USER GUIDE: SIPHER Synthetic Population for Individuals in Great Britain 2019-2021

Contents

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.

Acknowledgements

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. Understanding Society is an initiative funded by the Economic and Social Research Council (ESRC) and various Government Departments, with scientific leadership by the Institute for Social and Economic Research (ISER), University of Essex, and survey delivery by the National Centre for Social Research (NatCen) and Verian (formerly Kantar Public). The research data are distributed by the UK Data Service. We would like to thank Sadiq Rahman and Cristina Magder at the UK Data Service for stimulating discussion surrounding data properties, data sharing, and data documentation.

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 the creation of the SIPHER Synthetic population our analyses presented in this user guide.

Abbreviations

DZ – Data Zone (Scotland)

DWP - Department for Work and Pensions

FMF – Flexible Modelling Framework

GB – Great Britain

GMCA – Greater Manchester Combined Authority

LSOA/MSOA – Lower/Medium layer Super Output Area

NRS – National Records of Scotland

Ns-Sec – National Statistics Socio-economic classification

ONS – Office for National Statistics

SF-12 – 12-Item Short Form Survey

SIMD – Scottish Index of Multiple Deprivation

SIPHER Consortium - Systems Science in Public Health and Health Economics Research Consortium

UK – United Kingdom

1. Overview and Summary

This user guide provides a comprehensive overview of the SIPHER Synthetic Population for Individuals in Great Britain 2019-2021 (hereafter referred to as: Synthetic Population). Through a linkage with Understanding Society[1] survey data, the Synthetic Population allows for the creation of a survey-based full-scale synthetic population of Great Britain (GB). Drawing on data reflecting *"real"* survey respondents, the dataset allows us to reflect information for over 50 million synthetic (i.e. *"not real")* individuals. As the Synthetic Population is the outcome of a statistical creation process, all results obtained from this dataset should always be treated as *"model output"*, including basic descriptive statistics. While the Synthetic Population should not replace data reflecting actual (*"real"*) individuals in standard statistical analyses (e.g., regression analysis) typically performed with the Understanding Society survey, the dataset provides a novel source of data for understanding "*status quo*" or modelling *"what if"* scenarios (e.g., through static/dynamic microsimulations). In addition to this, the Synthetic Population can provide a dataset for "e*xploratory"* analyses and cases where a granular geographical resolution is required but no other data sources are available (e.g. estimation of indicators for small areas).

Once linked with survey data, the Synthetic Population reflects a synthetic version of the population aged 16 years and older in England, Scotland, and Wales. The resulting linkage is representative with respect to a number of key sociodemographic characteristics (hereafter referred to as: constraints) at the most granular non-disclosed area level reported publicly in the UK census 2011 (i.e. Lower layer Super Output Areas (LSOAs) for England and Wales and Data Zones (DZs) for Scotland). As a "*digital twin*" of the adult population in GB, the Synthetic Population supports a wide range of applications which require high-quality information on individuals and areas at a granular spatial resolution.

The Synthetic Population has been created through a spatial microsimulation approach called simulated annealing. For this approach, we combined data from the Understanding Society main survey with publicly available population statistics data describing LSOAs and DZs. The Synthetic Population harnesses the unique power of the Understanding Society survey - the UK's largest and longest running multi-topic panel

[1] https://www.understandingsociety.ac.uk/

study based on data collected from a large nationally representative sample of UK households with additional regional, ethnic minority and immigrant boost samples (Buck and McFall 2012). Here, the Synthetic Population builds directly upon the rich representation of different population subgroups and their heterogeneous and unique life courses across multiple life domains captured by the survey.

The Synthetic Population was developed by the SIPHER Consortium (Meier et al. 2019). With this upload, we aim to stimulate the independent usage of the Synthetic Population beyond the SIPHER Consortium. At the same time, this upload represents an update of previously published and peer-reviewed methodology for creating a full-scale synthetic population dataset for GB (Wu et al. 2022). While the Synthetic Population described in this user guide is based on the same seven core constraints previously used (i.e.: age/sex, highest qualification, ethnicity, marital status, economic activity, household type (composition), and household tenure), we have now also incorporated general health as an additional 8th constraint in this updated Synthetic Population.

2. Rationale for Creation

As other European countries, the countries of GB capture information on individuals and households in administrative registers. Similarly to other countries, the registers available for GB and its countries can be linked and accessed for different purposes such as research, planning of public services, and the routine reporting of population statistics. Despite this similarity, GB's registers tend to be less comprehensive – for example in comparison to the Nordic countries. Here, key factors contributing to differences are the decentralised nature of GB's registers, the dominance of a *"create and destroy"* approach for requested data linkages (contrasting the Nordic Countries' *"ready and re-use"* model), as well as the absence of a unified person identification system for cross-register linkages (see United Nations Economics Commission 2007 for further details).

The absence of a centralised and comprehensive register-based system limits opportunities for studying the interaction of aspects such as health, employment, education, benefit payments, or housing at the level of individuals and households in GB. At the same time, and as in all countries

with administrative registers, the data that exist for GB are strictly safeguarded. This means that GB's registers are not available outside of strictly safeguarded research environments. Especially when testing and modelling different policy scenarios (e.g. understanding *"status quo"* or modelling "*what-if"* scenarios), the sum of these limitations can present important hurdles to coproduction and collaborative work connecting researchers, policymakers, and other key stakeholders.

In some cases, surveys can provide a well-suited and swiftly available alternative to register-based data. However, survey data do typically not allow for a detailed spatial resolution. In some cases, special licence user agreement area-level linkages of surveys (such as Understanding Society), can enable a more granular spatial resolution. For example, it is possible to obtain information on the local authority of Understanding Society survey respondents throughout an area-level linkage. Despite an increase in geographical granularity, the resulting linkage comes with a reduced sample size for each local authority. As a result, estimates obtained for local authorities can be subject to a high level of uncertainty leading to wide confidence intervals surrounding point estimates (see Höhn et al. 2024 for an example). Disaggregating survey data even further (e.g., LSOA/DZ level) is often not meaningful at all due to the very small number of survey respondents across these areas, and as not all areas at such low levels are covered in the Understanding Society survey.

