www.nature.com/scientificdata

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A global dataset of 7 billion individuals with socio-economic characteristics

Marijn J. Ton[1]✉, Michiel W. Ingels[1], Jens A. de Bruijn[1,2], Hans de Moel[1], Lena Reimann[1], Wouter J. W. Botzen[1] & Jeroen C. J. H. Aerts[1,3]

In global impact modeling, there is a need to address the heterogeneous characteristics of households and individuals that drive different behavioral responses to, for example, environmental risk, socio-economic policy changes and spread of diseases. In this research, we present GLOPOP-S, the first global synthetic population dataset with 1,999,227,130 households and 7,335,881,094 individuals for the year 2015, consistent with population statistics at an administrative unit 1 level. GLOPOS-S contains the following attributes: age, education, gender, income/wealth, settlement type (urban/rural), household size, household type, and for selected countries in the Global South, ownership of agricultural land and dwelling characteristics. To generate GLOPOP-S, we use microdata from the Luxembourg Income Study (LIS) and Demographic and Health Surveys (DHS) and apply synthetic reconstruction techniques to fit national survey data to regional statistics, thereby accounting for spatial differences within and across countries. Additionally, we have developed methods to generate data for countries without available microdata. The dataset can be downloaded per region or country. GLOPOP-S is open source and can be extended with other attributes.

Background & Summary

In recent decades, several continental- to global-scale socio-economic impact assessment models have been developed to investigate the societal effects of, for example, diseases[1] (e.g., Balcan *et al.*[1]), transport systems[2], food security[3], energy consumption[4], water quality[5], and weather extremes[6]. By simulating the current and future projections of societal impacts, such large-scale models can be used to assess how societies may respond to (increasing) socio-economic and environmental risk. For example, integrated assessment models can provide insights into how sustainable development policies may reduce carbon emissions by providing a quantitative description of key processes in human and earth systems and their interactions[7].

In the majority of global models, impacts are simulated on the national level in global economic zones, focusing on policy responses by governments and the private sector. In these models, households and individuals are assumed to exhibit uniform behavioral responses to impacts and change[8]. For example, integrated assessment models of climate change typically assume a representative consumer of a single average global or regional consumer[9]. However, there is a need to address the heterogeneous characteristics of individuals and households on both the continental and global scales[8,10,11]. For example, the use of micro-simulation models addressing individual behavior during the COVID-19 pandemic has proven crucial in simulating the spread of a virus using sociopsychological factors that drive the behavioral responses of individuals across time and space[12]. Additionally, continental-scale agent-based models address interactions between households and the government concerning issues such as energy consumption[13] and climate risk adaptation[6]. While these models provide valuable insights, they assume generic household types, neglecting the heterogeneity of individuals and households under risk, and not considering how individual socio-economic and behavioral drivers influence adaptive responses.

To improve large-scale agent-based models, there is a need for a consistent synthetic global dataset on the attributes of households and individuals (age, income, education, etc.) that drive behavioral responses under environmental risk. In this context, synthetic means that the population in the dataset has socio-demographic

[1]Institute for Environmental Studies (IVM); Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [2]International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria. [3]Deltares, Delft, The Netherlands. ✉e-mail: marijn.ton@vu.nl

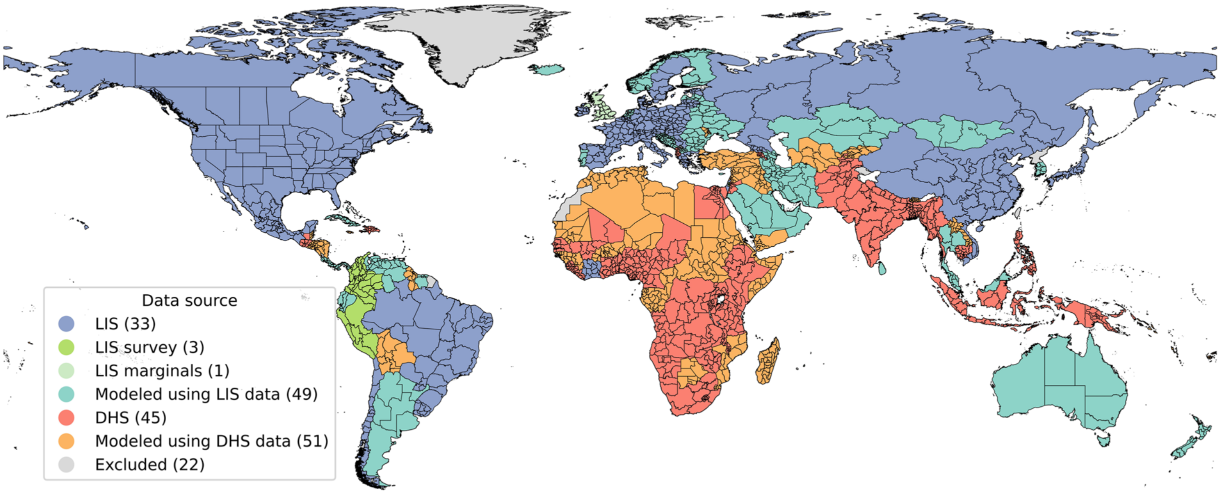


Fig. 1 Source of data for each country.

attributes that have a similar statistical distribution of characteristics to the real population[14]. To create a synthetic population that is representative of the actual population, a few key conditions should be considered:

