The case of population synthesis at the level of the households

Marija Kukic

Michel Bierlaire

TRANSP-OR, EPFL

September 2021

STRC

21th Swiss Transport Research Conference

Monte Verità / Ascona, September 12 – 14, 2021

Transport and Mobility Laboratory (TRANSP-OR), EPFL

The case of population synthesis at the level of the households

Marija Kukic, Michel Bierlaire
Transport and Mobility Laboratory
Ecole Polytechnique Fédérale de Lausanne
Station 18
1015 Lausanne
{marija.kukic,michel.bierlaire}@epfl.ch

September 2021

Abstract

Modern transportation science requires advanced demand models to predict the needs for the mobility of individuals and goods. In order to calibrate those models, we need data as an input. However, having in mind the data privacy constraints and the unavailability of that data, synthetically generated data is being used. Typically, generated data are either on the level of individuals or at the level of households. Although several different methodologies exist for accurately and efficiently generating synthetic population data at the level of the households, there are two main gaps. Firstly, in those approaches, the generation of individuals and their matching into households is done separately, through two sequential processes. Secondly, the state-of-the-art techniques might generate unrealistic observations due to high dependence on data and the lack of control within the generation process. This project aims to develop a methodology to integrate the generation of the agents and their matching into households in a one-step process. In this paper, we are presenting the first framework component for synthetic household imputation. By imputation, we imply the process of expanding the given dataset by adding synthetic people grouped into households using the information of a given individual. Another objective is to investigate the integration of real-world constraints and examine the amount of control we can embed within the generation process. The method is tested using census data from 2015 and mobile data from 2019 on the territory of Switzerland.

Keywords

Population synthetic, simulation, validation, activity-based models

1 Introduction

Modern transportation science requires advanced mobility and travel demand models to predict the complex needs for the mobility of individuals and goods. The models for predicting activity and travel related decisions of individuals and households are called Activity-Based Models (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 behavior. Such data are used to inform decision makers or to design accurate simulation models of travellers. Those data present an essential input to ABM, since they are required to calibrate the model. However, the problem lies in the confidentiality and the availability of such data. Traditional population census or travel survey datasets contain personal information about the individuals and households. Nevertheless, generally the datasets are not a complete representation of the whole population and the unprocessed data is 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. Additionally, the emergence of Big Data collection services used in conjunction with hand-collected survey data is highly detailed and has extensive coverage of the population. These data collection often come with strict data usage restrictions, or they are sanitized to remove identifying information.

To circumvent the aforementioned privacy and availability issues, synthetically generated data can be used. A good synthetic population has similar statistical properties as the real population of interest, but does not allow the identification of real individuals (in order to address the privacy issue) and compiles all the necessary data for the scientific analysis (in order to address the availability issue). Although several different methodologies exist for accurately and efficiently generating synthetic population data, several gaps in the literature can be identified:

Household generation: The objective of all synthetic generators is to produce data that will reflect the distribution of the selected key variables from the real population. The variables of interest can be at the individual or household level. For example, the variables which are of interest at the individual level are age, gender and driving license ownership. Household-relevant variables can be 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 the time, while multivariate characterize two or more. Usually, the existing synthetic generators are reproducing univariate distribution accurately. As a consequence, individual-level or household-level attributes can be generated by applying the same generation algorithm. The underlying issue is that, even if the household characteristics properly match the marginal distribution of the household variable, there would

be no links between the individuals and the specific households to which they belong (Ye *et al.* (2009)). Hence, applying the existing methodologies at the level of the households might lead to the generation of an unrealistic population. Authors have pointed out this gap by suggesting several techniques for mapping people into households using synthetic populations. Although several methodologies exist for assigning individuals to households, there is no methodology that combines the generation of individuals and their household assignment in a single step. The main disadvantage of two-steps methodologies is that they are performed in two sequential independent steps. In order to perform the matching step, the generated population is required beforehand. Usually, matching procedures are formed based on domain-knowledge assumptions. This means that interrelations between households and individuals can be omitted. Therefore, a simultaneous approach could be closer to capture correlations between households and individuals. Moreover, instead of generating two pools of agents, we would generate one, which might decrease complexity and computational time.

Capturing dynamics: The synthesis population methods work with a snapshot of a dataset at a specific moment in time. Once the initial synthetic population is generated, any changes in the reference data cannot be reflected upon the synthetic population. This means that for data released periodically, such as annual census data, or streaming data (e.g. travel activity data from mobile apps), population synthesis data generation doesn't keep track of changes between past, present, and future data. Because there are no relationships between successive iterations of the same population across time, capturing the most recent changes of the data requires re-generation of the whole population. The generation process is a one-shot process, and no methodologies exist for exploiting the evolution of the population.

Synthetic population validation: After the synthetic population's generation, the original data's representativity in statistical similarity should be verified. Despite the fact that the most recent synthetic generation approaches produce excellent correlation capture results, they are entirely data-driven. The absence of control throughout the generation process may result in population generation that does not satisfy real-world constraints. Usually, the representativity of the synthetic population is validated by comparing marginal distributions between real and synthetic data or by computing certain statistical tests. The assesses of the quality with existing methods consider the univariate distribution of each column separately. However, it is still a challenge to verify the plausibility of generated data due to the lack of metrics for validation of multivariate distributions (e.g. is a generated observation illogical and does it correspond to the real-world constraints). For example, even though synthetic data perfectly match the marginal age distribution of real data, it is possible to have a multivariate distribution of synthetic data

where 15 years old individual is retired.

In line with the aforementioned challenges, our broader research objective is to develop a simulation framework for synthetic household generation, which integrates the generation of the agents and relationships between them in a one-stage process. As the first step of the implementation, this paper seeks to develop a component for household imputation. The objective is to investigate the integration of domain-specific constraints (i.e. retired person doesn't go to primary school) and examine the amount of control we can embed within the generation process. This approach aims to assure consistency, representativity, and realism of generated households. We compare our methodology with the state-of-the-art approach using two criteria. The first criteria is the representativity of the marginals using existing metrics. Secondly, we analyse the synthetic population with a focus on the realism of the generated data by checking of the rules for the combination of columns for each observation. This methodology is tested on the case study as a part of a more comprehensive research study - "Multy-day and Multi-Person Activity Patterns and Schedules Owners". Given that, we will briefly present the common objectives of the collaborative project and how the developed component of the framework contributes in that context.

The document is organized as follows: Section 2 covers a detailed review of the previous research effort in this field; Section 3 introduces and formally specifies the methodology that we develop; Section 4 gives a detailed data description used in the specific case study. Through the preliminary results and comparative analysis with other methodologies, we discuss the implemented framework component. Based on the summary of results, Section 5 identifies future research directions.

2 Literature Review

The literature on synthetic population generation relying on statistical methods is vast. More recently, deep learning techniques have been explored. Miranda (2019) has done a systematic review covering several decades of synthetic population generation methods applied to transportation models. According to this study, we see that synthetic population generation methodologies evolved iteratively since each subsequent methodology addressed the limitations of the previous methods. Detail overview and analysis of the positive and negative aspects of each approach are given in Section 2.1.

In essence, all those research streams share the same objective. The population consists of individuals described by a set of discrete or continuous attributes X=(X^1,X^2,...,X^n). Those attributes have a unique joint distribution represented by $\pi(X)$, which is usually not available to the analyst due to privacy policy. As an alternative, a partial view, in the form of marginals, is used to draw samples from, as if we were drawing from the complete joint distribution (Farooq *et al.,* 2013). The realized form of the marginals should be as close as possible to the draws from $\pi(X)$.

Since generators are designed to generate joint distributions on chosen variables, in the context of the generation of sociodemographic characteristics, they can be used to generate of two types of attributes: individual-level (e.g. age, gender...) and household-level (e.g. household size, household type ...). In reality, the majority of the techniques were created to generate individual-level characteristics. Soon after, those approaches were expanded to be used in the context of household attribute generation. Although those methods were giving a good approximation of the marginals' distribution of each attribute separately, the population generated with these methods did not capture associations of the people within the same households. An overview of the methods for the creation of associations between households and individuals is given in Section 2.2.

2.1 Synthetic generation at the level of individual

With the popularity of activity-based models, population synthesis approaches began to receive a lot of attention (Miranda, 2019). The first methodology that appeared for synthetic population

generation 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 once at a time and change the sample's contingency table to reflect the aggregate property of the population (Ben-Akiva and Lerman, 1985). In the case of IPF, an increase of desired attributes causes exponential growth of the number of cells in the contingency table. As a consequence, there are many combinations of attributes with a low number of individuals, which leads to the presence of empty cells in the contingency table. It is proven that IPF fails to converge due to this so-called "zero issue" (Ben-Akiva *et al.,* 2021). Nowadays, datasets are described by the high number of dimensions and observations, which makes IPF insufficient to satisfy current needs in synthetic generation field. Another key issues of IPF which opened possibilities for the development of more satisfactory approaches are:

- the lack of a heterogeneous representative population due to cloning,
- scalability issues,
- deterministic realization of synthetic population.