Survey-based full-scale synthetic population datasets can help to bypass both, the highlighted limitations of administrative data in GB and available surveys. By providing readily available and attribute-rich micro data, survey-based full-scale synthetic populations can enable both: Full coverage and representativeness with respect to a set of utilised constraints obtained from external sources (e.g., UK Census), at a granular geographical resolution. In case of the Synthetic Population, both features were achieved through the use of population statistics data which informed the creation of the dataset in which survey respondents are assigned to synthetic zones (see Section 3 for more details). With increasing computational resources, survey-based full-scale synthetic population datasets have become increasingly popular - even within countries where comprehensive register-based systems are available (see Prédhumeau and Manley 2023 and Tozluoğlu et al. 2023 for further details).

Despite the Synthetic Population's full coverage and (aggregate-level) representativeness at a granular geographical resolution, we always

recommend using Understanding Society survey data for any studies seeking to understand associations or causal structures at the level of individuals in the UK, cross-sectionally or over time.

3. Creation of the Synthetic Population

The Synthetic Population was created by combining individual-level data from the Understanding Society main survey of wave 11 (also referred to as wave *"k"*) with aggregate-level population statistics data (hereafter referred to as: constraint tables). The utilised Understanding Society[2] survey data collection is available as standard safeguarded data from the UK Data Service, subject to the End User Licence Agreement terms and conditions.

Understanding Society, also known as the UK Household Longitudinal Study, is a large-scale, long-term panel survey that captures valuable information about the social and economic circumstances and attitudes of people living in the UK. It tracks changes over time, providing insights into the long-term effects of policy changes and economic cycles on British households and individuals. The study began in 2009 and is designed to represent the diversity of the UK population, offering a unique window into the dynamics of UK life.

Constraint tables were primarily obtained from the UK census 2011, which were provided by the Office for National Statistics (ONS) for LSOA's in England and Wales as well as National Records of Scotland (NRS) for Scottish DZs[3]. All utilised constraint tables are free and publicly available under an Open Government License for public sector information. At the time of creation, this Synthetic Population represents the best overlap of required and available constraint tables for small areas in England, Scotland, and Wales.

---

[2] 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, DOI: http://doi.org/10.5255/UKDA-SN-6614-18

[3] Please see **Table 2** for a detailed overview of all constraint tables and their sources.

The Synthetic Population was created via spatial microsimulation; in particular a method called simulated annealing (Wu et al. 2022). The underlying combinatorial optimisation algorithm (simulated annealing) is implemented in the JavaScript-based Flexible Modelling Framework (FMF) software. The algorithm has previously been described in full detail elsewhere (Harland 2013, Wu et al. 2022). In brief: The FMF software creates a synthetic population dataset by repeatedly assigning individuals from the Understanding Society main survey to LSOAs/DZs in GB. The repeated assignment of individuals is informed by the provided constraint tables, with the aim to resemble the aggregate-level patterns provided in these files as optimal as possible.

To allow for this algorithm-based assignment process, a preparatory step required an alignment of variables and their categories included in the Understanding Society survey of wave 11 (*"k"*) with information covered in the constraint tables. This process of alignment defines the domains with respect to which the resulting Synthetic Population dataset is representative of.

Table 1 provides a summary of all utilised variables from the Understanding Society survey and their exact source.

Table 1: Overview of all constraint dimensions and variables representing these dimensions in the Understanding Society survey of wave 11 ("k"), SN6614.

Constraint Dimension	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 (composition)	hhtype_dv	k_hhresp.tab

All constraint tables reflecting aggregate-level population statistics data for small areas were obtained via code-based web queries. This procedure ensures a maximum amount of replicability, reproducibility, and scalability. All resulting constraint tables, alongside the code we have utilised to query and modify raw data, are shared as a supplementary resource[4].

Table 2 provides a detailed overview of all utilised aggregate-level constraint tables and their exact source.

Table 2: Overview of all utilised aggregate-level population statistics data (constraint tables), the year reflected, and the exact source of each table.

Constraint Dimension	**Year**	**Source**
Age/sex	2020	NOMIS API[5]

[4] Available as a separate resource on ReShare:

http://doi.org/10.5255/UKDA-SN-856754

[5] Source for England & Wales: https://www.nomisweb.co.uk/query/construct/summary.asp?mode=construct&version=0&dataset=2010

Source for Scotland: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/small-area-population-estimates-2011-data-zone-based/time-series, file for year 2020: "sape-2020.xlsx"

Highest qualification	2011	Census 2011 tables QS501EW/SC - Highest level of qualification; NOMIs API ID 554 (for E&W only)[6]
Ethnicity	2011	Census 2011 tables LC6201EW/SC - Economic activity by ethnic group by age; NOMIs API ID 818 (for E&W only)[7]
Marital status	2011	Census 2011 tables LC6201EW/SC - Economic activity by ethnic group by age; NOMIS API ID 818 (for E&W only)[8]
Economic activity	2011	Census 2011 tables LC6201EW/SC - Economic activity by ethnic group by age; NOMIS API ID 818 (for E&W only)[9]

[6] Source for England and Wales:
https://www.nomisweb.co.uk/query/construct/summary.asp?reset=yes&mode=construct&dataset=554&version=0&anal=1&initsel=

Source for Scotland:
https://nrscensusprodumb.blob.core.windows.net/downloads/SNS%20Data%20Zone%202011%20blk.zip, file: "QS501SC.csv"

[7] Source for England and Wales:
https://www.nomisweb.co.uk/query/construct/summary.asp?reset=yes&mode=construct&dataset=818&version=0&anal=1&initsel=

Source for Scotland:
https://nrscensusprodumb.blob.core.windows.net/downloads/SNS%20Data%20Zone%202011%20blk.zip, file: "LC6201SC.csv"

[8] Source for England and Wales:
https://www.nomisweb.co.uk/query/construct/summary.asp?reset=yes&mode=construct&dataset=603&version=0&anal=1&initsel=

Source for Scotland:
https://nrscensusprodumb.blob.core.windows.net/downloads/SNS%20Data%20Zone%202011%20blk.zip, file: "KS103SC.csv"

