One-step simulator for synthetic households generation

Marija Kukić

Michel Bierlaire

STRC conference paper 2022 May 17, 2022

STRC 22nd Swiss Transport Research Conference
Monte Verità / Ascona, May 18-20, 2022

One-step simulator for synthetic households generation

Marija Kukić
Transport and Mobility Laboratory
Ecole Polytechnique Fédérale de Lausanne
Station 18, 1015 Lausanne
marija.kukic@epfl.ch

Michel Bierlaire
Transport and Mobility Laboratory
Ecole Polytechnique Fédérale de Lausanne
Station 18, 1015 Lausanne
michel.bierlaire@epfl.ch

May 17, 2022

Abstract

Transportation science today is tasked with predicting the complex mobility needs of individuals, which necessitates the use of advanced mobility and travel demand models. However, the quality of the model outputs depends on the data quality. In transport, data privacy and data availability are two limitations. Therefore, transportation scientists rely increasingly on the usage of synthetic populations. Typically, a synthetic population is generated either on the level of individuals or on the level of households using simulation or machine learning approaches. This paper presents a follow-up work on the existing simulation techniques for the generation of synthetic households, addressing several literature gaps. In existing methodologies, the generation of individuals and their matching into households is done separately, through two sequential processes. Although the marginal distributions of key generated attributes might show a perfect fit, the "two-step" household generator produces unrealistic households. The generation of illogical observations is caused by neglecting the dependencies between individuals while grouping them into households. In order to create realistic households, this paper suggests a "single-step" household simulator where relationships between individuals are considered simultaneously within the generation process by imposing various statistical constraints. However, as shown in the past, the simulation methods struggle to deliver accurate results in a reasonable time while dealing with high-dimensional datasets. We propose the so-called "Divide and Conquer Gibbs Sampler" that solves this problem by decomposing and parallelizing the generation process based on the level of correlation. This approach increases accuracy and efficiency, as highly correlated areas are isolated, enabling a better representation of less probable values. The case study compares the developed approach with state-of-the-art methodologies based on 2015 Swiss census data.

Keywords

Population synthesis, Markov Chain Monte Carlo simulation, Gibbs sampling, activity-based models

1 Introduction

Transportation science today is tasked with predicting the complex mobility needs of individuals, which necessitates the use of advanced mobility and travel demand models. The models for predicting activity and travel-related decisions of individuals are called Activity-Based Models (ABM). However, good models require good data to calibrate the model. Since the data quality affects the model's outputs, they are an essential input to ABM.

Highly-sensitive data such as the population census and travel activity information are extremely valuable in transportation science as they provide detailed insights into traveller behaviour. Such data are used to inform decision-makers or design accurate simulation models of travellers. However, the problem lies in the confidentiality and availability of such data. Traditional population census or travel survey datasets contain personal information about individuals and households. Nevertheless, the datasets do not fully represent the whole population, and the unprocessed data are not available due to privacy policies. Often they are either anonymized by removing attributes from the dataset or by extracting micro samples that represent a small subset of the entire data. Even though the emergence of Big Data collection services used in conjunction with hand-collected survey data is highly detailed and has extensive population coverage, data collection often comes with strict data usage restrictions. Additionally, those datasets are often sanitized before publication. This is a problem in transportation and travel activity modelling, as the model's quality in activity forecasting depends on the level of detail of the model descriptors.

In order to circumvent privacy and availability issues, synthetically generated data can be used. Synthetic data have similar statistical properties as the real population of interest. However, they do not allow the identification of individuals (to address the privacy issue) and compile all the necessary data for the scientific analysis, municipalities and other stakeholders who do not have access to raw data (to address the availability issue).

In the context of the synthetic population generation, the synthetic data can be at the level of individuals or households. The literature shows that synthetic population generators mostly support the generation of individual attributes only, which consequently affects the existing research efforts in ABM. The lack of household data might be the reason why activity-based modellers are usually focused on isolated individual behaviour analysis.

Without information on households, investigation of behavioural patterns is highly limited. Integrating household data into ABM methodologies would expand the model's capabilities to capture multi-individual decisions and understand mobility patterns by taking into account interactions and influence of the household .

Although several different methodologies exist for the accurately and efficiently generation of synthetic data, few gaps can be identified (see Section 2). In line with the challenges mentioned above and identified gaps, we try to improve the existing synthetic household generators by addressing the following research questions:

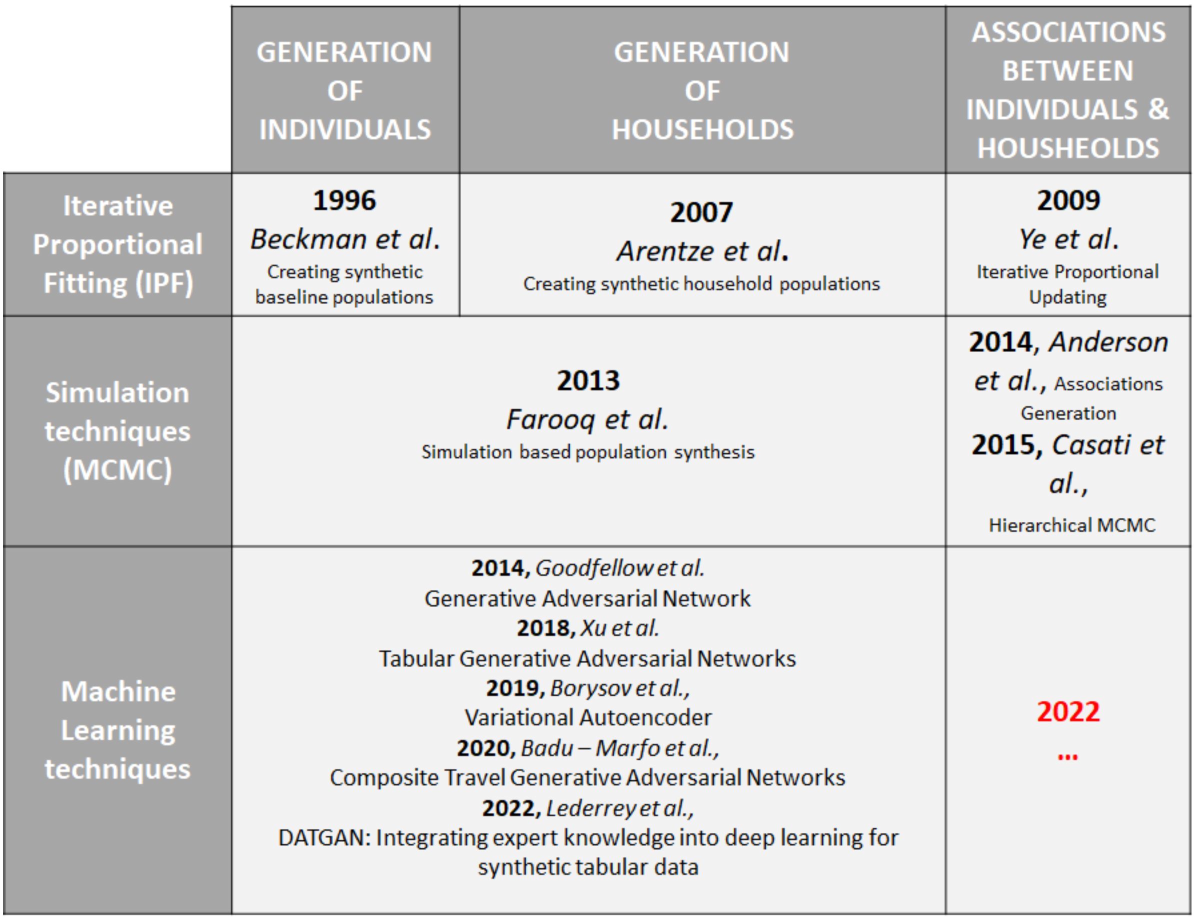
- How to develop a simulation framework for a synthetic household generation that integrates the individuals generation and their matching into households in a one-step simultaneous process?
- How much control can we embed into the generation process compared to other existing methodologies, in order to generate realistic households?
- How to redefine and improve the simulation approach to deal with the "curse of dimensionality"?

