GenSynthPop: Generating a Spatially Explicit Synthetic Population of Agents and Households from Aggregated Data

Marco Pellegrino
Intelligent Systems, Information and Computing Sciences, Utrecht University
Jan de Mooij (a.j.demooij@uu.nl)
Intelligent Systems, Information and Computing Sciences, Utrecht University
Tabea Sonnenschein
Department of Human Geography and Spatial Planning, Utrecht University
Mehdi Dastani
Intelligent Systems, Information and Computing Sciences, Utrecht University
Dick Ettema
Department of Human Geography and Spatial Planning, Utrecht University
Brian Logan
Intelligent Systems, Information and Computing Sciences, Utrecht University
Judith A. Verstegen
Department of Human Geography and Spatial Planning, Utrecht University

Research Article

Keywords: Synthetic Population, Spatial Heterogeneity, Sample-Free

Posted Date: October 9th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3405645/v1

Additional Declarations: No competing interests reported.

GenSynthPop: Generating a Spatially Explicit Synthetic Population of Agents and Households from Aggregated Data

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

[1*]Intelligent Systems, Information and Computing Sciences, Utrecht University, Princetonplein 5, Utrecht, 3584 CC, The Netherlands.
[2]Department of Human Geography and Spatial Planning, Utrecht University, Heidelberglaan 8, Utrecht, 3584 CS, The Netherlands.
[3]Global Public Health & Bioethics, Julius Centrum, University Medical Center Utrecht, Universiteitsweg 100, Utrecht, 3584 CG, The Netherlands.
[3]The School of Natural and Computing Sciences, University of Aberdeen, 32 Elphinstone Rd, Aberdeen, AB24 3EU, United Kingdom.

*Corresponding author(s). E-mail(s): m.pellegrino@uu.nl; a.j.demooij@uu.nl;
Contributing authors: t.s.sonnenschein@uu.nl; m.m.dastani@uu.nl; d.f.ettema@uu.nl; b.s.logan@uu.nl; j.a.verstegen@uu.nl;

Abstract

Synthetic populations are microscopic representations of actual citizens living in a specific area. They play an increasingly important role in studying and modeling citizens 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 agents or households from these. 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.

Keywords: Synthetic Population, Spatial Heterogeneity, Sample-Free

1 Introduction

A synthetic population is a representation of citizens living in a specific area that fits the spatial, socio-economic and demographic characteristics of a real-world population while also maintaining privacy, as no single entity in synthetic population represents a true citizen. 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 citizen-centric issues. Using synthetic populations in agent-based social simulations allows modeling each citizen as unique software agents, whose autonomous decisions in a complex social environment together can lead to emergent patterns that researchers and policy makers can use to study the effects of complex sociological dynamics.

If the behavior of individual agents in such simulation studies is to be guided 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. In reality, these demographics are highly heterogenic, both in terms of their values, and in terms of their geographic spread throughout the population. Moreover, if a change of demographics over time is to be studied, an accurate starting point becomes all the more important.

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 [1, 2], disaster management [3, 4] and epidemiology [5, 6].

The traditional approach of synthesizing such a population requires a detailed micro data 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 [7–9]. Synthetic citizens or households are then drawn from the micro data sample using the weights given by the estimated joint distribution. However, micro data may not always be available or affordable, which led to a new class of approaches often referred to as *sample-free* [10, 11]. The new challenge then is combining the aggregated data from multiple sources and levels of aggregation. Synthetic citizens or households are then directly drawn, one by one, from that distribution instead of from a sample. Both approaches can accurately match the joint frequencies of either individual level attributes or household level attributes, but struggle to match both.

This paper is the result of a data-driven and large-scale agent-based social simulation 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 heterogenic across the population. For the target area, no micro data 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 were in need of a method where new attributes could be added to an existing synthetic population without having to start over almost completely.

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 an R-package called *GenSynthPop* [12] to facilitate the data preparation and method implementation. The approach differs from existing approaches in that the citizens and households in the synthetic population directly represent the estimate of the true joint distribution, instead of being drawn from it. This is achieved by starting with a homogeneous population of citizens 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. Previously, we have used a probabilistic propensity to determine the attribute value for each citizen one by one, which caused the same skewing effects as random drawing. In a revised version instead, each group of candidate citizens 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. In the remainder of this paper, we describe both versions of *GenSynthPop* without disambiguating the two. 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 [13, 14], 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. In Section 3, we first present our methodology in general terms, before comparing a case study population generated using our method to known distributions in Section 4. We conclude our work in Section 5.

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 micro data (specifically, the US Public Use Microdata Sample, or PUMS) weighted by an estimate of the joint distribution of all relevant attributes occurring in that micro-data. 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) [15] to update each cell slightly such that the distributions margins move towards those reported in the aggregated data for that area. This is an iterative procedure since it is repeated until the relative changes are smaller than some predefined stopping criterion.