[9] Source for England and Wales:
https://www.nomisweb.co.uk/query/construct/summary.asp?reset=yes&mode=construct&dataset=818&version=0&anal=1&initsel=

Source for Scotland:
https://nrscensusprodumb.blob.core.windows.net/downloads/SNS%20Data%20Zone%202011%20blk.zip, file: "LC6201SC.csv"

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+.[10]
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+.[11]
Household type (composition)	2011	Census 2011 tables LC1109EW/SC - Household composition by age by sex; NOMIS API ID 849 (for E&W only)[12]

[10] Source for England and Wales: https://www.nomisweb.co.uk/query/construct/summary.asp?reset=yes&mode=construct&dataset=531&version=0&anal=1&initsel=

Source for Scotland: https://nrscensusprodumb.blob.core.windows.net/downloads/SNS%20Data%20Zone%202011%20blk.zip, file: "QS302SC.csv"

[11] Source for England and Wales: https://www.nomisweb.co.uk/query/construct/summary.asp?reset=yes&mode=construct&dataset=1403&version=0&anal=1&initsel=

Source for Scotland: https://nrscensusprodumb.blob.core.windows.net/downloads/SNS%20Data%20Zone%202011%20blk.zip, file: "QS403SC.csv"

[12] Source for England and Wales: https://www.nomisweb.co.uk/query/construct/summary.asp?reset=yes&mode=construct&dataset=849&version=0&anal=1&initsel=

4. Working with the Synthetic Population

4.1 Variable Description

The result of the spatial microsimulation process is the shared Synthetic Population, described in this User Guide. The filename is *"sp_ind_wavek_census2011_est2020_8cons.csv"*. Analogously to Understanding Society's area-level linkage files, the Synthetic Population file contains two key variables:

- ***"synthetic_zone":*** A variable of character type reflecting the assigned synthetic LSOA / DZ for each synthetic individual. These area codes are based on the classification the ONS has used for the 2011 UK census. Using lookup tables provided by the ONS, these codes can be used to build current and historic higher-level geographies of interest such as electoral wards, or local authorities.
- ***"pidp":*** A variable of numeric type which corresponds directly to the "*pidp*" person id number in the Understanding Society survey dataset for individuals, wave 11 ("k"). While "*pidp*" is a unique identifier in the Understanding Society survey dataset, this variable is not unique in the Synthetic Population.

In total, the two-variable Synthetic Population file covers approximately 52 million observations (rows) – one for each synthetic individual and its synthetic area.

4.2 Linkage with the Understanding Society Survey

Before we can work with the Synthetic Population, a preparatory step is required. During this preparatory step, we need to merge the Synthetic Population file *"sp_ind_wavek_census2011_est2020_8cons.csv*" with the Understanding Society survey datasets "*k_hhresp*" and/or "*k_indresp*". The linkage is performed analogously to other area-level linkages via the "*pidp*"

Source for Scotland: https://nrscensusprodumb.blob.core.windows.net/downloads/SNS%20Data%20Zone%202011%20blk.zip, file: "LC1109SC.csv"

variable, and can be achieved using any statistical software such as R, Python, Stata, or SPSS.

Please note that due to the size of the resulting linkage, it is not advised to link all variables captured in "*k_hhresp*" and "*k_indresp*" with all synthetic individuals covered in the Synthetic Population. To optimise computational performance and memory use throughout we recommend;

(1) to subset the Understanding Society survey datasets as early as possible to only include variables of interest for the intended linkage.

(2) to subset the Synthetic Population as early as possible to only reflect the geographical areas relevant for the application.

Figure 1 provides an illustration of the general Synthetic Population data structure and a resulting linkage example with the recommended Understanding Society survey datasets of wave *"k"*. As shown in **Figure 1**, the Synthetic Population was linked with the files *"k_indresp"* and *"k_hhresp"*.

ZoneID (LSOA / Datazone)	pidp (not unique)
E01004766	1
E01004766	2
E01004766	3
E01004766	4
E01004766	5
E01004767	1
E01004767	7
E01004767	4

Synthetic Population

Age	Sex	SF-12 Mental Health Score
20	Male	54.12
24	Female	47.69
34	Male	37.45
87	Female	51.71
49	Male	52.65
20	Male	54.12
54	Male	47.78
87	Female	51.71

k_indresp

HH has problems paying Council Tax
No
No
Yes
No
No
No
No
No

k_hhresp

Figure 1: Illustration of a linked Synthetic Population. Please note that these numbers do not reflect true survey respondents and are entirely hypothetic. Please note that the person identification number "*pidp*" is not unique in the Synthetic Population – while it is unique in the Understanding Society survey.

4.3 Applications

The dataset is primarily intended for understanding and modelling "*status quo*" or *"what if"* scenarios (e.g., through static/dynamic microsimulations). In addition to this, the Synthetic Population can provide a dataset for "e*xploratory"* analyses and cases where a granular geographical resolution is required but no other data sources are available (e.g. estimation of indicators for small areas). Due to the statistical processes underpinning the creation of the Synthetic Population, all insights gained from the dataset should always be interpreted with caution and treated as model output. This includes basic descriptive statistics obtained from the dataset.

As a readily available dataset, the Synthetic Population can provide a "*digital twin*" of the adult population aged 16 years and older in GB. The dataset harnesses and amplifies many strengths of the underlying Understanding Society survey, including its large sample size and multi-topic character. Being swiftly available outside of safe haven environments, the dataset is well suited for co-produced research connecting researchers, policymakers, and other stakeholders. In addition, the dataset provides a well-suited resource for teaching purposes - for example, as it can mimic large-scale multi-domain data reflecting total populations. The wealth of variables captured in the Understanding Society survey allows for a substantial level of intersectionality, while maintaining representativeness at a granular geographical resolution.

Once linked with the Understanding Society survey, the Synthetic Population enables a wide range of individual-level and aggregate-level applications. Potential applications for the Synthetic Population can range from descriptive statistics for assessing needs at the area-level, over testing of policy options using microsimulation models, to informing model parameters of Agent-Based Models. In addition to this, the Synthetic Population can provide a starting point for exploring probabilistic linkages at or look-alike modelling.