Many publications proposed incremental improvements of the IPF method until simulation-based methods became popular since they addressed the issues of IPF (Guo, 2007), (Arentze T, 2007), (Mohammadian and Zhang, 2010). The gold standard for population synthesis used in travel activity modelling and microsimulation today is Markov Chain Monte Carlo (MCMC) simulation developed by (Farooq *et al.,* 2013). This method implements Gibbs Sampling by drawing from pre-formed conditional distributions. This allows an approximation of the underlying joint distribution instead of focusing on the reproduction of marginals. Compared to IPF which has a deterministic realization of the synthetic population, simulation approaches enable the production of many realizations of a synthetic population. The main disadvantage of this method is that the complexity of constructing conditional distributions increases as the number of attributes increase. This leads to less accurate generated data, since this method depends on conditionals' quality.

Based on the aforementioned drawbacks of statistical generation methods while dealing with the creation of high dimensional datasets, those procedures typically fail to deliver high-quality data in the context of Big Data. With the availability of computing technology, new methodological views, such as deep learning and other artificial intelligence frameworks, are being established. One of the first applications of techniques developed to generate models with a much larger set of attributes was Variational Autoencoder (VAE), implemented by Borysov *et al.* (2019). An additional approach that has shown great success in generating high dimensional datasets in an accurate and computationally efficient way is Generative Adversarial Networks (GANs) specialized for tabular data (TGANs) proposed by (Xu and Veeramachaneni, 2018). GANs learn

the probability distribution of a dataset implicitly and may generate samples from it. It beats other techniques in terms of capturing column correlation and scaling up to large datasets. In the work of (Badu-Marfo *et al.,* 2020), it is stated that GANs outperformed the VAE in terms of performances. Hence, the use of neural networks may be considered as the state-of-the-art technique for population synthesis at the level of individuals at this moment.

However, the main limitation of GANs is that they are a data-driven technique, which makes it unable to include expert knowledge into the generation process. Lack of control during the generation process leads to the generation of illogical and biased observations. In this paper, we provide a comparison of TGANs focusing on verification the plausibility of the generated data, presented in Section 5.3.

2.2 Synthetic generation at the level of households

Individual information is crucial for analyzing and understanding travel behavior, but it should not be considered in isolation from the social and environmental context. Without information about households, investigation of behavioral patterns is highly limited. Furthermore, unlike trip-based models, ABM allow for the understanding of mobility patterns by taking into account interactions such as the influence of the household (Pougala *et al.,* 2021). Integrating household data into ABM methodologies would expand the model's capabilities at capturing multi-individual decisions and interactions, as opposed to single individual decisions.

Several authors focused on improving existing methodologies to solve the issue of generating multiple agent types (individuals and households). The major issue of existing methodologies is that they were ensuring that household attributes in the synthetic population closely matched the desired distributions. Nevertheless, they don't guarantee consistency for the person's attributes of interest.

In other words, the methodologies are designed in the way that household attributes are randomly drawn from the empirical data following the joint distribution of the chosen household-level attribute, in the similar way as it is done for individual-level attributes (Zhu and Ferreira, 2014). Even though the marginals of the generated household attributes might seem accurate, there is no guarantee of relationships between households and previously generated individuals. This is due to fact that existing methods do not impose any control to match individuals into households within the generation process. To address this problem, various studies have been published on how to create relationships between synthetic people and the households they belong to, as well as how to group them according to the constraints inherited from the real data (Anderson *et al.,*

2014), (Lenormand and Deffuant, 2013), (Ye *et al.,* 2009).

Those research streams can be divided based on two eligibility criteria:

- depending on the number of stages for association generation - long-based or fitted table methods,
- depending on the required form of initial sample - sample-free or sample-based methods.

The long-list refers to the collection of methods in which the pool of previously created agents is required before performing a matching procedure(Anderson *et al.,* 2014). It is a two-stage procedure in which people are assigned to households based on all of the previous combinations of people and homes. Because the number of people and households rises exponentially, this may lead to increased computing complexity.

The fitted table methods use a contingency table in the matching procedure to change and reallocate weights among households of a certain type until both household and individual-level attributes are matched (Kagho *et al.,* 2020).

The sample-based methods assume the availability of a disaggregated data sample. Contrary to that, sample-free methods don't rely on the structure of the initial sample (Wickramasinghe *et al.,* 2020). High dependence on initial real sample limits the wide use of sample-based methods since the disaggregate sample is rarely available due to privacy constraints.

Lenormand and Deffuant (2013) provide a detailed comparison of the two representatives of sample-free and sample-based approach. The chosen sample-based approach is Iterative Proportional Updating (IPU) (Ye *et al.,* 2009). IPU provides a high-performance level synthesis of the population by matching household and individual distributions simultaneously (Saadi *et al.,* 2016). It continues the work (Beckman *et al.,* 1996), where IPF is extended to estimate the joint distribution of household attributes. However, this method failed to match the known distribution of person within generated household in the case of low frequencies of the people in the contingency table's cells (Ye *et al.,* 2009). In the same way as IPF suffers from sparsity issue, IPU hinders the ability to match person-level in specific cases. The frequency in each cell decreases as the level of disaggregation grows, resulting in a reduced ability to replicate or fit to person-level joint distributions.

A more generic approach was desirable, which yield to development of sample-free methods. One of the sample-free methods is iterative semi-stochastic algorithm proposed by (Gargiulo *et al.,* 2010). The algorithm's goal was to generate an artificial population in which people were grouped into household while taking into account set of statistical constraints imposed by conditional distributions. The proposed idea is to randomly pick household from the pool of existing households described by size and type, and to gather individuals from the predefined pool of agents following the age distributions.

Despite the fact that all synthesis generators rely to some extent on the quality of the first dataset, the outcomes achieved by the sample-based technique are highly dependent on the initial sample. Not only the sample-free technique can be used in a broader range of cases, but it also provides superior fit to the reference distributions.

Those two methodologies can be compared based on one more criteria. IPU is representative of fitted table method, while iterative-semi stochastic algorithm belongs to the long-list methods. On the one hand, IPU performs matching of household-level and person-level distribution in the process of generation. On the other hand, the long-list algorithm requires generated pool of agents before it starts the matching procedure. Thus, one-phase generation is more desirable as it doesn't require definition and execution of generation and matching procedure separately.

3 Problem statement

The long-term objective of this research is to design a simulation algorithm for generation of the complete synthetic households in one-stage process. It would overcome the computational complexity of the long-list methods, and it would be adapted for general use since simulation technique is not conditioned by the form of the input. The benefit of this technique over others is that it tries to combine the generation and association processes into one, rather than establishing two separate procedures for individual generation and matching into households. Furthermore, we are concentrating on creating a representative, consistent, and realistic population. The consistency implies the setting of the rules which are arisen from real life and domain knowledge. Realism implies that the generated individual who satisfies real-world constraints is also someone who is a representative member of the population. For instance, consistency requires that children are not older than parents, and realism additionally requires that children and

parents are in certain range of ages.

In order to meet this goal, the first step was to investigate amount of control that we can integrate by usage of simulation techniques. The core of proposed methodology is Markov Chain Monte Carlo (MCMC) simulation technique, more precisely Gibbs sampler proposed by Farooq *et al.* (2013). Gibbs sampler iteratively draws from the probability distributions conditional to the chosen attribute. Moreover, the number of generated attributes using MCMC methods depends only on recognized relationships between them (Moeckel (2003)). Following this methodology, a certain level of control can be embedded into generation process by imposing different rules. The rules are translated into conditional distributions that ensure satisfaction of the real-world constraints by assigning different probabilities to certain events to happen.

The already existing simulation methodology for generation of individuals can be expanded to the level of the households. Instead of generation of the individuals described with a set of attributes X = (X^1,X^2,X^3,...,X^n) and grouping them into households, we are developing the framework for direct generation of households. Each individual is generated as a row of dataset which we can define as a vector of different attribute values. Conceptually, the household can be defined as a meta-individual which is described by a set of attributes of the household members. For example, instead of generation of several rows of individuals, we would generate one household row. By regulating the size of the household vector, information about individuals within generated household can be extracted. This approach requires a definition of the variables that characterize meta-individual and constructions of conditionals that we are drawing from.