Versions of this approach have since been widely applied (see e.g. [7–9, 16]). However, Guo and Bhat [17] observed that this approach 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 violate 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 micro data have also been developed, perhaps best known of which is Gen* [18].

Building on their work, Ye *et al.* [19] 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 micro-data 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 good enough.

Adiga *et al.* [7] 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* that 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 citizens co-locating in the same location any 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 micro data sample may not be available or affordable for every target region. While some authors have used surveys in their place [20], others have moved to a new class of approaches often referred to as *sample-free*. Gargiulo *et al.* [11], 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 than 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 citizens are taken (without replacement) from the generated set of individuals to match the corresponding attribute values of the selected household. However, 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, finding a suitable individual from the candidates only becomes harder when more person-level attributes are also considered, as this only further constraints the selection. Barthelemy and Toint [10], in contrast, generate individuals described by richer attributes which are drawn from known distributions.

In their approach, they first attempt to combine the various data sets into a single joint distribution, taking into account the level of spatial detail each data set provides. Like Gargiulo *et al.*, they then draw a household from the known household types and try to populate it with the previously created individuals. Lenormand *et al.* [21] have conducted a study to compare the accuracy of partitioning a synthetic population of individuals into households using the sample-free approach proposed by Barthelemy *et al.* [10] and the sample-based approach proposed by Ye *et. al* [19] and have concluded that while both approaches require similar amounts of compute power, the sample-free approach globally resulted in better matches, and has the additional advantage that it can be applied where no micro data sample is available.

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. The challenge for all approaches therefor is how to combine the available data into a single estimate of the true joint distribution. 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 estimating this distribution and then iteratively drawing individuals or households from it can be merged into one for the sample-free approaches, as no micro-data to draw from is available to start with.

3 Generating a Synthetic Population

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

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

Each such synthetic citizen $A_i = \langle v_1, \ldots, v_m \rangle$ is characterized as a vector of values for the m different attributes. What attributes are included depends on the design criteria of the study and the availability of data. Common attributes that are generally included are such attributes as *age*, *gender*, *education* and *income*. Spatial heterogeneity is achieved through the inclusion of spatial location in the citizens' attributes.

A synthetic population may optionally be augmented with synthetic households. Such a household $H = \langle \{A_i, \ldots, A_k\}, \langle v_1, \ldots, v_n \rangle \rangle$ too is a vector, where $A_i, \ldots, A_k \in S$ denote its constituent members, and $\langle v_1, \ldots, v_n \rangle$ is a vector of values for n 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 citizens, but the partitioning of citizens into households requires an additional step.

In this section, we detail our data-driven and sample-free approach for constructing the synthetic population of citizens 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.

The process starts by instantiating and locating the appropriate number of synthetic citizens in the region by creating n vectors of size $m = 1$ with the value of the single attribute representing the citizens location. Attributes V_{m+1} are then iteratively added to the synthetic population by assigning each citizen a value for the attribute $v_{m+1} \in V_{m+1}$ conditioned on the attributes $V_1, \ldots, V_m$ previously assigned.

Citizens 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 π that are published as k-way *contingency tables* with $k \geq 1$ categorical attributes $\mathcal{V}(\pi) = \{V_1, \ldots, V_k\}$. Such data sets provide the counts $\pi_{i,\ldots,k}$ for each combination of the levels, where i represents the value of the ith category, at the cells representing the intersections of those levels.

In other words, these data sets provide for each combination of attribute values the number of citizens within the population to which they apply. An example of a (2-way) contingency table is given in the top left of Figure 1. These 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 citizens) to which each combination of attributes applies. In this case, the value at the intersection of the levels is denoted $P(V_1 = v_1, \ldots, V_k = v_k)$ to denote the frequency of the combination of levels v_1 for the variable V_1 with the level v_2 for the variable V_2, etc. An example of these types of data sets is given in the bottom left of Figure 1. 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 levels of a single attribute are given. Figure 1 shows how joint (divide cell value by total number of citizens) and marginal distributions (sum row or column, depending on target attribute) can be derived from the contingency table. The contingency table can be reconstructed from the joint distribution if the total number of citizens is known, but neither the joint distribution nor the contingency table can be reconstructed from the marginal data, because inter-attribute dependencies are lost.

Note that these attributes are all assumed to be categorical. While some of these categories describe integer or real numbers (e.g., age, household income), these, too, are usually grouped into categories, such as the *age group* that is used in Figure 1. This is done both to maintain privacy and to keep data set sizes manageable. Of course, citizens or households can be assigned integer or real values from the assigned level.