Here, the unique strength of the Synthetic Population is its ability to enable statistical power at a granular spatial resolution. **Figure 2** illustrates this advantage by summarising (raw, not age-standardised) 12-Item Short Form Survey (SF-12) scores for more than 2 million synthetic individuals across the LSOAs of the Greater Manchester Combined Authority (GMCA).

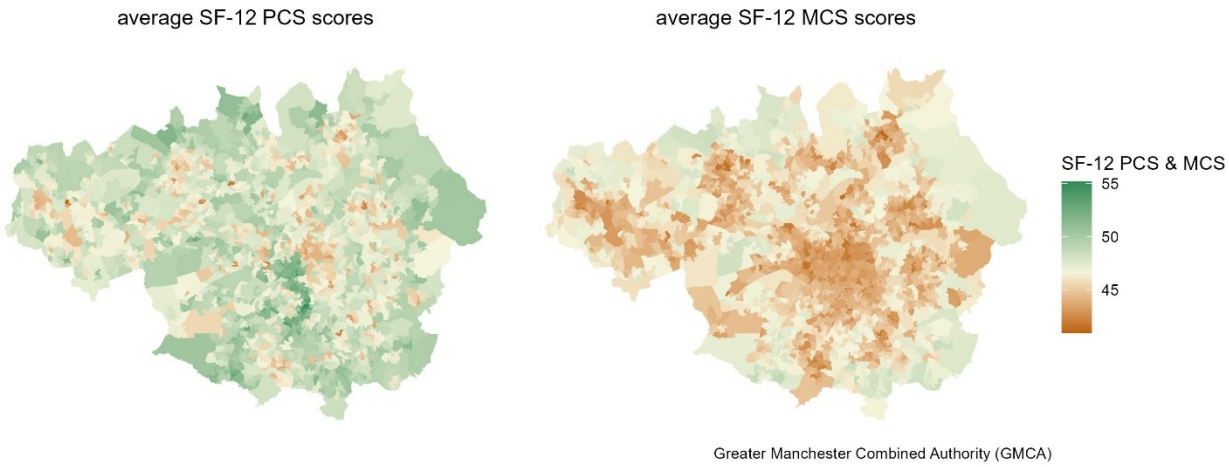

Figure 2: Raw 12-Item Short Form Survey (SF-12) scores for physical (PCS) and mental (MCS) health summary scores across the Lower layer Super Output Areas (LSOAs) of the Greater Manchester Combined Authority (GMCA).

5. Limitations and Levels of Confidence

Despite a high level of quality control throughout the creation and validation process, the Synthetic Population is subject to a number of limitations. We advise potential users of the Synthetic Population to fully acknowledge these conceptual limitations and read this section carefully.

Please note that the creation of the Synthetic Population does not consider any explicit individual-level information (e.g. area codes or names) on the *"true"* place of residence, such as those captured in special license area-level linkages. This means that the Synthetic Population does not reflect actual individuals and their *"true"* place of residence. The dataset should therefore not be seen as a like-for-like replacement for well-established data sources such as population-based registers or area-linked surveys. In addition to this, the Synthetic Population does not capture *"true"* household structures. While the Synthetic Population can be linked with information

capturing household-level characteristics, the dataset does currently not provide a reliable nesting of synthetic individuals within synthetic households.

The Synthetic Population builds upon the many unique strengths of the underlying Understanding Society survey. At the same time, the Synthetic Population is subject to very similar limitations as the survey, such as a certain degree of selection and recall bias. This means, for example, that the Synthetic Population is not well suited to study particular outcomes of marginalisation (e.g., illicit drug use or homelessness), details of routinely occurring service use (e.g., hospital treatments, medication use), or outcomes strongly associated with institutionalised populations (e.g. prisons, nursing homes).

Theoretically, all variables included in the Understanding Society survey data files k_indresp and k_hhresp can be linked with the Synthetic Population. This allows for a wide range of questions to be explored. At the same time, not all variables will allow for a meaningful application within a synthetic population setting. To guide new users of the Synthetic Population, we have suggested a classification system which aims to capture the level of confidence for specific variables.

As shown in **Table 3**, we suggest a 4-tier system to assess the suitability of variables intended for linkage with the Synthetic Population. The suggested levels of confidence range from (1) very high (as a variable was used as constraints) to (4) Unknown, likely problematic (as a variable captures very specific personality traits or area-level features which are difficult to validate).

Table 3: A 4-tier classification system suggested by the authors to evaluate the suitability of the linked Synthetic Population.

Level of Confidence	Explanation	Examples
(1) Very high	Survey variables which were aligned to constraint tables and have hence informed the spatial microsimulation. To ensure the highest level of confidence, we recommend using the aligned categories for these variables.[13]	The 8 constraint variables: age/sex, highest qualification, ethnicity, marital status, economic activity, general health, household tenure household type (composition).
(2) High, but caution	Survey variables which were not used as constraints, but which are either strongly associated with the utilised 8 constraints, and/or for which an external validation can be performed using reliable data sources.	Occupational group, financial hardship, health indicators and health risk factors, benefit payments, housing quality, neighbourhood safety, resources available to individuals and household.
(3) Unknown, potentially possible	Survey variables which were not used as constraints and for which the association with the utilised 8 constraints is not fully understood so that a careful assessment, and/or external validation (by proxy if required) is advised prior to usage.	The majority of variables captured in the Understanding Society survey datasets, including information on opinions and beliefs.
(4) Unknown, likely problematic	Survey variables which were not used as constraints and for which the association with the utilised 8 constraints is not fully understood, but which capture such specific traits of survey respondents and their place of residence that an external validation is not possible or not meaningful.	Variables capturing access or proximity to specific geographical features or engagement in specific leisure or recreational activities such as swimming, diving, mountaineering etc., or particular daily habits, commuting time etc.

[13] Code to reproduce this alignment process is part of a supplementary resource, available via ReShare: http://doi.org/10.5255/UKDA-SN-856754

As general rules, we recommend the following;

(1) Applications drawing on survey variables which were utilised as constraints, or survey variables which are strongly associated with the utilised constraints, can generally be considered the most meaningful.

