Comparing Methods for Generating a Two-Layered Synthetic Population

Boyam Fabrice Yameogo, Pascal Gastineau, Pierre Hankach, Pierre Olivier Vandanjon

Comparing Methods for Generating a Two-Layered Synthetic Population

This is the Authors post print version of the final paper published in Transportation Research Record (Nov 2020)
DOI: 10.1177/0361198120964734

Boyam Fabrice Yaméogo[1,2,3], Pascal Gastineau[4], Pierre Hankach[5] and Pierre-Olivier Vandanjon[6]

Abstract

Synthetic population is used in many transport models ranging from trip-based, hybrid trip, tour-based and activity-based models. As mobility decision depends on both individual's characteristics and family situation, generating a two-layered population that take into account not only the individual-level but also household-level is essential.

In the literature, three main categories of methods for two-layered population generation have been proposed. These categories are Synthetic Reconstruction (SR), Combinatorial Optimization (CO), and Statistical Learning (SL). SR and CO methods produce synthetic populations by means of replicating individuals, while SL methods generate a population following a joint probability estimation. However, selecting a generation process is not straightforward as it depends on input data and synthetic population characteristics. To the best of our knowledge, no clear methodology for selecting between these methods exists.

The main objectives of this paper is to provide 1) a detailed description of the available methods 2) a comparison between these methods and 3) a decision-making procedure for selecting between these methods. The description and comparison of the methods relies on different criteria: marginals availability, sample size, number of potential attributes that can be handled, population size to generate, possibility of zero-cell problem, etc. The advantages and shortcomings of each method are illustrated, and method performance is assessed. The decision-making procedure is carried out through the proposal of a decision tree.

Researchers and practitioners have now access to a comprehensive and unified framework to select the appropriate method depending on available data and features of their modeling purposes.

[1]AME-EASE, Univ Gustave Eiffel, IFSTTAR, F-44344 Bouguenais, France, Email: boyam-fabrice.yameogo@univ-eiffel.fr
[2]French Environment and Energy Management Agency 20, avenue du Grésillé - BP 90406 49004 Angers Cedex 01 France
[3]SNCF TER Mobilités Pays de la Loire, 44000 Nantes, France
[4]AME-EASE, Univ Gustave Eiffel, IFSTTAR, F-44344 Bouguenais, France, Email: pascal.gastineau@univ-eiffel.fr
[5]MAST-LAMES, Univ Gustave Eiffel, IFSTTAR, F-44344 Bouguenais, France, Email: pierre.hankach@univ-eiffel.fr
[6]AME-EASE, Univ Gustave Eiffel, IFSTTAR, F-44344 Bouguenais, France, Email: pierre-olivier.vandanjon@univ-eiffel.fr

Corresponding author:
Boyam Fabrice Yaméogo
Email: boyam-fabrice.yameogo@univ-eiffel.fr

Keywords

Synthetic Population Generation, Multi-level, Microsimulation, Simultaneous Control

Introduction

In recent years, Agent-Based Models (ABMs) have become more popular for simulating transport systems given their powerful capacity to simulate complex systems with a large number of agents (1; 2). For this reason, ABMs have also been applied to other fields, e.g. health (3; 4), economic policy evaluation (5; 6), geography (7).

In transport applications, disaggregated population information (individuals, households, etc.) is used as ABM input data. However, for privacy reasons, no comprehensive dataset containing the socio-demographic characteristics of individuals and households at a small geographical scale exists. Only the characteristics of a sample and aggregated statistics (marginal distributions) of socio-demographic variables of the actual population are known. In order to perform agent-based simulation, a necessary intermediate step therefore is the generation of a "synthetic population" representative of the actual population derived from available data. The resulting synthetic population constitutes a simplified microscopic representation of the actual population since only those variables of interest are to be reproduced (8).

Most of the approaches developed to generate a synthetic population (for transport or other applications) have emphasized generating either individual-centered or household-centered populations. For example, Iterative Proportional Fitting (IPF), which is by far the most widely used technique for generating a synthetic population, does not enable generating populations linking households with individuals (and introducing controls at both levels). In using this technique, a population of individuals can be generated but without linking any individual characteristics to household information. For transport modeling, this absence of a link is clearly a shortcoming, since an individual's mobility decision depends on both his/her characteristics and family situation (9; 10; 11). Highlighted herein is the need to generate synthetic populations that take into account not only the individual level but also household-level information. Therefore, only synthetic population generation methods that are capable of taking these two levels into account are considered in this paper.

The specific case of generating a two-layered synthetic population must satisfy certain conditions, namely:

- maintaining the hierarchical structure of the data by associating individual and household variables in the most optimal way possible;
- reflecting the heterogeneity of the distribution of households and individuals between geographical areas (12);
- reproducing the interdependencies among agents in the same household (13);
- demonstrating the ability to fit with marginals.

To tackle this two-layered (or multi-level) synthetic population generation issue, various approaches have recently been developed. Different classifications of these methods have been proposed (8; 13; 14). Our categorization is derived from Sun et al. (13). Three main categories of methods are thus considered: Synthetic Reconstruction (SR), Combinatorial Optimization (CO) and Statistical Learning (SL). SR and CO methods produce synthetic populations by means of replicating individuals, while

SL methods draw a population following a joint probability estimation. SR methods are deterministic whereas CO and SL are both stochastic.

Figure 1 provides an overview of these three categories, along with their primary associated methods for generating a synthetic population of individuals and households. Some of these methods have been implemented as commercial or open source software. All the implemented population synthesizers fall under the SR category. Transport practitioners most often use these population synthesizers without necessarily being aware of their suitability to their applications (input data, population size, sample size, etc.) and available alternatives.

The objective of this paper is to introduce and assess the various approaches based on the criteria listed above. A few studies have already carried out a review of these methods. Hermes and Poulsen (15) analysed generation methods but only at a single level: either household or individual; and do not take into account SL methods. Müller and Axhausen (16; 17) reviewed two-layered methods but only covered methods from the SR family. Chapuis and Taillandier (8) analysed two-layered generation methods classifying them in two main categories (grouping SR and SL methods into one category and maintaining CO methods in a distinct one). However the review was limited to articles published in one journal (Journal of Artificial Societies and Social Simulation) over a five-years period. None of the above reviews proposed neither a detailed comparison on identified criteria nor a decision-making procedure to help practitioners select between them.