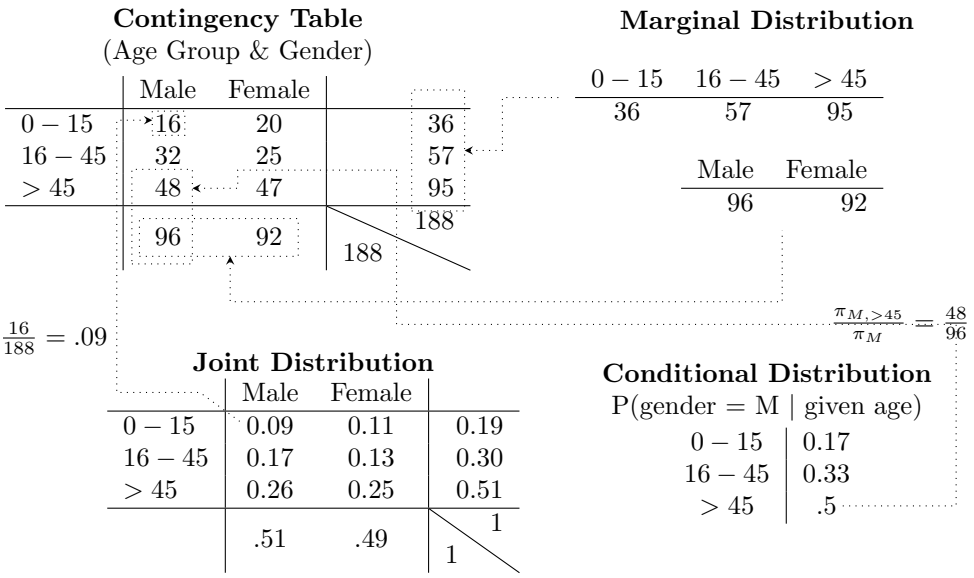

Fig. 1: Illustration of the relation between the contingency table, joint distribution and marginal data on a fictional population of 188 individuals.

A *conditional distribution* can be derived from both the contingency table and the joint distribution[1]. The conditional distribution gives the probability or frequency for each of the levels of a single variable given the presence of specific levels of one or more other variables. This conditional probability for each of the levels v_l of the variable V_l is given by equation 1 and illustrated for the conditional probability of the male gender giving the three age groups in the bottom right of Figure 1.

$$P(V_l = v_l | V_1 = v_1, \ldots, V_k = v_k) = \frac{P(V_1 = v_1, \ldots, V_k = v_k, V_l = v_l)}{P(V_1 = v_1, \ldots, V_k = v_k)} = \frac{\pi_{v_1,\ldots,v_k,v_l}}{\pi_{v_1,\ldots,v_k}} \tag{1}$$

We further assume that spatial information is encoded in these data sets. This may be through one of the categorical values, 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

[1] In the remainder, contingency table and joint distribution will be used interchangeably unless otherwise indicated, as either are only used to derive this conditional distribution.

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, in which case selection of a suitable data set is up to the creator of the synthetic population.

In general, our (iterative) methodology proposes to deal with the data sets one at the time. The sole exception is when there are only marginal data present for each of the smaller regions within the area of study, but at a higher level of aggregation, a data set with multiple attributes is available. In that case, we use the Iterative Proportional Fitting (IPF) procedure to create a new data set for each smaller region in which the joint distribution at the higher level of aggregation is fitted to the margins of that lower level of aggregation. 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.

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 and for each region in that level of aggregating determining the required number of citizens (e.g., the absolute number of individuals living there, but requirements differ per study) and creating a matching number of empty citizen vectors. The first attribute always is that region or a specific point within that region in which the citizen 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 citizens (which also applies to adding attributes to synthetic households) and give a more formal approach 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 with the attributes *age*, *gender* and *migration background* already added, to which we want to add the target attribute *education*. Consider we have found a contingency table or joint distribution containing the categorical values *education*, *age*, *gender* and *migration background*.

We first find all attributes that have previously been added to the synthetic population and also appear in the new data set – in our example *age* and *gender* – and derive the *conditional* distribution (Equation 1) of the target variable (*education*) conditioned on those other variables (*age* and *gender*). We then split the citizens into groups along the levels of those same co-occurring 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 we then use the derived conditional distribution to find the relative frequency of each of the levels of the target attribute, and further divide the members into subgroups whose relative sizes match the relative frequencies. For example, we may find that the data set specifies that 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 levels *low*, *middle* and *high* to those groups, respectively.

Note that even if the groups are very small, e.g., just 10 citizens, 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 levels, i.e., the example above would fail with a group of just 3 members). Note further that when we arbitrarily distribute the citizens within the age $20-30$ and gender *female* group across the three levels, it does not matter what citizen is placed in what group because, as far as all the previously included information can tell us, all citizens are functionally identical.

