Autonomous Agents and Multi-Agent Systems (2024) 38:48
https://doi.org/10.1007/s10458-024-09680-7

GenSynthPop: generating a spatially explicit synthetic population of individuals and households from aggregated data

Jan de Mooij[1] · Tabea Sonnenschein[2,3,4] · Marco Pellegrino[1] · Mehdi Dastani[1] · Dick Ettema[3] · Brian Logan[1,5] · Judith A. Verstegen[3]

Accepted: 22 September 2024

Abstract
Synthetic populations are representations of actual individuals living in a specific area. They play an increasingly important role in studying and modeling individuals and are often used to build agent-based social simulations. Traditional approaches for synthesizing populations use a detailed sample of the population (which may not be available) or combine data into a single joint distribution, and draw individuals or households from these. The latter group of existing sample-free methods fail to integrate (1) the best available data on spatial granular distributions, (2) multi-variable joint distributions, and (3) household level distributions. In this paper, we propose a sample-free approach where synthetic individuals and households directly represent the estimated joint distribution to which attributes are iteratively added, conditioned on previous attributes such that the relative frequencies within each joint group of attributes are maintained and fit granular spatial marginal distributions. In this paper we present our method and test it for the Zuid-West district of The Hague, the Netherlands, showing that spatial, multi-variable and household distributions are accurately reflected in the resulting synthetic population.

Keywords Synthetic population · Spatial heterogeneity · Sample-free data synthesis · Data disaggregation · Synthetic households · Iterative proportional fitting · Synthetic reconstruction

1 Introduction

A synthetic population is a representation of individuals living in a specific area that fits the spatial, socio-economic and demographic characteristics of a real-world population. This representation maintains privacy, as no single entity in synthetic population represents a true individual. Synthetic populations are often used to build agent-based social simulations to study and understand processes, test hypotheses, or make forecasts on a wide variety of social issues. Using synthetic populations in agent-based social simulations allows modeling each individual as unique software agents, whose autonomous decisions and interactions in a complex social and physical environment together can lead to emergent

Extended author information available on the last page of the article

patterns. Researchers and policymakers can use these patterns to study the effects of complex sociological or human-environment dynamics, such as school segregation [1], disaster simulation [2], environmental public health interventions [3] and, especially prevalent during the times of COVID-19, infectious disease dynamics [4–6]. Simulations that arose during that time incorporated socio-demographic attributes in a severely ad-hoc fashion, which highlights the relevance of robust population synthesis methodologies [7–9].

If the behavior of individual agents in such simulation studies is to be guided at least partly by their demographics, the quality of the modeled decision-making is dependent on the quality of the representation of those demographics in the target population. Demographics tend to be highly heterogeneous, with values unevenly distributed across social strata and geographic space. When studying demographic change over time, an accurate starting point becomes all the more pertinent. For this reason, the utility of synthetic populations is increasingly recognized, and they have been used in a wide variety of agent-based simulations, including urban mobility [10, 11], disaster management [12, 13] and epidemiology [7, 14].

The traditional approach of synthesizing such a population requires a detailed microdata sample of the actual population which includes all relevant attributes, and using a procedure called Iterative Proportional Fitting (IPF) to estimate the true joint distribution of one or more of the smaller regions in the sample's area to match the known margins of each of those attributes [15–17]. Synthetic individuals or households are then drawn from the microdata sample using the weights given by the estimated joint distribution. However, microdata may not always be available or within budget, which has led to a new class of approaches often referred to as *sample-free* or as *synthetic reconstruction* [18, 19]. These approaches—despite their less stringent data requirements—have been shown to perform on par with sample-based approaches [18, 20]. The challenge of sample-free approaches lies in combining the aggregated data from multiple sources and levels of aggregation. Synthetic individuals or households are then directly drawn, one by one, from that distribution instead of from a sample. Previous sample-free methods have focused either on: (1) granular spatial representativeness, ignoring the multi-variable joint distributions [21], or (2) the correlation structure, failing to spatially distribute the population representatively [22]. These methods struggle to simultaneously match both individual- and household level variable joint distributions together with granular spatial distributions [18, 19]. Our method tackles that gap by matching those distributions at different levels, allowing to maximize the data and variation used and fit the population to distributions on all those levels.

This paper is the result of a data-driven and large-scale agent-based social simulation project with the aim to study the effects of the future deployment of an on-demand bus service on modal choice of the population, and to investigate key interventions, or "nudging" policies for stimulating the use of healthier and more sustainable travel mode choices. The leading assumption for this study has been that the individuals' demographics, such as age, income, migration status, car ownership, etc, are key decision factors for modal choice, and that these attributes are highly heterogeneous across the population. For the target area, no microdata was available, but detailed and high quality aggregated data was. However, due to some attributes co-occurring relatively infrequently, we have found random drawing tended to skew the results in those smaller groups. Moreover, while going back and forth with stakeholders who could then come up with new requirements, we needed a method where new attributes could be added to an existing synthetic population without having to start from scratch.

In this paper, we propose a new methodology for generating a spatially explicit heterogeneous synthetic population from aggregated data without a sample. We have published

R- and Python packages called *GenSynthPop* [23, 24] to facilitate the data preparation and method implementation. The approach differs from existing approaches in that the individuals and households in the synthetic population directly represent the estimate of the true joint distribution, instead of being drawn from it. Moreover, the method combines spatial marginal distributions with contingency tables to allow for more granular spatial referencing. This is achieved by starting with a homogeneous population of individuals which are then iteratively disambiguated by the addition of a single attribute to the entire population at a time, conditioned on possible previously added attributes and fitting to both granular spatial marginal distributions and contingency tables. For each attribute, each group of candidate individuals in the synthetic population is split into groups such that the relative group sizes match the reported relative frequencies of the levels of the newly added attribute. These levels are then assigned to those groups respectively. The approach encourages reuse of synthetic populations, since attributes can be added without having to re-sample the entire synthetic population. The approach is also compatible with existing methods of assigning activity schedules [25, 26], which can help bootstrap realistic movement of agents across the study area.

The remainder of the paper is organized as follows. In the next section, we discuss the current state of the art and how our method relates to it. In Sect. 3, we first present our methodology in general terms, before comparing a case study population generated using our method to known distributions in Sect. 4. We discuss our methodology in Sect. 5 and finally conclude our work in Sect. 6.

2 Background

The concept of synthetic populations was introduced by Beckman et al. in 1996 as part of their work on the TRANSIM travel forecasting models. They proposed randomly drawing households from publicly available microdata (specifically, the US Public Use Microdata Sample, or PUMS) weighted by an estimate of the joint distribution of all relevant attributes occurring in that microdata. The joint distribution is estimated for each census tract or census block group sized area from the sample, by using a method known as Iterative Proportional Fitting (IPF) [27].

IPF has been employed by many population synthesis approaches where a joint distribution (e.g., a sample) provides the likely odds ratios between attributes in a population but margins of individual attributes are considered more accurate for the population at hand (see e.g. [15–17, 28]). The estimate of the true joint distribution is created by updating all cells in one dimension such that the dimension total matches the margins. This process is iteratively repeated for each dimension for which margins are available, until the relative changes are smaller than some predefined stopping criterion. Ireland and Kullback [29] have shown that IPF maximizes entropy under the provided marginal constraints.

Guo and Bhat [30] observed that drawing entire households does not preserve the joint personal level attributes, because the individuals are members of the selected households rather than drawn to match the joint personal level attributes directly. They further highlight the fact that, at the time, most synthetic populations were constructed for the specific application, limiting their reuse value. They proposed an extension of the approach in which each individual type (characterized by the combination of personal level attributes) is capped and drawn households are only placed in the synthetic population if adding its members does not exceed this cap. They published an implementation of their approach

in an Object-Oriented Programming (OOP) language, in which application specific details are abstracted away as much as possible. Later, other approaches to assist in the population synthesis based on microdata have also been developed, perhaps best known of which is Gen* [31].

Building on Guo and Bhat's work, Ye et al. [32] further refined the IPF procedure into something they called *Iterative Proportional Updating*, in which both the household-level and personal-level attributes can be fitted to the same data simultaneously. In their approach, all possible attributes appearing in the microdata sample related to both households and individuals are placed in a single table and are initially assigned equal *weights*. Instead of fitting the relative frequencies to the margins directly, the weights are iteratively (i.e. one attribute at the time, starting with household attributes and followed by person-level attributes) adjusted by the true counts reported in the marginal data. This approach is repeated until the goodness of fit is judged to be sufficient.

Adiga et al. [15] have used the approach of Beckman et al. to synthesize a population for the entire United States, but augmented the synthetic population with *activity schedules*, which give each individual a set of temporally consistent activities throughout the day, *location choice*, which assigns appropriate locations to each of those activities, and *contact estimation*, in which a realistic estimation of what other individuals co-locating in the same place an individual meets during their activities. This in turn means that the synthetic population is also an estimation of a *social contact network*.