It is proven that the accuracy of generated population on the output is highly influenced by the quality of created conditionals that are inputs (Farooq *et al.,* 2013). The difficulty is because Gibbs sampler requires conditional distribution of one attribute over all others, which is not always possible. With the increase of dimensionality, the creation of conditionals becomes more complex. However, there are possible simplifications to make this procedure more flexible. The main advantage of our approach compared to other methods is that we want to postulate methodology which is data independent. Data independence comes from the fact that the construction of conditionals can be done not only from the data but also from the assumptions, domain knowledge and models.

In this paper we are investigating the level of control that we can embed in simulation techniques

for generation of synthetic households. We will present the developed simplified subcomponent of the future framework tested for households imputation. By imputation, we imply the process of the expansion of the given dataset by adding synthetic people grouped into households. The synthetic individuals are generated conforming to the given row from one dataset and distributions from another dataset. The methodology is developed for the needs of the "Multi-day and Multi-Person Activity Patterns and Schedules Owners" research project described in Section 5.

4 Methodology: Household imputation

This methodology is designed by combining and extending approaches developed by Farooq *et al.* (2013) and Gargiulo *et al.* (2010). In the work of Gargiulo *et al.* (2010) they are generating empty households described by size and type, and fitting the generated individuals described by age into it, by imposing various rules to input associations between them. On the contrary, we are using a referent row from one dataset instead of designing the pool of synthetic individuals and households from scratch. Other members of the household are generated by sampling from conditionals formed from another dataset.
According to that, in our approach we can identify two different categories of attributes: generated or deterministic. The values of generated attributes are defined through stochastic extraction by drawing from conditional distributions. The deterministic values are either assigned based on domain knowledge or inherited under assumption that some information are shared within the household. A detailed description of the data is given in Section 5.1.

The consistency and realism of the generated agents is obtained by assigning different probabilities for a certain event to happen. With the usage of conditional distributions, we can ensure that list of defined rules will result in generation of realistic observation. For instance, we will construct conditionals by imposing zero probability for a child under certain age to have children, income and to be employed. The algorithm consists of several iterative steps, as it is shown through the pseudo-code given in Algorithms 1, 2 and 3.
The idea of the algorithm is to go through the dataset which should be expanded. In each iteration, one referent individual is chosen with specific attribute values. Conditional to the values of the chosen row, household will be created following distributions from another dataset. The conditionals must be constructed beforehand and provided as input. The values of generated attributes are sampled from distributions by applying the inverse transform.

The algorithm starts by picking the household size. Based on household size, household type is

defined. The one-member household is considered as a single household, while two-members household is considered as a couple without children. For all other household size values, household type is generated following the marginal distribution of households' types from another dataset. The possible values are: a couple with children, single parent with children, and non-family households. Depending on the household type, the algorithm can be divided into several steps:

- **Single household**: It is a household that contains one member i.e. original individual $X^{new} = X^{old}$ extended by household type and household role. It is assumed that this individual is the head of the single household.

- **Couple without children**: It is a household that consists of two people - original and synthetic individual. The gender of the partner is generated following the gender distributions of the couples from census data, conditional to the gender of the given individual. This is necessary since census data contains the percentage of homosexual couples. It is assumed that the older individual is the head of the household, while younger is the spouse. The values for language, household size and owning of the car are inherited from the original row since the partners share those characteristics. The age of the synthetic partner is generated conditional to the partner's age. It's important to mention that complexity of the construction of conditionals increase by increasing the number of dimensions. In order to simplify conditionals, we can bring assumptions by capturing correlation and assuming independence. If we assume independence, then chosen attributes are given uniform across other attributes. Considering the correlation between attributes, education is generated conditional to age, employment conditional to education and income conditional to employment. The before mentioned process is described in Algorithm 2.

- **Couple with children**: It is a household that consists of two parents and various numbers of children, depending on the household size. The parents are generated in the same way as it is done for couples without children. After the generation of the parents, children are generated following the procedure illustrated in Algorithm 3. Similar to two-member households, the values of households' attributes such as language, household size and presence of the car are inherited from the original row. In addition to the head and spouse, a new role emerges in this type of household. All younger members than head and spouse are assumed as children. Note that all generated attributes are designed following marginals and conditionals derived from the census sub-dataset that selects rows that belong to this type of the household. The gender is designed following marginal distributions of census gender children's distribution. In order to avoid the generation of children older than parents, the age of the first child is drawn conditional to the age of

the younger parent. Moreover, to obtain the realistic age difference between children, all younger children are generated conditional to the age of the older children. The education is chosen conditional to the age. The children are assumed to be unemployed without income.

- **Single parent with children**: It is a household that consists of one parent and various number of the children. Since all individuals in the reference dataset are adults, it is assumed that given person is the head of the household. The children are generated following the same procedure explained in Algorithm 3.

- **Non-family household**: In the formation of non-family household, none of the information of the given row is used for generated attributes. The chosen approach stems from the fact that any regularities in this household type cannot be identified. The relationships between household members are unknown due to the fact that members might share flat or might be relatives. Hence, the generation of age, gender, income, education and employment is done using Gibbs sampler. The necessary conditionals are designed based on the extracted subset of non-family households from census dataset. Based on household size required number of agents from generated pool of individuals is picked. The values for deterministic attributes such as household size, type, owning a car are assumed to be the same as they are in referent row.

5 Case Study

The household imputation methodology is tested on the real-world case study as a part of the "Multi-day and Multi-Person Activity Patterns and Schedules Owners" research project. The project aims to bring together expertise in activity-based modeling and quantitative sociology to enrich the current - and traditionally monodisciplinary - travel behavior approaches. Mixed methods from both fields are used (e.g., optimization techniques, pattern-recognition algorithms, latent-variable analysis, descriptive statistics, model interpretation, and multivariate regressions) to design a modeling framework destined to help understand mobility and all the intricacies it entails in a sustainable territory. The project is structured in three parts which are sorted in chronological order as it is shown in the Figure 1.

Firstly, Schultheiss (2021) investigates the operationalization of space and daily activity struc-

Algorithm 1: Household imputation

Data: X_{given}= (x^{age}_{given},x^{size}_{given},...,x^{n}_{given}) - the chosen row from the referenced dataset

n - number of the attributes of each individual

N - number of the individuals in referenced dataset

k - number of the processed rows

i - number of synthetic people in household

$\pi(X_i|X_j)$ - conditional distributions formed according to another dataset

Result: N * ($x^{size}_{given} - 1$) synthetic people grouped into N synthetic households

$k \leftarrow 0$

while $k \neq N$ **do**

  $i \leftarrow 0$

  while $i < x_{size}$ **do**

    initialize synthetic individual X_i= (x^{age}_i,x^{size}_i,...,x^{n}_i)

    if $x^{size}_{given} = 1$ **then**

      $x^{type}_i \leftarrow$ single;

      $x^{role}_i \leftarrow$ head;

      $X_i = X_{given}$

    else if $x^{size}_{given} = 2$ **then**

      $x^{nb\_children}_i \leftarrow 0$;

      generate_partner();

    else

      draw x^{type}_i following $\pi(\ X^{type}_k \mid X^{size}_{given} > 2)$;

      if x^{type}_i = *couple with children* **then**

        $x^{nb\_children}_i \leftarrow\ x^{size}_{given} - 2$;

        generate_partner();

        generate_children();

      else if x^{type}_i = *single parent with children* **then**

        $x^{nb\_children}_i \leftarrow\ x^{size}_{given} - 1$;

        generate_children();

      else

        generate_person();

      end

      $i \leftarrow i + 1$;

      $k \leftarrow k + 1$;

    end

  end

end

Algorithm 2: Generate partner

Data: X_{given}= (x^{age}_{given},x^{size}_{given},...,x^{n}_{given}) - the chosen row from the referenced dataset
n - number of the attributes of each individual
$\pi(X_i|X_j)$ - conditional distributions formed according to another dataset
Result: synthetic partner X_k = (x^k_{age}, x^k_{size}, ..., x^k_n), k = 1
initialize X_k
if x^{given}_{size} = 2 **then**
  $x^{type}_k \leftarrow$ couple without children;
else
  $x^{type}_k \leftarrow$ couple with children;
end
$x^{language}_k = x^{language}_{given}$;
$x^{size}_k = x^{size}_{given}$;
$x^{car}_k = x^{car}_{given}$;
Generate x^{age}_k, x^{gender}_k,$x^{employment}_k$,$x^{education}_k$,x^{income}_k using Inverse Transform on chosen conditional distribution $\pi(X_i|X_j = x_{given})$;
if $x^{age}_k > x^{age}_{given}$ **then**
  $x^{role}_k \leftarrow$ head;
else
  $x^{role}_k \leftarrow$ spouse;
end

tures, defining different activity-travel behavior metrics. Secondly, proposed metrics should be integrated into a multi-day activity scheduling model (as an extension of model proposed by Pougala *et al.* (2021)). In order to integrate coupling and interpersonal constraints - multi-person - into the multi-day activity scheduling model, information about households is needed. For these purposes, the methodology described in this paper plays an important role. Particularly, it is used to extend a given "one-person" dataset by generating and imputing synthetic households using the methodology described in this paper.