The main contribution of this paper is threefold. First, we provide a detailed description of the characteristics of the three approaches and the main methods of each approach. Second, we compare the methods using a set of criteria that illustrates the advantages and shortcomings of each one, as well as their applicability depending on available input data. Third, we propose a decision-making procedure, in the form of a decision tree, for selecting among these methods. The decision criteria include: sample representativeness, availability of marginal distribution, number of considered attributes, and population size.

The remainder of this paper is organized as follows. First, the approaches will be described according to their classification into three main categories: Synthetic Reconstruction (Section 2), Combinatorial Optimization (Section 3), and Statistical Learning (Section 4). Section 5 will discuss the three previous sections and provide both a comparison table and decision tree for selecting among the methods. The final section offers a conclusion and outlook.

Synthetic Reconstruction

This category of methods is the most widely used to generate synthetic populations and relies on a two-step procedure: fitting and allocation. The fitting step entails assigning positive weights to the individuals and households in the sample in such a way that the sum of these weights corresponds to the marginal sums. The resulting weights are typically non-integers (i.e. fractions of individuals). During the allocation step, these non-integer weights are converted into integer weights, and individuals are replicated in proportion to their weights. For example, an individual assigned a weight of 8 will be replicated 8 times in the synthetic population. Once this procedure has been completed, each member of the synthetic population will have clearly defined attributes. These two steps form the foundation of the synthetic reconstruction approach (14). Figure 2 shows a simplified flowchart of how these steps are implemented.

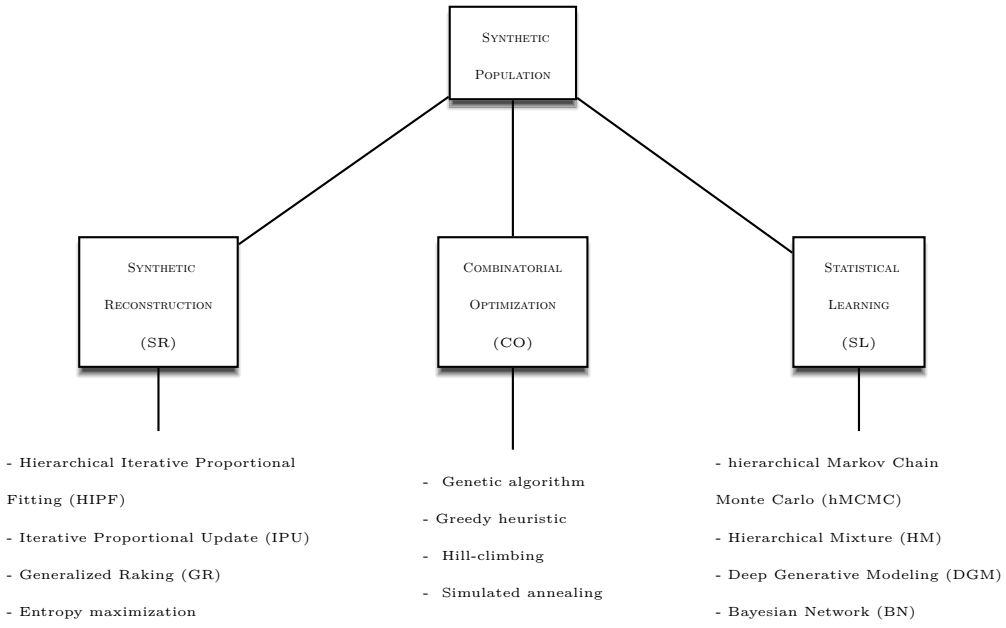

Figure 1: Different approaches to create a two-layered synthetic population (households and individuals)

SR methods are deterministic, which means that depending on the sample used, the weights obtained during the fitting stage never vary subsequently.

Most SR methods require having both a sample and aggregate data. The underlying assumption here is that the sample represents the true correlation structure among the attributes (18), and the interactions present in the sample are, to a great extent, preserved for the synthetic agents (16). The sample therefore needs to be consistent and representative while containing at least one observation for each type of individual in the actual population. This last constraint avoids the so-called "zero-cell problem", whereby in the absence of an observation in the sample of a specific individual present in the actual population, then SR methods cannot generate this specific individual.

Barthelemy and Toint (19) developed an alternative SR method that does not use a sample (and hence is called a sample-free method). Although this approach provides more flexibility in terms of data needs (only marginals are required), it ignores the dependence between household and individual levels without maintaining consistency between these two layers (20; 21). Therefore this method fails to satisfy the conditions we specified in the introduction for two-layered synthetic population generation and has not been considered in this paper.

Iterative Proportional Fitting (IPF) approach and its adaptations to the two-layered problem

A commonly used SR technique is Iterative Proportional Fitting (IPF). IPF adjusts a contingency table constructed from the sample in order to match marginal distributions; it minimizes discrimination information, and the contingency table is derived by iteratively adjusting the cell values so as to minimize deviation from the marginals of the attributes. Interested readers are referred to Beckman et al. (22) and Pritchard and Miller (23) for a complete description of this technique.