However, a representative microdata sample may not be available or within budget for every target region. While some authors have used surveys in their place [33], others have moved to a new class of approaches often referred to as *sample-free* or as *synthetic reconstruction*. Gargiulo et al. [19], for example, first generate a list of individuals that follows a known distribution, and then exhaustively generate a list of all possible household types characterized by the possible combinations of the relevant attributes values. Each of these household types is then assigned a probability defined as the independent combination of partial probabilities that come from the available data. Households are drawn from this list with this probability, and individuals are taken (without replacement) from the generated set of individuals to match the corresponding attribute values of the selected household. Lenormand et al. [20] have compared Garguilo et al's approach to the sample-based approach proposed by Ye et al. [32] and have concluded that while both approaches require similar amounts of computing power, the sample-free approach resulted in better global matches and is still applicable when no microdata sample is available. However, the method also has drawbacks. The only person-level attribute they consider is age. Moreover, sometimes a household cannot be filled, because no more individuals with the required attribute values exist. Presumably, the difficulty of finding a suitable individual from the candidates only increases when more person-level attributes are also considered, as this only further constrains the selection. Moreover, their method does not consider spatial heterogeneity in the distribution. The spatial distribution of the population is particularly relevant for spatial microsimulations.

Barthelemy and Toint [18], in contrast, generate individuals described by richer attributes which are drawn from known distributions. In their approach, they first estimate multiple joint distributions; one for each of the levels of aggregation they consider. Each of these only contains the attributes for which contingency tables are available for at minimum that level of aggregation. Next, an individual is drawn from the distribution with the lowest level of aggregation. The missing attributes are then drawn, conditioned on previously drawn attributes, from the first higher-level-of-aggregation distribution which contains them. This process is repeated until the required number of individuals is drawn. Like

Gargiulo et al., they then draw a household from the known household types and try to populate it with the previously created individuals. Barthelemy and Toint compare their method with Guo and Bhat [34], finding that it leads to more accurate distributions, with a particularly stark improvement of the joint distributions of household attributes. While their approach allows for spatial heterogeneity, the level of spatial detail of each attribute is limited to the one for which contingency tables are available. Instead of only combining multiple attributes from different levels of aggregation, using multiple levels of aggregation even for the same attribute could further enhance the level of detail. Moreover, their approach requires correcting inconsistencies in the margin—caused by sampling at different levels of aggregation—after the fact. In their approach, this is only possible for ordinal attributes.

We observe that all approaches highlighted here (both sample-based and sample-free) draw individuals or households from either the available sample or from the estimated joint distributions. This poses four major challenges. The first is how to combine the available data into a single estimate of the true joint distribution and overcoming varying levels of aggregation and multi-dataset inconsistencies. Secondly, if an attribute is added at a later stage, the estimated joint distribution changes, which requires redrawing all individuals and performing all subsequent steps again. Third, due to the stochastic nature of drawing, most approaches suggest generating a multitude of synthetic populations, and selecting the one that best fits, which comes with a significant computational penalty. Finally, combinations of attribute values that occur an infrequent but non-zero amount of times may never be drawn if each individual or household is considered individually and probabilistically.

Given that the distribution of attributes across the synthetic population itself is supposed to represent the true joint distribution, it seems the two steps of first estimating this distribution and then iteratively drawing individuals or households from it can be combined into one.

In this paper we introduce a new, iterative, sample-free, deterministic approach in which the individuals and households of the synthetic population themselves represent the increasingly detailed estimate of the joint distribution of attributes. Further, our method allows combining spatial marginal distributions, which are usually available at a more spatially detailed level, with higher level contingency tables for each attribute separately. The latter approach, moreover, captures relative distribution across spatial units better. Finally, the method allows intermediate validation at any stage, and captures infrequent values of attributes better by using deterministic assignment rather than probabilistic drawing. See Table 1 for an overview of differences to Barthelemy and Toint.

3 Generating a synthetic population

In this section we propose the methodology to build a synthetic population from multiple aggregated data sets. Where possible, these data sets come from attested institutes and have been designed to reflect the true distributions of the socio-demographic and geo-spatial attribute values of the population.

Formally, a synthetic population $S = \{A_1, \ldots, A_n\}$ is a spatially heterogeneous representation of the estimated joint distribution of a number of categorical attributes over some geographic region through n synthetic individuals. These individuals together represent the real individuals living in that region in such a manner that no specific synthetic individual can be linked to a real person, thus preserving privacy while still accurately representing

Table 1 Comparing GenSynthPop to Barthelemy & Toint's (2013) method

Method dimension	Barthelemy & Toint [18]	GenSynthPop
Spatial referencing	Determines per attribute by the lowest level of spatial aggregation for which a contingency table is available	Combines spatial marginal data with contingency tables at multiple levels of aggregation for each attribute
Integration of data of various aggregation levels	Merges contingency tables in distinct joint distributions per level of aggregation	Fits contingency tables to spatial marginal distributions of lowest level of aggregation for each attribute separately
Multi-dataset inconsistencies	Aggregate level distributions may be inconsistent with known margins; attributes are shifted between Individuals after drawing but only ordinal attributes can be corrected	Attributes' values are deterministically distributed over individuals using frequencies; the synthetic population intrinsically represents the best maximal-entropy estimate of the true distribution
Evaluation	Compare the total for the entire population against fitted joint distributions	Compare every available combination of attribute values against fitted joint distributions
Attribute assignment	Drawing from distribution	Deterministic assignment

the area of study. A synthetic individual $A \in S$ is characterized as a tuple of values for the m different categorical attributes $\mathcal{V}(S) = \langle V_1, \ldots, V_m \rangle$ of the synthetic population and represented as $A = \langle v_1, \ldots, v_m \rangle$, where v_i is one of the possible values of attribute V_i. The set of attributes that are included in the synthetic population depends on the design criteria of the study and the availability of data. Common attributes that are generally included in synthetic populations are for example *age*, *gender*, *education* and *income*. Spatial heterogeneity is achieved through the inclusion of spatial location in the individuals' attributes.

A synthetic population may optionally be augmented with synthetic households. Such a household $(H, \langle v_1^h, \ldots, v_l^h \rangle)$ is a pair, where $H \subseteq S$ denotes its constituent members, and $\langle v_1^h, \ldots, v_l^h \rangle$ is a tuple of values for l additional categorical attributes that apply to the constructed household. These attributes and their values are selected and distributed using the same approach as those of the individuals, but the partitioning of individuals into households requires an additional step. In this section, we detail our data-driven and sample-free approach for constructing the synthetic population of individuals and their partitioning into households. We propose an iterative approach for constructing a synthetic population by repeatedly adding a single attribute conditioned on previously added attributes (see Fig. 1).

The process starts by instantiating and locating the appropriate number of synthetic individuals in the region by creating n tuples of size one (i.e., $m = 1$) with the value of the single attribute representing the citizens location. New attributes are then iteratively added to the synthetic population by assigning each individual a value for the new attribute conditioned on the values of the previously added attributes.

Individuals are then partitioned into households, taking into account known household composition distributions, after which households and their constituent members can be further extended with additional attributes in the same manner. The exact process will be detailed after a general discussion on the types of data sets that can be used.

3.1 Data

We assume the availability of one or more data sets, where a data set π is a k-way *contingency tables* with $k \geq 1$ categorical attributes $\mathcal{V}(\pi) = \langle V_1, \ldots, V_k \rangle$. In a contingency table, each dimension enumerates the values for one of the attributes and each cell in the table provides the counts for how often the combination of attribute values that cell intersects occurs.

As an example, the top left of Fig. 2 shows a two-way contingency table for the attributes *Age Group* and *Gender*. The notation $\pi_{v_1,v_2,\ldots,v_k}$ is used to refer to the cell that intersects where $V_1 = v_1, V_2 = v_2, \ldots, V_k = v_k$. In the case of the example, the top left cell would be indicated as $\pi_{0\text{–}15,\text{Male}} = 16$

Note that these attributes are all assumed to be categorical in the sense that their values are categories such as being female or having a specific age group. So, while some of these categories describe integer or real numbers (e.g., age, household income), these are usually grouped into categories, such as the *age group* or *income level*. This is done both to maintain privacy and to keep data set sizes manageable. Of course, individuals or households can be assigned specific integer or real values from the assigned category and the number of categories for an attribute can be virtually unbounded where necessary.

The data sets are published by attested institutes but can differ in both presentation and their level of detail. In some cases, the data may be presented as a *joint distribution*. Such data sets provide the *relative* frequencies (i.e. fraction of all individuals) to which each combination of attributes applies. An example of these types of data sets is given

