Hidden Markov Model-based population synthesis

Ismaïl Saadi[a,*], Ahmed Mustafa[a], Jacques Teller[a], Bilal Farooq[b], Mario Cools[a]

[a] *University of Liège, ArGEnCo, Local Environment Management & Analysis (LEMA), Allée de la Découverte 9, Quartier Polytech 1, 4000 Liège, Belgium*

[b] *Department of Civil, Geotechnical, and Mining Engineering, Ecole Polytechnique de Montréal, 2500 Chemin de Polytechnique, Montréal, Canada*

Abstract

Micro-simulation travel demand and land use models require a synthetic population, which consists of a set of agents characterized by demographic and socio-economic attributes. Two main families of population synthesis techniques can be distinguished: (a) fitting methods (iterative proportional fitting, updating) and (b) combinatorial optimization methods. During the last few years, a third outperforming family of population synthesis procedures has emerged, i.e., Markov process-based methods such as Monte Carlo Markov Chain (MCMC) simulations. In this paper, an extended Hidden Markov model (HMM)-based approach is presented, which can serve as a better alternative than the existing methods. The approach is characterized by a great flexibility and efficiency in terms of data preparation and model training. The HMM is able to reproduce the structural configuration of a given population from an unlimited number of micro-samples and a marginal distribution. Only one marginal distribution of the considered population can be used as a boundary condition to "guide" the synthesis of the whole population. Model training and testing are performed using the Survey on the Workforce of 2013 and the Belgian National Household Travel Survey of 2010. Results indicate that the HMM method captures the complete heterogeneity of the micro-data contrary to standard fitting approaches. The method provides accurate results as it is able to reproduce the marginal distributions and their corresponding multivariate joint distributions with an acceptable error rate (i.e., SRSME=0.54 for 6 synthesized attributes). Furthermore, the HMM outperforms IPF for small sample sizes, even though the amount of input data is less than that for IPF. Finally, simulations show that the HMM can merge information provided by multiple data sources to allow good population estimates.

Keywords: Hidden Markov model, population synthesis, agent-based micro-simulation transportation modeling, multiple data sources, scalability

[*] Corresponding author: ismail.saadi@ulg.ac.be

Preprint submitted to Transportation Research Part B: Methodological *February 10, 2016*

1. Introduction

Following the improvements realized in terms of agent-based micro-simulation modeling, population synthesis, as a key input, has increasingly become a topic of great interest during the last decade. Agent-based micro-simulation models for transportation (Balmer et al., 2006; Bekhor et al., 2011; Rieser et al., 2007; Saadi et al., 2016) or land use (Tirumalachetty et al., 2013; Waddell, 2002) simulate the behavior of agents to determine the future states of a system when subjected to different constraints (e.g., space, time, congestion, agents' characteristics, etc.) and external factors (e.g., earthquakes, floods, etc.). As outlined by Hermes and Poulsen (2012), micro-simulation models have been used in many fields to study policy issues, population dynamics and econometric models. For example, Barthelemy and Toint (2015) developed a full activity-based model for Belgium using an existing synthetic population (Barthelemy and Toint, 2013) to improve the understanding of the mobility patterns and housing location decisions.

Activity-based models describe in great detail the activity-travel patterns of households throughout a period of time and for a specific study area. Within such models, a detailed description of the individuals and households in terms of socio-economic attributes is essential to guarantee the underlying behavioral aspects included within these models.

In particular, population synthesis is the sub-model ensuring the development of an estimated population. In this regard, it is necessary to capture the complex configuration of the population. When a complete census (i.e., survey of the entire population) is available to researchers and practitioners, a representative sampling could be extracted and used as input for agent-based micro-simulation models. However, for reasons of confidentiality and privacy, census is rarely available. In practice, only related micro-samples are provided by the government agencies. Despite potential biases in these micro-samples, we assume that the micro-samples used in this study are sufficiently representative of the true population. When micro-samples are not available, other alternatives such as travel surveys or workforce surveys can be considered. Besides micro-data, aggregate information derived from a census is also an essential input because of its overall reliability and stability. In this regard, the merger between multiple surveys, micro-samples (disaggregate information) and aggregate statistics is a way of capturing the complete heterogeneity of the true population as much as possible.

Another way of synthesizing populations consists of characterizing their related joint distributions $\Pi(X_i)$. Models fitting into such probabilistic frameworks have been proven to provide a good approximation of the underlying structure of the true population, using imputation techniques (Caiola and Reiter, 2010) and discrete choice or parametric models within a MCMC algorithm (Farooq et al., 2013) or Bayesian networks (Sun and Erath, 2015). The key challenge in such frameworks consists of identifying the correlations between the diversity of attributes among different subgroups of the population.

The main systematically encountered problem is related to the lack of data. Indeed, census surveys and micro-samples (PUMS) of populations are not published every year because of cost and other related privacy or confidentiality issues. In this context, different techniques have been developed to synthesize a disaggregated representation of the population related to a study area, taking into account a variety of available data sources.