The document is organized as follows: Section 2 covers a detailed review of the previous research in this field; Section 3 introduces and formally specifies the methodology; and Section 4 presents obtained results; Section 5 summarizes the research contributions and specifies all the necessary steps for future improvements.

2 Literature review

The existing methodologies in synthetic population generation can be divided based on two criteria. Firstly, the methods are classified based on the type of data they can generate, more precisely, whether they support the generation of individuals or complete households. The individuals are described by the set of attributes that we try to reproduce, while the households contain their own attributes (household size, household type) expanded with the set of associated individuals. The individuals generation ensures the reproduction of univariate distributions treating each attribute separately. The household generation is more complicated because, besides the univariate distributions, the multivariate distributions must be considered to capture relationships between household members correctly.

Secondly, the synthetic generation methods can be based on statistical (e.g. Iterative Proportional Fitting, Markov Chain Monte Carlo Simulation) or machine learning (e.g. Generative Adversarial Networks, Variational Autoencoder) approaches. Miranda (2019) has produced a systematic review covering several decades of synthetic population generation methods applied to transportation models. According to the findings of this study, in Figure 1, we illustrate the most popular methods for each category chronologically. In addition, we added the most recent research streams. We analyse all proposed approaches' positive and negative aspects and define the literature gaps we will try to fill.



	GENERATION OF INDIVIDUALS	GENERATION OF HOUSEHOLDS	ASSOCIATIONS BETWEEN INDIVIDUALS & HOUSHEOLDS
Iterative Proportional Fitting (IPF)	**1996** *Beckman et al.* Creating synthetic baseline populations	**2007** *Arentze et al.* Creating synthetic household populations	**2009** *Ye et al.* Iterative Proportional Updating
Simulation techniques (MCMC)	**2013** *Farooq et al.* Simulation based population synthesis		**2014**, *Anderson et al.*, Associations Generation **2015**, *Casati et al.*, Hierarchical MCMC
Machine Learning techniques	**2014,** *Goodfellow et al.* Generative Adversarial Network **2018,** *Xu et al.* Tabular Generative Adversarial Networks **2019,** *Borysov et al.,* Variational Autoencoder **2020,** *Badu – Marfo et al.,* Composite Travel Generative Adversarial Networks **2022,** *Lederrey et al.,* DATGAN: Integrating expert knowledge into deep learning for synthetic tabular data		2022 ...

Figure 1: The overview of existing synthetic population methods

A population consists of individuals described by a set of discrete or continuous attributes $X = (X^1, X^2, \cdots, X^n)$. In reality, these attributes have a unique joint distribution represented by $\pi(X)$. However, this is not available to the analyst since we have access only to the sample of the true population, which presents the partial view $\hat{\pi}(X)$ of the real unique joint distribution (Farooq *et al.*, 2013). The variables of interest in generating sociodemographic characteristics can be at the individual or household level. For example, those at the individual level are age, gender, and driving license ownership. Household-relevant variables can be the number of inhabitants and total income. Collecting this information from a sample of the existing population will form univariate and multivariate distributions. The univariate distribution characterizes one column at a time, while the multivariate characterizes two or more.

Although all the methods apply different algorithms, they have a common objective to produce data that share statistical properties of the real sample. Most of the methods were firstly focused on replicating marginal distributions of the selected key individual variables and later extended to support the generation of household variables.

The first methodology that appeared for the generation of synthetic individuals was an application of Iterative Proportional Fitting (IPF) (Beckman *et al.*, 1996). This approach is also known as a matrix fitting table. The concept behind the IPF is to take each marginal one at a time and change the sample's contingency table to reflect the aggregate property of the population. In the case of IPF, an increase in desired attributes causes exponential growth of the number of cells in the contingency table. Consequently, there are many combinations of attributes with a low number of individuals, leading to empty cells in the contingency table. It has been proven that IPF fails to converge due to this so-called "zero cell issue" (Ben-Akiva and Lerman, 1985). In addition to the scalability problem, there are other drawbacks of IPF, such as the lack of a heterogeneous representative population and the fact that it has a deterministic realization of synthetic population (Farooq *et al.*, 2013).

Even though IPF has revealed different flaws, many publications proposed incremental improvements to this method in order to handle household structures (Guo, 2007; Arentze *et al.*, 2007). The problem with these methodologies is that they are designed so that household attributes are randomly drawn from the empirical data following the joint distribution of the chosen household-level attribute, in a similar way to individual attributes. Even though the marginals of the generated household attributes might seem accurate, there is no guarantee of relationships between households and previously generated individuals (Zhu and Ferreira, 2014). In order to generate household members with all necessary relationships, the Iterative Proportional Updating was developed (IPU) (Ye *et al.*, 2009). IPU provides a synthesis of the population by matching household and individual distributions simultaneously (Saadi *et al.*, 2016). However, the main issues of IPU pointed out by several authors are that there is no theoretical proof of convergence, that it suffers from the same issues as IPF, and requires a disaggregated initial sample which is rarely available (Lenormand and Deffuant, 2013; Zhu and Ferreira, 2014). Nowadays, datasets are described by many dimensions and observations, making IPF insufficient to satisfy current needs in the generation of synthetic households. Also, the high dependence on the form of the initial real sample limits the wide use of IPF since the disaggregate sample is rarely available.

By overcoming the above-mentioned issues, the simulation outperformed IPF. The standard for population synthesis used in travel activity modelling and microsimulation today is the Markov Chain Monte Carlo (MCMC) simulation developed by Farooq *et al.* (2013). This method implements Gibbs Sampling by drawing from pre-formed conditional distributions. Based on the individual MCMC, various studies have proposed matching algorithms to group synthetic individuals into households in such a way as to follow the real-world constraints (Anderson *et al.*, 2014; Casati *et al.*, 2015). These approaches are so-called "two-step" procedures since they require a previously created sample of synthetic individuals to perform a matching procedure. Different "two-step" approaches propose different ways of assigning people to households. For example, in Casati *et al.* (2015), the first step consists of generating three pools of individuals independently through separated simulation runs, based on the household role (e.g. owner, spouse, others). In the second step, one individual is picked from each pool and placed in a specific household based on the household size. Although with "two-stage" simulation households methods, we can obtain a good approximation of the marginal distributions, there is no guarantee that relationships between different variables are preserved. By assuming independence while generating individuals, the multivariate distributions are ignored, resulting in the generation of a completely unrealistic population (e.g. the generation of a child who has children).

However, most authors have pointed out that the Gibbs sampler has difficulty in delivering acceptable results while working with high-dimensional datasets. The definition of the conditionals has a substantial impact on the accuracy of the results. The ideal case would be to consider all variables conditional to each other and capture all correlations and dependencies. However, in reality, this is not possible because Gibbs Sample suffers from the "curse of dimensionality" phenomenon. This means that the efficiency and accuracy drop with an increase in dimensionality. It happens because the algorithm performs for a long time in highly correlated areas, which slows down convergence and produces an under-representation of outliers.

As can be seen from Figure 1, most authors have recently switched to the application of Machine Learning techniques in this domain. Nevertheless, in this paper, we decided to use a simulation approach as a core methodology. In the following paragraphs, we justify our choice through the comparison between ML and simulation.

The technique, that might be considered a state-of-the-art for the generation of synthetic individuals, is Generative Adversarial Networks (GANs) specialized for tabular data (TGANs) proposed by Xu and Veeramachaneni (2018). This approach has shown great success in generating high dimensional datasets in an accurate and computationally efficient way. GANs learn the probability distribution of a dataset implicitly and may generate samples from it. It is made up of two neural networks called the generator and discriminator. The generator is trained to learn how to generate data from random noise to deceive the discriminator. The discriminator is trained to discriminate between the real and generated data (Goodfellow *et al.*, 2014).