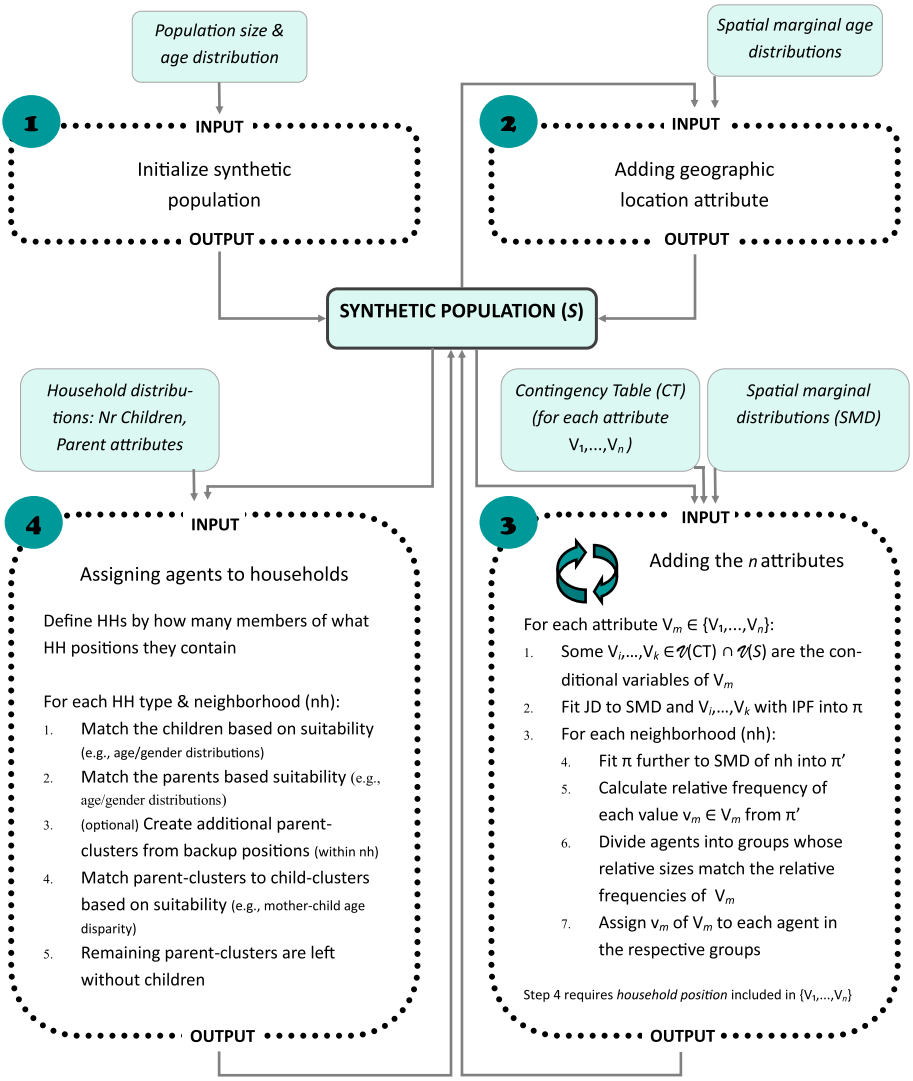

Fig. 1 Flowchart of the method

in the bottom left of Fig. 2. In some cases, only the *marginal data* are published. These types of data are essentially a special case of the k-way contingency table where $k = 1$, i.e., where only counts of the values of a single attribute are given. Figure 2 shows how joint and marginal distributions can be derived from the contingency table. The contingency table can be reconstructed from the joint distribution if the total number of individuals is known, but neither the joint distribution nor the contingency table can be reconstructed from the marginal data, because inter-attribute dependencies are lost.

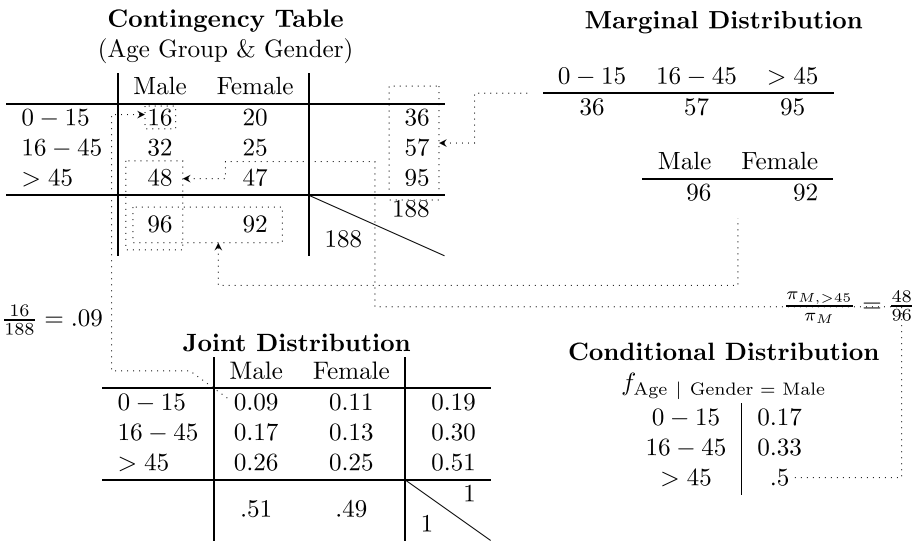

Fig. 2 Illustration of the relation between the contingency table, joint distribution and marginal data on a fictional population of 188 individual

A *conditional distribution* can be derived from both the contingency table and the joint distribution. The conditional distribution gives the relative frequency for each of the values of a single attribute given the presence of specific values of one or more other attributes. The relative frequency $f_{v_j \mid v_1,\ldots,v_k}$ of attribute V_j having value v_j, given specific values $V_1 = v_1, \ldots, V_k = v_k$ of all other attributes, is determined by Eq. 1. For example, the relative frequency of each of the three age groups for the male gender is given in the bottom right of Fig. 2.

$$f_{v_j \mid v_1,\ldots,v_k} = \frac{\pi_{v_1,\ldots,v_k,v_j}}{\pi_{v_1,\ldots,v_k}} \tag{1}$$

Any data for socio-economic and demographic characteristics also encodes spatial information. This may be explicitly through the value of one of the categorical attributes, but more often, the data are aggregated to a specific and well-defined region. In that case, any single data set is spatially homogeneous, but by combining data sets from multiple adjacent regions, spatial heterogeneity can be introduced. In line with the literature, we say that the smaller the region a data set describes, the lower its level of aggregation. For example, some data may be available on the level of single streets or postal codes, while other data may be aggregated to an entire city, province, or even country. To obtain maximum accuracy, when all else is equal, we encourage the use of data sets with more attributes over those with fewer attributes, and data sets with lower levels of aggregation over those with higher levels of aggregation. In practice, however, due to a variety of reasons such as privacy concerns or the cost of collecting the data, this often turns out to precisely be the trade-off, and data sets with lower levels of aggregation tend to be conditioned on fewer attributes. Data sets from different levels of aggregation can easily be combined using the IPF procedure when available. This has the benefit of maintaining inter-attribute

dependencies as well as possible, while still making use of the lower levels of aggregation for spatial heterogeneity. In general, our (iterative) methodology creates a synthetic population by dealing with the attributes one at the time. This means only one contingency table has to be fitted to the known margins at a time, which is considerably easier than estimating the full joint distribution of all categorical variables required for the synthetic population at once.

Whenever there is data available at lower levels or aggregation than the full contingency table we have estimated in this stage, the attribute is added one region distinguished by that level of aggregation at the time. The estimated contingency table is further fitted to the margins of that specific region as well. To emphasize this, we refer to the margins of each of those regions as the *spatial marginal data*.

For the method that we will present next, we assume that the values a categorical variable can take are exhaustive for the described population in the contingency table or spatial marginal data. In some cases, the data as published does not meet this requirement, in which case some manual processing is required. Processing varies from mapping the value names used in one data set to another, to estimating the count of a missing attribute value, to largely having to manually hunt for the data in various scattered publications. The way missing levels have to be addressed is very domain-sensitive, and the method presented in the next section does not provide a definitive solution. However, some example of how to deal with such a scenario will be discussed in Sect. 4 where we present a case study of the methodology.

3.2 Adding an attribute

The process of generating a synthetic population always starts with identifying the lowest level of aggregation that the available data permits. Then, for each region at that level of aggregation we instantiate the required number of individuals (e.g., the absolute number of individuals living there) by creating a matching number of empty individual tuples. The first attribute is always that region or a specific point within that region in which the individual is assigned. For the remaining iterative part of the procedure, we first sketch the general intuition of adding a single attribute to the population of synthetic individuals (which also applies to adding attributes to synthetic households) and give a pseudo algorithm later. Any attribute that is added to the synthetic population–safe for the first–will always be based on a data set related to that attribute. Consider, for example, a synthetic population for which the attributes *age*, *gender* and *migration background* are already added. Suppose now that we have a new data set containing a new attribute *education*, in addition to some of the previously added attributes such as *age* and *gender*.

Now, even if this contingency table represents the exact area of our synthetic population, its margins may be inconsistent with the data that was used to add the common attributes (age and gender) earlier, so we fit the table to those known margins first. Crucially, our synthetic population represents at any stage our best estimate of the true joint distribution of attributes we have already added, so we can derive the margins from the synthetic population itself if no (compatible) margins are published at the level of aggregation required.

We then derive the relative frequency for each value of the target attribute (education) in each group of the common attributes (age and gender) according to Eq. 1. We then split the individuals into groups along the values of those previously added attributes. This results in one group, for example, where all members have the age (or age group) of $20-30$

and gender *female*, and another where all members have age (group) $30-40$ and gender *female*.

Within each group, we then further divide the members into subgroups whose relative sizes match the relative frequencies. For example, we may find that the data set specifies that of the individuals within the age group $20-30$ and gender *female*, 20% has a *low* level of education, 30% has a *middle* level of education and 50% has a *high* level of education. Let's assume our selected group consists of 100 individuals. We then divide those 100 individuals into three groups with 20, 30 and 50 individuals, respectively. Finally, we add the attribute values *low*, *middle* and *high* to those groups, respectively.

Note that even if the groups are very small, e.g., just 10 individuals, the relative frequencies between the categorical attribute values should still be preserved using this method (provided the group has enough members to assign to each of those values, i.e., the example above would fail with a group of just 3 members). Note further that when we arbitrarily distribute the individuals within the age $20-30$ and gender *female* group across the three levels, it does not matter what individual is placed in what group because, as far as all the previously included information can tell us, all individuals are functionally identical.