(2) Applications which draw on survey variables for which the association with the utilised constraints is unknown and cannot be validated, or survey variables reflecting very distinct experiences of individuals (e.g., commuting time), are likely to be less meaningful.

(3) We always recommend an external validation of variables prior to usage. Examples of external validation are presented in the Appendix of this user guide. All presented examples are provided in a reproducible format as part of a separate resource[14].

We advise all users of the Synthetic Population to carefully assess which category is the most likely to apply – and ideally to always externally validate the assumptions made. If no *"direct"* external validation is possible, at least an *"indirect"* external validation should be considered – for example by utilising proxy indicators, previous findings, or expert opinion.

[14] Available as a separate resource on ReShare:

http://doi.org/10.5255/UKDA-SN-856754

FAQ

What is the SIPHER Synthetic Population for Individuals in Great Britain 2019-2021 (referred to as Synthetic Population throughout this user guide)?

The Synthetic Population is the output of spatial microsimulation, hence all results obtained from it should be understood as *"model output"*. The Synthetic Population allows for the creation of a *"digital twin"* of the adult population (aged 16 years and older) in England, Scotland, and Wales. The dataset itself includes only two variables ("*pidp*" and "*synthetic_zone*"). The Synthetic Population is intended for linkage with Understanding Society survey data, analogously to area-level linkages. Once merged with Understanding Society survey data (in particular: *"k_indresp"* and *"k_hhresp"*), the Synthetic Population provides a unique multi-topic dataset, reflecting the lives of over 50 million synthetic individuals in Great Britain at a granular spatial resolution.

Why did we create the Synthetic Population?

Comprehensive population-based register data reflecting the lives of individuals across multiple domains are not readily available for all of Great Britain. While surveys are often the next-best alternative, they typically do not allow for a granular geographical resolution. Building upon the unique multi-topic features of the Understanding Society survey, the Synthetic Population allows us to navigate some of these limitations. This makes the Synthetic Population an easily accessible alternative for a range of different applications involving co-production and collaborative work such as research, planning, or teaching with a particular focus on individuals and areas. Here, the dataset is well suited to be used for understanding *"status quo"* or modelling *"what if"* scenarios (e.g. through static and dynamic microsimulation).

Does the Synthetic Population reflect actual individuals?

No, the Synthetic Population does not reflect actual individuals and their *"true"* place of residence. However, the synthetic individuals represented by the Synthetic Population are based on *"true"* Understanding Society survey

participants. During the creation process, these survey respondents were repeatedly and randomly assigned to small geographical areas in order to represent the adult population in Great Britain as closely as possible, given a range of area-level information available.

Why is the person identification number *"pidp",* which is unique in the Understanding Society survey, not unique in the Synthetic Population?

This is because Understanding Society survey participants were randomly and repeatedly assigned to small areas, with the purpose of representing the entire adult population in Great Britain. No information on the actual place of residence (e.g. area names or codes) was utilised when creating the Synthetic Population. Due to this, and the sheer number of duplications, it is not possible to ever determine a survey respondent's *"true"* place of residence from the Synthetic Population, given the information on the *"synthetic"* place of residence. In other words: It is impossible to make any inferences whether one of the assigned synthetic areas is the *"true"* place of residence. On average, every survey respondent was duplicated and assigned to approximately 2,000 different areas (with a minimum of 200+ duplications and assignments to different areas).

How do I link the Synthetic Population with Understanding Society survey data?

The Synthetic Population is intended for a linkage with the individual- and household-level response files of wave 11 ("k"): k_indresp and k_hhresp. This linkage can be performed similarly to other, well-established, area-level linkages. No further pre-requisites are required prior to linkage. However, the size of the resulting linkage should be considered. The linkage can be established using a variety of different software packages such as SPSS, Stata, R, or Python.

Can the Synthetic Population be linked to other waves of Understanding Society?

Generally, we always recommend the usage of Understanding Society main survey data for any longitudinal studies that require the linkage of

multiple waves. However, the Synthetic Population can, in theory, be linked to other waves of Understanding Society. While theoretically possible, such linkages will likely result in situations where some *"pidp"* numbers included in the Synthetic Population cannot be linked with Understanding Society survey data. If a longitudinal perspective is required for the Synthetic Population, we recommend the usage of variables which capture retrospective questions (if possible).

Which types of analyses does the Synthetic Population enable?

While the Synthetic Population draws directly on data provided by *"real"* Understanding Society survey respondents, it does only ever reflect synthetic (*"not real"*) individuals. In addition, it is important to keep in mind that the Synthetic Population is the outcome of a statistical creation process (i.e. spatial microsimulation). Therefore, all results obtained from this dataset should always be treated and understood as model output - even basic descriptive statistics. Hence, the Synthetic Population should not be seen as a replacement of any *"real"* data in standard statistical analyses (e.g., regression analysis) which are typically performed with the Understanding Society survey. However, the dataset provides a great source of data for understanding *"status quo"* and modelling *"what if"* scenarios (e.g., through static/dynamic microsimulations), as well as for exploratory analyses (e.g., when no other data available). Here, the Synthetic Population can be used for both, area-level and individual-level studies. At the same time, results obtained from the Synthetic Population can be utilised in a variety of external applications, such as informing parameters of external models, including Agent-Based Models.

Are all variables included in Understanding Society survey data files suitable for usage with the Synthetic Population?

In theory, all variables captured in the Understanding Society survey data files k_indresp and k_hhresp can be linked with the Synthetic Population. At the same time, not all variables will allow for a meaningful application. Applications which draw upon variables which we have utilised as part of the 8 constraints, as well as variables strongly associated with these constraints, can generally be considered more meaningful. In contrast, applications which require variables for which the association with the

utilised constraints is (a) unknown and cannot be validated, or (b) is likely to reflect very distinct experiences of individuals (e.g., commuting time) are likely to be less meaningful. As a general rule, we always recommend an external validation of variables prior to usage. To guide potential users of the Synthetic Population, we have provided a classification system which aims to capture the level of confidence for specific types of variables → **Table 3.**

What is an external validation, and why is it important?