ML methods have recently become popular because simulation methods typically fail to deliver high-quality data in the context of Big Data. However, it is interesting to notice that none of the Machine Learning techniques is adapted to be used in the domain of the household generation. The potential reason might be that it is difficult to impose rules of expert knowledge due to the data-driven nature of the ML approach. On the contrary, the simulation approach is more model-driven, allowing us to control the generation process and embed rules. Recently, the first trial to incorporate expert knowledge into ML models in the context of synthetic data generation has taken place (Lederrey *et al.*, 2022). With this model, it is possible to specify the relationships between variables through Directed Acyclic Graph, that is later treated by Tabular GAN (DATGAN). Although this model better represents multivariate distributions, it still does not support dealing with hierarchical structures such as households.

In this paper, we propose an extension of individual MCMC for complete one-step household generation by addressing several gaps:

- Contrary to IPU, the simulation methods are sample-free, which means they can work with aggregated or disaggregated initial samples. Although several methodologies exist for assigning individuals to households, no sample-free methodology combines the generation of individuals and their household assignment in a single step. The main disadvantage of two-step methodologies is that they are composed of two sequential independent steps. This way, some interrelations between individuals are neglected, resulting in the generation of illogical observations. Therefore, a simultaneous approach could be closer to capturing correlations between households and individuals.

- In the generation of hierarchical structures, such as the generation of households with all constituent individuals, the consideration of multivariate distributions is crucial if we want to replicate realistic relationships between household members. The model-driven structure of Gibbs Sampler will allow us to enforce the realistic relationship between different household members.
- To address the "curse of dimensionality", we propose a new algorithmic approach that decomposes a full dataset based on the correlation rate between different attributes.

3 Methodology

This section presents a new simulator for generating complete synthetic households with all necessary relationships between individuals. This approach relies on the idea of existing individual Markov Chain Monte Carlo simulation (Farooq *et al.*, 2013).

The main goal of our methodology is to ensure the generation of realistic households simultaneously in one stage. To preserve logical relationships between individuals, we must generate each person conditional to other household members. The difference between our and the existing "two-step" approach is illustrated in Figure 2. Instead of independently generating agents and mapping them into corresponding households, we generate a synthetic dataset where each row corresponds to one household with all information about individuals. We propose the re-definition of the existing modelling framework and investigate the amount of control we can achieve during the generation process. The description of the existing methodology and proposed modifications are presented in Section 3.1.

The simulation approach is prone to failure while working with multi-dimensional datasets. Since the household generation problem includes many attributes, it is inevitable to face

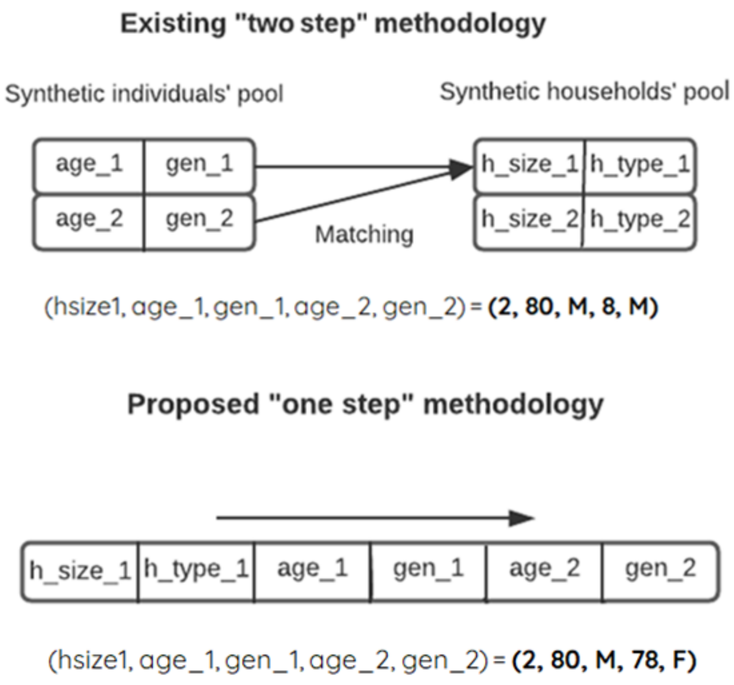

Existing "two step" methodology
Synthetic individuals' pool
Synthetic households' pool
age_1
gen_1
age_2
gen_2
Matching
h_size_1
h_type_1
h_size_2
h_type_2
(hsize1, age_1, gen_1, age_2, gen_2) = (2, 80, M, 8, M)
Proposed "one step" methodology
h_size_1
h_type_1
age_1
gen_1
age_2
gen_2
(hsize1, age_1, gen_1, age_2, gen_2) = (2, 80, M, 78, F)

Figure 2: Comparison between existing approach and one-stage simulator methodology

the “curse of dimensionality” phenomena. To overcome this problem, we decompose a generation process based on the level of correlation among variables, as described in Section 3.2.

3.1 Modeling

The core of the proposed methodology is the Markov Chain Monte Carlo (iMCMC) simulation technique, more precisely Gibbs Sampler proposed by Farooq et al. (2013). In technical terms, the sequence of individuals that are generated is called a Markov chain, while the Gibbs sampling is an algorithm that is used to draw the attribute values from the given distributions.

The synthetic individuals are random variables characterized by discrete and continuous attributes $(X_1, X_2, ..., X_n)$ that create a unique joint distribution $\pi(X)$. However, we only have a partial insight into this unique distribution derived from the real reference sample, denoted as $\hat{\pi}(X)$. In each iteration, Gibbs Sampler draws a value of one attribute from the probability distributions conditional to the fixed values of other attributes, using inverse transform. In the case of discrete variables, the algorithm draws from probability mass function, and in the case of continuous from probability density function. The conditional distributions are calculated beforehand and provided to Gibbs Sampler as input. In our case, the conditional probabilities vectors are computed from data using contingency tables by counting categories of one attribute given all others. In general, conditionals can be derived from data, models or assumptions.

Hypothetically, all correlations could be captured by integrating all attributes into conditionals. However, in practice, it is not possible to consider full conditionals because it requires dealing with a lot of attributes which produces the multi-polarization that makes Gibbs Sampler extremely slow (Casati *et al.*, 2015). Because of this, the authors usually simplify conditionals by assuming the independence between specific attributes. For example, if we want to generate the education of the person, ideally, the full conditional would contain information on all other attributes (e.g. age, household size). However, based on expert knowledge, we can assume that education depends more on age than household size. According to that, instead of drawing from $\pi(X_{education}|X_{age}, X_{hsize})$, we would draw from the simplification $\pi(X_{education}|X_{age})$. Assuming that education is uniform across the household size, we obtain almost the same accuracy while improving efficiency. The simplifications might be a reasonable decision while working with several attributes. The more attributes we have, the more simplifications we need, which complicates a modelling process and might influence accuracy.

The logic of generation follows the existing individual MCMC explained in Section 3.1.1. Compared to iMCMC, we define a household as a meta-individual characterized by household attributes and attributes of household members. In other words, instead of generating individual vectors described with a set of features $X = (X1, X2, ..., Xn)$ and grouping them into households afterwards, in each iteration, we form a vector of households. In the current version of our methodology, this vector consists of 4 household attributes (household size, household type, number of cars within the household, and household income) and an array of household members. This last attribute helps us to model the hierarchy between individuals and household. The individuals are described with their own set of attributes, in our case, by age and gender. The sequence of individuals has a variable size that is determined based on household size. Note that this methodology is generic, and it can be expanded with a various number of attributes.