Two main families of population synthesis techniques can be distinguished: (a) fitting methods and (b) re-weighting methods. During the last few years, a third family of techniques has emerged, i.e., Markov process-based methods. In this paper, a Hidden Markov Model (HMM)-based approach is presented, which can serve as an efficient alternative to the existing methods. A HMM is a Markov process, where the underlying internal states are supposed to be hidden from the observer. Hypotheses related to the number of states of the system and the state-transition probabilities are assumed to be known. Therefore, every state of the Markov Chain is characterized by two parameters: the symbol emission describing the emission probabilities of each state and the transition probabilities corresponding to the probability to change to another state (Ibe, 2013). First, the results presented in this paper show that for small sample sizes (<25%), the HMM-based approach improves the accuracy of the synthetic population in comparison with the standard IPF. In addition, we show that integrating information provided by an unlimited number of micro-samples facilitates solving problems related to data quality, data availability and variables through different data sets. The new HMM approach is tested in the context of Belgium, but is easily transferable to other regions.

The remainder of this paper is organized as follows. In Section 2, an overview of the main existing population synthesis methods with their respective characteristics is presented. Then, in Section 3, the Hidden Markov model (HMM) is described from both theoretical and practical perspectives. Consequently, the main data issues are commented on in Section 4.1 to enable a better understanding of the data quality and preparation. In Section 4, numerical simulations are realized to generate a synthetic population. The performance of the model are assessed using two statistical indicators. Finally, the results are discussed in Section 5 to address the main advantages and limitations of HMM-based patterns as well as some further improvements that could be made.

2. Literature review

Population synthesis techniques can be classified according to two main categories: fitting methods, including data matching, data fusion, IPF, and reweighting methods, such as deterministic reweighing (e.g., adapted IPF) and combinatorial optimization (CO) (Voas and Williamson, 2000; Williamson et al., 1998). Some techniques consist of a combination of the previous two different approaches (Hermes and Poulsen, 2012).

The basic idea behind reweighing procedures lies in systematically using survey micro-data, including a detailed set of individuals characterized by specific attributes and constraint data (benchmarks), for providing more general information (e.g., socio-demographics from a census). Then, individuals or households from the micro-sample are simply reweighted such that the constraints are matched (Hermes and Poulsen, 2012).

Among the precursors of such concepts, Beckman et al. (1996) applied the Iterative Proportional Fitting (IPF) method to create baseline synthetic populations by coupling a census survey with a public-use micro-data sample (PUMS). The proportions of households were generated according to PUMS at the census tracts and on a block group basis. First, a multivariate demographic table of proportions is estimated using IPF. Then, as a second step, a synthetic population of households is drawn from the PUMS in such a way that it matches the proportions of the above estimated table.

IPF was largely used during the last decade to synthesize household data sets (Duguay et al., 1976; Pritchard and Miller, 2012; Rich and Mulalic, 2012). As a first step, contingency tables initialized with micro-samples (e.g., PUMS) are estimated using an iterative procedure such that the deviation between the estimated and observed marginal distributions is minimized. The inter-connections in between attributes are supposed to be saved through the iterations. IPF can only be used with discrete variables containing a limited number of categories otherwise there may be a significantly greater effect to the zero-cell problem (Farooq et al., 2013). Fitting a high number of attributes makes the computational process relatively costly. A lot of efforts were made to design more efficient algorithms (Badsberg and Malvestuto, 2001; Denteneer and Verbeek, 1985; Endo and Takemura, 2009; Jiroušek and Přeučil, 1995). A recurrent problem regarding the matching between household and individual attributes appears in most of the population synthesis approaches. In most cases, the focus is directed towards synthesizing individual or household attributes, although some techniques have been developed to match both individual and household attributes (Pritchard and Miller, 2012; Ye et al., 2009). Note that this paper does not discuss the latter point but rather presents a new and more efficient methodology for synthesizing agents' attributes.

Standard techniques, including IPF as a sub-module for fitting cross-tabulations, are not capable of generating individual attributes with corresponding household attributes and their related joint distributions. To fill this gap, Ye et al. (2009) proposed a heuristic approach where both synthesized households and individual attributes match the real population. The algorithm adjusts and reweights iteratively the different kinds of households. Convergence is reached once both individual and household attributes are matched. This technique has been applied for small geographical units in the case of Maricopa County of Arizona, USA (Ye et al., 2009).

The other popular paradigm employs Combinatorial Optimization (CO) techniques to perform micro-data reconstruction (Voas and Williamson, 2000). The CO approach consists of selecting a combination of households extracted from samples of anonymized records (extracts of census) to reproduce, as closely as possible, the characteristics of a geographical unit (e.g., district). The iterative process starts from a random initial set of households (originally from SAR). Then, after replacing a selected household by a fresh one from the SAR, effects of this change are observed to assess the goodness of fit. If an improvement occurs, a swap is made; otherwise, the same households remain. The process is repeated many times until the best fit is found between the data and the corresponding sample extracted from SAR households. As outlined by Voas and Williamson (2000), this methodology requires a robust statistical technique to assess the goodness of fit every time a replacement is performed.