Algorithm 1 An algorithm to assign the levels of a new target attribute V_m to the individuals in the existing synthetic population S conditioned on the distribution specified in a data set π

Require: Data set π with categorical attributes $\mathcal{V}(\pi)$
Require: A synthetic population S with categorical attributes $\mathcal{V}(S) = \langle V_1, \ldots, V_n \rangle$
Require: $V_m \in \mathcal{V}(\pi) \setminus \mathcal{V}(S)$ is the new attribute in π, but not in S
1: **procedure** ASSIGN(π, S, V_m)
2: $V_i, \ldots, V_k \leftarrow \mathcal{V}(S) \cap \mathcal{V}(\pi)$ ▷ Attributes that appear in both S and π
3: **for all** $v_i, \ldots, v_k \in V_i \times \cdots \times V_k$ **do**
4: $S' \leftarrow \{\langle v'_1, \ldots, v'_n \rangle \in S \mid v'_i = v_i, \ldots, v'_k = v_k\}$
5: **for all** $v_m \in V_m$ **do**
6: $f \leftarrow \frac{\pi_{v_i,\ldots,v_k,v_m}}{\pi_{v_i,\ldots,v_k}}$ ▷ Determine relative frequency of subgroup
7: $S \leftarrow S \setminus S''$ s.t. $S'' \subseteq S \wedge |S''| = |S'| \cdot f$ ▷ Take fraction of S (without replacement) corresponding to f
8: **for all** $\langle V_1 = v_1, \ldots, V_n = v_n \rangle = s_i \in S''$ **do**
9: $s_i \leftarrow \langle V_1 = v_1, \ldots, V_n = v_n, V_m = v_m \rangle$ ▷ Update in-place
10: **end for**
11: **end for**
12: **end for**
13: **end procedure**

The formal approach for adding a target attribute V_m to the existing set of attributes $\mathcal{V}(S) = \langle V_1, \ldots, V_n \rangle$ is given in Algorithm 1. The approach starts with finding a suitable data set π that contains the target attribute (i.e., $V_m \in \mathcal{V}(\pi)$).

This pseudo algorithm requires as input a data set π, a synthetic population S to which a (possibly empty) subset of attributes in the data set π has already been added previously, and a new attribute V_m that is in $\mathcal{V}(\pi)$, but not yet added to the synthetic population S. The goal of this algorithm is to add the new attribute V_m to the synthetic population. Line 2 determines the set of attributes from π which is already added to the synthetic population

S. For each possible combination of the values of those intersecting attributes (line 3), we identify all synthetic individuals that are characterized by exactly that combination (line 4). We then iterate over all values of the target attribute V_m (line 5) and calculate the relative frequency of that attribute value using Eq. 1 (line 6). Finally, from the identified individuals, take a fraction corresponding to the relative frequency of the attribute value, and assign them that value (lines 7–10). We make sure to only select individuals who were not already assigned a value for this attribute earlier. The entire process is repeated for each additional attribute that is added to the synthetic population, and stops when no more attributes are required.

3.3 Households

Different data sets are not always congruent, even when they come from the same provider, which is why, so far, we have ignored the true reported number of individuals and instead only used the (conditional) relative frequencies. We apply the same approach to households, because the number and size of reported households is similarly incongruent with the reported number of individuals in the same area. This incongruency causes existing approaches (Sect. 2) to struggle finding individuals that meet the household criteria before all individuals are associated with a household. Instead, we propose to procedurally determine the number of households required to exactly place each individual, while maintaining only the relative frequencies–instead of absolute counts–from the known household distributions.

We provide here a general approach for determining the number of households required to house the known number of individuals such that household members follow real inter-household distributions such as for gender and age.

We first assign household positions as individual level attributes using the same approach as in Sect. 3.2. We then specify the household types in terms of what household positions compose them, and how many of each are required. We iterate over those household types, and cluster the required number of children (if present in the household type) together. Next, adults are clustered together and (again only if children are present) matched with children clusters.

We note that our approach for creating the synthetic individuals starts with the required number of individuals. While we make no claims as to what this required number is (each study will have their own requirements), one possible method is to instantiate the reported number of households of each type, and start the population synthesis by initializing the reported number of individuals in each of those households. This would allow skipping the household partitioning process entirely, but is otherwise functionally identical to the methodology described here.

Formally, the process starts with a contingency table of household position on the level of individuals. Realistically, this information will likely be available at the household level instead of the individual level. The latter can be calculated from the former, but how depends on data availability. For example, the number of singles households requires the same number of individuals assigned the household position single, while two individuals are required to fill a household consisting of a couple without children. The level of detail included in this artificial table is up to the synthesizer, but crucially, a distinction between children who live with at least one adult and the rest of the population needs to be present. In the case that the individual-level household position data is available, that may be used as is, or enhanced with household level data.

If a realistic distribution of the number of children in a household is important, the assigned household position of a child should indicate the number of children that live in their household. This number, if not available in the contingency table directly, can be derived from the n total number of children and a distribution of the number of children per household. Each household will at least have 1 child and the distribution may specify up to c^+ children per household. Then h_c with $c \in 1 \ldots, c^+$ represents the number of households with c children and n_c the number of children living in a c-person household. The total number of households should therefore satisfy the constraint that $n = \sum_{c=1}^{c^+} h_c \cdot c$, but likely, there are either too few or too many children for this constraint. Instead, we derive the required values for h_c for each c, by calculating the number n_c as follows:

$$n_c = n \cdot \frac{c \cdot h_c}{\sum_{c'=1}^{c^+} c' \cdot h_{c'}} \tag{2}$$

Once the household position contingency table is in place, we assign the household position to the individuals as a normal individual level attribute, conditioned on other data and, where possible, including spatial marginal data (Sect. 3.2).

Next, we specify household types in terms of which household positions compose them, and how many of each of those positions should be included in a household. Each household position must occur in one household exactly once. We then iterate over each of these household types, starting with the households with children, and ending with the singles households. Within each household that contains children, we first create the child clusters such that each cluster contains the specified number of children. A cluster is created by selecting a random child first. The remaining children are scored with a suitability function that takes into account age and gender disparity, or any other sibling data available. The child that minimizes the score is selected and this is repeated until the specified number of children is reached. The final cluster may contain fewer than c children, if the total number of candidates is not divisible by c.

The parents (or childless adults) are clustered in the same manner, where the suitability score now uses data on partner distributions. If no suitable partner can be found with the specified household position, or if the number of parent-clusters is insufficient to house all the child-clusters, a candidate partner is switched with a similar individual from a backup household position or, in the case that there are too few individuals, taken from a backup position without switching. To avoid assigning an individual to multiple households, the backup positions should be selected such that the household type they are part of is constructed later in the process. Household types without children or singles households are suitable candidates for backup positions.

Finally, the child clusters are matched to the parent clusters where the suitability score represents information about child-parent distributions.

With that, the partitioning of individuals into households is complete, and each household now looks like a tuple $(H, \langle\rangle)$, where the second element is a–for now–empty tuple. This tuple is extended one attribute at a time using Algorithm 1, where new attributes can be conditioned on existing attributes in that tuple, as well as on attributes present in any of the individuals in that household. Person-level attributes can still be added after the household partitioning as well, and can now additionally be conditioned on attributes assigned to their household.

4 Case study

We now report on a case study using the methodology proposed in Sect. 3 applied to the the Zuid-West (South-West) district of The Hague in The Netherlands.[1] The Netherlands allows for a perfect case study for this approach, as detailed and reliable statistics are published by Statistics Netherlands (CBS) [35] (largely) for free, while detailed census samples are not available and usage of other micro-data is heavily regulated and subject to high premiums. While the quality and extent of availability of data may differ, these types of data sets are common in many other European countries as well.

The studied Zuid-West district was reported to have 84880 inhabitants in 2019 [36].[2] The lowest level of aggregation for which data were available in the studied district were the 14 neighborhoods. However, the majority of attributes is only provided (marginally or jointly) on the level of the entire municipality of The Hague. This is a larger area than that of our study, which is very illustrative of data availability in general. By conditioning those municipality-level attributes on common attributes that are available on the neighborhood level, some sense of spatial heterogeneity is still achieved for those attributes, especially when those data are fitted to each neighborhood separately.

4.1 Individual generation

The process of generating the synthetic population was started by creating, for each neighborhood, a number of tuples matching the reported number of individuals living in that neighborhood, and adding that neighborhood to those tuples as the first value (thus locating those synthetic citizens). Next, we identified the first attributes to be added. As most available contingency tables seemed to include at least age (or age *groups*), closely followed by gender, those two attributes were added first.

Age groups ($0-15, 15-25, 25-45, 45-65, >65$) are provided as marginal data per neighborhood [36], so within each neighborhood, citizens were assigned age groups with relative proportions matching those of the 5 reported age group sizes within the neighborhood's marginal data. Within each age group, integer *age* was then assigned following the marginal age distribution within each age group [37] (at the municipality level). Gender is only available marginally at the neighborhood level [36] or jointed over age at the municipality level [37], so the IPF procedure was used to first fit the joint distribution of age and gender to the margins of the age groups in the synthetic population as a whole and the margin totals over all neighborhoods for gender, and then further fitted to the marginal age and gender data of each neighborhood. Gender (binary) was then assigned, conditioned on age, to the citizens according to the neighborhood proportions given by the fitted data. The same procedure was then applied for migration background with values 'Dutch' (both parents born in the Netherlands), 'Western' (Europe, North-America or Oceania, excluding Turkey) and 'Non-Western', but using a different data set conditioned on both age group and gender [38] (at municipality level).

[1] Available as Open Source at https://github.com/A-Practical-Agent-Programming-Language/Synthetic-Population-The-Hague-South-West

[2] 2019 was the most recent data not collected during the COVID-19 pandemic which we have opted to disregard as the pandemics influence on the relevant socio-demographic and spatial attributes is unclear.