Algorithm 1 An algorithm to assign the levels of a new target attribute V_{m+1} to the existing synthetic population S conditioned on the distribution specified in a data set π

Require: Data set π with categorical attributes $\mathcal{V}(\pi)$ and s.t., $V_{m+1} \in \mathcal{V}(\pi)$
Require: A synthetic population S with categorical attributes $\mathcal{V}$
1: **procedure** ASSIGN(π, S)
2: $V_i, \ldots, V_k \leftarrow \mathcal{V} \cap \mathcal{V}(\pi)$ ▷ Attributes that appear in both synthetic population and target data set
3: **for all** $v_i, \ldots, v_k \in V_i \times \cdots \times V_k$ **do**
4: $A \leftarrow \mathcal{A}_{v_i,\ldots,v_k}$
5: $a \leftarrow |A|$ ▷ Number of citizens in this group
6: **for all** $v_{m+1} \in V_{m+1}$ **do**
7: $f \leftarrow P(V_{m+1} = v_{m+1} \mid V_i = v_i, \ldots, V_k = v_k)$
8: $A' \leftarrow \binom{A}{a \cdot f}$ ▷ Choose fraction corresponding to f
9: **for all** $\langle V_1 = v_1, \ldots, V_m = v_m \rangle = A_i \in A'$ **do**
10: $A_i \leftarrow \langle V_1 = v_1, \ldots, V_m = v_m, V_{m+1} = v_{m+1} \rangle$
11: **end for**
12: $A \leftarrow A \setminus A'$ ▷ Avoid assigning more than once
13: **end for**
14: **end for**
15: **end procedure**

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

Some (possibly empty) set of attributes in the new data set π will already have been added to the synthetic population previously, denoted $\mathcal{V} \cap \mathcal{V}(\pi)$. For each possible combination of the levels of those intersecting attributes (line 3), we identify all synthetic citizens that are characterized by exactly that combination (line 4). We then iterate over all levels of the target attribute V_{m+1} (line 6) and calculate the conditional relative frequency of that attribute level using Equation 1 (line 7). We then, from the identified individuals, take a fraction corresponding to the relative frequency of the attribute level, and assign them that level (lines 8-11). Finally, we remove the citizens

that have already been assigned a new level before restarting from line 6 to ensure each citizen is assigned exactly one level (Line 12). 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 (Section 2) to struggle finding individuals that meet the household criteria before all citizens are associated with a household. Instead, we propose to procedurally determine the number of households required in such a way that all citizens can be placed, while maintaining only the relative frequencies – instead of absolute counts – from the known distributions.

We provide here a general approach for determining the number of households required to house the known number of citizens such that household members follow real inter-household distributions such as for gender and age. We first determine the number of households that have children and then determine what number of those household are one and two-parent households respectively. Next, we determine the number of two-person households without children. Finally, what is left of the population will inform the number of one-person and many-person households without children. We note that our approach for creating the synthetic citizens starts with the required number of citizens. 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.

The first step is to determine the number of households with children. We have previously classified n of the synthetic citizens as children (e.g., using an age threshold, data specifying the number of frequency of children, or data specifying the number of citizens living with at least one parent), and we assume a data set exists that provides the relative frequency distribution $P(C = c)$ of the number of children per household (or absolute numbers, from which $P(C = c)$ can be derived). From this frequency distribution we then obtain a probability distribution $P'(C = c)$ that any child lives in a c-children household, by weighing the frequency of each household size with the number of children in that household size, and normalizing again to the range $[0, 1]$, as per Equation 2:

$$P'(C = c) = \frac{P(C = c) \cdot c}{\sum_{c' \in C} P(C = c') \cdot c'} \tag{2}$$

Then, to accommodate the n synthetic children in total, the number of households $h(c)$ with c children is given in Equation 3:

$$h(c) = \lfloor \frac{n \cdot P'(C = c)}{c} \rfloor \tag{3}$$

Now the children could arbitrarily placed in the households following $P'(C = c)$. If data availability permits, however, the number of children in a household can be considered an attribute, and a primary child can be assigned to each household (i.e., be assigned one of the levels of this attribute), conditioned on other attributes such as their age, gender, etc (Algorithm 1). The remaining children can then be assigned any of the multi-child households based on e.g. age and gender disparity with the first child.

The second step consists of assigning parents or caregivers to the household. The fraction of the already created households that have only one versus two parents is again informed by the data. In both cases, each household is assigned a single parent conditioned on age disparity with the oldest child. All two-parent households are then assigned a partner conditioned on age and gender disparity between partners.