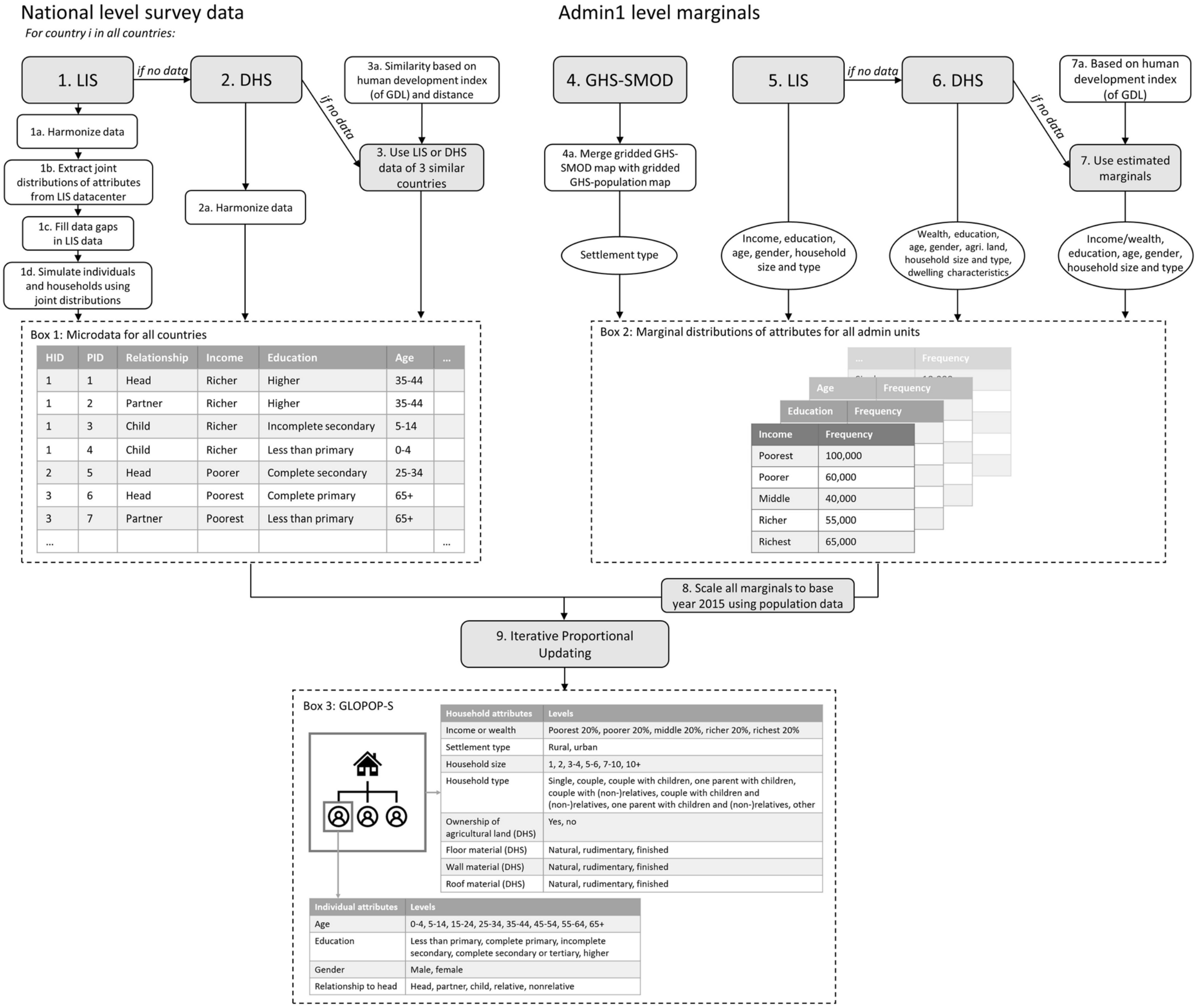
- Reflecting the heterogeneity of the distribution of households and individuals across geographic areas[15]
- Reproducing the interdependencies among agents in the same household[16]
- Ensuring data confidentiality[17]
- Avoiding pure replication of the underlying sample[17]

Recent studies have shown that it is possible to create large-scale synthetic population datasets with individual characteristics; examples include the works of Prédhumeau & Manly[14] and Wu *et al.*[18]. In their study, Prédhumeau & Manly[14] generated a synthetic population dataset for Canada by integrating various sources of microdata and sub-country census data using proven statistical methods, such as the Iterative Proportional Fitting algorithm[19,20].

In this research, we present GLOPOP-S[21], the first global synthetic population dataset of 1,999,227,130 households and 7,335,881,094 corresponding individuals for the year 2015 that is consistent with population statistics (e.g., distributions of age, gender, income, etc.) at an administrative unit 1 level, thus addressing spatial heterogeneity within and across countries[15]. We focus on the following attributes: age, education, gender, income or wealth, settlement type (urban/rural), household size and household type (e.g., single, couple, couple with children). We additionally include the feature ownership of agricultural land and dwelling characteristics, but these attributes are not available for all countries. The choice of these variables is motivated by the need to understand behavior related to environmental risk. Variables such as education and age are often found to be related to environmental (risk) awareness[22], whereas income or wealth of a household influences the capacity of households to act against environmental risk[23]. Gender has been shown to influence environmental risk perceptions[24], while impacts of environmental risk often differ between genders. Settlement types (urban versus rural) influence the type of adaptation measures that are applicable in the context of addressing environmental risk[25,26]. Household size and type may also influence preferences for different protection options and the ability to act against environmental risks[27].

To create the dataset, we apply state-of-the-art statistics (e.g., Iterative Proportional Updating) and parallel computing power to integrate and upscale two key databases on microdata. The Demographic and Health Surveys[28] (DHS) provide specific household indicators for 90 different countries, including demographics, health, psychological and economic indicators. The Luxembourg Income Study[29] (LIS) focuses on harmonizing income microdata for 50 countries over the last 50 years, including countries in Europe, North America, Latin America, Africa, Asia, and Australia. We have considered using other databases, such as EU microdata, UNICEF's Multiple Indicator Cluster Surveys (MICS) and Integrated Public Use Microdata Series (IPUMS) data, but we preferred LIS and DHS. Despite restricted access to LIS data, we favored it over EU microdata, because the latter database does not allow linking individuals to their respective households. MICS and IPUMS have similar coverage to DHS, but DHS covers more countries, has more recent data compared to IPUMS, and includes dwelling characteristics.

Figure 1 shows which database is available for each country. The two databases complement each other well, with DHS data mostly available for low-income countries and LIS data predominantly available for high- and middle-income countries. After processing the data, the number of countries with available LIS or DHS survey data and regional statistics (i.e., marginal distributions) is 78, covering 81.4% of the global population. There are a few countries for which we either have survey data or marginals. For example, when the microdata lacks the variable indicating the region, we can still use these data to create the households and individuals, but we cannot construct the regional marginals. Alternatively, in the case that the rural/urban indicator is missing, we cannot use the survey data, but we can still construct the marginals of the other variables. For 99 countries, covering


National level survey data
For country i in all countries:
1. LIS
if no data
2. DHS
if no data
3a. Similarity based on human development index (of GDL) and distance
1a. Harmonize data
1b. Extract joint distributions of attributes from LIS datacenter
1c. Fill data gaps in LIS data
1d. Simulate individuals and households using joint distributions
2a. Harmonize data
3. Use LIS or DHS data of 3 similar countries
Admin1 level marginals
4. GHS-SMOD
5. LIS
if no data
6. DHS
if no data
7a. Based on human development index (of GDL)
4a. Merge gridded GHS-SMOD map with gridded GHS-population map
7. Use estimated marginals
Settlement type
Income, education, age, gender, household size and type
Wealth, education, age, gender, agri. land, household size and type, dwelling characteristics
Income/wealth, education, age, gender, household size and type
Box 1: Microdata for all countries
Box 2: Marginal distributions of attributes for all admin units
...
Frequency
Age
Frequency
Education
Frequency
8. Scale all marginals to base year 2015 using population data
9. Iterative Proportional Updating
Box 3: GLOPOP-S

HID	PID	Relationship	Income	Education	Age	...
1	1	Head	Richer	Higher	35-44	
1	2	Partner	Richer	Higher	35-44	
1	3	Child	Richer	Incomplete secondary	5-14	
1	4	Child	Richer	Less than primary	0-4	
2	5	Head	Poorer	Complete secondary	25-34	
3	6	Head	Poorest	Complete primary	65+	
3	7	Partner	Poorest	Less than primary	65+	
...						...

Income	Frequency
Poorest	100,000
Poorer	60,000
Middle	40,000
Richer	55,000
Richest	65,000

Household attributes	Levels
Income or wealth	Poorest 20%, poorer 20%, middle 20%, richer 20%, richest 20%
Settlement type	Rural, urban
Household size	1, 2, 3-4, 5-6, 7-10, 10+
Household type	Single, couple, couple with children, one parent with children, couple with (non-)relatives, couple with children and (non-)relatives, one parent with children and (non-)relatives, other
Ownership of agricultural land (DHS)	Yes, no
Floor material (DHS)	Natural, rudimentary, finished
Wall material (DHS)	Natural, rudimentary, finished
Roof material (DHS)	Natural, rudimentary, finished

Individual attributes	Levels
Age	0-4, 5-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65+
Education	Less than primary, complete primary, incomplete secondary, complete secondary or tertiary, higher
Gender	Male, female
Relationship to head	Head, partner, child, relative, nonrelative

Fig. 2 Methodological approach in nine steps of creating a global synthetic database of households and individuals with attributes (GLOPOP-S).