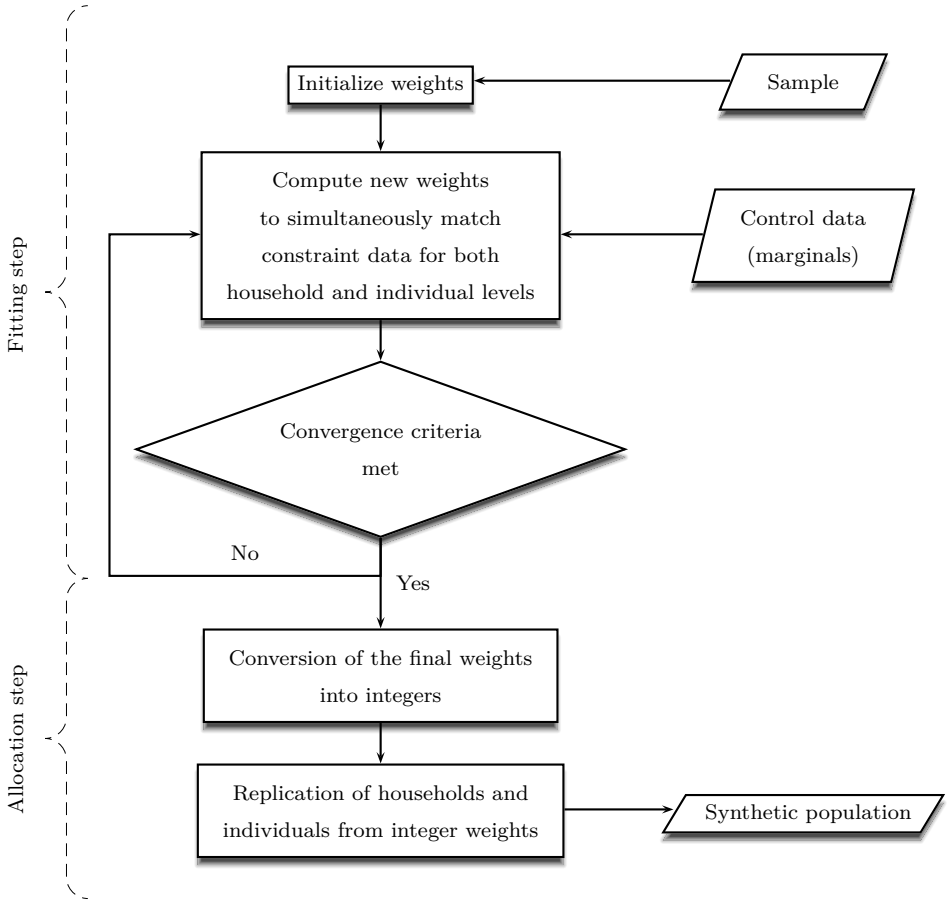

Figure 2: Simplified flowchart of Synthetic Reconstruction methods to generate a two-layered synthetic population (households and individuals)

The IPF technique has several advantages, the first being its guarantee that marginals will be fit and the correlation structure will be preserved (24). As a second advantage, the distribution of variables not used to match marginal distributions are reasonably estimated (22). IPF is simple, fast, accurate (25) and moreover offers a pervasive method to estimate socioeconomic variables(26). In its original formulation however, IPF cannot simultaneously estimate both household and individual level attributes.

Some IPF-based approaches have sought to deal with both household and individual attributes. Arentze et al. (27) proposed a two-step procedure: in the first step, marginal distributions of individuals are converted into a marginal distribution of households by using a relation matrix; in the second step, the resulting household marginal distribution serves to constrain a multiway table of household counts. Guo and Bhat (28) and Auld and Mohammadian (29) estimated the joint distributions of attributes for households and persons separately by means of IPF techniques. Household samples can then be drawn into the synthetic population iteratively depending on how well the types of households and their members fit to the estimated joint distributions at both the household and person levels (20). According to Pritchard and Miller's generation process (23), the person population is first fitted with IPF. These authors selected a set of shared attributes between the person and household levels and performed a

second IPF at the household level in considering the shared attributes as marginals. At the generation stage, households and persons are matched through these shared attributes, and a conditional Monte Carlo method is employed to assign persons to households (20).

In all these studies, the joint distribution of household and individual-level attributes is fitted either separately or sequentially, which does not guarantee consistency between these two levels.

Two-layered generation methods

Iterative Proportional Update (IPU), Hierarchical Iterative Proportional Fitting (HIPF), entropy maximization (ent) and Generalized Raking (GR) all generate an intrinsically two-layered population. As such, they generate populations of persons grouped into households by computing household-level weights that satisfy the marginals at both the household and person levels.

Iterative Proportional Update (IPU) (30) and Hierarchical IPF (HIPF) (17; 31) are intended to simultaneously estimate both household and individual-level attributes. These methods respectively compute the factors and weights linked to individual and household records using an iterative approach with a cross-categorization of individual types into household types (8). The HIPF approach constantly switches between the household domain and the person domain, through employing an entropy-optimizing adjustment step (32). In the IPU approach, weights are adjusted to satisfy household-level constraints first and then updated to satisfy person-level constraints.

The population synthesis problem can also be formulated as a constrained optimization problem. The goal of this formulation is to find a weight for each household such that the distributions of characteristics in the weighted sample match the exogenously given distributions in the population, for both household and person characteristics (33). This optimization problem can be solved using two approaches: the entropy-optimizing procedure, and the calibration estimation.

The entropy-optimizing procedure is based on the thermodynamic concept of entropy and seeks the most likely configuration of elements within a constrained situation (34). This method directly optimizes an information-based similarity metric of the weights and therefore introduces the least possible amount of new information, just like IPF (31). The technique can offer an alternative to IPU and HIPF since it also generates weights associated with individual and household records.

The aim of a calibration estimation is to weight a sample of individuals using information available on certain variables, called calibration variables, which are the control variables or marginals obtained from the aggregated data. Deville et al. (35) called these methods "generalized raking" (GR) as they generalize the classical statistical raking ratio method. This is done by minimizing a distance measure between initial weights and final weights, subject to calibration equations. For example, for a qualitative variable, the weighted counts of the categories (attributes) of this variable in the sample, after calibration, will be equal to the known counts of the population.

Different distance functions can be used to solve this optimization problem: linear method, multiplicative method (or raking ratio method), logit method, and truncated linear method. Generalized raking directly adjusts weights to satisfy both individual and household-level constraints and is a viable alternative to the multi-level fitting methods like HIPF or IPU.

Almost all population synthesizers fall under the SR category. For example, PopulationSim (36) and PopSynIII (37) are both based on the entropy maximization procedure, TransCAD (38) uses an enhanced approach of IPU. For a more in-depth knowledge of the characteristics of some synthetic population generators, the interested readers are referred to (31) (chapter 2).

Combinatorial Optimization

The second category of approaches refers to Combinatorial Optimization (CO) techniques. CO-based techniques are two-layered since they can directly generate a list of households and persons (39).

Similar to SR methods, CO requires information on both the sample and marginal level, and the synthetic population is obtained by a replication of individuals (without explicitly determining the joint distribution across all controlled attributes). But unlike SR methods, CO is a stochastic process. Another difference is that the CO methods result in entire individuals or households being allocated to each zone, whereas SR methods yield fractions of individuals or households being allocated to each zone (40). One key advantage of CO methods is that data requirements remain less restrictive than those for SR methods (14). However, a major drawback of CO methods is their computational complexity for large population sizes. As population size grows, CO methods cannot guarantee the optimal solution and require too much computational time (34). For this reason, CO-based methods are not widely used.