To demonstrate this approach, In Figure 3, we discuss one instance of the problem. Assume that we want to generate a household registered as a couple. The spouse's generation must be conditional to the age of the owner in order to avoid the generation

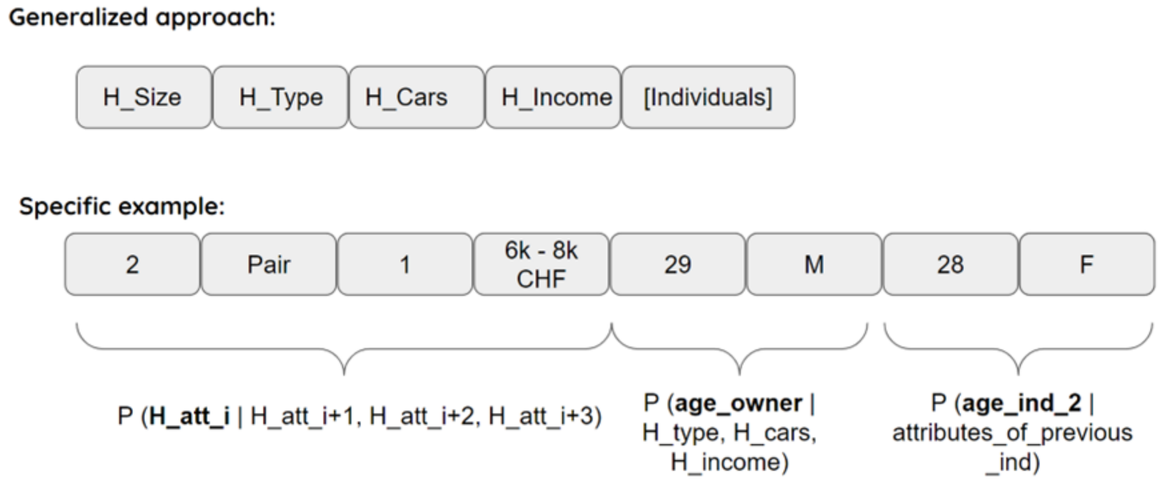

Generalized approach:
H_Size
H_Type
H_Cars
H_Income
[Individuals]
Specific example:
2
Pair
1
6k - 8k CHF
29
M
28
F
P (**H_att_i** | H_att_i+1, H_att_i+2, H_att_i+3)
P (**age_owner** | H_type, H_cars, H_income)
P (**age_ind_2** | attributes_of_previous _ind)

Figure 3: One-step household generation with an instance of the problem

of an unreasonable age difference. On the other hand, the realistic generation of gender has to consider the distribution of heterosexual and homosexual couples. By design of conditionals, we enforce the satisfaction of expert knowledge rules as explained in detail in Section 3.1.2.

3.1.1 Existing methodology

The Gibbs Sampler consists of four phases: initialization, warm-up, generation and skipping phase (Ben-Akiva *et al.*, 2021). Starting from random initial states, we simulate several chains of draws simultaneously. After a certain number of iterations, these sequences should converge to a common unique joint distribution noted $\hat{\pi}(X)$ (Gelman *et al.*, 2013). In the warm-up phase, we run the simulator until it reaches a steady-state. All draws generated in this phase are discarded. This way, we can guarantee that only draws from $\hat{\pi}$ are accepted. The indicators used to identify the stationarity are explained in Section 3.2.1.

Given the number of attributes k, the algorithm picks a random number $r = U(0, k-1)$ in each iteration. The number k defines the index of the attribute that will be generated in the current iteration. The result of one iteration is a household vector with one changed attribute, while all other values are kept from the previous iteration. This may produce a sequence of similar draws. To avoid the generation of similar observations, we skip the specified number of observations in the post-processing phase.

3.1.2 Description of conditionals

The construction of conditionals is based on modelling assumptions, and it should capture the essential correlations that are valid independently of data. However, the assumptions can be verified through the correlation analysis obtained from data.

Generation of households attributes The household size is one of the most informative attributes in the household generation since it is used as an indicator of how many individuals should be generated. As shown in Table 1, for some categories, household size has a strong correlation with household type. Independently on the dataset, the one-member household is considered a single-type, while the couple-type household always implies two inhabitants. For these categories, considering other information such as income or number of cars would only add unnecessary complexity. Thus, the values which can be assumed (such as household size = 1 when household type = 'Single') are deterministically assigned, not stochastically generated. None of the specific rules can be identified for households with more than three members, so the values of household size or household type must be stochastically drawn from the probability vector obtained from the conditionals.

In the current version of the methodology, the household attributes are generated using following conditionals:

- $\pi(X_{size}|X_{type} = x_{type}, X_{cars} = x_{cars}, X_{income} = x_{income})$,
- $\pi(X_{type}|X_{size} = x_{size}, X_{cars} = x_{cars}, X_{income} = x_{income})$,
- $\pi(X_{cars}|X_{size} = x_{size}, X_{type} = x_{type}, X_{income} = x_{income})$,
- $\pi(X_{income}|X_{size} = x_{size}, X_{type} = x_{type}, X_{cars} = x_{cars})$.

Generation of individual attributes

The correct generation of age attribute is crucial for a generation of the realistic household since it is very informative and reveals a lot of information about the household. For example, the people live in a specific type of household at a certain

	1	**2**	**3 or more**
Single	1	0	0
Couple	0	π_{11}	0
Couple with children	0	0	π_{21}
Single with children	0	π_{12}	π_{22}
Non-family	0	π_{13}	π_{23}

Table 1: The relationship between household size and household type

age. This is a consequence of the cycle of life events that usually happen. In general, there are some patterns such as children living with their parents up to some age, then they move from their parent's place to study or establish their own family. Also, depending on the age, the individual can get a driving licence or start to contribute to the household financially. Because of this, the first individual is generated from the selected subset based on household type given number of cars, income and role (i.e. $\pi(X_{age_1}|X_{cars} = x_{cars}, X_{income} = x_{income}, X_{role} = x_{role})$). In the preprocessing phase, the role of individual is assigned based on the age and household type. That way, we guarantee that the first individual (owner) is the oldest. All other individuals are generated considering the age difference of previously generated agents. The generation of individuals always includes the generation of the whole sequence. This simplifies the generation of other individuals because only the conditional for the owner generation takes into account household attributes such as number of cars and income. For all others, it is assumed the they share the same household attributes as the owner. According to that, the age of other individuals is generated from $\pi(X_{age_i}|X_{type} = x_{type}, X_{age_{i-1}} = x_{age_{i-1}})$. The gender of individuals depends on the type of household and the gender of previously generated individuals. For example, in the case of generating couples, the gender of the spouse is drawn from the distribution of heterosexual and homosexual couples given the gender of the owner.

3.2 Algorithm

The "curse of dimensionality" means that with an increase in the number of attributes, the algorithm's efficiency drops. In some specific cases, the algorithm even fails to execute. So far, the authors have dealt with this issue by making assumptions to simplify conditionals. However, we need a more general solution to this issue. This behaviour is a consequence of the fact that with an increase in dimensions, there are a lot of correlations that must be taken into account, which produces a multi-polarization (Casati *et al.*, 2015). Since the algorithm iteratively picks the values from probability vectors, it has to perform for a long time to leave the “high probability” region. Furthermore, in the case of “1-1” correlation, it might end up in a degenerative state. This means that the algorithm always chooses the same value and cannot proceed to another state. We propose the so-called “Divide and Conquer” (DAC) Gibbs Sampler to avoid this situation. This procedure decomposes the generation process based on correlations between different attributes.

The main motivation for the development of the Divide and Conquer Gibbs Sampler is to:

- prevent the algorithm from being stuck in the highly correlated areas by isolating the generation of strongly-correlated values,
- avoid unnecessary stochastic generation for values that can be assumed (e.g. one-member-single-household) which results in saving time,
- maximize the tread-off between accuracy and efficiency,
- avoid under-representing weakly correlated areas and over-representing highly correlated areas.

On the one hand, with decomposition, we improve efficiency since we work with the small batches in parallel. On the other hand, we improve the accuracy since the highly correlated areas are isolated. This enforces the generation of weakly correlated values and better representation of outliers because Gibbs Sampler does not spend a lot of time choosing only high probability values. Some of the attribute values are assumed based on expert knowledge to prevent the algorithm from staying stuck in a degenerative state (e.g. in the case of “1-1" correlation). By assuming values, we can impose a perfect fit more efficiently.