In the third step, the number of two-person households is determined, either from absolute numbers, or as a proportion to the number of households with children. The first household member is selected conditioned on age distributions within this group of two-person households and the second person again on age and gender disparity.

Lastly, all remaining citizens are either each placed in a single-person household, or, if data availability permits, a group household (e.g., communal living, student living, etc.).

With that, the partitioning of citizens into households is complete, and each household now looks like a tuple $H = \langle \{A_1, \ldots, A_n\}, \langle\rangle \rangle$, where the second element is a – for now – empty vector. This vector is extended one attribute at a time using Algorithm 1, where new attributes can be conditioned on existing attributes in that vector, as well as on attributes present in any of the citizen vectors 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 Section 3 applied to the the Zuid-West (South-West) district of The Hague in The Netherlands[2]. The Netherlands allows for a perfect case study for this approach, as detailed and reliable statistics are published by Statistics Netherlands (CBS) [22] (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[3] [23]. 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

[2] Available as Open Source at https://www.github.com/marcopellegrinoit/DHZW_synthetic-population

[3] 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

provided (marginally or jointly) on the level of the entire city of The Hague. This is a larger area than that of our study, but still considered sufficiently representative lacking more fine-grained data.

4.1 Citizen Generation

The process of generating the synthetic population was started by creating, for each neighborhood, a number of vectors matching the reported number of individuals living in that neighborhood, and adding that neighborhood to those vectors as the first value (thus locating those synthetic citizens). Next, we identified the first attributes to be added. As most available joint data seemed to be conditioned on 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 [23], 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 [24] (at the municipality level). Gender is only available marginally at the neighborhood level [23] or jointed over age at the municipality level [24], so the IPF procedure was used to fit a joint distribution of age and gender to the marginal gender data of each neighborhood. Gender (binary) was then assigned, conditioned on age, to the citizens according to the proportions given by the fitted data. The same procedure was then applied for migration background ('Dutch', 'Western' or 'Non-Western'), but using a different data set conditioned on both age group and gender [25] (at municipality level). For *current education* ('low', 'middle', 'high' or none) no marginal data at the neighborhood level is available, so this attribute was added directly from the joint distribution at the municipality level [26] conditioned on all three previously added attributes (but not on location). *Education attainment* (using the same categories) was then added from marginal data on the neighborhood level [26] using manually engineered constraints based on current education (typically one level lower) and age (e.g., no education attainment for individuals younger than 9 years) if citizens are currently enrolled in education, and using the established approach for the rest. Finally *car license ownership* and *moped license ownership* was added based on a distribution jointed with age [27] (at the province level) and (to determine if a citizen should be considered a child) *living with parents* was added based on a distribution jointed over both age and gender [28] (at the municipality level).

4.2 Household Partitioning

The synthetic citizens were then partitioned into households in line with Section 3.3. First marginal data on the level of the municipality [29] was used to find the relative frequencies of households with 1, 2 or 3 (or more) children and the number of households was determined such that all individuals marked as children in the synthetic population could be distributed across those households to match those frequencies (Equation 3). The children were then randomly placed into those households accordingly. Available inter-household child age distributions (at the municipality level) [30]

were not considered in this case study. Each of those households was then assigned a parent conditioned on the age difference with the oldest assigned child using marginal data on the level of the municipality [31]. The relative frequency of those households with two parents was found from marginal data on the municipality level [28] and that fraction of households was assigned a partner to the first parent conditioned on age and gender disparity between the both parents, based on marginal data at the municipality level [32]. From the same data set, the proportion of households with couples or singles *without* children was found. The number of each of those households to instantiate was extrapolated from the fraction *with* children compared to the actual number of households with children instantiated. As before, one individual was assigned to each household randomly, and couple households without children were then assigned a partner following the same gender and age disparity figures. Finally, remaining individuals are placed in their own single households.

The households were then annotated with *standardized income group* [33] (municipality level) conditioned on household composition using the same approach as for adding citizen attributes. The *car ownership* [27] 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

We evaluate the resulting synthetic population by comparing the relative frequencies of attribute values in our synthetic population to those reported by the published data. For this evaluation, we either compare the attribute margins in each neighborhood and largely ignore the effect of inter-attribute dependencies, or compare the joint frequencies on a larger area. Both should match the provided data well, but using both simultaneously would just reproduce our method without offering additional insight. We report the difference between these two frequencies multiplied by a hundred as the percentage points difference.

Safe for two of the 14 neighborhoods, the margins for gender matched exactly (Figure 2), i.e., the relative frequencies in the synthetic population matched the margins for each neighborhood. In the other two regions, there was a difference of 1.8 percentage points, with one region tending towards slightly more female and the other towards slightly more male citizens. Note that these are the two least populated neighborhoods with just 55 and 155 individuals respectively, compared to 15.000 individuals in the most densely populated neighborhood and about 6000 individuals in each neighborhood on average.