15.7% of the world population, where no survey data were available, we generated a synthetic population based on either DHS or LIS data from other countries. A small number of countries (~1% world population) does not have a synthetic population due to a lack of information about the settlement type (urban/rural) or because they are (disputed) countries/regions, such as North Korea, Taiwan and Western Sahara.

To construct the dataset, we follow the procedure depicted in Fig. 2 (for details see Methods). Starting at the top left of Fig. 2, we obtain national-level household data from survey data (i.e., LIS and DHS) for each country (steps 1–3). Then, we categorize and harmonize the DHS and LIS data; see Table S1 in the Supplementary Information, where we list the original categories in the survey data and the corresponding categories in the synthetic dataset. Next, we apply data processing steps (e.g., imputing missing data and merging datasets) to make the data comparable and consistent (See Box 1 in Fig. 2 for an example of the processed survey data). In steps 4 to 8, we construct the regional marginal distributions of household and individual attributes. These marginal distributions indicate the number of households and individuals within each category of the attributes. For example, the marginal distribution of the age attribute indicates the number of individuals aged 0–4, 5–14, 15–24, 25–34, and so forth, for each region (see Box 2 in Fig. 2 for an example of the processed marginal distributions). After preparing the survey data (Box 1) and the marginal distributions (Box 2), the Iterative Proportional Updating algorithm is applied to calculate weights for the households in the survey data such that the weighted households match the marginal distributions on the regional level (step 9). By doing so, we account for differences in socio-economic characteristics across regions within a country.

The global dataset presented in this paper will be a valuable starting point for conducting research on societal impact assessment and risk adaptation, helping to better understand the adaptive capacity or social vulnerability of households to risk[30]. GLOPOP-S is available for download at the regional to the global scale and provides a unique input in micro-simulation models addressing individual decision-making under climate and other types of risk. Additionally, we provide a data file with the marginal distributions of agent attributes at a regional level for all countries.

Household ID	Relationship to head	Income	Wealth	Settlement type	Age	Gender	Education	Household type	Household size	Ownership of agricultural land	Floor material	Wall material	Roof material	Data source
1	1	1	−1	0	3	0	1	1	1	−1	−1	−1	−1	1
2	1	5	−1	0	4	1	4	2	1	−1	−1	−1	−1	1
2	2	5	−1	0	3	0	4	2	2	−1	−1	−1	−1	1
3	1	2	−1	1	5	0	2	3	3	−1	−1	−1	−1	1
3	3	2	−1	1	2	1	1	3	3	−1	−1	−1	−1	1
3	3	2	−1	1	2	0	2	3	3	−1	−1	−1	−1	1

Table 1. Example of the data in GLOPOP-S.

Attributes	HH/I	Levels
Income	Individual	1: poorest 20%, 2: poorer 20%, 3: middle 20%, 4: richer 20%, 5: richest 20%, −1: unavailable for country
Wealth	Individual	1: poorest 20%, 2: poorer 20%, 3: middle 20%, 4: richer 20%, 5: richest 20%, −1: unavailable for country
Settlement type	Household	0: urban, 1: rural
Age	Individual	1: 0–4, 2: 5–14, 3: 15–24, 4: 25–34, 5: 35–44, 6: 45–54, 7: 55–64, 8: 65+
Gender	Individual	1: male, 0: female
Education	Individual	1: less than primary, 2: complete primary, 3: incomplete secondary, 4: complete secondary or tertiary, 5: higher
Household type	Household	1: single, 2: couple, 3: couple with children, 4: one parent with children, 5: couple with (non-) relatives, 6: couple with children and (non)-relatives, 7: one parent with children and (non-) relatives, 8: other
Household ID	Household	1, …,
Relationship to head	Individual	1: head, 2: partner, 3: child, 4: relative, 5: non-relative
Household size	Household	1: 1, 2: 2, 3: 3–4, 4: 5–6, 5: 7–10, 6: 10+
Ownership of agricultural land (DHS)	Household	1: yes, 2: no, −1: unavailable for country
Floor material (DHS)	Household	1: natural, 2: rudimentary, 3: finished, −1: unavailable for country
Wall material (DHS)	Household	1: natural, 2: rudimentary, 3: finished, −1: unavailable for country
Roof material (DHS)	Household	1: natural, 2: rudimentary, 3: finished, −1, unavailable for country
Source	Country	1: LIS, 2: LIS survey, 3: LIS marginals, 4: Modeled by LIS data, 5: DHS, 6: Modeled by DHS data

Table 2. Attributes in GLOPOP-S.

In Table S1 in the Supplementary Information, we list the variables used in the LIS and DHS databases and we show how the values and categories in the databases correspond to the categories in the synthetic population database. For a quick overview of the variables and categories in the synthetic population database, see Box 3 in Fig. 2 or Table 2 in the Data Records section.

Methods

In this section, we describe our methods following the steps in the flow diagram shown in Fig. 2.

Survey data from LIS database (step 1).

We begin with a more detailed explanation of the processing procedures for the LIS data. The LIS database contains microdata from around 50 middle- and high-income countries spanning a wide range of years. Since the base year of GLOPOP-S is 2015, we only use LIS surveys that originate from 2015, and if not available, within the time span 2013–2017. The majority of surveys are from the year 2015 (22/33), a high number of the surveys are from 2014 or 2016 (7/33)and a small number of the surveys originate from 2013 or 2017 (4/33). Even though the base year of GLOPOP-S is 2015, we prefer to maintain some flexibility here. We favor using survey data from 2013–2017 rather than replacing it with survey data from 3 similar countries.

As the LIS data are only accessible through a remote execution system, there are some challenges regarding extracting the attributes from this database. In particular, there is no access to individual records in the microdata but only to aggregated data (marginal and joint distributions). Therefore, we developed a procedure that extracts the joint distributions from the LIS server of both households and individual variables of our interest for each country (Step 1a in Fig. 2), which preserves data confidentiality[17]. These joint distributions, also referred to as contingency tables or frequency tables, provide the number of households for each combination of attributes. For example, the joint distribution gives us the number of households that are a couple with a child (household type), belong to the poorer income group, live in an urban area, and where the household head has finished primary school, is aged between 35 and 44 and is male.

When the number of surveyed households is limited, the probability of encountering gaps (zeros) in the joint distribution of household attributes increases. This phenomenon, commonly known as the ‘zero-cell problem’, occurs when households are missing from the collected sample, but do exist in the actual population[31]. These zeros in the joint distribution lead to a lack of heterogeneity and potential biases in the synthetic population because the iterative proportional updating algorithm cannot generate new households that were not part of the sample data.