5.1 Data

This research project leverages two datasets. The first dataset is the Swiss Mobility and Transport micro census data (MTMC), "Swiss census data" collected by the Federal Office for Spatial Development (ARE) and the Federal Statistical Office (FSO). The second dataset is a mobile phone record dataset of travel history "MOBIS data" collected by the Institute for Transport Planning and Systems (IVT) group at ETH Zurich.

Algorithm 3: Generate children

Data: X_{given}= (x^{age}_{given},x^{size}_{given},...,x^{n}_{given}) - the chosen row from the referenced dataset
$\pi(X_i|X_j)$ - conditional distributions formed according to another dataset
Result: synthetic children $X_k = (x^k_{age}, x^k_{size}, ..., x^k_n)$
initialize X_k
$x^{type}_k \leftarrow$ couple with children;
$x^{language}_k = x^{language}_{given}$;
$x^{size}_k = x^{size}_{given}$;
$x^{car}_k = x^{car}_{given}$;
$x^{role}_k \leftarrow$ child;
Generate x^{gender}_k draw from marginal distribution $\pi(X^{gender})$;
if $first\_child = True$ **then**
    Generate x^{age}_k using Inverse Transform on
    $\pi(X^{age\_child}|X^{age\_parent} = x_{age\_of\_younger\_parent})$;
else
    Generate x^{age}_k using Inverse Transform on
    $\pi(X^{age\_child}|X^{age\_parent} = x_{age\_of\_older\_sibiling})$;
end
Generate $x^{education}_k$ using Inverse Transform on $\pi(X^{education}|X^{age} = x^{age}_k)$;
Generate $x^{employment}_k$ using Inverse Transform on $\pi(X^{employment}|X^{education} = x^{education}_k)$;
Generate x^{income}_k using Inverse Transform on $\pi(X^{income}|X^{employment} = x^{employment}_k)$;

The Swiss nationwide survey collected the Swiss census data to gather insights on the mobility behaviors of local residents (OFS, 2015). Respondents provide their socio-economic characteristics and the other household members, information on their daily mobility habits, and detailed records of their trips during a reference period (1 day). The 2015 edition of the MTMC contains 163,843 individuals grouped in 57,090 households, with a record of 43,630 trip diaries. The MOBIS data were developed from the MOBIS study (Molloy et al., 2020) to allow the longitudinal study of travel behavior activities in greater detail using passively collected data. These data were collected from a combined "travel survey" and "mobile phone traces" method from 3,700 participants over 8 weeks in fall/winter 2019. The data contain information about the socio-demographics of individuals and their trips. The primary application of this project is on socio-economic data.

This approach takes referent individuals from the MOBIS dataset, and based on the list of rules and assumptions introduced in 4, generates and groups new individuals into households,

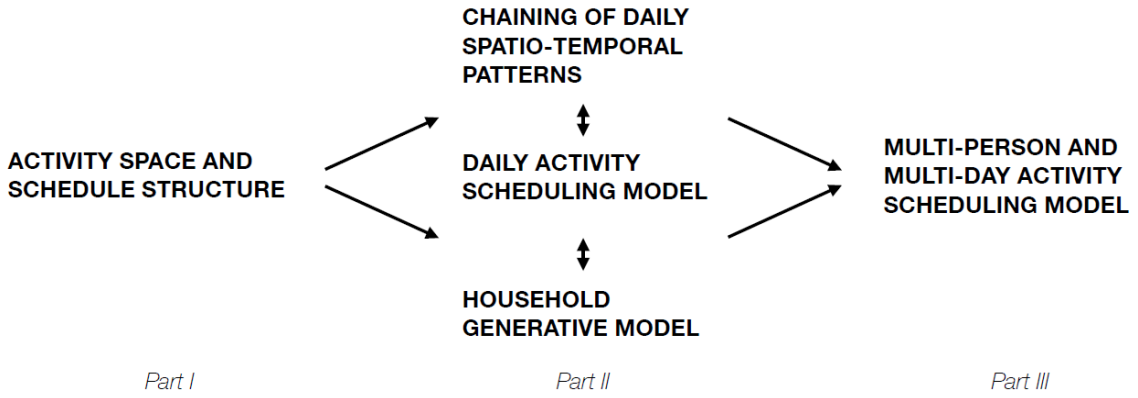

Figure 1: Different phases of "Multi-day and Multi-person Activity Patters and Schedules Owners" framework

following distributions from the Swiss census data from 2015. These two datasets also present a unique opportunity for them to be merged or used simultaneously in the synthesis process since the MOBIS dataset was collected from the same territory. Moreover, some attributes (age, education, employment, etc.) are similar across both datasets. It is assumed that during the time between surveys, population kept the similar statistical properties. It's acceptable to bring this assumption since Switzerland belongs to the stable low growth according to the classic model of demographic transition model.

Data preprocessing is a significant part of this project since the quality of created conditionals formed from the data heavily influence quality of results. For the creation of conditionals identification of correlation between different features must be taken into account. For those purposes, we used patterns identified through the dataset and assumptions. Moreover, the assumptions are used to perform necessary simplifications to reduce the complexity of conditionals while keeping essential correlations. The preprocessing phase was including dealing with the missing and unknown values, conversion in desired type, discretion of chosen attributes (e.g. age), encoding of categorical data, comparing and adjusting of the different categories of the attributes. Every individual in the generated dataset is described by 11 attributes. They can be separated into two categories: generated (e.g. gender, age, income, education, employment) and deterministic (e.g. household size, owning a car, household type, household role, number of children, language). The data description for both datasets is given in Table 1 and Table 2, respectively.

Individual level attributes	
Attribute	**Values**
Gender	[male, female]
Age	[<15,15-24,25-34,35-44,45-54,55-64,65-74,>=75]
Income	[<=4000, <=8000, <=12000, <=16000, >16000]
Education	[mandatory, secondary, higher]
Employment	[full time, part time, in education, unemployed]
Language	[german, italian, french]
Households level attributes	
Household size	[1-6]
Household type	[single, pair with children, non family households, pair without children]
Household role	[head, spouse, child, other]
Marital status	[single, married, widow, unmarried, divorced, partnership]
Number of children	[0-15]
Number of cars	[0 - 3]

Table 1: Swiss census 2015 data description

Individual level attributes	
Attribute	**Decription**
Gender	[male, female]
Age	[19 - 66]
Income	[<=4000, <=8000, <=12000, <=16000, >16000]
Education	[mandatory, secondary, higher]
Employment	[apprentice, employed, student, self-employed, retired, unemployed]
Language	[german, italian, french]
Households level attributes	
Household size	[1,6]
Owning car	[yes, no]

Table 2: MOBIS 2019 data description

5.2 Results

We tested our algorithm using two datasets - MOBIS and census. We selected one instance from the MOBIS dataset whose values we used to draw from pre-constructed distributions formed on the census dataset. Before the imputation, MOBIS was counting 3,700 individuals. Given

the household size, the algorithm generates, adds, and groups the required number of synthetic individuals to the MOBIS dataset. In the Figure 2, the results before and after imputation are shown. The final dataset counts 10,736 individuals following household size distribution, grouped into 3,700 households. Every generated row is described by 11 attributes (6 at the individual level, and 5 at the household level).

At the beginning of the study, we wanted to verify that the developed generator accurately reflects desired distributions. The algorithm was used to generate several individual attributes given the conditional distributions from the census dataset. The comparison between real and generated data for the age attribute is shown in Figures 3, 4 and 5. In order to verify the quality of the generated sample, standard validation techniques (SRMSE and R^2) are used. SRMSE represents the most used metric for quantifying how close datasets are. It is suitable for discrete attributes, and zero value means perfect match (Müller and Axhausen (2011)). The R^2 shows variation in the real population that is not reproduced in the generated population (Farooq *et al.* (2013)). Synthetic data better fit real data if the value of R^2 is closer to 1. As we can notice, the generated data replicate real distribution with acceptable precision.