External validation describes the process of comparing patterns obtained from the Synthetic Population for individuals and areas with an external data source which can be considered a reliable reference. This process allows us to make a judgement whether the usage of the Synthetic Population is appropriate for the given context and related variables of interest. We always recommend an external validation of variables prior to usage. Examples of external validations are provided in the **Appendix** of this user guide.

Appendix

Validation: Rationale

All newly created synthetic data sources require a thorough validation process. This is to ensure that the resulting synthetic dataset;

(1) has been created correctly based on the provided input data sources (internal validation)

(2) can be considered an appropriate representation of the population it is intended to reflect (external validation)

This version of the Synthetic Population is based on the Understanding Society wave 11 (*"k"*) and a total of 8 constraints. A previous version was based on Understanding Society wave 9 (*"i"*) and 7 constraints (Wu et al. 2022). In addition to updates of survey data and constraint tables, this most recent Synthetic Population also captures general health as an additional 8th constraint dimensions.

In line with the validation of the previous synthetic population dataset described by Wu et al. 2022, we have fully quality controlled and validated this Synthetic Population. All examples presented in the Appendix are fully reproducible and are available as a separate resource[15]. Please note that this section provides a selection of examples. These examples might, or might not be transferable to different contexts or applications. In addition, a wealth of new or additional requirements for validation can arise as the dataset is used in different applications.

Internal Validation

In a first validation step, we compared the number of unique areas in the Synthetic Population with the number of unique areas specified in ONS UK census 2011 output area classification. This allowed us to identify whether

[15] Available as a separate resource on ReShare:

http://doi.org/10.5255/UKDA-SN-856754

there were any areas in the 2011 UK census that were not included in this version of the Synthetic Population. The ONS UK census 2011 output area classification contained a total number of 41,729 unique LSOAs and DZs. Of these, 41,726 areas were matched. This means that three areas did not appear in the Synthetic Population.

Table A-1 lists all areas which did not appear in the Synthetic Population. In all cases, we were able to identify the reason for this mismatch. The mismatch resulted as these areas had a population size of zero based on the queried 2020 population estimates - but were non-zero-population areas in the 2011 UK census. Due to the small scale of the problem, no further steps were taken to address this mismatch.

Table A-1: Overview of unmatched areas, all of which were Data Zones in Scotland. These areas are included in the 2011 ONS census lookup, but are not included in the Synthetic Population due to a population size of zero in 2020 population estimates.

Data Zone	Intermediate Zone	Local Authority	
S01010226	S02001925	S12000049	Glasgow City
S01010227	S02001925	S12000049	Glasgow City
S01010206	S02001921	S12000049	Glasgow City

We then compared the number of created synthetic individuals within each LSOA and DZ, obtained from the Synthetic Population, with the expected number of individuals. The expected number of individuals was obtained separately for each of the 8 utilised constraint dimension and represents the total number of individuals within each LSOA/DZ according to the constraint table. Differences in the number of created synthetic individuals vs. expected individuals can occur for two reasons: (1) a misalignment of Understanding Society variables to reflect information captured in the utilised constraint tables, or (2) poor performance of the spatial microsimulation algorithm when populating areas in the Synthetic Population.

Figure A-1 provides a high-level summary of differences in the total number of individuals within all LSOA and DZs. Ideally, differences should be as small as possible; close to and normally distributed around a median/mean of zero. As shown in **Figure A-1**, the Synthetic Population

achieved an excellent fit in terms of number of individuals for each of the utilised constraint tables. This indicates an excellent alignment of Understanding Society variables with the information provided in the constraint tables, as well as a well-performing spatial microsimulation algorithm.

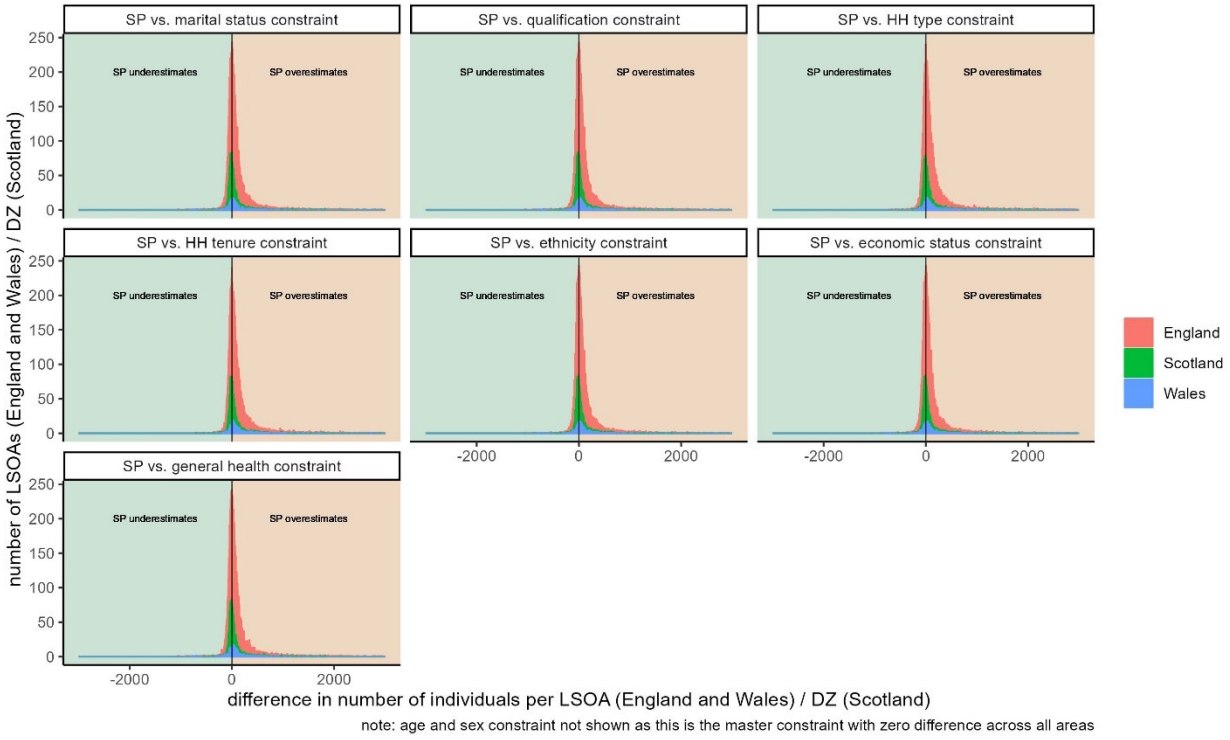

Figure A-1: Overview of differences in the total number of individuals across areas: Synthetic Population ("observed") vs utilised constraint tables ("expected").