In the LIS data, the number of surveyed households per country ranges from 2,352 to 232,219 (6,237 to 783,131 individuals) (see Table S2 in the Supplementary Information). Especially in the lower end of this range, we observe significant gaps in the joint distributions of countries. Several methods have been proposed to fill the gaps in the contingency table (i.e., joint distribution) (e.g., Mueller & Axhausen[32]; Choupani & Mamdoohi[33]). The simplest solution is to replace the zeros with a small number[19], for example, 0.1 or 0.01, so that there is a small probability that these households are generated in the synthetic population. However, we do not adopt this solution as it does not consider the correlation structure among variables.

Instead, we take the following steps to reduce the number of zeros in the joint distribution and to address any remaining zeros, while preserving the interdependencies among household and individual characteristics and considering regional differences:

1. First, we decrease the size of the joint distribution tables by reducing the number of variables or by reducing the number of categories per variable[33–35]. For example, we divide the age variable into eight groups mostly spanning 10 years, whereas Prédhumeau & Manly[14] created 18 age groups, each spanning five years, in their synthetic population dataset for Canada.
2. Additionally, we extract the frequency tables at the national level instead of the regional level, benefitting from the larger sample size at the national level, and thus a more realistic population (e.g., more heterogeneity among households). However, a downside of this approach is that it could potentially lead to spatialization errors because it increases the homogeneity of households across regions. In other words, a household residing in region A may be assigned to region B in the synthetic population. Khachman *et al.*[36] describes this dilemma of aggregating or disaggregating the spatial resolution in detail, and we find that a more aggregate approach is favorable given the small sample sizes. Besides, we still account for regional differences because we are fitting the national survey data to regional statistics (marginal constraints).
3. We fill the remaining zeros by factorizing the joint distribution into a conditional and marginal distribution (Step 1b in Fig. 2). Considering three attributes, A1, A2, and A3, we marginalize the joint distribution as follows: J(A1, A2, A3) = C(A1, A2 | A3) * M(A3). Then, we assume that the joint distribution of A1 and A2 is independent from A3, i.e., C(A1, A2 | A3) * M(A3) = J(A1, A2) * M(A3). By marginalizing the joint distribution and assuming independence, we have essentially reduced the number of attributes in the contingency table, as proposed by Auld *et al.*[35] and Choupani & Mamdoohi[33]. However, we want to avoid assuming independence for a specific attribute, which, in the example, is A3. Therefore, we iteratively exclude one of the attributes from the joint distribution and count the number of households given by the reduced joint distribution multiplied by the marginal distribution of the excluded attribute. Then, we take the mean across all these counts and impute the missing value from this mean. Since this procedure increases the total number of households in the joint distribution, we correct all the frequencies in the distribution by multiplying them with the fraction "old number of households"/"new number of households". We perform this procedure to fill all gaps in the joint distribution of household characteristics, except for the zeros that are structural zeros[37]; for example, the probability of a household being a couple and having a household size of one is always zero. By applying this approach, we consider the correlation structure between the attributes rather than simply imputing the missing values with a small, fixed number.

Thus far, we have only discussed joint distributions of the characteristics of the household and the household head. Depending on the household type and size, we need to add a partner and/or children, relatives, or non-relatives to the household (Step 1c in Fig. 2). Starting with the partner, we extract joint distributions of the age, gender and education of the household head and the age, gender and education of the partner. We merge the partner's age conditional on the head's age, the partner's gender based on the head's gender, while the partner's education is assigned based on the head's education. To add the characteristics of children, relatives and non-relatives to the households, we follow a similar method to the previously described approach (see Supplementary Information S3). In S3, we give a detailed description of how individuals are merged using the joint distributions obtained from the LIS database.

Survey data from DHS database (step 2).

The DHS database is a collection of more than 320 household surveys in 90 middle- and low-income countries, which provide information on a wide range of monitoring and impact evaluation indicators in the areas of population, health, and nutrition. The DHS data contain a wealth index that categorizes individuals into five equal groups, ranging from the poorest 20% to the richest 20% of the individuals. The index is not related to a household's income, but based on a household's assets, dwelling materials, access to water and sanitary facilities. In low-income countries, this asset-based index is more commonly used as a measure for a household's economic situation than a household's income because reliable income data are difficult to obtain. For a large portion of the population in those countries, it is difficult to express income in monetary terms for various reasons; for example, because they earn their income in the informal sector[38,39]. Hence, there is no income data in the DHS surveys, so the economic situation of households is proxied by the wealth index. On the other hand, for developed countries, wealth data are hard to obtain and disposable household income is the most common proxy for economic well-being[40]. Therefore, we have two columns in the synthetic population of each country, income and wealth. For synthetic populations derived from LIS survey data, the income column contains data, whereas the wealth column is empty. For synthetic populations derived from DHS survey data, the wealth column has data, whereas the income column is empty. Since the sample sizes are generally larger in the DHS surveys than in the LIS surveys, we do not impute missing combinations of household and individual characteristics. However, we still use survey data at the national level to ensure a larger heterogeneity of households within the region, instead of using survey data at the regional level. The number of countries with available DHS data is 45, where 12 out of 45 surveys come from 2015, and the other 33 surveys are within a 2 year time range.

Countries lacking survey data (step 3). The number of countries with available LIS or DHS survey data that contain our variables of interest is 33 and 45, respectively, and the population in these 78 countries covers 81.4% of the world population. In step 3, we focus on countries lacking survey data, for which we use the processed LIS or DHS data (as outlined in steps 1 and 2) of three demographically similar countries. These three similar countries are selected using the Human Development Index (HDI) of the Global Data Lab (GDL)[41]. The index is calculated based on a few indicators: mean years of schooling of adults aged 25+, expected years of schooling of children aged six, life expectancy at birth, and the gross national income per capita. For each country without any survey data, we first select 10 countries with the most similar HDI index and combine the survey data of the three countries that are geographically closest ('nearest distance') to substitute the missing survey data. To avoid combining LIS and DHS survey data, we replace the missing survey data with data from either the LIS or DHS database. We select data from the three countries within the respective database that have the most similar HDI index and shortest distance. An important reason for not combining LIS and DHS survey data is that LIS contains an income variable, whereas DHS contains a wealth index and we do not want to mix the two variables. To equally account for the characteristics of the three similar countries, we correct for differences in population size; without this correction, there is a risk that the features of a country with a large population will dominate in the synthetic population. To correct for population size, we multiply the household weights of the two smallest countries by a factor such that the sum of the household weights is equal to the sum of the household weights of the country with the largest population size.

Presumably, the survey data from the identified three similar countries do not perfectly represent the population of the country lacking survey data. However, this is a minor concern because the replacement survey data should only show a comparable correlation structure between the household attributes in the country lacking microdata. For example, the probability of earning a certain income given a specific education level should be similar in the country lacking survey data and the countries classified as similar.

Marginal distributions at administrative unit 1 level (steps 4–8). The marginal distributions indicate the number of households or individuals within each category of the attributes. For example, the marginal distribution of the age attribute indicates the number of individuals aged 0–4, 5–14, 15–24, 25–34, and so forth, for each region. The settlement type (rural/urban) marginal distribution is created by overlaying an SMOD-GHS grid with a GHS population grid to count the number of people in the rural and urban grid cells per region for all countries (step 4)[42,43]. The advantage of this approach is that we do not need to estimate the urban/rural marginal for countries lacking survey data, as the GHS data provide the number of individuals residing in rural and urban areas for each region in the world.