The idea of the algorithm is given in Figure 4. From the real sample, we identify the correlation coefficients among the attributes (e.g. using Pearson’s correlation). Based on the analysis, we extract the subset of the observations that contain the strongly correlated values. For example, as shown in Figure 5, we can identify that there is some positive correlation between household size and household type. We expect that the categories that mainly produce strong correlations are one-member-single and two-member-couple households based on expert knowledge. Indeed, by extracting the subsets from these categories, we verify our assumption as shown in Figure 6. Instead of using the real sample

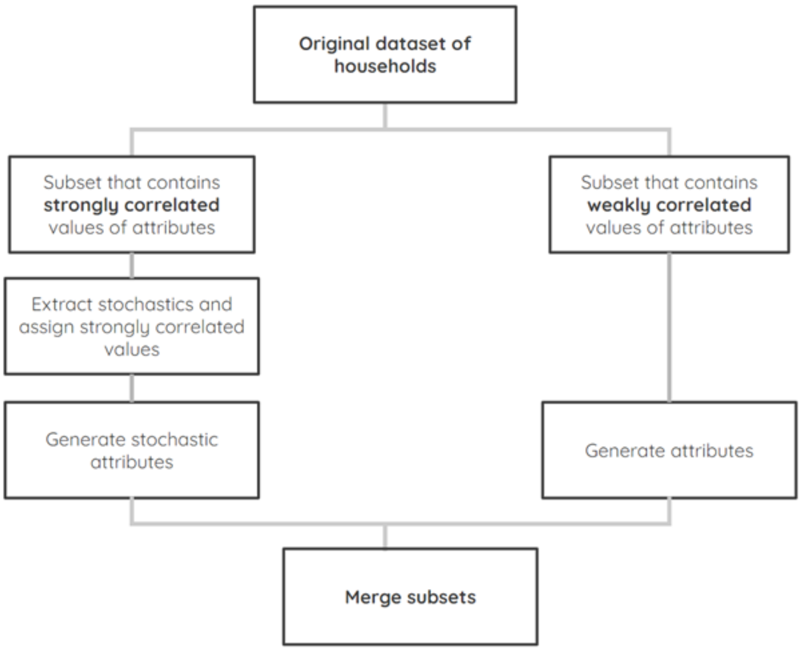

Original dataset of households
Subset that contains strongly correlated values of attributes
Subset that contains weakly correlated values of attributes
Extract stochastics and assign strongly correlated values
Generate stochastic attributes
Generate attributes
Merge subsets

Figure 4: Algorithmic framework of Divide and Conquer Gibbs Sampler

as a reference for the generation, we run two separate generation processes in parallel using those two subsets as references. In the end, we merge synthetic subsets into one. This approach can be generalized as illustrated in Algorithm 1.

Algorithm 1 Divide and conquer one-step simulator of synthetic households

```
Investigate correlation matrix of attributes two by two
if $corr\_coefficient \geq threshold_1$ then

    Create a frequency matrix between two correlated attributes
    Assign correlation coefficients for attributes' values based on counts from the table
    if $abs(corr\_coeff\_of\_values) \geq threshold_2$ then
        Assume their value based on the data
    end if

end if
```

As shown in Section 4, the DAC Gibbs gives a better capture of correlations than the general Gibbs Sampler. Also, this approach decreases computational time. However, the efficiency improvement should be formally verified by metrics that indicate a convergence of the algorithm, as explained in Section 3.2.1.

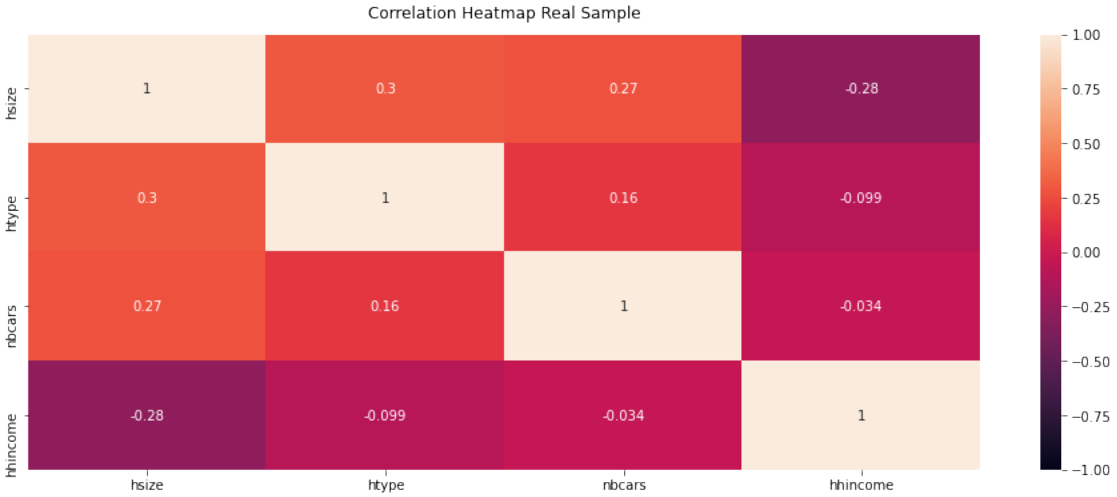


Figure 5: The correlation matrix of real data sample

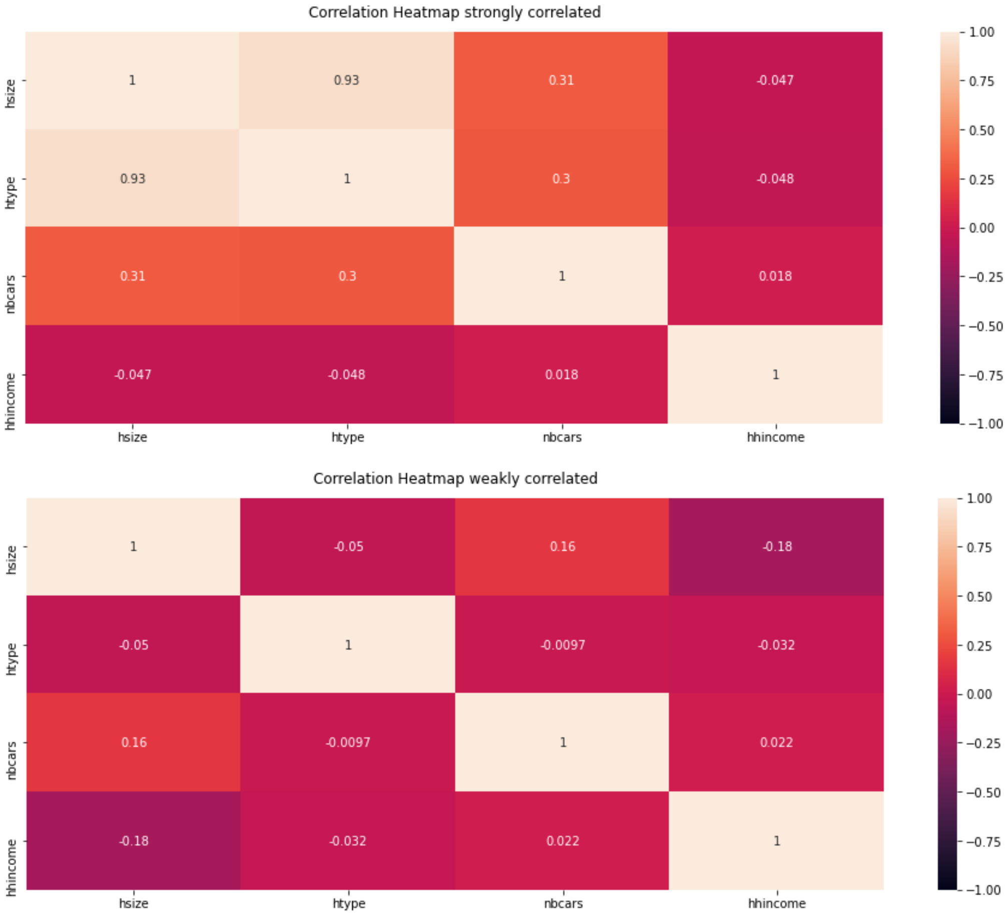


Figure 6: The identification of strongly and weakly correlated subsets

3.2.1 Investigation of convergence

As stated before, to reach stationarity, several chains of draws should be created simultaneously. Each chain contains draws of one attribute. After a certain number of iterations, the sequences should converge to a common target joint distribution noted as $\hat{\pi}(X)$. In order to identify the suitable number of iteration we monitor the convergence by comparing variation 'between' and 'within' simulated sequences. The variation of 'within' should be almost equal to 'between' variation. To estimate this, we calculate two metrics: potential scale reduction factor and the effective sample size proposed by Gelman *et al.* (2013). Once we proved that each simulated sequence is close to the distribution of all the sequences mixed together, we can treat the draws as sample from the target distribution.