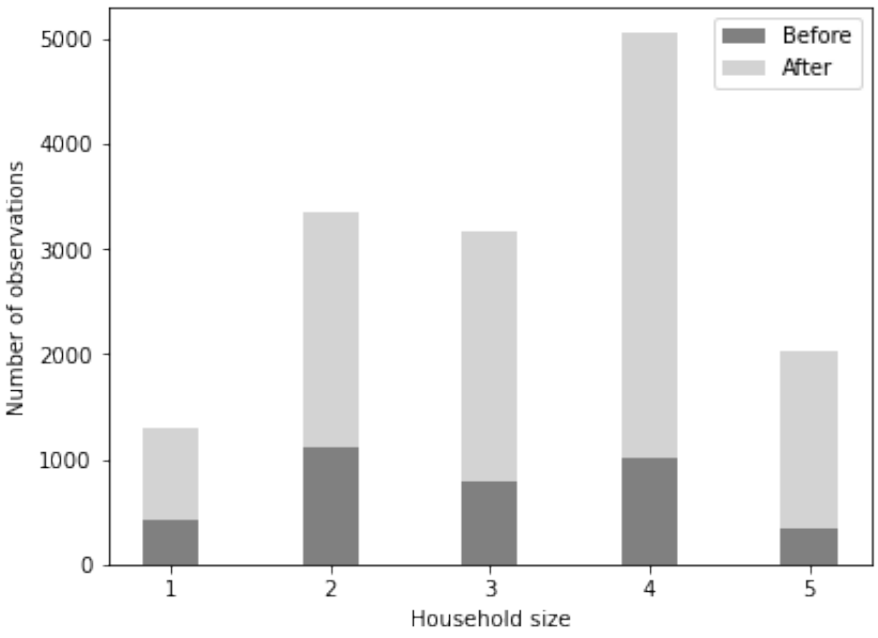

Figure 2: MOBIS dataset before and after imputation

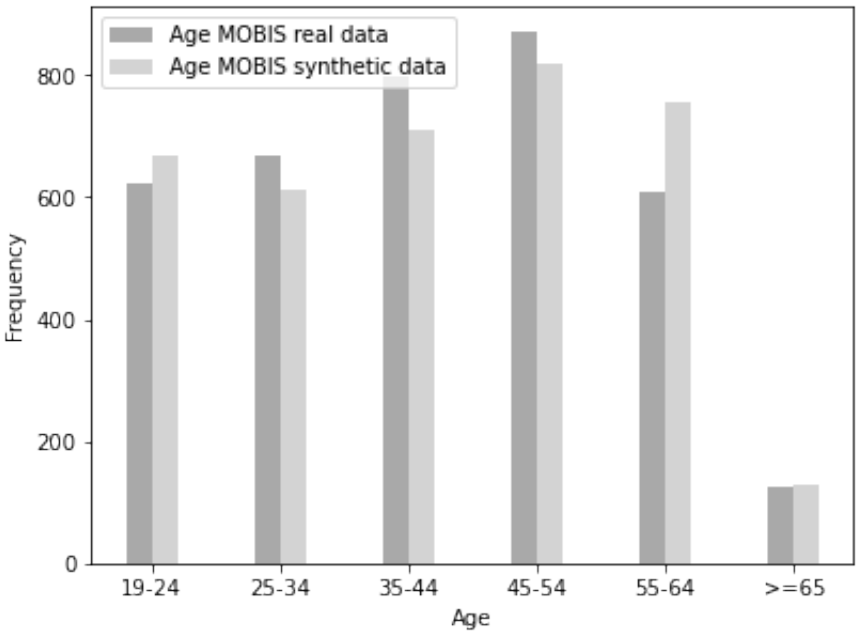

Figure 3: Generated and real age distribution - MOBIS dataset

X	Generated sample	Real sample	Empirical Probability	Actual Probability
1	667	621	0.180759	0.168293
2	612	669	0.165854	0.181301
3	709	796	0.192141	0.215718
4	820	870	0.222222	0.235772
5	754	607	0.204336	0.164499
6	128	127	0.034688	0.034417

Figure 4: Generated and real age distribution statistics - MOBIS dataset

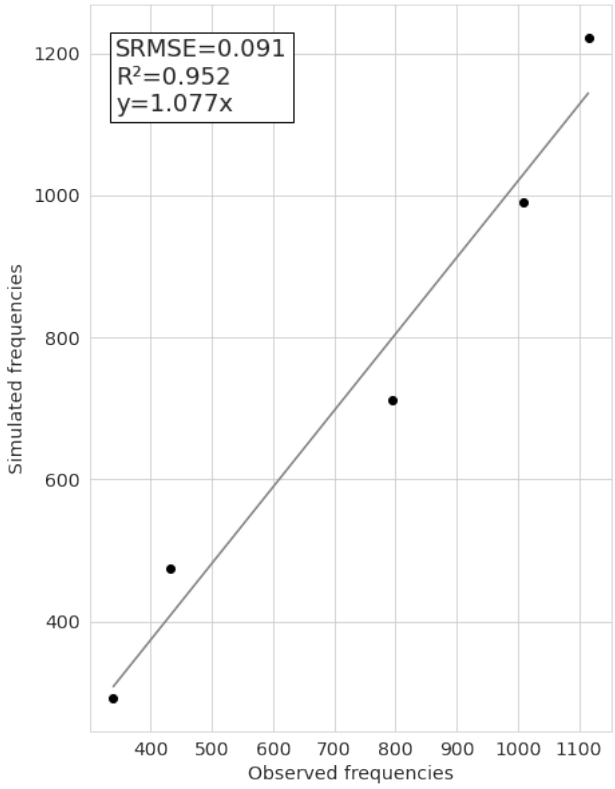

Figure 5: Standardized Root Mean Square error - age attribute

As it was stated in Section 4, we can differentiate two types of attributes: generated in a stochastic way and deterministically assigned. The example shown in Figure 6 shows that the household type is set in line with household size information for the single and the couples without children (deterministic part). In contrast, the households with household size above 3, follow marginal census topology (stochastic part).

If we compare the marginal gender distribution of couples without children, we can notice that the bars of real and generated data are inverse depending on the gender for most of the age groups. Although homosexual couples were taken into account during the generation of couples without children, Figure 13 shows the significant generation of heterosexual couples.

In contrast to generation from the empty households, the imputation method will not reflect just one distribution. It is important to note that we are mixing information from two datasets, resulting in keeping some characteristics of both distributions, as is shown in Figure 7. We are drawing attribute values from conditional distributions formed based on one dataset, conditional to the value from another dataset. Based on this, we could not expect a complete match with the marginals of census data. We can notice that the census has a more significant number of older

people than the MOBIS dataset. This leads to the result that in generated data category '>=65' will be overrepresented and '55-64' underrepresented. For the same reason, category '19-24' is underrepresented while '25-34' is overrepresented. For rest of categories, the distribution will reflect MOBIS dataset.

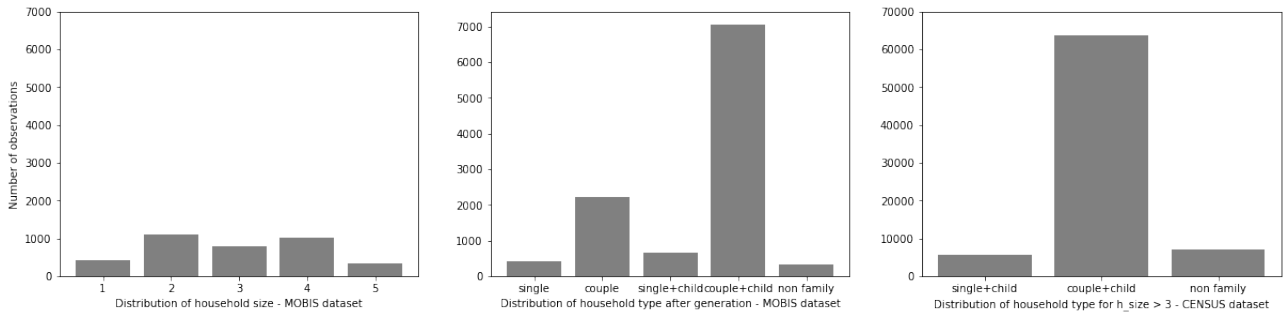

Figure 6: Relationship between household size and household type

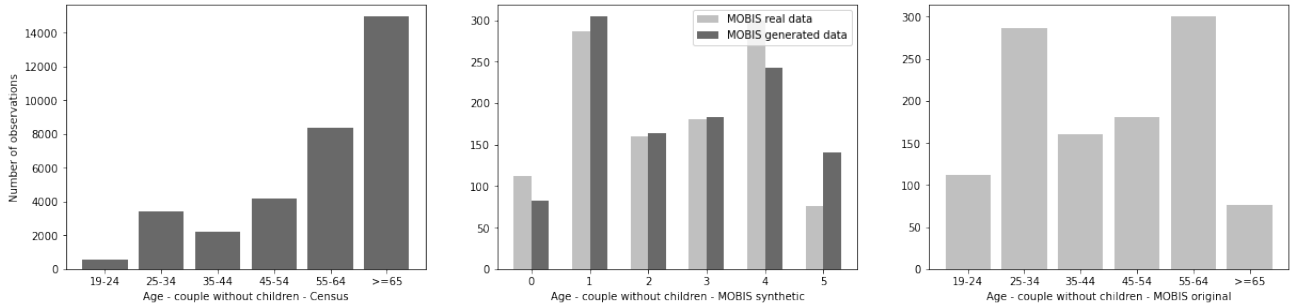

Figure 7: Generated and imputed data - age distribution for 'couple without children' household