A description of the method has been given by (41) and (14). A simplified flowchart of how CO techniques function is presented in Figure 3. Starting from a sample of households and individuals, the sample is first divided into exclusive groups (e.g. small geographic areas), for which some control variables (marginals) are available. The process then becomes iterative:

- an initial number of households is randomly chosen from the sample in each small geographic area in order to form an appropriately sized group; next, the fitness of the selected population to the marginals for the newly formed group is estimated;
- •one household is either added, replaced or swapped with another from the sample, and the goodness of fit is recalculated. If the replacement improves the fit, this household is kept; otherwise, this household will be removed from the zone. The process is repeated until a certain goodness-of-fit threshold or an arbitrarily fixed number of iterations has been reached.

An example of a statistical measure of fit that could be used is the overall relative sum of squared Z scores (RSSZ) (42; 43). Various optimization procedures have been proposed to estimate the goodness-of-fit measure, i.e.: genetic algorithm (44), greedy heuristic (45), hill-climbing (46), and simulated annealing (47).

Most studies using CO methods have focused on synthesizing a population with a handful of attributes, typically either individual or household. For example, Ryan et al. (43)) showed that CO methods produce more accurate populations than IPF, although their study was conducted on a relatively small population with just three individual level variables. Few studies have sought to simultaneously synthesize populations with controls at both the household and person levels (46; 39).

The former uses a hill-climbing algorithm to resolve the combinatory problem, while the latter employs its own algorithm, called the fitness-based synthesis (FBS).

Statistical Learning

The third approach that may be used to generate a two-layered synthetic population is statistical learning (SL), also known as a simulation-based approach. SL focuses on the joint distribution of all attributes in the sample by directly estimating a probability for each combination, including those not observed in the sample (13). In fact, in the actual population, individuals are characterized by discrete or continuous attributes, which feature a unique joint distribution. Due to privacy concerns and data availability, only partial views (samples, marginals or conditional-marginal distribution) of this unique joint distribution are available. The aim of SL methods is to use these partial views to

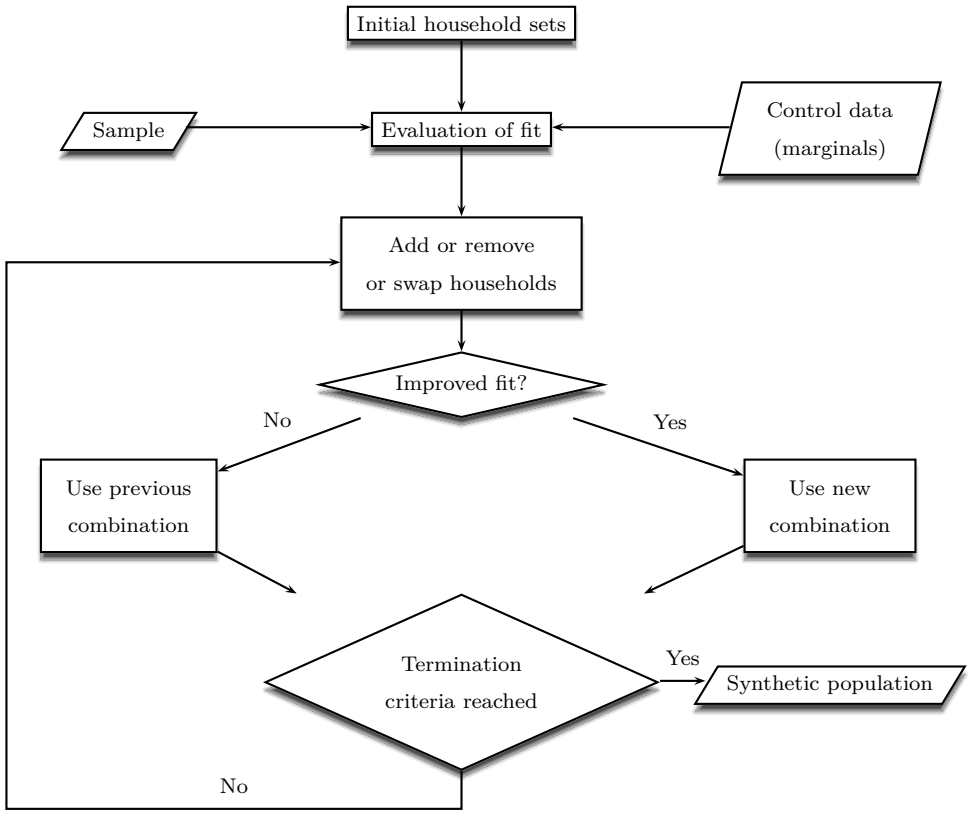

Figure 3: Simplified flowchart of Combinatorial Optimization methods to generate a two-layered synthetic population (adapted from (15))

construct a synthetic population, whereby the empirical distribution of this synthetic population is as close as possible to the unique joint distribution in the actual population. The basic premise behind this approach is given by Farooq et al. (18) and Sun et al. (13).

SL methods offer greater flexibility in terms of data requirements and data sources; in general, they demonstrate good performance in dealing both with the lack of heterogeneity problem raised in both SR and CO (13) and with small sample sizes. Also, these methods provide a more systematic way of imputing or interpolating data and are capable of reproducing agents that do not exist in the sample. Another advantage with SL is that these methods can generate a synthetic population from sample data only, when marginal information at the zonal level is not available.

However, a major drawback of SL methods is that they fail to satisfy the conditional distributions while satisfying the marginal distributions of all variables simultaneously. During the population synthesis process, when marginals are available, it is indeed necessary to precisely match the observed marginal distributions with the generated population at the zonal level. Sun and Erath (21) and Sun et al. (13) both recommend applying SR methods as a post-processing step (when marginals are available) on a synthetic sample of the population generated from SL methods so as to create a population matching the marginals. In some configurations, combining SL and SR methods actually proves to be the most relevant means for generating a synthetic population. If sample data are not comprehensive (do not contain at least one observation for each type of individual in the actual population) and representative (valid description of the study area's population(17)) of population data, yet marginals are available, then SL methods can be applied first to construct a suitable sample. Afterwards, SR

methods can be introduced to create a population matching the marginals. Figure 4 displays a simplified flowchart of how SL-based methods can be used.