The potential scale reduction $\hat{R}$ indicates whether we should stop or continue the simulation runs, and it is calculated using the following Eq. (1):

$$\hat{R} = \sqrt{\frac{\frac{n-1}{n} \cdot W + \frac{1}{n} \cdot B}{W}} \tag{1}$$

where n is number of simulation draws, W is within-sequence variances, and B is between-sequence variance. The numerator estimates the marginal posterior variance for each chain.

Additional to $\hat{R}$, we calculate effective sample size n_{eff} to get an idea of the simulations precision using Eq. (2):

$$n_{eff} = \frac{m \cdot n}{1 + 2 \cdot \sum_{t=1}^{\infty} \rho_t} \tag{2}$$

where m is number of sequences, and ρ_t is autocorrelation of each sequence at lag t.

4 Results

The objective of our analysis is to compare the accuracy and efficiency between the developed algorithm (DAC Gibbs) and the state-of-the-art technique (DATGAN) (Lederrey *et al.*, 2022). DATGAN model is used through the python module available at https://github.com/glederrey/DATGAN. We examined both graphical and statistical validation tests to determine the fit between generated and the real sample.

4.1 Data description

This research project leverages two datasets, on both individual and household levels. The data were collected by the Federal Office for Spatial Development (ARE) and the Federal Statistical Office (FSO) by conducting a nationwide survey. This data sample so-called the Swiss Mobility and Transport micro census data (MTMC) gather insights into the mobility behaviours of residents (Office fédéral de la statistique and Office fédéral du développement Territorial, 2017). Respondents provide their socio-economic characteristics (e.g. age, gender) and list the other household members, information on their daily mobility habits, and detail records of their trips during a reference period (1 day). The disaggregated sample contains information on 163,843 individuals living in 57,090 different households. From all collected attributes, only six attributes have been selected for our experiments. Two attributes are chosen from the individual dataset (age and gender) and four from the households dataset (household size, type, income and number of cars). All of the variables are discrete except for age, which is continuous. Each household is characterized by a unique identifier shared between the household members. This attribute defines a hierarchal structure between individuals and households, which we use to merge two datasets. The quality of obtained results using the proposed algorithm depends on the conditionals provided as input. The construction of conditionals requires the identification of correlations between different attributes. Since this can be achieved by investigating data patterns, the pre-processing phase presents a significant part of the project and includes: dealing with the missing and unknown values, encoding categorical data, and comparing and adjusting the different categories of the attributes. For any attribute where the frequency of missing values is less than 1%, the entire row containing missing value of that attribute is eliminated. The 36% of household income values are unknown, so all these values are kept as such. The final sample used in all experiments contains 163,322 observations described by attributes, as shown in Table 2.

Individual attributes	
Attribute	**Values**
Age	[0 - 103]
Gender	{M, F}
Household attributes	
Size	{1,2,3,4,5,6,7,8,9,10,11,12,13,14,17}
Type	{Single, Single with children, Pairs, Pairs with children, Non-family}
Income	{person below 18, prefer not to say, unknown, less than 2000 CHF, CHF 2000 - 4000, CHF 4001 - 6000, CHF 6001 - 8000, CHF 8001 - 10000, CHF 10001 - 12000, CHF 12001 - 14000, CHF 14001 - 16000, CHF 16001 - 18000, more than 18000 CHF}
Number of cars	{0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,20,51}

Table 2: Data description

4.2 Case study

The purpose of the graphical analysis is to visualize the marginal and conditional distributions. In our experiment, we compare the original data sample with two synthetic samples generated by DAC Gibbs and DATGAN. The marginal distribution analysis for all discrete household attributes is shown in Figure 7. Moreover, we provide an evidence of marginal fit for continuous variables as shown in Figure 8. As can be seen from Figure 7, both DAC Gibbs and DATGAN show almost a perfect fit of marginals. The interesting aspect of this graph is that the DAC Gibbs algorithm gives a better result while reproducing outliers, particularly for household size variables. The observed behaviour is expected, given the idea of the Gibbs's "divide and conquer" part explained in Section 3. The main motivation of the dataset division based on the level of correlation was to enforce the generation of weakly correlated variables and avoid overrepresentation of the highly correlated variables.

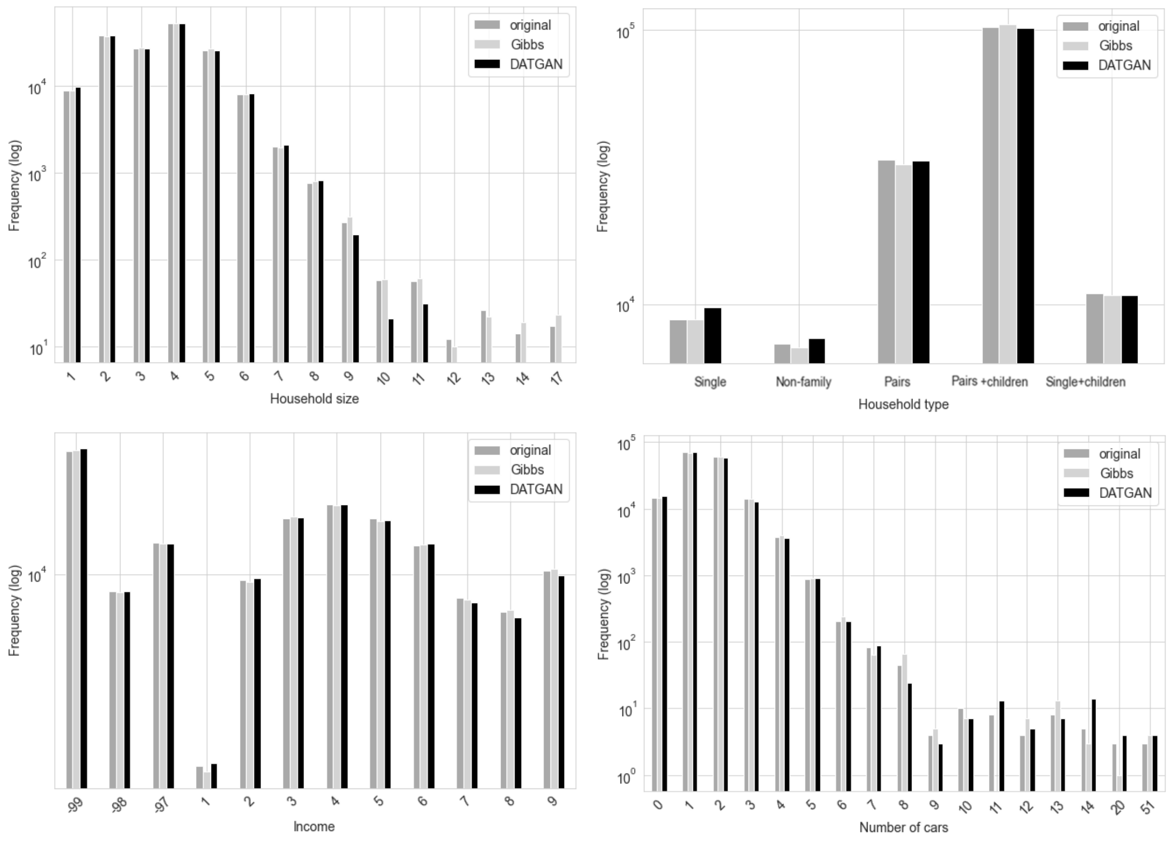
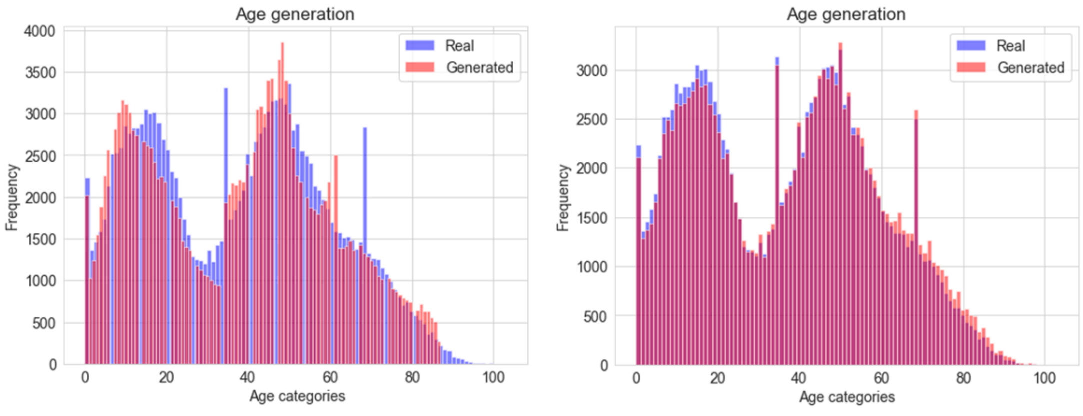