Although the marginals show that data fit precisely into the desired distribution, plausibility at the individual row level is not guaranteed. The principal idea of the algorithm was to impose consistency and realism within the generation process by the correct construction of conditionals. The verification of plausibility is explained through the example of a "couple with the children" generation. If we try to generate individuals using the conditional distribution of one attribute over all the other attributes, unrealistic observations might appear. Hence, we have precisely to choose and construct conditionals to avoid those occurrences. The example of conditionals used for generation of the age between partners is presented in Figure 9. Figure 8 shows that the generation of unrealistic observations (e.g., the mother is younger than the child) is possible if conditionals are not correctly constructed. One possible way to avoid such a behavior is to generate children's age concerning the age difference of the younger parent. However, due to the stochasticity of the algorithm, we can't wholly rely on the satisfaction of this rule. In order to ensure generation under all imposed rules, all unfeasible observations are discarded and regenerated. After these steps have been carried out, the results shown in Figure 10 are

obtained. Moreover, results may be considered realistic since there is no significant number of households with old children living with their parents. As expected, the distribution of generated children is reflecting the shape of the census data.

Surprisingly, in some generated households, the mother was in the same age group as the children. In Figure 12, we extracted and analysed this phenomenon. If we take a detailed look, we can identify that the head is older than a spouse for all of those observations, meaning that most likely, children are from another marriage. This assumption can be verified by analysing the correlation between partners' characteristics and marital status, which is not included in our study.

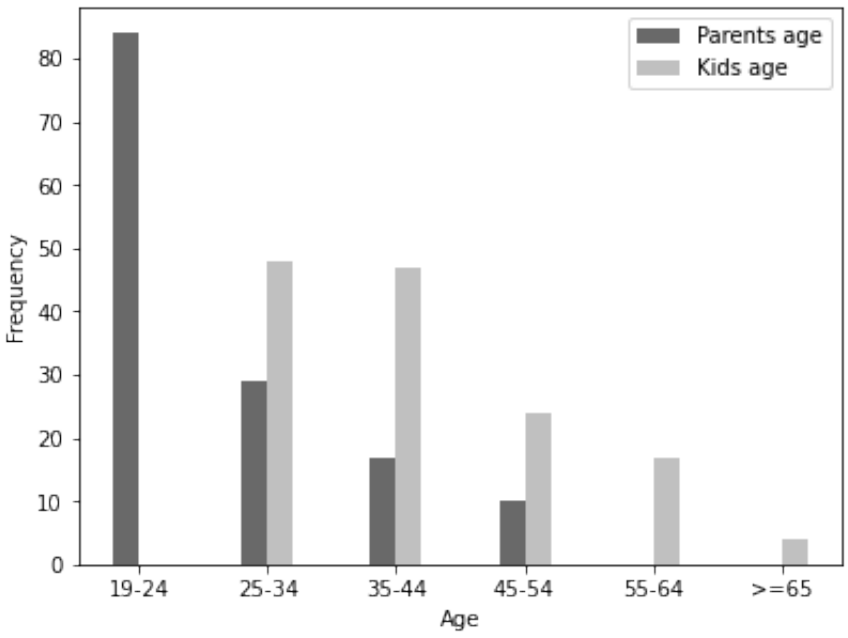

Figure 8: Unrealistic observations

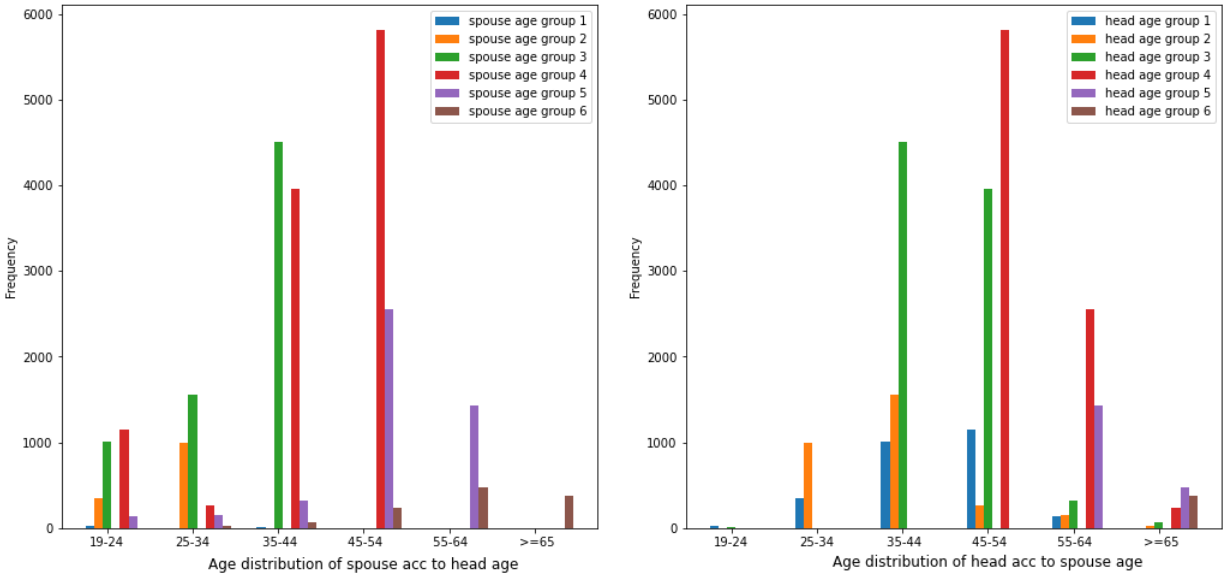

Figure 9: Conditional distributions for age generation of partners in 'couple with children' household

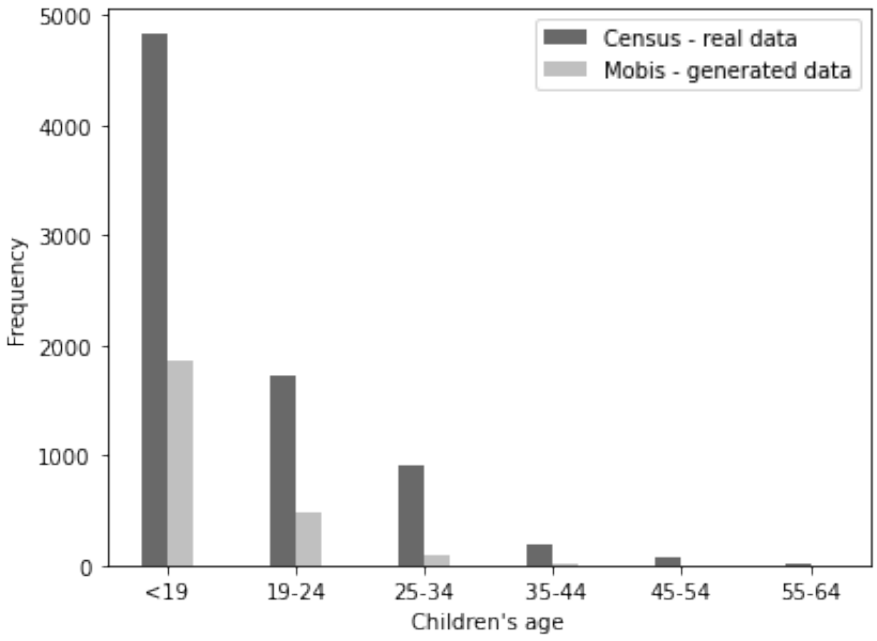

Figure 10: The comparison of the children's age distribution between Census and MOBIS datasets