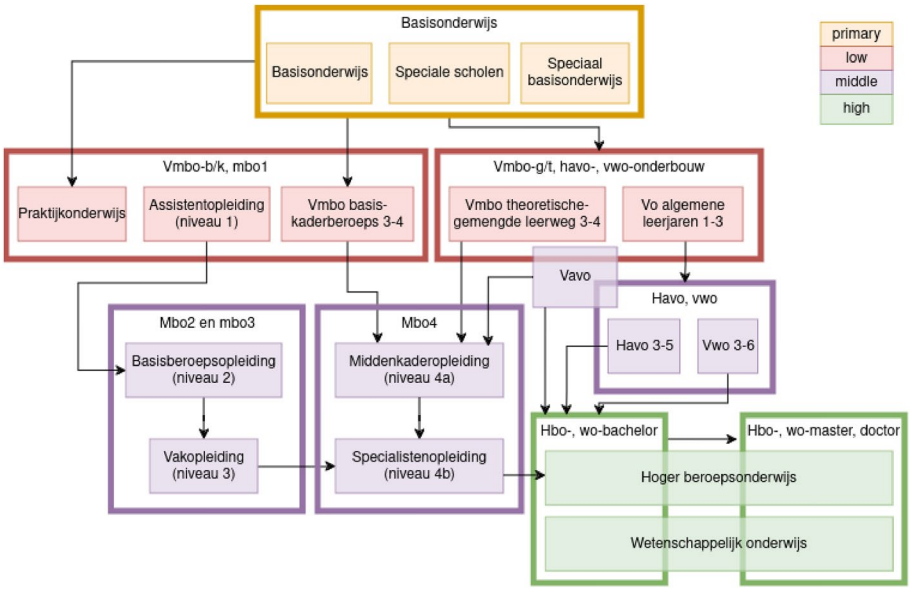

Fig. 3 The relation between education levels as used in various data sources. Outlined boxes indicate categories used for education attainment, the filled boxes show the categories used for current education, and the colors indicate what education level (primary education, low, medium or high education) the more detailed categories belong to. The arrows show how an individual may move from one type of education to the next in the Dutch education system

Education attainment was available with 8 different education levels on the municipality level [39] conditioned on age group, gender and migration background, but only as the levels 'low', 'medium' and 'high' on the neighborhood level. Moreover, this data is only reported for ages 15 and beyond. Current education was available in two different data sets. The first [40] reports the number of people enrolled in three types of primary school conditioned on age (up to 18) and gender. The second [41] reports 14 different levels of education conditioned on age, gender and migration background. These two data sets overlap for the ages 10–14 years. For current education, no marginal data is available. Because attained and current education are very likely related, all three data sets were first combined, to give an exhaustive joint distribution of both current and attained education on the level of the municipality using domain knowledge, which is visualized in Fig. 3. The categories from the attained education (outlined boxes) were mapped to that of absolved education (filled boxes). The arrows indicate how a student may progress through the Dutch education system. The reported counts from the current education data were used to proportionally distribute the reported counts from education attainment into the connected groups of current education.[3] When this was done, two contingency tables, one for current education and one for education attainment were derived from the combined table, both conditioned on age, gender and migration background. First education attainment was added to the synthetic population, conditioned on age, gender and migration background

[3] The process has been documented in the repository of this case study for readers interested in the details.

and on the neighborhood level on the marginal values (‘low‘, ‘middle‘ and ‘high‘). Next, current education was added, conditioned on age, gender and migration background.

Finally, *car license ownership*, *motor cycle license ownership* and *moped license ownership* were added based on a distribution jointed with age [42] (at the province level).

4.2 Household partitioning

The synthetic citizens were then partitioned into households in line with Sect. 3.3. Household positions were derived from data on the municipality level [43] conditioned on age and gender. This data distinguishes married and unmarried couples with or without children, single parents, singles and children. Next, the number of 1, 2 or 3-children households were estimated using Eq. 2 for single-parent and two-parent households respectively. The household counts were provided by marginal data on the municipality level [44]. Following the relative frequencies, the children's household positions were then replaced with labels indicating if they live in a single- or two-parent household and with how many children in total. The positions for couples with children and single-parents were similarly replaced. The resulting contingency table was then used to add the household position attribute to the synthetic individuals, further conditioned on three types of households (single-person, with children or without children) available marginally at the neighborhood level [36].

First, all households with children (married couples, unmarried couples and single parents) were created. Sibling suitability was determined based on closeness in integer age, except where integer ages are the same, which was arbitrarily given equal suitability as a 10-year age gap. Partner suitability (in couple households) was determined based on gender [45] and age- [46] disparity. Parent–child cluster suitability was scored based on parent–child age disparity [47]. Where too few parents were available, backup candidates were found from individuals classified as couples without children or as singles in case of couples households, and only from singles in case of single-parent households. Where too many parents were available, parent-clusters that had the worst suitability match with any child-cluster were left without children.

The households were then annotated with *standardized income group* [48] (municipality level) conditioned on household composition using the same approach as for adding citizen attributes. The *car ownership* [42] attribute was further added conditioned on household composition and standardized income group using a data set on the national level.

Lastly, each household was assigned a randomly weighted postcode from the neighborhood it was located in.

4.3 Results

The synthetic population is evaluated through four metrics suggested by Voas and Williamson [49], who examined the nature of goodness-of-fit tests and the consequences of their application to synthetic microdata. The measures they propose are designed to reflect specific peculiarities associated with synthetic microdata and are aimed at improving consistency in reporting the goodness-of-fit between synthesis methods. The scores of the person-level attributes are presented in Table 2 in Appendix A. The published packages include methods to facilitate determining these scores for a synthetic population.

The table lists in the first column the attribute for which the fit is evaluated. The second column contains all the other attributes over which the target attribute was jointed in the contingency table it was evaluated against. For example, the row with 'gender' in the first column and 'age group' in the second column reports how well the number of individuals with each combination of the two values of 'gender' and five values of age group (10 different groups) match the reported counts. In the case of an empty value in the second row, the scores indicate how well the number of individuals with each attribute value corresponds to the contingency table's marginal totals, and when the second column lists 'neighborhood', the scores indicate how well the number of individuals with each attribute value corresponds to the spatial margins of each of the 14 individual neighborhoods.

Before commenting on these scores, we will discuss the selected metrics in more detail.

The first metric is the traditional Pearson's goodness-of-fit score. Contingency tables are treated as single-variable tables to make them compatible with this test through the perspective that an individual either *is* or is *not* a 35 years old female with a high level of education and the Dutch migration background. The traditional test score is calculated per Eq. 3. Here O_i is the observed count at cell i (i.e., the number of individuals in the synthetic population with the combination of attribute-values associated with cell i), and E_i is the expected count (i.e., as found in the data).

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

0-values in the denominator are replaced with 1. According to Voas and Williamson's analysis, this makes the test suitable even where cell values are very small.

The second metric applies the binomial test to each cell individually. The unique stated advantage of the binomial is that it takes into account the relative difference between cells. This is relevant, because an absolute difference of 1 should have a larger penalty if the expected value is 2 than if it is 200. Let t_i be the observed proportion $\frac{E_i}{\sum E}$ at the ith cell and p_i the expected proportion $\frac{E_i}{\sum E}$ if $E_i \neq 0$ and $\frac{1}{\sum E}$ otherwise. Then the binomial score Z_i for the ith cell is calculated as per Eq. 4.

$$Z_i = \frac{(t_i - p_i) \pm \frac{1}{2 \cdot \sum E}}{\sqrt{\frac{p_i \cdot (1 - p_i)}{\sum E}}} \tag{4}$$

The quantity $\frac{1}{2 \cdot \sum E}$ is called the *continuity correction factor* and is subtracted from a non-zero difference between t_i and p_i or added if that difference is negative, and serves to counter-act the fact that small differences in small values are penalized too severely.

The Z^2 score is given by $\sum Z_i^2$ and can–just like the X^2–be tested against the χ^2 distribution with the number of degrees of freedom equal to the number of cells in the table. This value is equal to the number of cells (instead of 1 less) because neither the observed nor expected totals are constrained.

A χ^2 test is a hypothesis test for the null hypothesis that observed and expected values come from the same distribution. The critical value depends on the confidence level (conventionally $\alpha = 0.95$) and the degrees of freedom (DoF). Conventionally, the alternate hypothesis (the values do *not* come from the same distribution) is only accepted when the null hypothesis is rejected, i.e. when the X^2 or Z^2 score exceeds this critical value. In this case, the statistical probability of having wrongly rejected the null-hypothesis is $p < 1 - \alpha$. Increasing the confidence level thus decreases the statistical probability of falsely rejecting the null-hypothesis. Model testing such as applied here is an exception in that it proposes to accept the null hypothesis (the synthetic data is a good fit) whenever it is not rejected, i.e., the obtained Z^2 or X^2 scores are less than the critical value, and thus when $p > 1 - \alpha$. Changing the value of α does not formally change the confidence level in this case. For this reason, we report the X^2 and Z^2 directly along with the Degrees of Freedom (DoF) and p values at which each score would be equal to the critical value. We leave the decision of whether the results are statistically convincing to the reader, with the one note that under this interpretation, a larger p value suggests a closer fit between the observed and expected values and that $p = 1$ would therefore be the best case scenario.