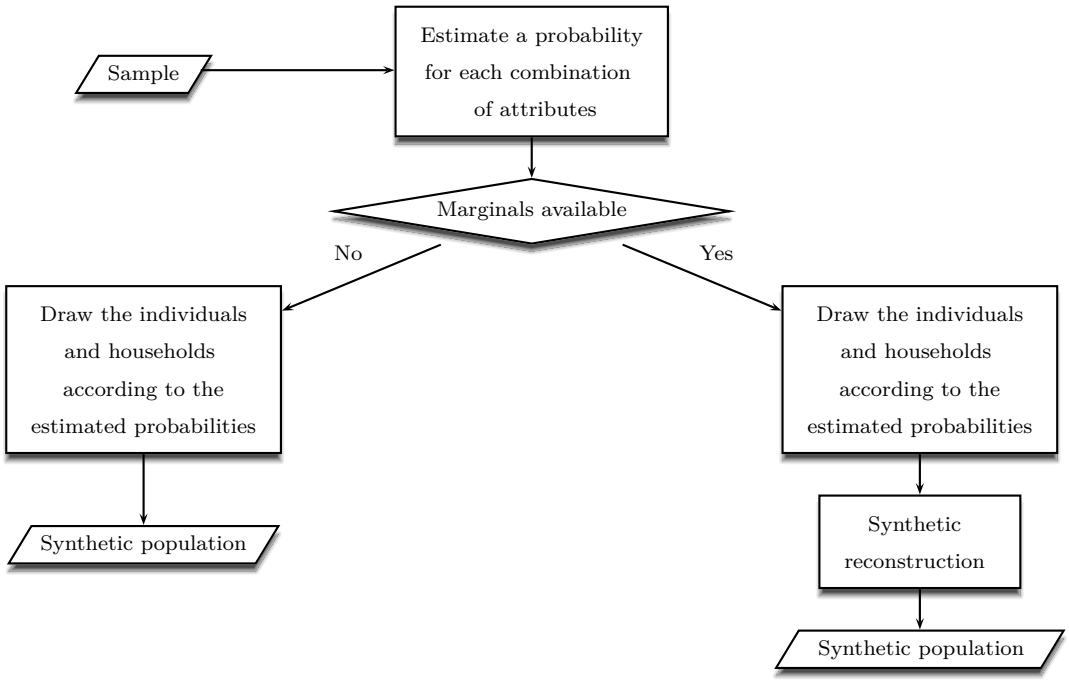

Figure 4: Simplified flowchart using Statistical Learning methods to generate a two-layered synthetic population (households and individuals)

Many SL-based methods for synthetic population generation have been proposed. Farooq et al. (18) suggested implementing Gibbs sampling and a Markov Chain Monte Carlo method (MCMC) to generate the synthetic population. This method uses the conditional-marginal distribution to simulate drawing from the joint distribution in the actual population. Farooq et al. (18) used discrete choice models to construct these conditionals. Once the conditionals are obtained, the Gibbs sampler is run and a synthetic population obtained by drawing the number of individuals/households corresponding to the required population size.

Saadi et al. (48) presented a Hidden Markov Model (HMM) to generate a synthetic population. These authors considered that each attribute represents a state. The number of states is known and each state of attribute i is connected to all states of the subsequent attribute i + 1. An HMM is a Markov process characterized by two parameters, i.e.: the symbol emission describing the emission probabilities of each state, and the transition probabilities of changing to another state. The basic idea behind this process is to position the attributes in a certain sequence (by descending order of the number of attributes) and then estimate these probabilities. However, both Markov process-based methods (MCMC and HMM) only pay slight attention to the hierarchical household structure (48; 13). In order to generate a two-layered synthetic population, Casati et al. (49) proposed an extension of Farooq et al.'s MCMC model (18) to account for the hierarchical structure of households and individuals; they named this process hierarchical MCMC (or hMCMC). hMCMC entails ordering the agents living in the same household according to their household roles. Three roles are thus defined by a rule-based approach: owners (persons with the highest income in their household or chosen by virtue of another selection criterion), intermediate types (spouses and children), and others. To generate a household, all owners are first generated according to the methodology used in (18), thus taking into account the variables

characterizing the household (as all members of a household share the same household attributes). If the household size is greater than 1, then the other household members are generated as follows:

1. the conditionals of the spouse are generated by incorporating certain variables relating to the owner;
2. the conditionals of a child are generated by incorporating certain variables relating to both the owner and spouse;
3. the conditionals of others solely depend on their owners and intermediate types.

In this manner, individual and household attributes are simultaneously combined, and the association between individuals and households is respected. However, as noted by the authors, the main flaw in this method pertains to the fact that the number of categories of an attribute characterizing a certain agent type may differ when this variable is involved in the generation of its agent type and when it influences the generation of subsequent types. It may often be unclear which of the owner's variables should be used to generate intermediate types or other person types; another uncertainty lies in the implications of choosing one particular set over another.

Another SL-based method applied to population synthesis is the Bayesian Network (BN) (21; 50). A BN is a probabilistic graphical representation of the factorization of the joint probability distribution into conditional distributions, presented by means of a directed acyclic graph (DAG) (51). Sun and Erath (21) argued that a BN, by abstracting the structure of population systems using a DAG and local conditional probabilities, is able to capture and reproduce the complex dependence and higher-order interactions among a large set of variables. According to the DAG, nodes represent random variables, while arrows show the probabilistic dependencies between these variables. From the definition given by Zhang et al. (50), BN consists of two main parts: (1) structural learning to define the network structure that describes the conditional independence of the random variables through a scoring approach, and (2) parameter estimation to learn a conditional distribution of random variables given this DAG structure. The synthetic population can then be generated accordingly, by sampling from the obtained Bayesian Network. With this BN method, the hierarchical structure of the data (households and individuals) is respected. Nevertheless, learning a Bayesian structure is a computationally challenging task, particularly when a large number of attributes for each agent must be considered. As pointed out by Sun and Erath (21), a network with six nodes (where each node represents an attribute) contains some 3 million possible DAGs. Zhang et al. (50) suggested keeping the Bayesian Network as simple as possible and creating a "whitelist" defining a set of relationships in the structural learning procedure that are to be preserved and guaranteed present in the final graph. This procedure involves building the network based on expert knowledge; yet despite this, mistakes can still be made.