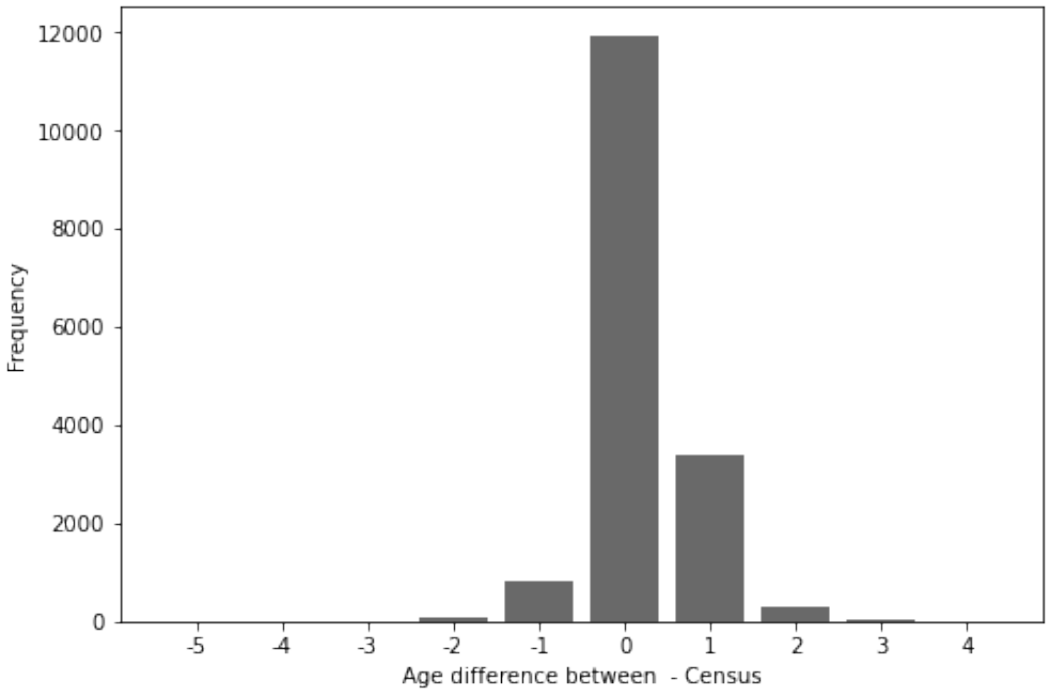

Figure 11: Age difference between partners

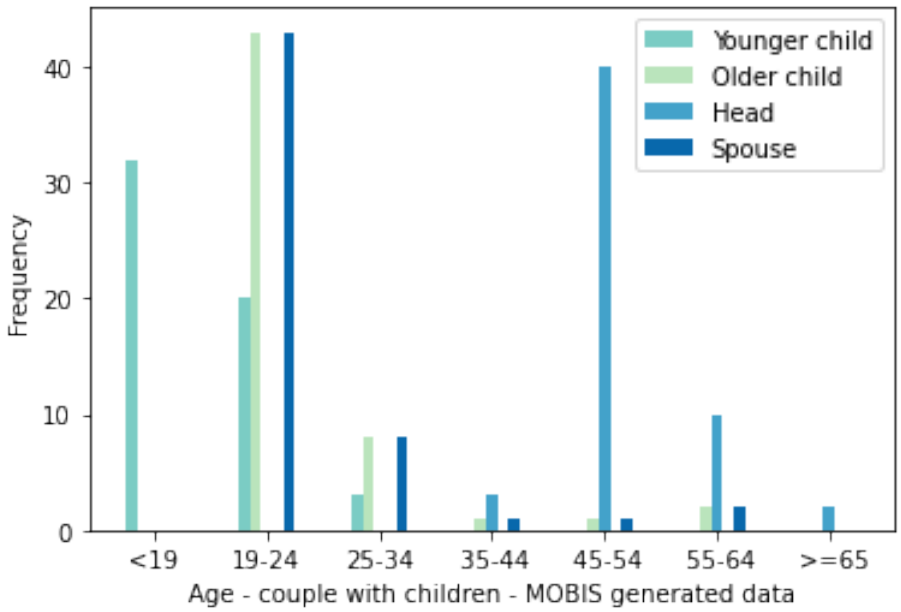

Figure 12: The age distribution of couples with children where spouse is in the same age group as children

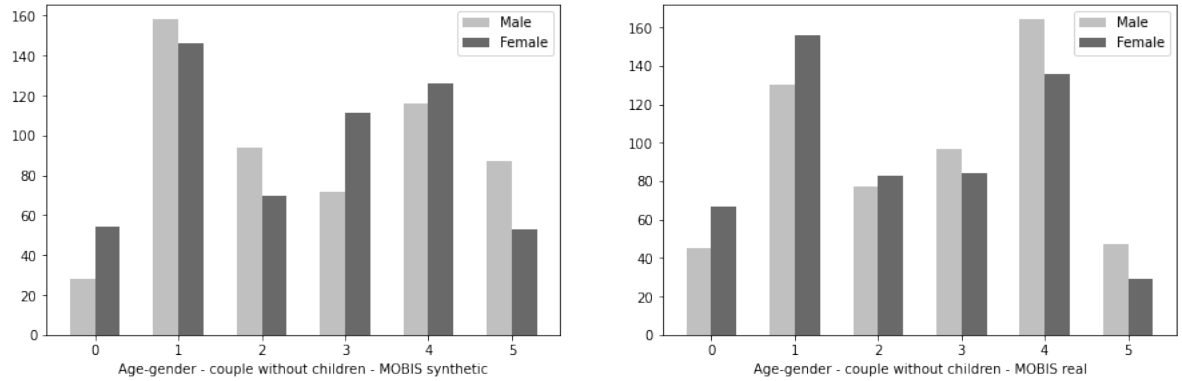

Figure 13: The comparison of the age-gender distributions within couple without children household

5.3 Comparison to other techniques

GANs may be considered as state-of-the-art technique for synthetic generation at the moment. It is made up of two neural networks, the generator and the discriminator. The generator accepts as inputs basic random variables and generates new data. The discriminator uses real and generated data to create a classifier by discriminating between them. The generator's objective is to deceive the discriminator (raise classification error by mixing as much created data as possible with true data), while the discriminator's goal is to distinguish between true and generated data (Xu and Veeramachaneni, 2018).

On the one side, we tested TGAN by using the model developed by Xu and Veeramachaneni (2018), and on another, we used our methodology. The experiments are performed using Swiss census dataset 2015. The objective of comparison was to validate quality of the generated data in terms of consistency and realism. In Figure 15 we are showing how both of methodologies match marginal distribution of the real population. To evaluate the marginals' representativity, statistical analysis is performed to compare draws against the real data, by standardized root mean squared error (SRMSE) Müller and Axhausen (2011) or R^2 goodness of fit Farooq *et al.* (2013). The results confirm the fact that TGAN outperforms Gibbs sampling, and it gives more accurate replication of the distribution at the column-level as it is shown in Figure 16. Except for a better fit, it's important to mention that with TGAN we generated population described with the full set of attributes, while with Gibbs sampler we generated just two. It has impact on the interpretation of SRMSE, because those results could be comparable just if we generated the same number of attributes. However, TGANs already shows better performances even though it generated more attributes. An increase of generated attributes in the case of Gibbs sampler

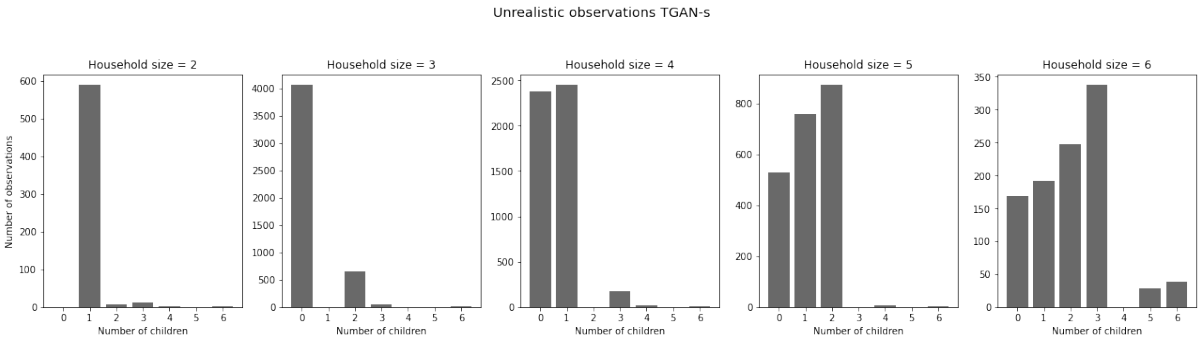

Figure 14: Unrealistic observations generated by TGANS

might lead to worse result.

Except for column analysis, analysis on the observation level is performed. The test was including satisfaction of real-world constraints in order to verify consistency. One example of it is the validation of the relationship between three correlated attributes: household size, household type, and the number of children. In Figure 14 unrealistic observations of TGANs in the combination of household size and number of children are presented. For instance, if a couple without children has a child, this row is infeasible. With the usage of our method, none of the unrealistic rows occurred due to the fact that generated agent is discarded if it doesn't satisfy the real-world constraint.