Huang and Williamson have reported on the performance of both a sample-free and sample-based approach [50]. Because both algorithms are stochastic, they were each run 100 times for 7 different tables. Their Table 9 reports how many of the resulting models did not fit the source tables. They only report how many models fitted below the critical value of the χ^2 distribution at the $\alpha = 0.95, p = 0.05$ level, which makes a direct comparison of Z-scores impossible. However, a small number of the resulting models did exceed the critical value even at this level. The Z-scores reported for our synthetic population are well below this critical value. Moreover, the assignment of attributes is deterministic in terms of how many individuals are assigned each attribute value, so repeated runs would result in the same scores.

The remaining two metrics are both expressed in terms of the absolute error between expected and observed counts for attribute combinations, which Voas and Williamson call a purely descriptive statistic.

The *total* absolute error as given in the second-to-last column of Table 2 is given by Eq. 5

$$TAE = \sum |O_i - E_i| \tag{5}$$

Since the magnitude of TAE depends on the total number of individuals, it is not helpful in comparing approaches where different populations are modeled. They thus proposed the *standardized* absolute error as $SAE = \frac{TAE}{N}$, where N is the total expected count for the table. The lower the TAE and SAE, the better the synthetic population fits the contingency tables. The SAE is close to the absolute percentage difference (ADP) reported by Barthelemy and Toint [18]. In their paper, the ADP is reported as between 0.00 and 0.005, which is also the range for most of the SAE scores in our synthetic population. A small number of our attributes does exceed this range slightly. We believe this is largely explained by groups where an exact match is not possible. For example, Fig. 4 shows the absolute percentage-points difference between the number of females in each neighborhood in our synthetic population and the reported data.

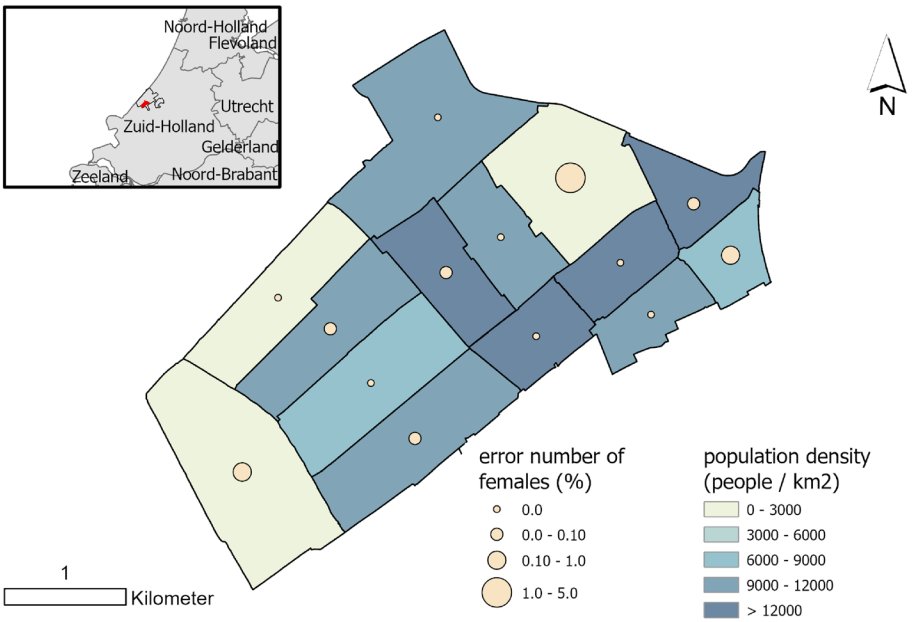

Fig. 4 Error between the synthetic population and the data in number of females (%pt) per neighborhood in Zuid-West (South-West) district (indicated in red in the inset map) of The Hague (city indicated by the black outline within Zuid-Holland). Population density in the neighborhood is shown to visualize the correlation between error and population density

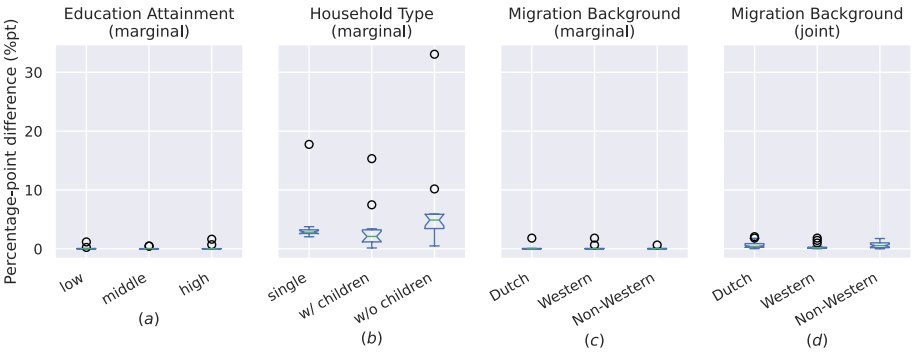

Fig. 5 Box plots of percentage difference in synthetic population and spatial marginal data for the values of three attributes **a**–**c**, repeated with the percentage difference in joint data for one of the attributes **d**

This difference is exactly 0 in each neighborhood where the total number of reported individuals matches that of our synthetic population. Even though the values come from the same data source [36], however, the number of reported male and female individuals in each neighborhood does not always sum to the reported neighborhood totals, because

each value is arbitrarily rounded up or down to a multiple of 5 for privacy considerations. In this situation, the number of male and female individuals cannot be matched simultaneously, which automatically leads to a (small) difference contributing to the *SAE*. The neighborhoods that show the largest percentage-points difference are also the smallest neighborhood (55 total in the top right, 155 total in the bottom left). The largest absolute difference in any of these neighborhoods is 3, i.e., at or below the rounding level of the source data.

A similar effect occurs in the contingency tables. After these are fitted using IPF, the reported counts can be non-integer values, while the synthetic population can only produce integer value counts. We do not round these values before scoring to avoid unfairly favoring our results, so any expected non-integer value also contributes towards the *SAE*.

The same can be observed in Fig. 5. This plot shows the percentage-points difference in each of the 14 neighborhoods, for each of the values of the 'education attainment' (5a), and 'migration background' (5d) attributes that were available at that level. The differences are close to zero. The true outliers, once again, represent the smallest neighborhoods. In this case, too, the largest absolute difference is 3.

Figure 5d shows the percentage-points difference for migration background, but this time conditioned on both 'age' and 'gender'. This data is not available at the neighborhood level, so the difference between each combination of 'migration background' × 'age' × 'gender' is shown for the synthetic population as a whole. The differences are very close to the neighborhood margins.

Finally, Fig. 5b shows the percentage difference in household types in each neighborhood. These differences are considerably larger, but this is not an individual-level attribute. Our methodology does not attempt to match the reported number of households, but rather prioritizes creating the number of households required to exactly house all the individuals. The outliers are once again caused by the smallest neighborhoods. For example, for the neighborhood with 55 individuals, the youngest person is reported as between 25 and 45 years old, while there are also 5 households with children reported. Moreover, 50 of the 55 individuals are reported as male, while half of the households are reported as containing couples. When also incorporating the known gender distributions, matching the individuals into the reported numbers of household types thus is non-trivial. A manual tweak could be a solution to this specific problem, but is beyond the scope of our methodology.

5 Discussion

The approach we have proposed here improves on existing methodology by integrating spatial marginal distributions with other variable joint distributions at various levels of aggregation for each attribute separately, immediately assigning that attribute's values to the individuals in the population deterministically, before moving on to the next attribute. Our methodology comes with several benefits. First, because the individuals are constructed iteratively instead of drawn from the estimated joint distribution in which all attributes are already present, additional attributes can be added to an existing synthetic population at any time. This removes the need to redo all steps that normally take place after estimating the source distribution, which typically includes drawing or constructing

all individuals and households (i.e., a large portion of the work). Secondly, while random drawing of individuals can skew distributions in smaller subgroups, the proposed methodology explicitly and deterministically replicates known proportions between all dependent attributes in the synthetic population. Thirdly, because spatial marginal distributions are available at a higher spatially granular level than contingency tables and capture relative spatial distributions better, our approach allows for higher spatial representativeness.

Our approach shares a number of limitations with other sample-free approaches. The first is that only attributes that have explicitly been added are incorporated. A sample-based approach ensures all attributes included in the sample will be included in the synthetic population as well—even in the absence of marginal data—because households and individuals are drawn from the sample with all included attributes. By merit of their co-occurrence with attributes that appear in the marginal data, they are even likely to be somewhat realistically distributed. Secondly, each additional attribute will require at least some extra data preparation in finding contingency tables and/or spatial marginal distributions and, where possible, combining the sources from various levels of aggregation. In the best case scenario, where all data sets use the same values for the attribute, the effort is minimal. In the worst case, as shown by the addition of current and attained education in Sect. 4.1, a large amount of domain knowledge is necessary to get the data to a usable state. When no sufficiently low level of aggregation data is available, the benefit our approach provides over other sample-free approaches is largely diminished. Some spatial heterogeneity will still occur simply by conditioning on existing attributes, but when no such data is available at all, this benefit cannot be enjoyed.

Another limitation comes from how we evaluate the synthetic population; counts of all groups in the synthetic population are compared against the prepared contingency tables, which have already been fitted to the known margins of the synthetic population, and which are used as direct input for the synthesis algorithm. This is defensible, since comparing two distributions with incongruent margins all but ensures a poor fit under the used metrics, but does warrant a comment. We assume as an integral part of our methodology that the spatial marginal data at lower levels of aggregation is more accurate than the contingency tables at higher levels of aggregation. IPF is designed for precisely this scenario, where the contingency table fitted to the known true margins is considered the best estimate of the true distribution. If the original and fitted distributions are not the same, then that means the original cannot be the true distribution, and thus should not be used for evaluation. Barthelemy and Toint also evaluate against their own estimate of the true distribution for the same reason [18].

