in dynamic microsimulations. For example, the data set is used in a project examining the impact of the Energy Price Cap Guarantee and Energy Bill Support Scheme Policies, introduced by the UK Government in response to the current cost-of-living crisis, via the dynamic microsimulation framework MINOS (Clay et al 2023).

In another application, a subset for the Scottish population is used to quantify the magnitude of population health improvements emerging from meeting the child poverty targets as outlined by the current Scottish Government (Scottish

Government 2022). As both applications progress and evolve further, we will provide finalised results of this co-produced research in summer 2024.

5. Conclusion

SIPHER's Synthetic Population for Individuals is a validated, high-quality, full-scale synthetic population data set reflecting the adult population in GB. It will soon become available for all registered users of the UK Data Service. The data is well-suited to examine the interaction of health and socioeconomic domains among individuals at a small-area level. Despite careful validation, care is recommended when studying individual-level characteristics for which the association with the utilised constraint variables has not yet been evaluated or when residual categories are being examined.

References

Benzeval, M., Bollinger, C. R., Burton, J., Crossley, T. F., & Lynn, P. (2020). The representativeness of understanding society. Institute for Social and Economic Research, 2020-08.

Clay, R., Archer, L., Heppenstall, A., & Lomax, N. (2023). Using the Dynamic Microsimulation MINOS to Evidence the Effect of Energy Crisis Income Support Policy (Short Paper). In 12th International Conference on Geographic Information Science (GIScience 2023). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Edwards, K. L., & Tanton, R. (2012). Validation of spatial microsimulation models. In Spatial microsimulation: A reference guide for users (pp. 249-258). Dordrecht: Springer Netherlands.

Harland, K., Heppenstall, A., Smith, D., & Birkin, M. H. (2012). Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. Journal of Artificial Societies and Social Simulation, 15(1).

Harland, K. Microsimulation Model User Guide (Flexible Modelling Framework). Working Paper, NCRM https://eprints.ncrm.ac.uk/id/eprint/3177 (2013).

Lomax, N., Clay, R., Archer, L., Rice, H. and Heppenstall, A., 2023. Briefing note: A dynamic model of disposable income impacts on mental health. No 2679v, OSF Preprints, Centre for Open Science, 10.31219/osf.io/2679v.

Meier, P., Purshouse, R., Bain, M., Bambra, C., Bentall, R., Birkin, M., et al. (2019). The SIPHER consortium: Introducing the new UK hub for systems science in public health and health economic research. Wellcome open research, 4.

Scottish index of multiple deprivation (SIMD) 2020: Ranks https://www.gov.scot/publications/scottish-index-of-multiple-deprivation-2020v2-ranks/ last accessed: 2023-08-08.

Scottish index of multiple deprivation (SIMD) 2020: Technical notes https://www.gov.scot/publications/simd-2020-technical-notes/ last accessed: 2023-08-08.

Scottish Government 2022: Tackling Child Poverty Delivery Plan 2022-26. https://www.gov.scot/news/tackling-child-poverty-delivery-plan-2022-26/

University of Essex, institute for social and economic research. (2022). Understanding society: Waves 1-12, 2009-2021 and harmonised BHPS: Waves 1-18, 1991-2009. [Data collection]. 17th edition. UK data service. SN: 6614, http://doi.org/10.5255/UKDA-SN-6614-18. (n.d.).

Wu, G., Heppenstall, A., Meier, P., Purshouse, R., & Lomax, N. (2022). A synthetic population dataset for estimating small area health and socio-economic outcomes in Great Britain. Scientific Data, 9(19).