An even better match was obtained for migration background (third box plot in Figure 3) and age group (not plotted), where there was no difference in any of the neighborhoods. For the latter, this is expected, since this was the first attribute that was added to the synthetic population directly from those known margins, and not conditioned on anything else. For migration background, these margins were also used, but conditioned on both age and gender. To further dissect this result, we have further split the citizens by age group and gender and for each neighborhood compared

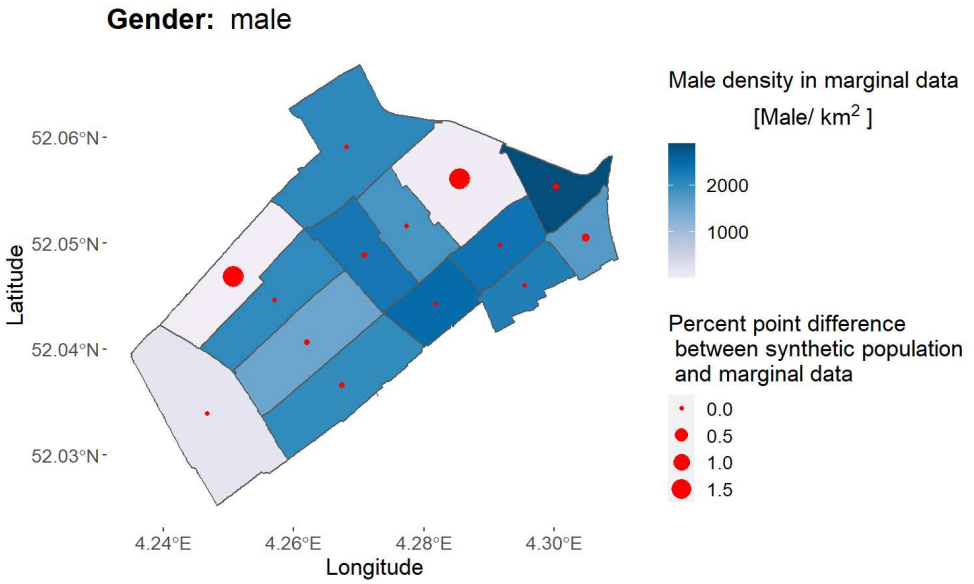

Fig. 2: A plot of the population density (darker means denser) in the 14 neighborhoods together with the percentage difference between the proportion of male citizens in the synthetic population and the target marginal distributions.

the frequencies of those groups to the known frequencies in the joint distribution of the entire region (the lowest level of aggregation available) as well (last box plot in Figure 3). Here, the percentage point differences become larger. However, this is a slightly disingenuous comparison: while the conditional frequency distribution of migration background over age group and gender was only available for this larger area, the assignment of this attribute was conditioned on the age group and gender that had already been added to the synthetic population, and for which data *was* available for each individual neighborhood. In other words, spatial heterogeneity was introduced based on all the available information, which is not reflected in the data set we compare against. Note that, even if though the relative conditional frequency of migration background is maintained within each neighborhood, the relative frequency of the neighborhoods combined is not necessarily maintained if the conditional frequency distribution of age group and gender in the individual neighborhoods does not match the distribution in the larger data set. In other words, in this case, we had various distributions which included age group and gender, at different levels of aggregation, which were not congruent with each other, resulting in these larger differences.

A similar effect (in a similarly disingenuous comparison) can be seen for education attainment (first box plot in Figure 3). In this case, we compare the margins of each of the 14 individual neighborhoods to a fixed set of margins for the entire municipality, which fails to reflect possible spatial heterogeneity the same as above. For this

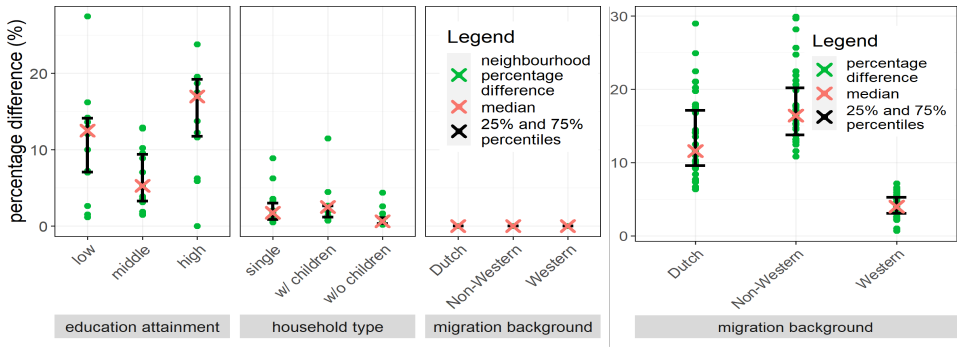

Fig. 3: Box plots of percentage difference in synthetic population and marginal data for the values of three attributes (left), repeated with the percentage difference in joint data for one of the attributes (right).

attribute, however, additional constraints were devised as well, all based on current education (which was available at the neighborhood level). In both cases, if the municipality margins were used without conditioning on other attributes or constraints, the frequencies in each neighborhood would be expected to match both each other, and the data from which they were derived. These results have been included despite their negative appearance to show that combining multiple attributes can introduce heterogeneity even for attributes for which that data is not available.