Sun et al. (13) proposed an SL method that relies on a hierarchical mixture (HM) modeling framework. According to this approach, the authors considered the existence of latent classes at the household level. For each household-level latent class, latent classes also exist at the individual level. The interaction between these two levels can thus be captured by using a conditional distribution on the latent class labels. The proposed framework integrates three component models (probabilistic tensor factorization, multi-level latent class model, rejection sampling) to reproduce the underlying statistical distribution between the two levels. However, as the number of attributes increases, the HM method becomes very challenging to apply. Another challenge with this approach pertains to its robustness with respect to the selection of latent variables (51).

More recently, Borysov et al. (51) proposed a deep generative modeling (DGM) approach based on a Variational Autoencoder (VAE). According to these authors, each person is represented by a vector of random variables. The objective of such a VAE (which is an unsupervised generative model based

on a deep artificial neural network) is to learn the joint distribution of all individuals in the sample. A consistent synthetic population of households and individuals with a large number of attributes can thus be generated. Implementation of this method however requires fairly advanced computer science skills.

Comparison of methods

The methods that serve to generate a representative synthetic population of households and individuals differ depending on their underlying assumptions, the approach employed or the data required. According to Sun et al. (13), let's sort these methods into three categories: synthetic reconstruction (SR), combinatorial optimization (CO), and statistical learning (SL).

The SR approach combines information from sample data and aggregate statistics (marginals) to compute the weights that reflect each sample agent's representativeness in a given zone (8). The CO methods also use the sample and aggregate data to select an appropriate combination of households that best fit the marginals. SL methods solely consider the sample and focus on the joint distribution of all attributes by estimating a probability for each combination.

The main characteristics, advantages and shortcomings of these approaches are summarized in Table 1, with a discussion of these characteristics given below:

- In order to operate, CO and SR methods require both sample data and aggregate statistics (marginals), while SL methods can produce a population from sample data only, as needed.
- The output from CO and SR methods directly fits to marginals, whereas the SL method necessitates a post-processing step in order to achieve this step.
- SL methods can produce consistent results even if the sample size is small. In contrast, sample size is critical for both CO and SR. In fact, controlling the marginal statistics of attributes with infrequent categories requires larger reference samples. For example, in their case study, Sun and Erath (21) showed that for IPF results to be similar to BN results, the sample size needs to be greater than 20%.
- SR methods have been proven to be scalable. Many papers report on SR applications for generating large synthetic populations with a large number of attributes and categories (30; 31; 33; 34).

  CO methods are hindered by their computational complexity, which winds up compromising their scalability when generating large synthetic populations. Most applications of this method have been restricted to small population sizes (46; 39). Some SL methods are hindered by the scalability problem when the number of attributes is large. For Bayesian Networks and Hierarchical Mixtures, it is preferable for the number of attributes to be small, as pointed out by Borysov et al (51) and Zhang et al. (50). In fact, a Bayesian network with six nodes (corresponding to attributes) contains some 3 million possible DAGs, and this number becomes 1.1 billion with a network of seven nodes (21).
- SR and CO methods generate a synthetic population by reproducing a reference sample. Consequently, combinations of attributes that are not part of the sample cannot be generated; this restriction is known as the "zero-cell problem" and can occur more frequently when dealing with small geographic zones. The zero-cell problem is inherently present in both SR and CO methods. A pre-processing procedure to eliminate the zero cell problem is often necessary. SL methods, by virtue of estimating a joint population distribution from the sample, effectively treat this issue.

Table 1: A brief summary of different methods for generating a two-layered synthetic population.

Characteristics	**Synthetic Reconstruction** IPU, HIPF, GR, entropy optimization	**Combinatorial Optimization**	**Statistical Learning**			
			hierarchical Markov Chain Monte Carlo	Bayesian Network	Hierarchical Mixture	Deep Generative Modeling
Main processing algorithms	A deterministic algorithm using a two step procedure: fitting and allocation.	A stochastic algorithm based on a combination of individuals and households to satisfy a fitness criterion.	Definition of different groups of agent types and computation of conditional distribution from a Gibbs sampling for each type of agent to estimate the joint probability distribution.	Learning both model structure and model parameter to estimate the joint probability distribution.	A framework integrating three models to estimate the joint probability distribution.	A method using a Variational Autoencoder (VAE) to estimate the joint probability distribution.
			+ Generation of a pool of individuals respecting the joint distribution			
Data preparation	Sample and aggregate data on both the individual and household levels.	Sample and aggregate data on both the individual and household levels.	Only a sample (both the individual and household levels).			
Fitting marginals	Yes	Yes	Requires a post-processing step using SR methods like GR to fit marginals.			
Sample size	Produces good results if the sample size is sufficiently large	Produces good results if the sample size is sufficiently large	Can be used with a small-sized sample			
Number of potential attributes that can be handled	Large	Large	Large	Few		Large
Zero-cell problem	Possible	Possible	No zero-cell problem			
Dissemination	Widely used	Not extensively used	Rarely used	Not extensively used	Rarely used: these methods have been implemented recently and only tested by their designers	

Decision-making procedure

The choice of appropriate method for generating a two-layered synthetic population is critical to transport modelers. This choice is partially driven by the amount, type and quality (representativeness and comprehensiveness) of available data (52). In order to facilitate this process and in addition to the method comparison presented herein, the decision tree shown in Figure 5 identifies the most suitable methods for generating a two-layered synthetic population, based on the following parameters: available input data (sample and marginals), number of attributes to be considered, and synthetic population size.