For the other marginal distributions, we use LIS (step 5) and DHS (step 6) data and count the number of households and individuals within each category of the attributes. Both LIS and DHS contain a variable that specifies the region of a household, but the subnational division of regions in the two databases is not consistent. As GDL is available globally for all subnational regions, we use GDL as our baseline. Since almost all DHS surveys are supplied with coordinates per cluster of households, we can link the households to GDL regions based on the households' coordinates. For the LIS surveys (and DHS surveys without coordinates), we assign each household to a GDL region manually by name and location to match the subnational regions of GDL.

As some regions lack marginal distributions in both LIS and DHS, we developed the following procedure to estimate the marginal distributions of those regions: for each attribute (i.e., education, household size, age, etc.), we estimate the marginal distributions by taking the average over the marginals of the *k* regions (with LIS or DHS data) that have the most similar Subnational Human Development Index (SHDI) as provided by GDL. The number of regions, *k*, which varies for each of the attributes and for each of the two databases, is determined by minimizing the difference between the mean of the *k* marginals and the true marginals given by the LIS or DHS data. After we have determined the *k* for each attribute, which is in the range of 50–250 regions (see Table S3 in the Supplementary Information), we can predict the marginals of the regions without LIS or DHS data. Again, we do not combine LIS and DHS data. For each country without available data, the marginals are estimated using either LIS or DHS data depending on which database was selected for replacing the missing survey data. Since income and wealth are measured in percentiles, we employ a slightly different procedure to construct the income and wealth marginals. Instead of matching regions with a similar HDI index, we match regions based on their relative HDI index within the country, meaning the region's HDI relative to the country's HDI. First, we divide the regions into *n* equally sized groups based on their relative HDI. This approach results in *n* groups of regions with a similar level of (economic) development within the country. Next, we calculate the average percentages for the five income or wealth groups (i.e., % poorest, % poorer, % middle, % richer, and % richest) across each of the *n* groups. For regions without observed marginal data, we calculate the relative HDI index and we assign the average percentages of the corresponding group.

Finally, all marginals are scaled to match the region's population size in 2015 as provided by GHS[43] (step 8). This means that the shape of the marginal distributions remains the same, regardless of the original year of the data. We aim to use survey data from 2015, but if data from that year are missing, we select data from other years. To minimize the possible error caused by this procedure, we only use data from years closest to 2015, and not from more than 2 years prior or after.

2.5 Iterative Proportional Updating algorithm (step 9). Various methods have been developed that can generate a synthetic population that is statistically representative of the actual population using sample data and aggregated statistics (marginal distributions). The methods can be categorized into three groups: synthetic reconstruction, combinatorial optimization, and statistical learning[31,44]. The most popular and conventional population synthesis technique is Iterative Proportional Fitting (IPF), which belongs to the synthetic reconstruction category. In this approach, a contingency (frequency) table based on survey data is fitted to the marginal constraints from aggregated population data in an iterative manner. IPF produces weights for all households in the survey data, which represent the probabilities of drawing the household for the synthetic population.

The method has been adapted by Ye *et al.*[45] to allow for simultaneously fitting household characteristics and the characteristics of individuals within those households. This novel IPF method is referred to as Iterative Proportional Updating (IPU). In each iteration of the algorithm, the households are first reweighted based on the household's marginal constraints and then reweighted according to the marginal constraints of the individual attributes. In this way, households with similar attributes but consisting of individuals with different characteristics result in different household weights. We apply the IPU algorithm to scale the national survey data to the regional marginals using the R package mlfit (https://rdrr.io/github/krlmlr/mlfit/man/).

However, IPU creates fractional weights for the agents, whereas an ABM requires an integer number of households and individuals. To convert the weights to integers, we apply the 'truncate, replicate, and sample' (TRS) algorithm developed by Lovelace & Ballas[46]. They showed that their method performs better than other integerization methods, such as simple rounding or the threshold approach.

A drawback of IPU is that it replicates the households in the sample data and cannot generate "new" households that are not present in the sample[31]. This could result in a lack of heterogeneity in the synthetic population. Nevertheless, for countries derived by LIS data, we expect that this issue is limited because we imputed missing households with rare combinations of characteristics. For countries covered by DHS data, the sample sizes are generally larger than those for LIS countries, which would ensure sufficient heterogeneity in the synthetic population. In addition, for countries modelled using DHS or LIS data, we combine survey data from three similar countries, thereby increasing the sample size and enhancing heterogeneity in the synthetic population.

Data Records

The dataset of households and individuals with socio-economic characteristics is public and freely available on Harvard Dataverse: https://doi.org/10.7910/DVN/KJC3RH[21]. The data can be downloaded per country, whereby the folders are named according to the country's ISO codes, and administrative unit level 1, corresponding to GDL regions (see *Nr_individuals_data_availability.csv* for a csv of the GDL regions). The number of individuals per region ranges from 1,285 to 219,533,849, hence, it is recommended to download data per region for large countries, such as China and India. To save memory, the data are stored in binary format and should be read following the code in *read_synthpop_data.R* or *read_synthpop_data.py* on GitHub: https://github.com/VU-IVM/GLOPOP-S/. Table 1 provides an example of how the data look like. Note that we have not included a column referring to the GDL region in the data since it is already specified in the filename. Additionally, it is important to be aware that household IDs are unique per region, not per country. Furthermore, we want to emphasize that the income or wealth index is not cross-country comparable, as the poorest 20% in country x does not have the same income or wealth as the poorest 20% in country y.

Table 2 lists the different variables and describes the values each attribute can take.

Technical Validation

Following the discussions by Sun and Erath[47], Zhou *et al.*[44] and Prédhumeau & Manly[14], we evaluate the goodness of fit of the synthetic population by quantifying how well the synthetic population resembles the observed population in the sample data. We do this by comparing the frequencies of all pairs of agent characteristics in the synthetic population with the sample data. Additionally, we evaluate the goodness of fit of our procedures to create synthetic agents for countries lacking survey data and marginals by applying these methods to the countries for which we have survey data and marginals. The goodness of fit is calculated by the formula in Eq. 1:

$$\text{Sum of squared errors} = \sum_{r=1}^{R}\sum_{c=1}^{C}\sum_{h=c+1}^{C}\sum_{i=1}^{I_c}\sum_{j=1}^{J_h}\left(f_{ijr}^{sample} - f_{ijr}^{synth}\right)^2 \quad (1)$$

In Eq. 1, f_{ijr} represents the share of households or individuals with two specific characteristics, i and j, for example, a rural household in the highest income group, in a region, r. f_{ijr}^{sample} is the share of households or individuals with characteristics i and j in the sample data and f_{ijr}^{synth} is the share in the synthetic population. The set C represents the different attributes (age, education, household size, etc.) and the sets I_c and J_h represent the categories (e.g., for age we have the categories 0–4, 5–14, 15–24, etc.) within the attribute c and h. For all possible pairs of characteristics, we take the sum of the difference between the share in the sample data and the synthetic population for all regions, R. When the sum of squared errors is zero, we have a perfect fit.