Example of External Validation: Ns-Sec Groups

The primary aim of synthetic data sources is to provide appropriate representations of real-world (in some cases even non-existing) data. To explore whether the Synthetic Population can provide appropriate representations of real-world patterns, we examined the performance of the Synthetic Population when reflecting a range of non-constraint information. Such external validations can be performed for a wide range of variables. Harnessing the multi-topic features of the Synthetic Population, we have validated our dataset for a range of different dimensions.

The following section provides two tangible examples of performed external validations: occupational groups and employment deprivation across DZs in Scotland. Both examples are fully reproducible as part of the supplementary resource mentioned earlier.

This Synthetic Population is based on 8 constraints reflecting the sociodemographic components and health status of individuals. While occupational group is not part of the utilised constraints, we can assume that the concept should be well represented in the Synthetic Population – due to its strong association with the 8 included constraints. We explored this by comparing area-level patterns in National Statistics Socio-economic classification (NS-SeC) categories across data zones in Scotland. For this purpose, area-level NS-Sec patterns were obtained from a linked Synthetic Population, and compared with area-level NS-Sec patterns reported in the 2011 UK census for data zones in Scotland[16].

As shown in **Figure A-2**, the Synthetic Population is able to provide a reliable representation of a 5-category based NS-SeC classification. Despite this good representation of area-level patterns for major NS-SeC classification groups, a good amount of care is required for residual categories of variables (e.g., *"all other"*) which were not utilised as constraints.

---

https://nrscensusprodumb.blob.core.windows.net/downloads/SNS%20Data%20Zone%202011%20blk.zip, file: "LC6201SC.csv"

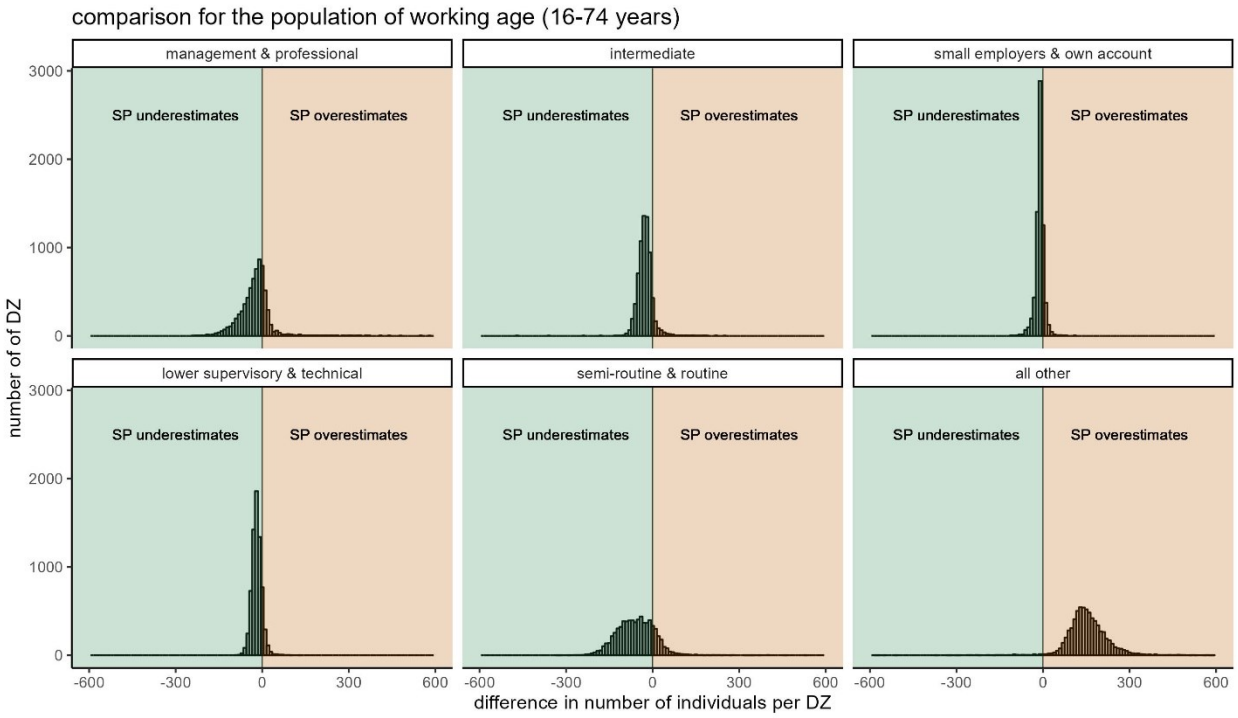

Figure A-2: Comparison of area-level patterns in National Statistics Socio-economic classification (NS-SeC) across data zones (DZs) in Scotland: Synthetic Population vs. UK census 2011 data for Scotland. Note: No NS-SeC information has been used when creating the Synthetic Population, but its distribution within Scottish DZs is known as a point reference in this comparison from census data.

Example of External Validation: Employment Deprivation

In another external validation, we explored whether the Synthetic Population is suitable for capturing the concept of employment deprivation as specified in the Scottish Index of Multiple Deprivation (SIMD) 2020. We chose this comparison as, despite a substantial degree of intersectionality, the concept of employment deprivation can be operationalised precisely using the Understanding Society survey.