Figure 7: The representation of marginal distributions for discrete variables



DATGAN **DAC Gibbs**

Figure 8: The representation of marginal distributions for continuous variables

It is worth mentioning that the marginal distributions prove only that aggregated properties of the real sample are well reproduced. Although the synthetic data at the column level show almost a perfect fit, it is not guaranteed that the generated observations are logical and realistic. In order to test the realism of the generated sample, we have to verify the satisfaction of rules derived from expert knowledge. We investigate these rules by drawing the one attribute's value conditionally to the other, as shown in Figure 9.

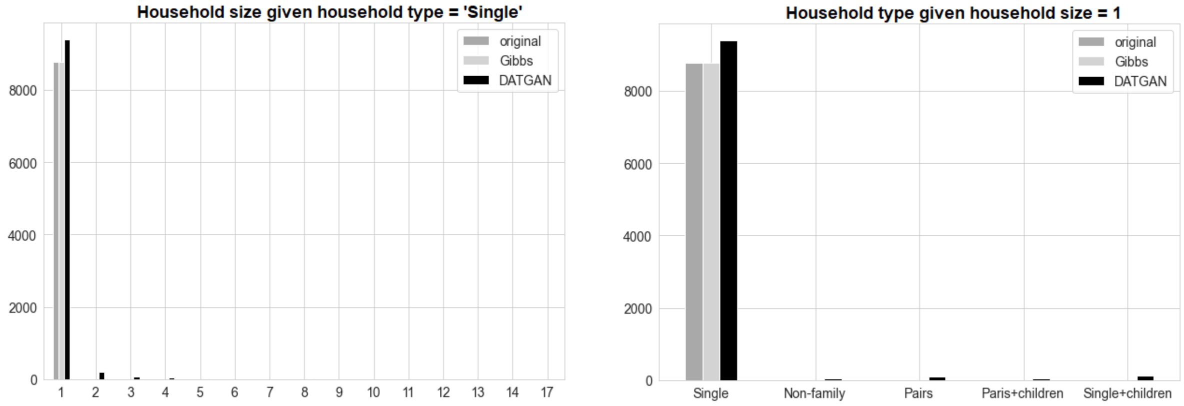


Figure 9: Conditional distributions of household size = 1 and household type ='Single'

According to Figure 9, one-member households comprise the same number of individuals from this category as in the real sample. We know that a one-member household is always qualified as a single household from expert knowledge. We expect the "one-to-one" correlation between one member and a single household. However, if we analyze sub-distributions of the household type given a household size, we can observe that it is not the case in data generated with DATGAN. DATGAN produces size 1 synthetic households for other categories such as non-family, pairs, pair with children and single with children. This result is an evidence that DATGAN follows a data-driven paradigm, while DAC Gibbs is more model-driven. The latter approach allows us to embed control within the generation process in order to impose some domain-specific rules. As explained in Section 3, we assign some values instead of generating them, which is the reason for the perfect fit between one member and a single household. The categories of household size and household type, which are not considered as highly correlated (e.g. households with more than 3 members), are generated stochastically from constructed conditionals (Figure 10).

Statistical tests should confirm the results obtained by the visualization. We use the existing validation approach that unites five metrics typically used in the transportation field for synthetic data assessment (Lederrey *et al.*, 2022). This approach relies on a frequency matrix that contains counts of observations for each column of real and

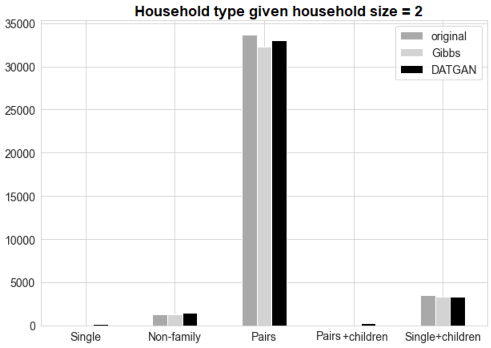


Figure 10: Conditional distributions of household type given household size 2

First-order test		Second-order test	
Rank and method	**Score**	**Rank and method**	**Score**
1. DAC Gibbs	2.81e-02 ± 1.53e-02	1. DATGAN	1.39e-01 ± 9.52e-02
2. DATGAN	4.32e-02 ± 3.41-02	2. DAC Gibbs	2.32e-01 ± 3.21e-01

Table 3: First and second order validation tests

synthetic samples. Based on the number of compared columns, we can identify two ways of validating data: the first and second order tests. The first-order statistical test assesses the aggregated statistical properties of synthetic data. The formulation of the first-order method includes the comparison column by column, which results in assessing only univariate distributions.

On the other hand, the second-order tests assesses the multivariate distributions by checking the relationship between pairs of columns. Both of methods provide a score taking into account all the columns and interpretation by ranking the different methods based on the score. From Table 3, we can conclude that the results are not completely aligned with the results obtained by visualization. As expected, DAC Gibbs gives a better score for a first-order statistical test since it accurately reproduces outliers. Surprisingly, DATGAN shows a better capture of the logic in the data based on the second-order test. This outcome is counterintuitive given the sub-distribution presented in Figures 9 and 10. However, it might be a case that for other attributes the conditionals that we were using while drawing from distributions failed to capture all the correlations among the attributes. Based on this result, it would be interesting to revise the modelling part described in Section 3 and improve the construction of conditionals.

In the last phase of the case study, we justify the advantages of DAC Gibbs Sampler compared to the general Gibbs Sampler. At first, we tried to run the regular Gibbs Sampler in order to generate synthetic households using the entire dataset. Due to strongly correlated variables, it was impossible to perform the algorithm. To address this problem, we apply DAC Gibbs Sampler. We divide the dataset into two subsets and select all observations that cause strong correlation (i.e. one-member and two-member households) into one batch, and all others to another. The results of separated generation are shown in Figure 11. Given that some values are deterministically assigned, we prevent the algorithm's failure since it never encounters the degenerative state. In Figure 12 we compared the correlation matrix between the original dataset and the merged generated dataset. As observed, the correlations are well represented, which means that we achieved desired accuracy more efficiently.