Compared to the TGANs which are completely data- driven in capturing of correlation between attributes, simulation methodology does not need to know the details about data collection details and aggregation process. This is due to the fact that conditionals can be created not only from data but also from the assumptions, domain knowledge and models (Discrete Choice Model, Machine Learning). Since the main motivation for developing the synthesis population is unavailability of the data, the high dependence on the initial sample is the factor that makes certain method extremely limited. The main reason why we chose simulation technique over state-of-the art is because we are focusing on the generation of realistic households, which gives an advantage to the simulation approach where we can embed complete control to generation process.

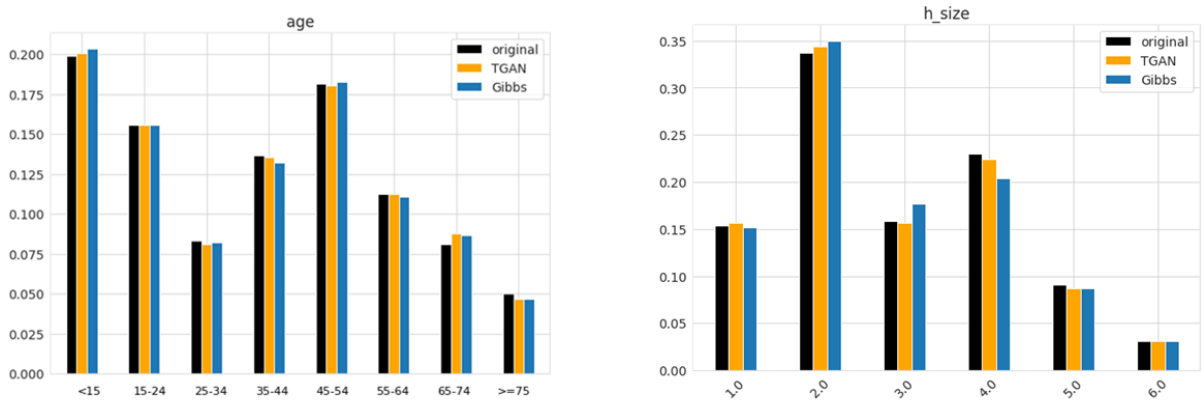

Figure 15: Comparison between GANS and GibbsSampler on census data

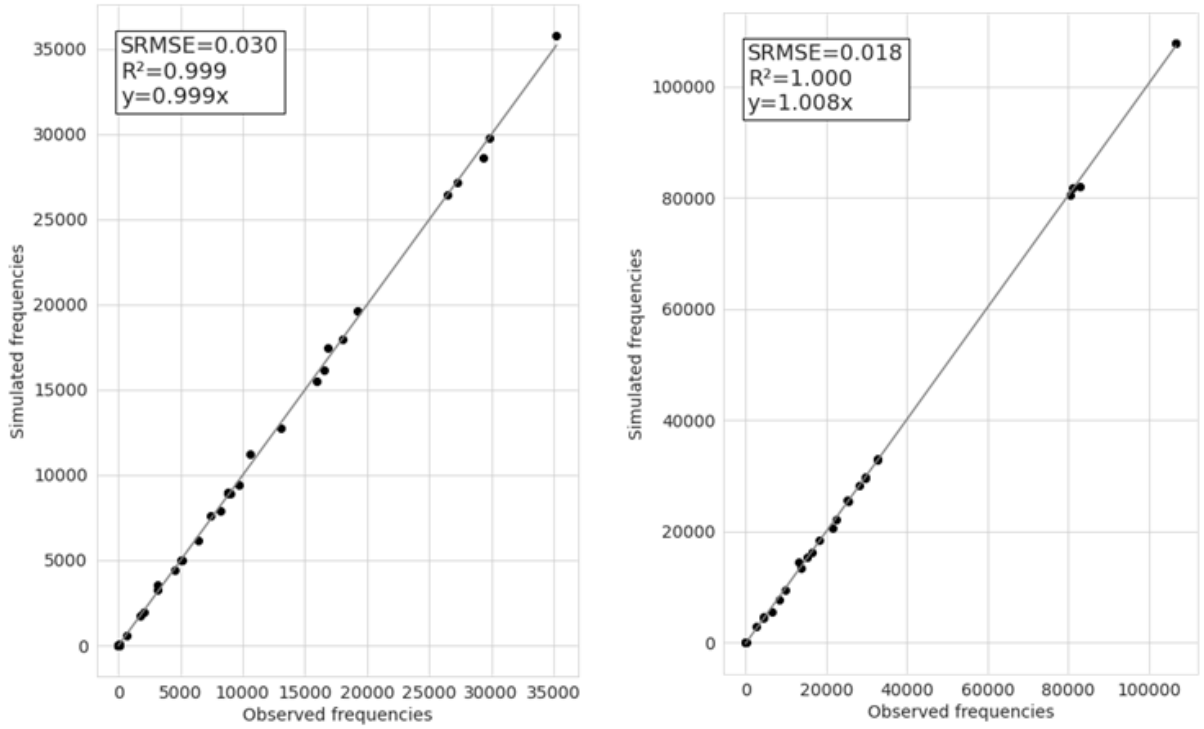

Figure 16: SRMSE of TGANS on the population of individuals (left) and the population of the households (right)

6 Conclusion

In this paper, we investigated the possibilities of simulation techniques for synthetic imputation of households. It is shown that even though the algorithm's complexity increases with the increase of dimensionality, it is possible to ensure a certain level of control within the generation process. Compared to other techniques, our method is focused on the generation of realistic observations with respect to the imposed rules. This research is part of the wider research which aims to develop a solid methodology for the synthetic household generator in one stage. The

main goal is to reduce computational time by merging generation and association processes in one. Moreover, we will investigate is it possible to enhance our methodology to capture dynamics from the evolving data, which has not been possible in state-of-the-art methods. The main challenge will be to investigate how to keep track of frequent changes to the underlying data and having the ability to apply new changes into the synthetic data more efficiently.

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

Badu-Marfo, G., B. Farooq and Z. Paterson (2020) Composite Travel Generative Adversarial Networks for Tabular and Sequential Population Synthesis, *arXiv:2004.06838 [cs, stat]*, April 2020. ArXiv: 2004.06838.

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.

Ben-Akiva, M. E., M. Bierlaire, D. McFadden and J. L. Walker (2021) *Discrete Choice Analysis*, MIT Press, Cambridge.

Ben-Akiva, M. E. and S. R. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge.

Borysov, S. S., J. Rich and F. C. Pereira (2019) How to generate micro-agents? A deep generative modeling approach to population synthesis, *Transportation Research Part C: Emerging Technologies*, **106**, 73–97, September 2019, ISSN 0968090X.

Farooq, B., M. Bierlaire, R. Hurtubia and G. Flötteröd (2013) Simulation based population synthesis, *Transportation Research Part B: Methodological*, **58**, 243–263, December 2013, ISSN 01912615.

Gargiulo, F., S. Ternes, S. Huet and G. Deffuant (2010) An Iterative Approach for Generating Statistically Realistic Populations of Households, *PLoS ONE*, **5** (1) e8828, January 2010, ISSN 1932-6203.

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

Kagho, G. O., A. Ilahi, M. Bala? and K. W. Axhausen (2020) Synthetic population of Greater Jakarta: An iterative proportional updating approach, 14 p. Artwork Size: 14 p. Medium: application/pdf Publisher: ETH Zurich.

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, 17.

Müller, K. and K. W. Axhausen (2011) Population synthesis for microsimulation: State of the art, paper presented at the *90th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2011.

Moeckel, R. (2003) Creating a Synthetic Population, 18.

Mohammadian, J. and Zhang (2010) Synthetic household travel survey data simulation, *Transportation Research Part C: Emerging Technologies*, 19.

Pougala, J., T. Hillel and M. Bierlaire (2021) Capturing trade-offs between daily scheduling choices, 27.

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, August 2016, ISSN 01912615.

Schultheiss, M.-E. (2021) Sustainable urban territories: unravelling spatial familiarity and multy-day mobility motifs, 12.

Wickramasinghe, B. N., D. Singh and L. Padgham (2020) Building a large synthetic population from Australian census data, *arXiv:2008.11660 [cs, stat]*, August 2020. ArXiv: 2008.11660.

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

Ye, X., K. Konduri, R. M. Pendyala, B. Sana and P. A. Waddell (2009) A methodology to match distributions of both household and person attributes in the generation of synthetic populations, paper presented at the *88th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 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, January 2014, ISSN 0361-1981, 2169-4052.