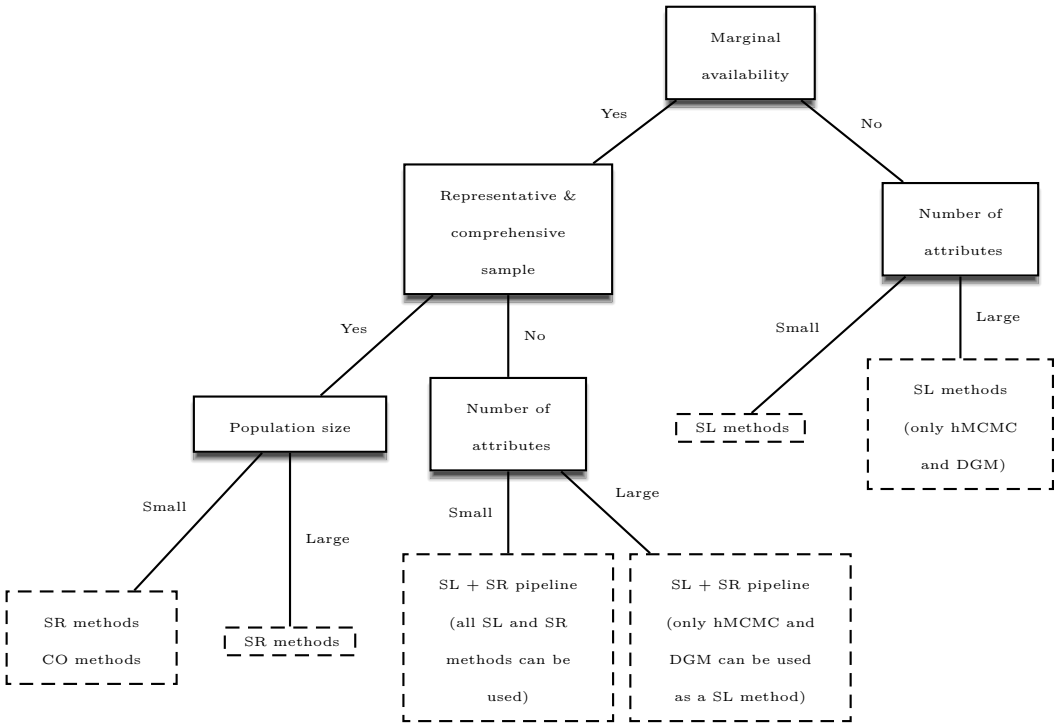

Figure 5: Decision Tree to identify the most suitable methods for generating a two-layered synthetic population.

Our starting assumption is that there is a sample with both household and person characteristics (census or household travel surveys data) as it is the case in most studies.

When marginal data are available and the sample is representative (at least 5% of the population), SR methods are preferred for generating a large population size (millions of persons and households with a high number of attributes as in (32; 37; 11)).

CO methods are limited to the generation of small size populations. This is due to the computational complexity of these methods. For example, Hafezi and Habib (53) and Ma and Srinivasan (39) generate populations with fewer than 60.000 persons and 30.000 households.

When the sample size is very small (less than 5%) but marginal data are available, combining the SL and SR methods is more relevant. SL methods can be applied first to construct a sample of suitable size, afterwards SR methods can be applied to create a population matching the marginals. Following this line, Borysov et al. (51) in Denmark and Sun et al. (13) in Singapore generated a population from household travel survey samples corresponding to 2.5% and 1% of the total population, respectively.

If marginals are not available, only SL methods can be applied to generate the population.

Conclusion

For transport modeling, the ability to generate a two-layered population (i.e., persons and households) is essential since an individual's mobility decisions depend on both his/her characteristics and family situation. Generating such a population is one of the most challenging problems in population synthesis (11). This paper has provided a state of practice of the main methods available for generating a two-layered synthetic population. It has also proposed a decision tool (in the form of a decision tree) for selecting between these methods. To the best of our knowledge, no other study has provided such a detailed analysis of these methods.

The various methods employed to tackle the two-layered (or multi-level) synthetic population generation problem have been reviewed. A classification of these methods, into three categories (Synthetic Reconstruction (SR), Combinatorial Optimization (CO) and Statistical Learning (SL)) based on Sun et al. (13), has been provided, along with a detailed description of the characteristics of methods found in each of the three categories. Different criteria have been used to draw a method comparison, namely: processing approach, output format, solution quality (fitting marginals), input data characteristics, mandatory sample size, number of attributes potentially handled, data preparation, the zero-cell problem, and dissemination. This comparison illustrates the advantages and shortcomings of each method, as well as its applicability depending on available input data. However, the paper does not deal with the cases where no initial sample is available. In order to facilitate the choice of an appropriate method for generating a two-layered synthetic population, a decision tree is proposed. This choice is driven by the amount, type and quality (representativeness and comprehensiveness) of available data and synthetic population characteristics. Decision criteria include: marginals availability, number of attributes to be considered, and synthetic population size. Facing the problem of generating a synthetic two-layered population, transport modelers have now access to a comprehensive and unified framework to select the appropriate method depending on available data and features of their modeling purposes. We here only consider the generation of synthetic population based on census or surveys data. We are nevertheless witnessing the emergence of heterogeneous data sources derived from big data that may be exploited in the context of synthetic population generation. The next step would be to look for a methodology to integrate this new type of data. The deep learning methods seem to be a promising way to do it.

References

[1] Kickhöfer B and Kern J. Pricing local emission exposure of road traffic: An agent-based approach. *Transportation Research Part D: Transport and Environment* 2015; 37: 14–28.

[2] Hörl S, Balac M and Axhausen KW. A first look at bridging discrete choice modeling and agent-based microsimulation in matsim. *Procedia computer science* 2018; 130: 900–907.

[3] Tomintz MN, Clarke GP and Rigby JE. The geography of smoking in leeds: estimating individual smoking rates and the implications for the location of stop smoking services. *Area* 2008; 40(3): 341–353.

[4] Edwards KL and Clarke G. Simobesity: combinatorial optimisation (deterministic) model. In *Spatial Microsimulation: A reference guide for users*. Springer, 2013. pp. 69–85.

[5] Avram S, Figari F, Leventi C et al. The distributional effects of fiscal consolidation in nine eu countries. Technical report, Euromod working paper, 2013.

[6] Sutherland H and Figari F. Euromod: the european union tax-benefit microsimulation model. *International Journal of Microsimulation* 2013; 6(1): 4–26.

[7] O'Sullivan D. Geographical information science: agent-based models. *Progress in Human Geography* 2008; 32(4): 541–550.

[8] Chapuis K and Taillandier P. A brief review of synthetic population generation practices in agent-based social simulation. In *SSC2019, Social Simulation Conference*.

[9] Loo BP and Lam W. A multilevel investigation of differential individual mobility of working couples with children: a case study of hong kong. *Transportmetrica A: Transport Science* 2013; 9(7): 629–652.

[10] Kalter MJO and Geurs KT. Exploring the impact of household interactions on car use for home-based tours: a multilevel analysis of mode choice using data from the first two waves of the netherlands mobility panel. *European journal of transport and infrastructure research* 2016; 16(4): 698–712.

[11] Fournier N, Christofa E, Akkinepally AP et al. Integrated population synthesis and workplace assignment using an efficient optimization-based person-household matching method. *Transportation* 2020; : 1–27.

[12] Münnich R and Schürle J. On the simulation of complex universes in the case of applying the german microcensus. *DACSEIS Research Paper Series No 4* 2003; .