For countries lacking survey data, it is not possible to calculate the sum of squared errors, because we do not have information on f_{ijr}^{sample}. To validate the performance of the methods for replacing the survey data and estimating the marginals, we apply our procedure from step 3 for countries with data as if the survey data were not available. For these countries, we assess the goodness of fit of the synthetic population under three conditions:

1. *Estimated survey and estimated regional marginals*: both microdata and (regional) marginals are missing. In this case, the microdata are copied from the three most similar countries (based on HDI; see step 3) and regional marginals are based on the 50–250 most similar regions (depending on the attribute; see steps 4–8).
2. *Estimated survey and observed regional marginals*: for a few countries, the survey data lack the urban/rural indicator and, therefore, could not be used. In this case, survey records are copied from the three most similar countries (based on HDI; see step 3). However, by augmenting the missing urban/rural information with SMOD-GHS data (see step 4–8), regional marginals can be determined from the survey and are used.
3. *Observed survey and estimated regional marginals*: if there are microdata on individuals but the survey data do not specify the regions from which the individuals originate, the regional marginals cannot be determined. In this case, the survey data are used, but the marginals are determined using the 50–250 most similar regions (see step 4–8).

In addition to the described variants of the synthetic population related to incomplete data, we determine synthetic population for two hypothetical situations to illustrate the value of considering spatial heterogeneity and correlations between characteristics. These are:

1. *Benchmark 1 No survey data and observed regional marginals*: here we look at the added value of using joint probabilities for the individual/household characteristics of agents instead of working with (uncorrelated) aggregated data for a whole population. To calculate this error, we determine the sum of squared error by replacing f_{ijr}^{synth} with $(f_{ir}^{sample} \times f_{jr}^{sample})$ in Eq. 1, where f_{ir}^{sample} is the frequency of characteristic i in region r independent of characteristic j, and vice versa.
2. *Benchmark 2 Observed survey data and observed national marginals*: to show the added value of using regional marginals, we calculate the frequencies of all pairs of characteristics at the national level. The error is calculated by replacing f_{ijr}^{synth} in Eq. 1 with the frequencies given by the sample data at the national level, i.e., replacing f_{ijr}^{synth} with f_{ij}^{sample}. Hence, subscript r is missing in f_{ij}^{sample}, indicating that the share of households with characteristics i and j is the same over all regions. Therefore, we refer to this error as the national error.

Figures 3, 4 show the sum of squared errors for the synthetic and hypothetical variants mentioned above.

Figures 3, 4 reveal that when we substitute the survey data with survey data from three similar countries, the error only increases slightly (*Estimated survey and observed regional marginals*). More precisely, the median error increases by roughly a factor two for DHS countries and a factor three for LIS countries, see Figures S2, S3 in the Supplementary Information, where we plot the hypothetical errors relative to the synthetic error. When marginals need to be estimated (*Observed survey and estimated regional marginals*), the error increases by a factor ten for DHS countries and a factor 12 for LIS countries. When both the survey data and marginals are replaced (*Estimated survey and estimated regional marginals*), errors increase by a factor ten for DHS countries and a factor 13 for LIS countries. These results show that, particularly, the estimation of (regional) marginals is difficult and predominantly contributes to an increase in errors compared to the substitution of survey data.

The results of Benchmark 1 further illustrate the importance of using survey data to account for correlations between characteristics, instead of merely using regional marginals. The errors increase by a factor 50 for countries with LIS data and a factor three for countries with DHS data. In case of Benchmark 2, when regional marginals are not used and the population is synthetically generated at the national level, the error increases considerably for LIS countries (by a factor 150) and less so for DHS countries (a factor six). It is interesting to note that the errors of the benchmarks are a lot higher for LIS countries than DHS countries. This suggest that correlations between characteristics are stronger in countries in the Global North than in the Global South. In addition, spatial differences seem to be larger in countries in the Global North than in the Global South.

Usage Notes

GLOPOP-S can be used in global, national and regional impact models where individuals or households are identified as agents that can make decisions given a set of attributes and their environmental or geographical context. Agent-based models are typically applied to study decision-making behavior, and the GLOPOP-S dataset can be used as input for such models. We provide household and individual characteristics at administration level 1, and data specific to geographic regions can be downloaded by identifying the GDL geographical IDs in our dataset. GLOPOS-S covers the entire world, but there are some uncertainties when microdata were not available for a country. As Figs. 3, 4 show, the synthetic population does not resemble the true population so well when there is no information on aggregated statistics at the subnational level (i.e., no regional marginals). Hence, the data for these countries should be handled with care. However, these are also generally countries that can be considered data-scarce. Also, we want to emphasize that the income or wealth index are not comparable across countries, i.e., the poorest 20% individuals in country x have different economic resources than the poorest 20% individuals in country y.

Additionally, GLOPOP-S can be extended with other attributes using other survey data. When the additional survey data includes variables present in GLOPOP-S, a regression model can be estimated. In the regression model, the regressors should align with the attributes in GLOPOP-S, such as age of the household head, household size, education level of the partner, and the dependent variable represents the new variable a researcher aims to include.

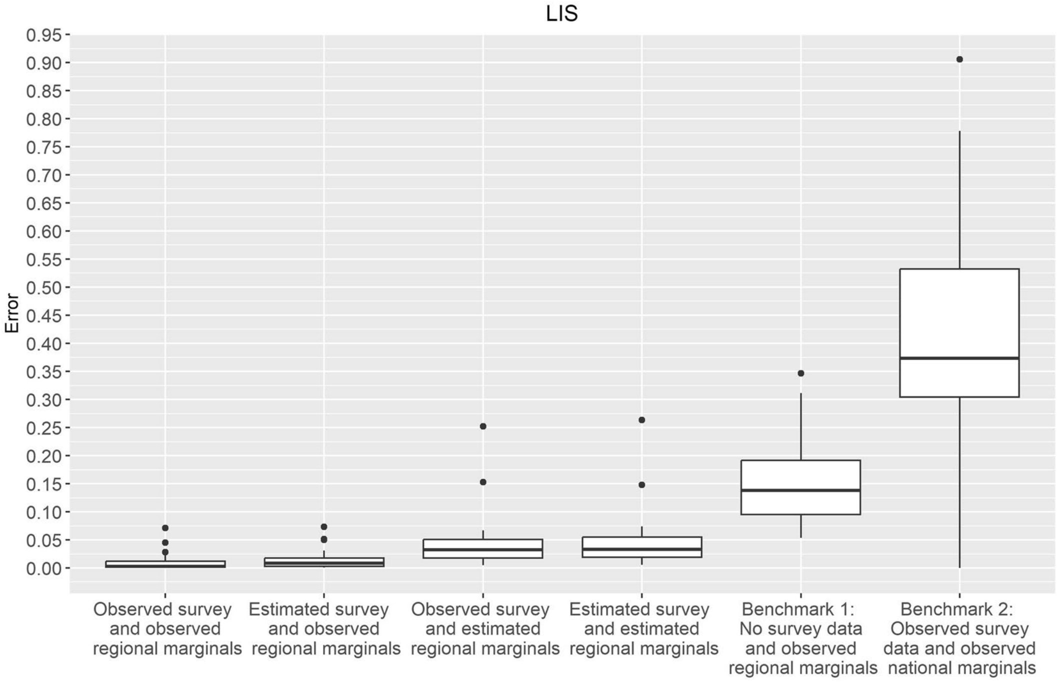


Fig. 3 Size of the errors for countries with LIS data under different conditions.