Lastly, ignoring outliers, the frequencies for household distribution matched the neighborhoods' margins within 5% (second box plot in Figure 3). This is perhaps the result most indicative of the performance of the proposed methodology, as these reported margins were not used for the synthesis at all, but the attribute can be expected to be correlated to other attributes which were in fact used. The numbers of households with each composition type in our synthetic population are a consequence of how many households of each type were required to fit all individuals in the synthetic population in such a way that the frequencies of number of children in each household matched the reported data. The relative frequencies of household types themselves were not considered, so any correspondence found here is a consequence of other correlated attributes affecting the household type distribution in a similar way in our synthetic population as in the real population that the reported data describes.

5 Conclusion

We have proposed a new methodology for generating a spatially heterogeneous synthetic population of citizens and households from aggregated data only, without a detailed micro data sample. The approach improves on existing methodology by combining the estimation of the joint distribution with the generation of citizens and households, which in previous sample-free approaches are drawn from the estimate of the joint distribution. This methodology comes with two main benefits. First, because the citizens are not drawn from the estimated joint distribution, additional attributes can be added to an existing synthetic population at any time, removing the need to

redraw all the citizens and households after the estimated joint distribution has been updated with the new attribute. This allows for both finer iteration when additional attributes are required for the target study, and for reuse, as others can extend the synthetic population with the required additional attributes while reusing work that has already been done. Secondly, while drawing agents randomly can skew distributions in smaller subgroups, the proposed methodology explicitly replicates known proportions between all dependent attributes in the synthetic population. We published an open-source R-package called GenSynthPop implementing the proposed approach.

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 but also show that errors in the data can propagate when conditioning on other flawed data. More generally, the methodology cannot overcome limitations of the source data. 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.

The case study is limited compared to the proposed methodology in two significant ways. First, the proposed method for using inter-household child age distributions has not yet been applied in the current case study, nor is migration background taken into account for agents sharing households, despite the availability of such data, and should be incorporated in future work. Secondly, in the case study, the combination of joint distributions at higher levels of aggregation with the marginal data of the lower levels of aggregation still employed the old propensity scoring method to assign each citizen an attribute value independently. In this case especially, we observed that relative conditional frequencies could be skewed in smaller group sizes. In the extended methodology, we have proposed first estimating the joint distribution for each of the regions using the Iterative Proportional Fitting procedure, and then jointly assign the attribute values to subgroups of citizens in exactly the same manner as we do when using existing joint distributions (Section 3.2). Where data sets could be used directly, this limitation does not apply. Thirdly, the household assignment method could consider additional variables for estimating the conditional propensity of individuals to be part of a specific household composition (e.g. income, level of education and migration background as determinants for having a child, being in a household of x size).

Finally, in future work, we intend to collaborate with institutions with detailed micro data of the target area to compare the accuracy of our results to those known distributions.

References

[1] Basu, R. *et al.* Automated mobility-on-demand vs. mass transit: a multi-modal activity-driven agent-based simulation approach. *Transportation Research Record* **2672**, 608–618 (2018).

[2] Martinez, L. M. & Viegas, J. M. 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 (2017).

[3] Barrett, C. *et al. Planning and response in the aftermath of a large crisis: An agent-based informatics framework*, 1515–1526 (IEEE, 2013).

[4] Lewis, B. *et al.* A simulation environment for the dynamic evaluation of disaster preparedness policies and interventions. *Journal of public health management and practice: JPHMP* **19**, S42 (2013).

[5] Ferguson, N. M. *et al.* Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand (2020).

[6] Barrett, C. L., Bisset, K. R., Eubank, S. G., Feng, X. & Marathe, M. V. *EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks*, 37:1–37:12 (2008).

[7] Adiga, A. *et al. Generating a synthetic population of the united states* (2015).

[8] Namazi-Rad, M.-R., Mokhtarian, P. & Perez, P. Generating a dynamic synthetic population – using an age-structured two-sex model for household dynamics. *PLOS ONE* **9**, 1–16 (2014). URL https://doi.org/10.1371/journal.pone.0094761.