[13] Sun L, Erath A and Cai M. A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological* 2018; 114: 199–212.

[14] Templ M, Meindl B, Kowarik A et al. Simulation of synthetic complex data: The r package simpop. *Journal of Statistical Software* 2017; 79(10): 1–38.

[15] Hermes K and Poulsen M. A review of current methods to generate synthetic spatial microdata using reweighting and future directions. *Computers, Environment and Urban Systems* 2012; 36(4): 281–290.

[16] Müller K and Axhausen KW. Population synthesis for microsimulation: State of the art. *Arbeitsberichte Verkehrs-und Raumplanung* 2010; 638.

[17] Müller K and Axhausen KW. Multi-level fitting algorithms for population synthesis. *Arbeitsberichte Verkehrs-und Raumplanung* 2012; 821.

[18] Farooq B, Bierlaire M, Hurtubia R et al. Simulation based population synthesis. *Transportation Research Part B: Methodological* 2013; 58: 243–263.

[19] Barthelemy J and Toint PL. Synthetic population generation without a sample. *Transportation Science* 2013; 47(2): 266–279.

[20] Zhu Y and Ferreira Jr J. Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. *Transportation Research Record* 2014; 2429(1): 168–177.

[21] Sun L and Erath A. A bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies* 2015; 61: 49–62.

[22] Beckman RJ, Baggerly KA and McKay MD. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice* 1996; 30(6): 415–429.

[23] Pritchard DR and Miller EJ. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation* 2012; 39(3): 685–704.

[24] Rich J and Mulalic I. Generating synthetic baseline populations from register data. *Transportation Research Part A: Policy and Practice* 2012; 46(3): 467–479.

[25] Lovelace R and Ballas D. 'truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems* 2013; 41: 1–11.

[26] Choupani AA and Mamdoohi AR. Population synthesis using iterative proportional fitting (ipf): A review and future research. *Transportation Research Procedia* 2016; 17: 223–233.

[27] Arentze T, Timmermans H and Hofman F. Creating synthetic household populations: problems and approach. *Transportation Research Record* 2007; 2014(1): 85–91.

[28] Guo JY and Bhat CR. Population synthesis for microsimulating travel behavior. *Transportation Research Record* 2007; 2014(1): 92–101.

[29] Auld J and Mohammadian A. Efficient methodology for generating synthetic populations with multiple control levels. *Transportation Research Record* 2010; 2175(1): 138–147.

[30] Ye X, Konduri K, Pendyala RM et al. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In *88th Annual Meeting of the Transportation Research Board, Washington, DC*.

[31] Müller K. *A generalized approach to population synthesis*. PhD Thesis, ETH Zurich, 2017.

[32] Müller K and Axhausen KW. Hierarchical ipf: Generating a synthetic population for switzerland. *paper presented at the 51st Congress of the European Regional Science Association* 2011; .

[33] Bar-Gera H, Konduri K, Sana B et al. Estimating survey weights with multiple constraints using entropy optimization methods. In *88th Annual Meeting of the Transportation Research Board, Washington, DC*.

[34] Lee DH and Fu Y. Cross-entropy optimization model for population synthesis in activity-based microsimulation models. *Transportation Research Record* 2011; 2255(1): 20–27.

[35] Deville JC, Särndal CE and Sautory O. Generalized raking procedures in survey sampling. *Journal of the American statistical Association* 1993; 88(423): 1013–1020.

[36] Paul BM, Doyle J, Stabler B et al. Multi-level population synthesis using entropy maximization-based simultaneous list balancing. In *97th Annual Meeting of the Transportation Research Board, Washington, DC*.

[37] Vovsha P, Hicks JE, Paul BM et al. New features of population synthesis. In *94th Annual Meeting on Transportation Research Board, Washington, DC*.

[38] Balakrishnaa R, Sundarama S and Lam J. An enhanced and efficient population synthesis approach to support advanced travel demand models. In *99th Annual Meeting of the Transportation Research Board, Washington, DC*.

[39] Ma L and Srinivasan S. Synthetic population generation with multilevel controls: A fitness-based synthesis approach and validations. *Computer-Aided Civil and Infrastructure Engineering* 2015; 30(2): 135–150.

[40] Lovelace R and Dumont M. *Spatial microsimulation with R*. Chapman and Hall/CRC, 2016.

[41] Voas D and Williamson P. An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography* 2000; 6(5): 349–366.

[42] Voas D and Williamson P. Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling* 2001; 5(2): 177–200.

[43] Ryan J, Maoh H and Kanaroglou P. Population synthesis: Comparing the major techniques using a small, complete population of firms. *Geographical Analysis* 2009; 41(2): 181–203.

[44] Birkin M, Turner A and Wu B. A synthetic demographic model of the uk population: Methods, progress and problems. In *Regional Science Association International British and Irish Section, 36th Annual Conference*.

[45] Srinivasan S, Ma L and Yathindra K. Procedure for forecasting household characteristics for input to travel-demand models. project report of university of florida, gainesville; florida department of transportation. Technical report, TRC-FDOT-64011-2008, 2008.

[46] Abraham JE, Stefan KJ and Hunt JD. Population synthesis using combinatorial optimization at multiple levels. In *91th Annual Meeting of the Transportation Research Board, Washington, DC*.

[47] Harland K, Heppenstall A, Smith D et al. Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation* 2012; 15(1).

[48] Saadi I, Mustafa A, Teller J et al. Hidden markov model-based population synthesis. *Transportation Research Part B: Methodological* 2016; 90: 1–21.

[49] Casati D, Müller K, Fourie PJ et al. Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. *Transportation Research Record* 2015; 2493(1): 107–116.

[50] Zhang D, Cao J, Feygin S et al. Connected population synthesis for transportation simulation. *Transportation Research Part C: Emerging Technologies* 2019; 103: 1–16.

[51] Borysov SS, Rich J and Pereira FC. How to generate micro-agents? a deep generative modeling approach to population synthesis. *Transportation Research Part C: Emerging Technologies* 2019; 106: 73–97.

[52] Rich J. Large-scale spatial population synthesis for denmark. *European Transport Research Review* 2018; 10(2): 63.

[53] Hafezi MH and Habib MA. Synthesizing population for microsimulation-based integrated transport models using atlantic canada micro-data. *Procedia Computer Science* 2014; 37: 410–415.