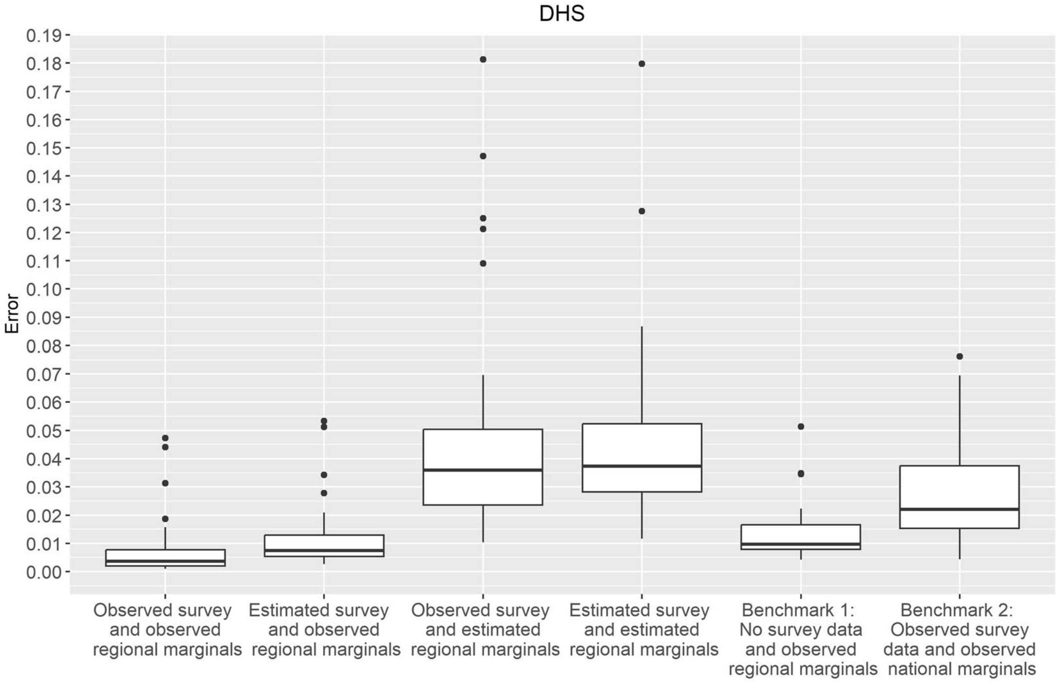


Fig. 4 Size of the errors for countries with DHS data under different conditions.

Code availability

The synthetic population database is developed in R using the following packages: dplyr, stringr, tidyr, zoo, reshape2, tibble, wrswoR, data.table, mlfit and readstata13, and MASS. The scripts are publicly and freely accessible on GitHub: https://github.com/VU-IVM/GLOPOP-S/. The code was run on the VU BAZIS cluster, where we used up to 500GB of RAM to generate the synthetic populations of the largest countries (China and India).

Received: 21 December 2023; Accepted: 11 September 2024;
Published online: 07 October 2024