In detail, the SIMD 2020 defines employment deprivation based on a combination of the following characteristics:[17]

- Working age recipients of Jobseeker's Allowance
- Working age recipients of Incapacity Benefit, Employment and Support Allowance, or Severe Disablement Allowance
- Working age recipients of Universal Credit not in employment

As shown in **Figure A-3**, the Synthetic Population is well-suited to capture the concept of employment deprivation. A particularly good fit can be achieved among areas with low to medium-high levels of employment deprivation. However, the Synthetic Population might slightly underestimate the number of individuals experiencing employment deprivation across outlier areas, characterised by an exceptionally high prevalence of employment deprivation. In **Figure A-3** this is shown by an increasing deviation between the fitted red line and the 45-degree black line.

These differences, particularly among areas with exceptionally high levels of employment deprivation could arise from selection effects with respect to participating in the Understanding Society main survey as well as differences in the years used for this comparison. For example, the SIMD 2020 employment deprivation data are based on 2017 data from the Department for Work and Pensions (DWP), and have been subject to some degree of statistical modelling. In contrast to this, survey and constraint table data used for the creation of the Synthetic Population span different points in time between 2011 and 2021.

17 https://www.gov.scot/binaries/content/documents/govscot/publications/statistics/2020/01/scottish-index-of-multiple-deprivation-2020-indicator-data/documents/simd_2020_indicators/simd_2020_indicators/govscot%3Adocument/SIMD%2B2020v2%2B-%2Bindicators.xlsx

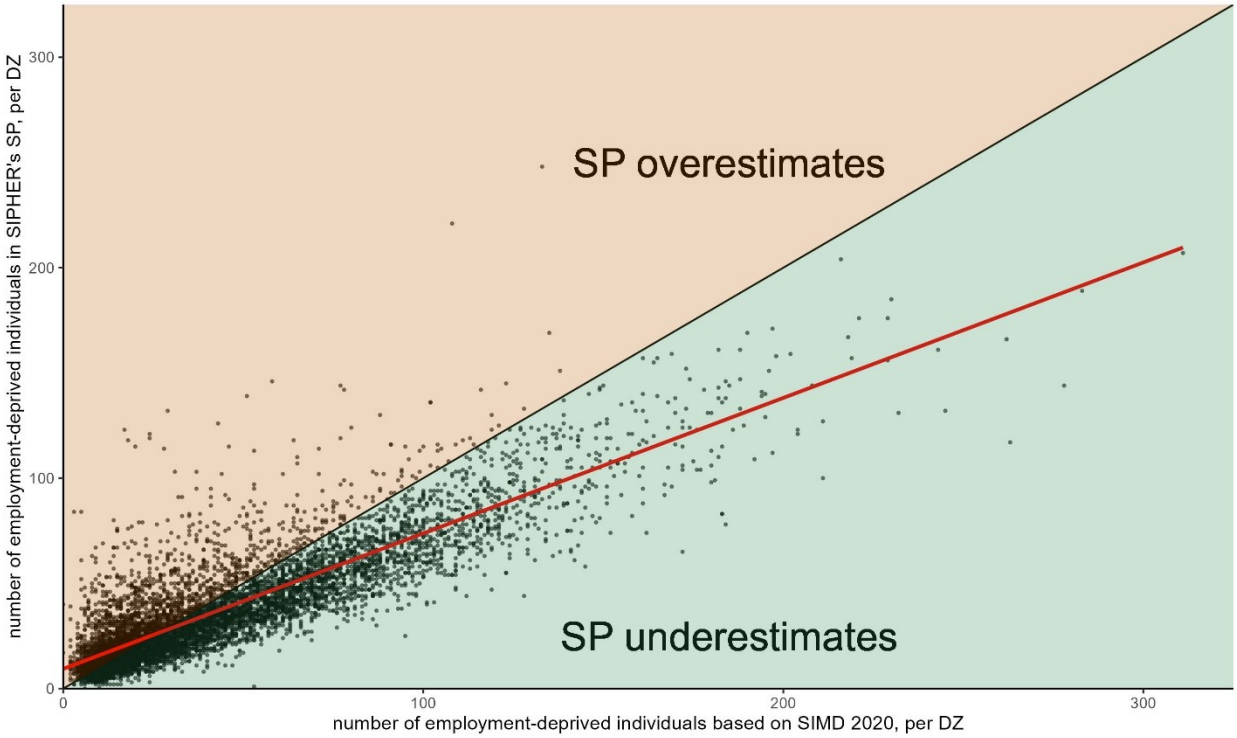

Figure A-3: Comparison of the absolute number of individuals experiencing employment deprivation within data zones (DZs) in Scotland: Synthetic Population vs. Scottish Index of Multiple Deprivation (SIMD) 2020 employment deprivation domain (which is based on data from the Department for Work and Pensions (DWP) from 2017).

References

Buck, N., & McFall, S. (2012). Understanding Society: design overview. Longitudinal and Life Course Studies, 3(1), 5-17.

Harland, K. (2013). Microsimulation model user guide (flexible modelling framework).

Höhn A, Lomax N, Rice H, et al (2024). Estimating quality-adjusted life expectancy (QALE) for local authorities in Great Britain and its association with indicators of the inclusive economy: a cross-sectional study. *BMJ Open* 2024;14:e076704.

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*.

Prédhumeau, M., & Manley, E. (2023). A synthetic population for agent-based modelling in Canada. *Scientific Data*, *10*(1), 148.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available Online: https://www.R-project.org/. Accessed on 2023-12-20.

Scottish index of multiple deprivation (SIMD) 2020: Ranks. Available Online: https://www.gov.scot/publications/scottish-index-of-multiple-deprivation-2020v2-ranks/. Accessed on 2023-12-20.

Scottish index of multiple deprivation (SIMD) 2020: Technical notes https://www.gov.scot/publications/simd-2020-technical-notes/. Accessed on 2023-12-20.

Tozluoğlu, Ç., Dhamal, S., Yeh, S., Sprei, F., Liao, Y., Marathe, M., et al. (2023). A synthetic population of sweden: Datasets of agents, households, and activity-travel patterns. *Data in Brief*, *48*, 109209.

United Nations Economics Commission (2007). Register-based statistics in the Nordic countries Review of best practices with focus on population and social statistics. Available Online: https://unstats.un.org/unsd/dnss/docViewer.aspx?docID=2764. Accessed on 2023-12-20.

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

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).