[9] Yameogo, B. F., Vandanjon, P.-O., Gastineau, P. & Hankach, P. 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 (2021). URL http://jasss.soc.surrey.ac.uk/24/2/5.html.

[10] Barthelemy, J. & Toint, P. L. Synthetic population generation without a sample. *Transportation Science* **47**, 266–279 (2013). URL https://doi.org/10.1287/trsc.1120.0408.

[11] Gargiulo, F., Ternes, S., Huet, S. & Deffuant, G. An iterative approach for generating statistically realistic populations of households. *PLOS ONE* **5**, 1–9 (2010). URL https://doi.org/10.1371/journal.pone.0008828.

[12] Sonnenschein, T. TabeaSonnenschein/GenSynthPop: R-package for Generating Representative Spatially Explicit Synthetic Populations (2023). URL https://doi.org/10.5281/zenodo.7582110.

[13] Hörl, S. & Balac, M. 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 (2021). URL https://www.sciencedirect.com/science/article/pii/S0968090X21003016.

[14] Hajduk, P., Roncoli, C. & Pihlatie, M. *Data-based Synthetic Population Generator for Activity Based Transport Models*, 58–59. No. 7 in Traficom Research Reports (Liikenne- ja viestintävirasto Traficom, Finland, 2020). URL https://traconference.eu/overview/. 8th Transport Research Arena, TRA 2020 - Conference cancelled, TRA 2020 ; Conference date: 27-04-2020 Through 30-04-2020.

[15] Deming, W. E. & Stephan, F. F. 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 (1940).

[16] Lin, Y. & Xiao, N. Generating small areal synthetic microdata from public aggregated data using an optimization method. *The Professional Geographer* 1–11 (2023).

[17] Guo, J. Y. & Bhat, C. R. Population synthesis for microsimulating travel behavior. *Transportation Research Record* **2014**, 92–101 (2007).

[18] Chapuis, K., Taillandier, P., Gaudou, B., Amblard, F. & Thiriot, S. Ahrweiler, P. & Neumann, M. (eds) *Gen*: An Integrated Tool for Realistic Agent Population Synthesis*. (eds Ahrweiler, P. & Neumann, M.) *Advances in Social Simulation*, 189–200 (Springer International Publishing, Cham, 2021).

[19] 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* (2009).

[20] Fosset, P. *et al.* Exploring intra-urban accessibility and impacts of pollution policies with an agent-based simulation platform: Gamirod. *Systems* **4** (2016). URL https://www.mdpi.com/2079-8954/4/1/5.

[21] Lenormand, M. & Deffuant, G. Generating a synthetic population of individuals in households: Sample-free vs sample-based methods. *Journal of Artificial Societies and Social Simulation* **16**, 12 (2013). URL http://jasss.soc.surrey.ac.uk/16/4/12.html.

[22] Central Bureau of Statistics. Online portal (2023). URL https://www.cbs.nl/en-gb.

[23] CBS kerncijfers wijken en buurten 2019. URL https://www.cbs.nl/nl-nl/cijfers/detail/84583NED.

[24] CBS bevolking op 1 januari en gemiddeld; geslacht, leeftijd en regio. URL https://www.cbs.nl/nl-nl/cijfers/detail/03759ned.

[25] CBS bevolking; migratieachtergrond, generatie, lft, regio, 1 jan; 2010-2022. URL https://www.cbs.nl/nl-nl/cijfers/detail/84910NED.

[26] CBS leerlingen en studenten; onderwijssoort, woonregio. URL https://www.cbs.nl/nl-nl/cijfers/detail/71450ned.

[27] CBS huishoudens in bezit van auto of motor; huishoudkenmerken, 2010-2015. URL https://www.cbs.nl/nl-nl/cijfers/detail/81845NED.

[28] CBS huishoudens; personen naar geslacht, leeftijd en regio, 1 januari. URL https://www.cbs.nl/nl-nl/cijfers/detail/71488NED.

[29] CBS huishoudens; samenstelling, grootte, regio, 1 januari. URL https://www.cbs.nl/nl-nl/cijfers/detail/71486NED.

[30] CBS huishoudens; kindertal, leeftijdsklasse kind, regio, 1 januari. URL https://www.cbs.nl/nl-nl/cijfers/detail/71487NED.

[31] CBS geboorte; kerncijfers vruchtbaarheid, leeftijd moeder, regio. URL https://www.cbs.nl/nl-nl/cijfers/detail/37201.

[32] CBS marriages and partnership registrations; key figures. URL https://www.cbs.nl/nl-nl/cijfers/detail/37772eng.

[33] CBS inkomen van huishoudens; huishoudenskenmerken, regio (indeling 2021). URL https://www.cbs.nl/nl-nl/cijfers/detail/85064NED.