Another limitation is that the household distribution depends on the inclusion of household position as an individual level attribute. While the proposed methodology attempts to provide for the largest amount of possible circumstances by being nonrestrictive on how this data is obtained, this is also the biggest drawback, as there is no unambiguous method for this step. There is too much variation in the types of households included in published data and the way those data sets are organized, so considerable effort and some level of creativity is required of the population synthesizer to complete this step.

The method for distributing individuals across households ensures all individuals are placed in a suitable household. Other sample-free methods that we have discussed attempt to match a predetermined number of households and individuals, which almost inevitably stops when the pool of individuals is exhausted before all households are filled, or when the pool of households is empty before all individuals are placed. In most cases, it is unclear how the remaining number of individuals or households should be treated. Our approach does not

suffer from this limitation, as the number of households is a function of how many individuals need to be placed in each household type, but the trade-off is that the distribution of household types across the neighborhoods is not as good as that of the personal-level attributes or as some rival methods.

Finally, the proposed matching of individuals into and between parent- and child clusters using a suitability score is very naive. While this method allows incorporating all the data that is available, it iterates over the pool of candidates one by one and selects the best suited individual from the remaining candidates. The first few clusters that are matched in this manner will have the best possible fit, but this decreases as the pool of remaining candidates get smaller.

6 Conclusion and future work

We have proposed a new methodology for generating a spatially heterogeneous synthetic population of individuals and households from aggregated data only, without a detailed microdata sample. The approach allows combining data from different levels of aggregation for each attribute, which improves spatial heterogeneity and resolution. Each attribute is added to the entire population at once in a deterministic manner, maintaining the relative frequencies of its values. Individuals jointly represent the best estimate of the true joint distribution instead of being drawn from such an estimate. As such, the population can easily be extended later, thus encouraging reuse.

We have published open-source packages called *GenSynthPop* implementing the proposed approach in both R and Python.

We have applied the proposed methodology to a small case study of the Zuid-West region of The Hague, The Netherlands. Our results demonstrate the true frequency distributions can accurately be reflected. We are working on an agent-based simulation that integrates the generated synthetic population to study the effects of the introduction of an on-demand bus service on modal choice of the population in this area, and to investigate key interventions, or “nudging" policies for stimulating the use of healthier and more sustainable travel mode choices.

Future work should investigate if a more standardized approach for determining household position can be developed. Further, future work should investigate if the household type distribution across neighborhoods can be improved without compromising on placing all individuals. Additionally, future work should improve the matching in- and between parent- and child clusters. One possible direction is to investigate if the entropy maximization and subsequent Tabu-search used by Barthelemy and Toint [18] are suitable alternatives for the naive suitability score. Finally, in future work, we intend to collaborate with institutions with detailed microdata of the target area to compare our results to their known distributions.

Appendix A: Individual attribute results

See Table 2

Table 2 Performance of GenSynthPop on individual-level attributes

		DoF	Z^2		X^2		Absolute error	
			Score	p value	Score	p value	Total	Standardized
Age group	Neighborhood	70	0.160028	1.000000	0.016845	1.000000	20.000000	0.000236
Gender	Neighborhood	28	0.168613	1.000000	0.185536	1.000000	26.000000	0.000306
	Age group	10	0.612648	0.999983	0.579719	0.999987	150.559563	0.001773
Integer age		106	2.144096	1.000000	3.601348	1.000000	124.606355	0.001468
	gender	212	26.086832	1.000000	31.153039	1.000000	1183.331065	0.013938
Migration background	Neighborhood	42	0.191206	1.000000	0.500781	1.000000	62.000000	0.000730
	Age group	60	26.784031	0.999936	27.521184	0.999898	1175.066312	0.013844
	Gender	6	1.421525	0.964536	1.140259	0.979733	296.952058	0.003498
	Age group × gender	120	27.329810	1.000000	30.074286	1.000000	1195.071783	0.014080
Absolved education		9	0.522445	0.999963	0.490611	0.999972	151.258870	0.001782
	Neighborhood	42	0.066162	1.000000	0.237732	1.000000	43.868289	0.000517
	Age group	171	69.205262	1.000000	71.689507	1.000000	1392.723678	0.016408
	Gender	18	3.668375	0.999874	3.583914	0.999894	460.845914	0.005429
	Age group × gender	342	73.542488	1.000000	78.444162	1.000000	1438.030022	0.016942

Table 2 (continued)

		DoF	Z^2		X^2		Absolute error	
			Score	p value	Score	p value	Total	Standardized
Current education		18	5.271613	0.998370	5.559259	0.997681	241.257163	0.002842
	Age group	738	400.828726	1.000000	88.304314	1.000000	344.817221	0.004062
	Gender	36	5.406962	1.000000	6.937750	1.000000	265.580831	0.003129
	Migration background	54	13.529318	1.000000	9.604675	1.000000	255.267884	0.003007
	Absolved education	162	6.639501	1.000000	6.341646	1.000000	241.481890	0.002845
	Age group × gender	1476	707.310600	1.000000	122.133588	1.000000	415.442018	0.004894
	Age group × migration background	2214	853.547746	1.000000	108.171522	1.000000	392.444181	0.004624
	Age group × absolved education	6642	464.316168	1.000000	92.685205	1.000000	360.126489	0.004243
	Gender × migration background	108	20.653331	1.000000	13.456209	1.000000	281.248106	0.003313
	Gender × absolved education	324	7.692929	1.000000	8.568989	1.000000	271.231784	0.003195
	Migration background × absolved education	486	19.424479	1.000000	12.676304	1.000000	258.486825	0.003045
	Age group × gender × migration background	4428	1574.162354	1.000000	157.970713	1.000000	487.830608	0.005747
	Age group × gender × absolved education	13,284	933.428873	1.000000	130.189879	1.000000	443.881705	0.005230
	Age group × migration background × absolved education	19,926	1093.462434	1.000000	113.877827	1.000000	416.662545	0.004909
	Gender × migration background × absolved education	972	29.108628	1.000000	21.881322	1.000000	305.042419	0.003594
	Age group × gender × migration background × absolved education	39,852	2194.267693	1.000000	167.084847	1.000000	527.015116	0.006209
Car license		2	0.000165	0.999917	0.000159	0.999921	3.595921	0.000042
	Age	26	0.005861	1.000000	0.006811	1.000000	14.370524	0.000169

Table 2 (continued)

		DoF	Z^2		X^2		Absolute error	
			Score	p value	Score	p value	Total	Standardized
Motor cycle license		2	0.002777	0.998612	0.001954	0.999024	6.370590	0.000075
	Age	26	0.875771	1.000000	1.484794	1.000000	15.663175	0.000185
Moped license		2	0.000220	0.999890	0.000196	0.999902	3.975836	0.000047
	Age	26	0.003083	1.000000	0.010500	1.000000	13.535124	0.000159
	Car license	4	0.002277	0.999999	0.003920	0.999998	3.975836	0.000047
	Age × car license	52	0.027725	1.000000	0.075585	1.000000	13.535124	0.000159
Household position		21	1.803299	1.000000	1.847752	1.000000	250.398797	0.002950
	Neighborhood	42	1.802755	1.000000	2.501504	1.000000	146.000000	0.001720
	Age group	420	249.202608	1.000000	126.578850	1.000000	754.259661	0.008886
	Gender	42	2.699896	1.000000	2.918825	1.000000	336.681633	0.003967
	Age group × gender	840	334.941882	1.000000	182.929638	1.000000	896.190081	0.010558

The first column shows the target attribute, the second column the conditional attributes (in the empty case this is the correspondence of the marginal of the target attribute, in the case of 'neighborhood' this is the correspondence in each individual neighborhood