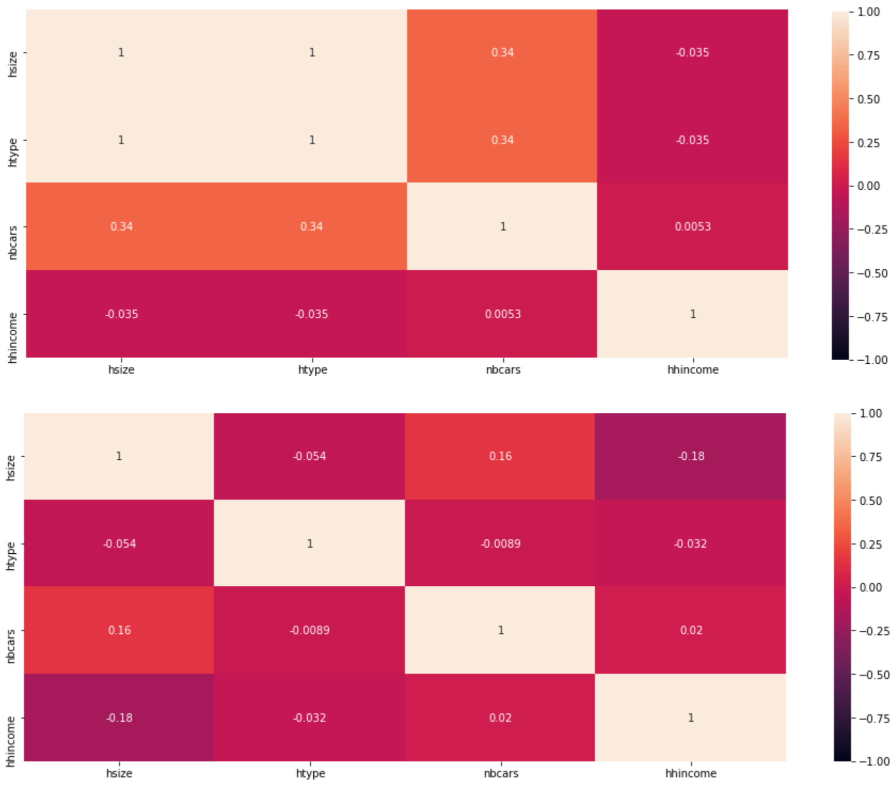


Figure 11: The correlation matrix of strongly correlated subset (top) and weakly correlated subset (bottom)

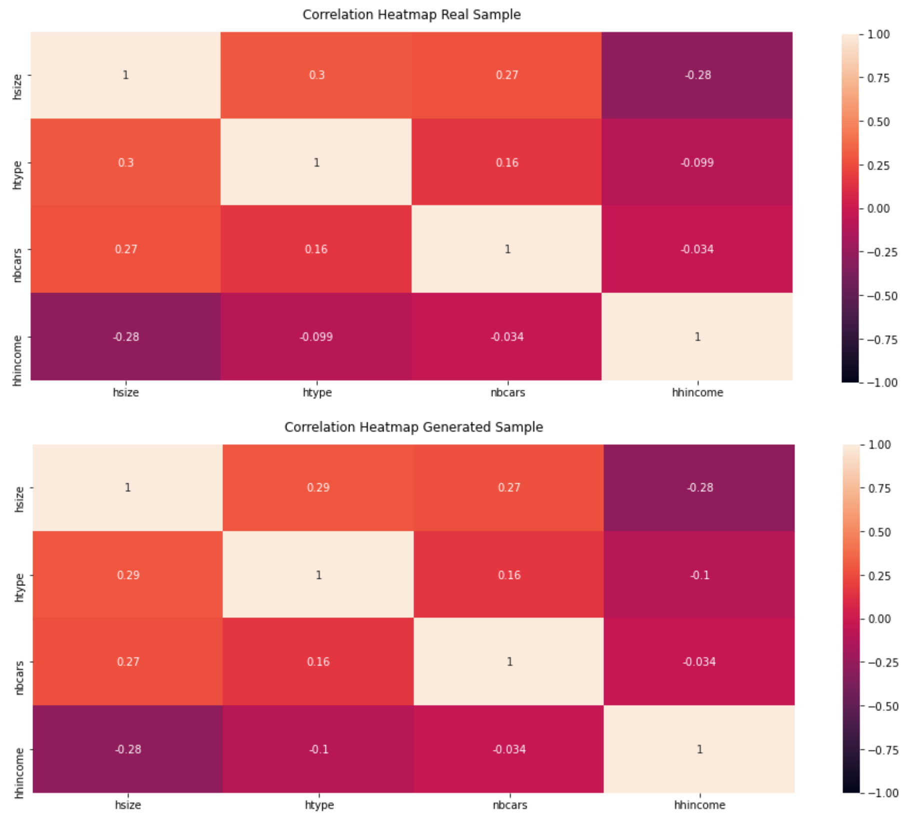


Figure 12: The correlation matrix original (top) and generated dataset (bottom)

5 Conclusion

This paper introduces the new methodological approach for a synthetic household generation. It unites the hierarchical generation process into the one-step procedure since the associations among individuals are formed within the generation process. This way, no additional matching procedure is needed. To integrate two steps into one, we re-define the modelling of conditionals in existing methods. Furthermore, we modify an existing algorithm and show that decomposition of the generation process can improve accuracy and efficiency.

However, this approach should be expanded to generate more attributes needed in activity-based modelling (etc. driving licence, occupation). At the moment, we only consider correlations among the attributes available from the data sample. By including some other attributes that are more informative we could simplify the generation of specific attributes. For instance, instead of generating the number of cars given household size, type and income, we can generate it only by using a number of driving licences in the household.

To formally prove the increase of efficiency by using DAC Gibbs Sampler, it remained unexplored how sensitive is the computational cost and steady-state achievement to the unit increase in a number of attributes. Since the purpose of DAC Gibbs Sampler is to generate realistic households, the trade-off between utility and privacy should be examined.

6 References

Anderson, P., B. Farooq, D. Efthymiou and M. Bierlaire (2014) Associations Generation in Synthetic Population for Transportation Applications.A Graph-Theoretic Solution, 23.

Arentze, T., F. Hofman and H. Timmermans (2007) Creating synthetic household populations: Problems and approach, *Transportation Research Record*, 6.

Beckman, R. J., K. A. Baggerly and M. D. McKay (1996) Creating synthetic baseline populations, *Transportation Research Part A: Policy and Practice*, **30** (6) 415–429, ISSN 0965-8564.

Ben-Akiva, M., M. Bierlaire, J. Walker and D. McFadden (2021) *Discrete choice analysis*, Cambridge, MA: MIT Press.

Ben-Akiva, M. and S. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*, MITPress, cambridge-ma.

Casati, D., K. Müller, P. J. Fourie, A. Erath and K. W. Axhausen (2015) Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking, *Transportation Research Record*, **2493** (1) 107–116.

Farooq, B., M. Bierlaire, R. Hurtubia and G. Flötteröd (2013) Simulation based population synthesis, *Transportation Research Part B: Methodological*, **58**, 243–263, ISSN 0191-2615.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari and D. Rubin (2013) *Bayesian Data Analysis (3rd ed.)*, A Chapman and Hall Book,CRC Press, London.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio (2014) Generative adversarial networks.

Guo, B. (2007) Population synthesis for microsimulating travel behavior, *Transportation Research Record*, 9.

Lederrey, G., T. Hillel and M. Bierlaire (2022) Datgan: Integrating expert knowledge into deep learning for synthetic tabular data, `https://arxiv.org/abs/2203.03489`.

Lenormand, M. and G. Deffuant (2013) Generating a Synthetic Population of Individuals in Households: Sample-Free Vs Sample-Based Methods, *Journal of Artificial Societies and Social Simulation*, **16** (4) 12, ISSN 1460-7425.

Miranda, D. F. (2019) Reviewing synthetic population generation for transportation models over the decades.

Office fédéral de la statistique and Office fédéral du développement Territorial (2017) Comportement de la population en matière de transports. Résultats du microrecensement mobilité et transports 2015, *Technical Report*, Neuchâtel, Berne.

Saadi, I., A. Mustafa, J. Teller, B. Farooq and M. Cools (2016) Hidden Markov Model-based population synthesis, *Transportation Research Part B: Methodological*, **90**, 1–21, ISSN 01912615.

Xu, L. and K. Veeramachaneni (2018) Synthesizing Tabular Data using Generative Adversarial Networks, *arXiv:1811.11264 [cs, stat]*. ArXiv: 1811.11264.

Ye, X., K. Konduri, R. Pendyala, B. Sana and P. Waddell (2009) Methodology to match distributions of both household and person attributes in generation of synthetic populations, 01 2009.

Zhu, Y. and J. Ferreira (2014) Synthetic Population Generation at Disaggregated Spatial Scales for Land Use and Transportation Microsimulation, *Transportation Research Record: Journal of the Transportation Research Board*, **2429** (1) 168–177, ISSN 0361-1981, 2169-4052.