References

1. Balcan, D. *et al*. Modeling the spatial spread of infectious diseases: the GLobal Epidemic and Mobility computational model. *J. Comput. Sci.* **1**(3), 132–145, https://doi.org/10.1016/j.jocs.2010.07.002 (2010).
2. Yeh, S. *et al*. Detailed assessment of global transport-energy models' structures and projections. *Transport. Res. D-Tr. E.* **55**, 294–309, https://doi.org/10.1016/j.trd.2016.11.001 (2017).
3. Van Meijl, H. *et al*. Modelling alternative futures of global food security: Insights from FOODSECURE. *Glob. Food Secur.* **25**, https://doi.org/10.1016/j.gfs.2020.100358 (2020).
4. Ahmad, T. & Zhang, D. A critical review of comparative global historical energy consumption and future demand: The story told so far. *Energy Reports* **6**, 1973–1991, https://doi.org/10.1016/j.egyr.2020.07.020 (2020).
5. Jones, E. R. *et al*. DynQual v1.0: a high-resolution global surface water quality model. *Geosci. Model Dev.* **16**(15), 4481–4500, https://doi.org/10.5194/gmd-16-4481-2023 (2023).
6. Haer, T., Husby, T. G., Botzen, W. J. W. & Aerts, J. C. J. H. The Safe Development Paradox: An Agent-Based Model for Flood Risk under Climate Change in the European Union. *Glob. Environ. Change* **60**, 1–12, https://doi.org/10.1016/j.gloenvcha.2019.102009 (2020).
7. Nordhaus, W. D. Revisiting the Social Cost of Carbon. *PNAS* **114**(7), 1518–1523, https://doi.org/10.1073/pnas.1609244114 (2017).
8. Arneth, A., Brown, C. & Rounsevell, M. Global models of human decision-making for land-based mitigation and adaptation assessment. *Nature Clim. Change* **4**, 550–557, https://doi.org/10.1038/nclimate2250 (2014).
9. Weyant, J. Some Contributions of Integrated Assessment Models of Global Climate Change. *REEP* **11**(1), 115–137 (2017).
10. Lippe, M. *et al*. Using agent-based modelling to simulate social-ecological systems across scales. *Geoinformatica* **23**, 269–298, https://doi.org/10.1007/s10707-018-00337-8 (2019).
11. Giarola, S., Sachs, J., d'Avezac, M., Kell, A. & Hawkes, A. MUSE: An open-source agent-based integrated assessment modelling framework. *Energy Strat. Rev.* **44**, https://doi.org/10.1016/j.esr.2022.100964 (2022).
12. De Mooij, J. *et al* A Framework for Modeling Human Behavior in Large-Scale Agent-Based Epidemic Simulations. *SIMULATION* **99(12)**, https://doi.org/10.1177/00375497231184898 (2023).
13. Rai, V. & Henry, A. Agent-based modelling of consumer energy choices. *Nat. Clim. Change* **6**, 556–562, https://doi-org.vu-nl.idm.oclc.org/10.1038/nclimate2967 (2016).
14. Prédhumeau, M. & Manley, E. A synthetic population for agent-based modelling in Canada. *Sci. Data* **10**, 148, https://doi-org.vu-nl.idm.oclc.org/10.1038/s41597-023-02030-4 (2023).
15. Münnich, R., Schürle, J. On the simulation of complex universes in the case of applying the German Microcensus. DACSEIS research paper series No. 4, University of Tübingen (2023).
16. Sun, L., Erath, A. & Cai, M. A hierarchical mixture modeling framework for population synthesis. *Transport. Res. B-M* **114**, 199–212, https://doi.org/10.1016/j.trb.2018.06.002 (2018).
17. Alfons, A., Kraft, S., Templ, M. & Filzmoser, P. Simulation of close-to-reality population data for household surveys with application to EU-SILC. *SMA* **20**, https://doi.org/10.1007/s10260-011-0163-2 (2011).
18. Wu, G., Heppenstall, A., Meier, P., Purshouse, R. & Lomax, N. A Synthetic Population Dataset for Estimating Small Area Health and Socio-Economic Outcomes in Great Britain. *Sci. Data* **9**, https://doi.org/10.1038/s41597-022-01124-9 (2022).
19. Beckman, R. J., Baggerly, K. A. & McKay, M. D. Creating synthetic baseline populations. *Transport. Res. A-PP* **30**(6), 415–429, https://doi.org/10.1016/0965-8564(96)00004-3 (1996).
20. Lovelace, R., Birkin, M., Ballas, D. & van Leeuwen, E. Evaluating the performance of iterative proportional fitting for spatial microsimulation: new tests for an established technique. *JASSS* **18**(21) (2015).
21. Ton, M.J. *et al*. GLOPOP-S. *Harvard Dataverse*, V4. https://doi.org/10.7910/DVN/KJC3RH (2023).
22. Hino, M., Fiel, C. B. & Mach, K. J. Managed retreat as a response to natural hazard risk. *Nat. Clim. Change* **7**(5), 364–370, https://doi.org/10.1038/nclimate3252 (2017).
23. Berhe, M. *et al* The effects of adaptation to climate change on income of households in rural Ethiopia. *Pastoralism* **7**, https://doi.org/10.1186/s13570-017-0084-2 (2017).
24. Ehsan, S., Begum, R. A., Nizam, K., Maulud, A. & Mia, S. Assessing household perception, autonomous adaptation and economic value of adaptation benefits: Evidence from West Coast of Peninsular Malaysia. *Advances in Climate Change Research* **13**(5), 738–758, https://doi.org/10.1016/j.accre.2022.06.002 (2022).
25. Chai, L., Han, Y., Han, Z., Wei, J. & Zhao, Y. Differences in disaster preparedness between urban and rural communities in China, *IJDRR* **53**, https://doi.org/10.1016/j.ijdrr.2020.102020 (2022).
26. Lindersson, S. *et al*. The wider the gap between rich and poor the higher the flood mortality. *Nat. Sustain.* **6**, 995–1005, https://doi.org/10.1038/s41893-023-01107-7 (2023).
27. Nixon, R. *et al* The relationship between household structures and everyday adaptation and livelihood strategies in northwestern Pakistan. *Ecol. Soc.* **28**(2), https://doi.org/10.5751/ES-14026-280231 (2023).
28. The Demographic and Health Surveys (DHS) Program. https://dhsprogram.com/ (2022).
29. Luxembourg Income Study Database (LIS). https://www.lisdatacenter.org/ (multiple countries: November 2022 – September 2023). Luxembourg: LIS. (2022).
30. Rufat, S., Tate, E., Burton, C. G. & Maroof, A. S. Social vulnerability to floods: Review of case studies and implications for measurement. *IJDRR* **14**(4), 470–486, https://doi.org/10.1016/j.ijdrr.2015.09.013 (2015).
31. Yaméogo, B. F., Gastineau, P., Hankach, P. & Vandanjon, P. O. Comparing Methods for Generating a Two-Layered Synthetic Population. *Transport. Res. Rec.* **2675**(1), 136–147, https://doi.org/10.1177/0361198120964734 (2020).
32. Mueller, K. & Axhausen, K. Population synthesis for microsimulation: State of the art. In: Proceeding of Transportation Research Board 90th Annual Meeting. (2010).
33. Choupani, A. A. & Mamdoohi, A. R. Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research. *Transport. Res. Proc.* **17**, 223–233, https://doi.org/10.1016/j.trpro.2016.11.078 (2016).
34. Guo, J. & Bhat, C. Population Synthesis for Microsimulating Travel Behavior. *Transport. Res. Rec.* **1**, 92–101, https://doi.org/10.3141/2014-12 (2007).
35. Auld, J., Mohammadian, A. K. & Wies, K. Population synthesis with control category optimization, paper presented at *the 10th International Conference on Application of Advanced Technologies in Transportation*, Athens, Greece (2008).
36. Khachman, M., Morency, C. & Ciari, F. Impact of the Geographic Resolution on Population Synthesis Quality. *ISPRS Int. J. Geo-Inf* **10**(790), https://doi.org/10.3390/ijgi10110790 (2021).
37. Agresti, A. Categorical Data Analysis. New York: John Wiley (1990).
38. Chi, G., Fang, H., Chatterjee, S. & Blumenstock, J.E. Microestimates of wealth for all low- and middle-income countries. *PNAS* **119**(3), https://doi.org/10.1073/pnas.2113658119 (2022).
39. Vyas, S. & Kumaranayake, L. Constructing socio-economic status indices: how to use principal components analysis. *Health Policy Plan.* **21**(6), 459–468, https://doi.org/10.1093/heapol/czl029 (2006).
40. OECD. Economic well-being. *OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth*, OECD Publishing, Paris. https://doi.org/10.1787/9789264194830-en (2013).
41. Smits, J. & Permanyer, I. The Subnational Human Development Database. *Sci. Data* **6**, 190038, https://doi.org/10.1038/sdata.2019.38 (2019).

42. Pesaresi, M., Florczyk, A., Schiavina, M., Melchiorri, M. & Maffenini, L. GHS settlement grid, updated and refined REGIO model 2014 in application to GHS-BUILT R2018A and GHS-POP R2019A, multitemporal (1975-1990-2000-2015), R2019A. European Commission, Joint Research Centre (JRC) https://doi.org/10.2905/42E8BE89-54FF-464E-BE7B-BF9E64DA5218 (2019).
43. Freire, S., Halkia, M. & Pesaresi, M. GHS population grid, derived from EUROSTAT census data (2011) and ESM R2016 - OBSOLETE RELEASE. European Commission, Joint Research Centre (JRC) [Dataset] PID: http://data.europa.eu/89h/jrc-ghsl-ghs_pop_eurostat_europe_r2016a (2016).
44. Zhou, M., Li, J., Basu, R. & Ferreira, J. Creating spatially-detailed heterogeneous synthetic populations for agent based microsimulation. *Comput. Environ. Urban. Syst.* **91**, 101717, https://doi.org/10.1016/j.compenvurbsys.2021.101717 (2022).
45. Ye, X., Konduri, K., Pendyala, R. M., Sana, B. & Waddell, P. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In: 88th Annual Meeting of the Transportation Research Board (2009).
46. Lovelace, R. & Ballas, D. "Truncate, Replicate, Sample": A Method for Creating Integer Weights for Spatial Microsimulation. *Comput. Environ. Urb. Syst.* **41**, 1–11, https://doi.org/10.1016/j.compenvurbsys.2013.03.004 (2013).
47. Sun, L. & Erath, A. A Bayesian network approach for population synthesis. *Transport. Res.C-ET* **61**, 49–62, https://doi.org/10.1016/j.trc.2015.10.010 (2015).

Acknowledgements

We are very grateful to be able to use Luxembourg Income Study (LIS) data in this study. This research would not have been possible without access to the data resources provided by LIS. In addition, we kindly acknowledge sponsoring by the VU HPC Council and the IT for Research (ITvO) BAZIS Linux computational cluster at the Vrije Universiteit Amsterdam. The study has been funded by the EU ERC COASTMOVE project grant nr. 884442.

Author contributions

M.T.: conceptualization, methodology, software, validation, visualization, writing - original draft, review and editing. M.I.: methodology, software, writing – review and editing. J.B.: conceptualization, methodology, writing – review and editing, supervision. H.M.: writing - review and editing, supervision. L.R.: writing - review and editing. W.B.: writing - review and editing. J.A.: conceptualization, writing - review and editing, supervision, project administration, funding acquisition.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03864-2.

Correspondence and requests for materials should be addressed to M.J.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.