Author contributions J.M., T.S and M.P wrote the main manuscript text. All authors contributed various paragraphs and have made improvements to various parts of the manuscript. T.S. designed and implemented the original algorithm described. M.P., J.M., M.D. and B.L. contributed to the design of the second iteration. A.J. and M.P. implemented the second iteration of the algorithm and described and prepared the figures. A.J. has performed the evaluation. All authors reviewed the manuscript.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Michailidis, D., Tasnim, M., Ghebreab, S., & Santos, F. P. (2024). Tackling school segregation with transportation network interventions: An agent-based modelling approach. *Autonomous Agents and Multi-agent Systems, 38*, 1–22. https://doi.org/10.1007/s10458-024-09652-x
2. Parikh, N., Hayatnagarkar, H. G., Beckman, R. J., Marathe, M. V., & Swarup, S. (2016). A comparison of multiple behavior models in a simulation of the aftermath of an improvised nuclear detonation. *Autonomous Agents and Multi-agent Systems, 30*, 1148–1174.
3. Sonnenschein, T., Scheider, S., de Wit, G. A., Tonne, C. C., & Vermeulen, R. (2022). Agent-based modeling of urban exposome interventions: Prospects, model architectures, and methodological challenges. *Exposome, 2*, 1–26. https://doi.org/10.1093/exposome/osac009/6754814
4. de Mooij, J., et al. (2023). A framework for modeling human behavior in large-scale agent-based epidemic simulations. *Simulation, 99*, 1183–1211.
5. Ozik, J., Wozniak, J. M., Collier, N., Macal, C. M., & Binois, M. (2021). A population data-driven workflow for COVID-19 modeling and learning. *The International Journal of High Performance Computing Applications, 35*(5), 483–499.
6. Bissett, K. R., Cadena, J., Khan, M., & Kuhlman, C. J. (2021). Agent-based computational epidemiological modeling. *Journal of the Indian Institute of Science, 101*, 303–327.
7. Ferguson, N. M., et al. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. *Imperial College London, 24*, 456. https://doi.org/10.25561/77482
8. Dignum, F. (2021). *Social simulation for a crisis*. Cham: Springer.
9. Gaudou, B., et al. (2020). Comokit: A modeling kit to understand, analyze, and compare the impacts of mitigation policies against the COVID-19 epidemic at the scale of a city. *Frontiers in Public Health, 8*, 587. https://doi.org/10.3389/fpubh.2020.563247
10. Basu, R., et al. (2018). Automated mobility-on-demand vs. mass transit: A multi-modal activity-driven agent-based simulation approach. *Transportation Research Record, 2672*, 608–618.
11. Martinez, L. M., & Viegas, J. M. (2017). Assessing the impacts of deploying a shared self-driving urban mobility system: An agent-based model applied to the city of Lisbon, Portugal. *International Journal of Transportation Science and Technology, 6*, 13–27.
12. Barrett, C. et al. (2013). Planning and response in the aftermath of a large crisis: An agent-based informatics framework. In R. Pasupathy, S. -H. Kim, & A. Tolk (Eds.), *Proceedings of the 2013 winter simulation conference: Simulation: Making decisions in a complex world, WSC '13* (pp. 1515–1526). IEEE Press.
13. Lewis, B., et al. (2013). A simulation environment for the dynamic evaluation of disaster preparedness policies and interventions. *Journal of Public Health Management and Practice: JPHMP, 19*, S42.
14. Barrett, C. L., Bisset, K. R., Eubank, S. G., Feng, X. & Marathe, M. V. (2008). Episimdemics: An efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In IEEE Staff Corporate Author (Ed.), *Proceedings of the 2008 ACM/IEEE conference on supercomputing, SC '08* (pp. 37:1–37:12). IEEE Press. https://dl.acm.org/doi/10.5555/1413370.1413408

15. Adiga, A. et al. (2015). Generating a synthetic population of the united states. In Technical Report, network dynamics and simulation science laboratory. https://arifuzzaman.faculty.unlv.edu/paper/synth_popu15.pdf
16. Namazi-Rad, M.-R., Mokhtarian, P., & Perez, P. (2014). Generating a dynamic synthetic population–using an age-structured two-sex model for household dynamics. *PLOS ONE, 9*, 1–16. https://doi.org/10.1371/journal.pone.0094761
17. Yameogo, B. F., Vandanjon, P.-O., Gastineau, P., & Hankach, P. (2021). Generating a two-layered synthetic population for French municipalities: Results and evaluation of four synthetic reconstruction methods. *Journal of Artificial Societies and Social Simulation, 24*, 5.
18. Barthelemy, J., & Toint, P. L. (2013). Synthetic population generation without a sample. *Transportation Science, 47*, 266–279. https://doi.org/10.1287/trsc.1120.0408
19. Gargiulo, F., Ternes, S., Huet, S., & Deffuant, G. (2010). An iterative approach for generating statistically realistic populations of households. *PLOS ONE, 5*, 1–9. https://doi.org/10.1371/journal.pone.0008828
20. Lenormand, M., & Deffuant, G. (2013). Generating a synthetic population of individuals in households: Sample-free vs sample-based methods. *Journal of Artificial Societies and Social Simulation, 16*, 12.
21. Harland, K., Heppenstall, A., Smith, D., & Birkin, M. (2012). Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation, 15*, 1–24.
22. Chapuis, K., & Taillandier, P. (2019). A brief review of synthetic population generation practices in agent-based social simulation. In *Social simulation conference*.
23. Sonnenschein, T. (2023) TabeaSonnenschein/GenSynthPop: R-package for generating representative spatially explicit synthetic populations, v1.0.0. https://doi.org/10.5281/zenodo.7582109
24. de Mooij, J. et al. (2024). Gensynthpop-python, v2.0.1. https://doi.org/10.5281/zenodo.12200893
25. Hörl, S., & Balac, M. (2021). Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. *Transportation Research Part C: Emerging Technologies, 130*, 103291.
26. Hajduk, P., Roncoli, C. & Pihlatie, M. Lusikka, T. (2020). Data-based synthetic population generator for activity based transport models. In T. Lusikka (Ed.), *Proceedings of TRA2020, the 8th transport research Arena, no. 7 in Traficom research reports* (pp. 58–59). Liikenne- ja viestintävirasto Traficom, Finland.
27. Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics, 11*, 427–444.
28. Lin, Y., & Xiao, N. (2023). Generating small areal synthetic microdata from public aggregated data using an optimization method. *The Professional Geographer, 75*, 1–11.
29. Ireland, C. T., & Kullback, S. (1968). Contingency tables with given marginals. *Biometrika, 55*, 179–188.
30. Guo, J. Y., & Bhat, C. R. (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record, 2014*, 92–101.
31. Chapuis, K., Taillandier, P., Gaudou, B., Amblard, F., Thiriot, S., Ahrweiler, P., & Neumann, M. (2021). Gen*: An integrated tool for realistic agent population synthesis. In P. Ahrweiler & M. Neumann (Eds.), *Advances in Social Simulation* (pp. 189–200). Cham: Springer.
32. Ye, X., Konduri, K. C., Pendyala, R. M., Sana, B. & Waddell, P. (2009). Methodology to match distributions of both household and person attributes in generation of synthetic populations. In Transportation Research Board (Eds.), *88th Annual meeting of the transportation research board*, Washington, DC, USA. https://trid.trb.org/view/881554
33. Fosset, P., et al. (2016). Exploring intra-urban accessibility and impacts of pollution policies with an agent-based simulation platform: Gamirod. *Systems, 4*, 5.
34. Guo, J. Y., & Bhat, C. R. (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record, 2014*, 92–101.
35. Central Bureau of Statistics. (2023). Online portal. https://www.cbs.nl/en-gb
36. CBS kerncijfers wijken en buurten 2019. https://www.cbs.nl/nl-nl/cijfers/detail/84583NED
37. CBS bevolking op 1 januari en gemiddeld; geslacht, leeftijd en regio. https://www.cbs.nl/nl-nl/cijfers/detail/03759ned
38. CBS bevolking; migratieachtergrond, generatie, lft, regio, 1 jan; 2010–2022. https://www.cbs.nl/nl-nl/cijfers/detail/84910NED
39. Bevolking; hoogst behaald onderwijsniveau en herkomst. https://opendata.cbs.nl/CBS/nl/dataset/85453NED/table?dl=9EDBE
40. (cbs) (speciaal) basisonderwijs en speciale scholen; leerlingen, schoolregio. https://opendata.cbs.nl/statline/CBS/nl/dataset/71478NED/table?dl=9E57F
41. CBS leerlingen en studenten; onderwijssoort, woonregio. https://opendata.cbs.nl/CBS/nl/dataset/71450NED/table?dl=9E581
42. CBS huishoudens in bezit van auto of motor; huishoudkenmerken, 2010-2015. https://www.cbs.nl/nl-nl/cijfers/detail/81845NED

43. CBS huishoudens; personen naar geslacht, leeftijd en regio, 1 januari. https://opendata.cbs.nl/CBS/nl/dataset/71488ned/table?dl=9D241
44. CBS huishoudens; samenstelling, grootte, regio, 1 januari. https://opendata.cbs.nl/CBS/nl/dataset/71486ned/table?dl=A68AA
45. CBS marriages and partnership registrations; key figures. https://www.cbs.nl/nl-nl/cijfers/detail/37772eng
46. CBS groom usually older than bride. https://www.cbs.nl/en-gb/news/2019/07/groom-usually-older-than-bride
47. CBS geboorte; kerncijfers vruchtbaarheid, leeftijd moeder, regio. https://opendata.cbs.nl/#/CBS/nl/dataset/37201/table?dl=A68B5
48. CBS Inkomen van huishoudens; huishoudenskenmerken, regio, 2021. https://www.cbs.nl/nl-nl/cijfers/detail/85064NED
49. Voas, D., & Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling, 5*, 177–200. https://doi.org/10.1080/13615930120086078
50. Huang, Z., & Williamson, P. (2001). *A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata*. Liverpool: Department of Geography, University of Liverpool.

Authors and Affiliations

Jan de Mooij[1] · Tabea Sonnenschein[2,3,4] · Marco Pellegrino[1] · Mehdi Dastani[1] · Dick Ettema[3] · Brian Logan[1,5] · Judith A. Verstegen[3]

✉ Jan de Mooij
a.j.demooij@uu.nl

Tabea Sonnenschein
t.s.sonnenschein@uu.nl

Marco Pellegrino
m.pellegrino@uu.nl

Mehdi Dastani
m.m.dastani@uu.nl

Dick Ettema
d.f.ettema@uu.nl

Brian Logan
b.s.logan@uu.nl

Judith A. Verstegen
j.a.verstegen@uu.nl

[1] Intelligent Systems, Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands

[2] Exposome and Planetary Health, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands

[3] Department of Human Geography and Spatial Planning, Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, The Netherlands

[4] Institute of Risk Assessment Sciences, Utrecht University, Yalelaan 2, 3584 CM Utrecht, The Netherlands

[5] The School of Natural and Computing Sciences, University of Aberdeen, 32 Elphinstone Rd, Aberdeen AB24 3EU, UK