Another CO technique is simulated annealing, which includes a probabilistic reweighing approach because of its random sampling (Williamson et al., 1998). It has been mentioned that selecting a sample randomly from survey micro-data as input provides a more optimal selection of households (Voas and Williamson, 2000; Williamson et al., 1998). To some extent, it is admitted that CO-related methods are more successful in generating synthetic populations while maintaining an acceptable level of goodness of fit (Hermes and Poulsen, 2012). In this regard, when Williamson (2013) compared synthetic reconstruction and combinatorial optimization methodologies, he concluded that the latter perform better in terms of NFC, NFT and PFC over 100 runs.

To overcome the limitations of standard fitting techniques, Barthelemy and Toint (2013) presented a synthetic reconstruction method following three steps: (i) the generation of individuals, (ii) the estimation of household joint distributions, and (iii) the generation of households by grouping the individuals. Moreover, the proposed methodology does not require a disaggregate sample, which typically serves as the seed for IPF. By opting for a sample-free approach, costs related to the data collection of disaggregate data, as well as privacy and consistency concerns related to such data, are avoided.

In a comparative study between a sample-free (Gargiulo et al., 2010) and a sample-based approach (Ye et al., 2009), Lenormand and Deffuant (2012) concluded that results from the sample-free approach were better than those from the sample-based approach. However, they acknowledged that further research is needed to validate this conclusion. Moreover, other sample-based approaches emerged (Caiola and Reiter, 2010; Farooq et al., 2013; Sun and Erath, 2015) and showed important improvements with respect to the fit between simulated and observed populations. In particular, Farooq et al. (2013) proposed a Markov Chain Monte Carlo (MCMC) simulation-based approach enabling the emergence of Markov Process-based methods (MPBM). This technique overcomes the shortcomings related to the previous presented methodologies (e.g.. multiple solutions for matching contingency tables, loss of heterogeneity inherent in the micro-data, and low scalability regarding the number of synthesized attributes). It is stated that different data sources that are synonyms of the partial views of the joint distribution can be used to draw the marginal distributions related to the true population. In a case study, the Swiss census was used to assess the performance of the presented technique (Farooq et al., 2013). The MCMC method provided better results in terms of matching the true population, when compared to IPF. Indeed, by using MCMC, a SRMSE=0.35 was obtained in the worst case, while IPF provided a SRMSE=0.65 in the best case (Farooq et al., 2013). Some studies (Geard et al., 2013; Namazi-Rad et al., 2014) also included dynamic effects to propagate the effects of population growth over time.

3. Methodology

3.1. Problem formulation and notations

For a given spatial area and a period of time, a true population $\delta(\mathbf{X})$ exists that includes a certain number of agents. Every agent is associated with a set of specific attributes $\mathbf{X} = \{X_1, X_2, ..., X_i, ..., X_N\}$, where N is the number of attributes to be synthesized (e.g., age, income, etc.). In this paper, the challenge consists of building the joint distribution $\delta(\mathbf{X})$ by generating a set of sequences of observations from a HMM describing the structure of the true population through available micro-samples and an initial aggregate marginal distribution. The synthetic population generation process is regarded as a variant of the standard decoding problem. In the standard decoding problem, the state sequences are supposed to be unknown. In this regard, the Maximum Likelihood (ML) estimators related to the transition states are determined through the Viterbi algorithm (Ibe, 2013). However, in the synthetic population problem, both the state sequences and emission symbols are known. Thus, the model estimation run time is smaller and more efficient. The following notations indicate how the HMM structure can be used in the context of the synthetic population problem.

Let $\mathbf{h} = \{h_n, n = 1, ..., N_h\}$ be a Markovian chain process and $\mathbf{m} = \{\mu_m, m = 1, ..., N_m\}$ be a function of $\mathbf{h}$ such that $\mathbf{m} = f(\mathbf{h})$. Then, it is possible to observe the sequence of Markovian hidden state processes $\mathbf{h}$ throughout $\mathbf{m}$. A HMM is generally defined by five different parameters $(\mathbf{h}, \mathbf{m}, \mathbf{T}, \mathbf{Z}, \mathbf{\Pi})$, where we have the following: $\mathbf{h} = \{h_n, n = 1, ..., N_h\}$ is a set of N_h states that represent the total number of levels within all the attributes of the synthetic population; $\mathbf{m} = \{\mu_m, m = 1, \ldots, N_m\}$ is a set of N_m different possible symbols that indicate which level is represented by the considered state (e.g., age could have 100 states to emit 1, 2, . . . , 100; gender has two states to emit 1 or 2, etc.); $\mathbf{T} = \{t_{ij}\}$ is a set of state-transition probabilities, where t_{ij} represents the probability to move from state h_i to state h_j; $\mathbf{Z} = \{\xi_{h_i}(\mu_k)\}$ represents the observation probabilities, where $\xi_{h_i}(\mu_k)$ is the probability of emission of μ_k at state h_i with $k \in \Omega_{h_i}$, a set of possible symbols within h_i; and $\mathbf{\Pi} = \{\pi_i\}$ is the initial set of probabilities before generating sequences of attributes. Based on this distribution, the starting state can be selected.

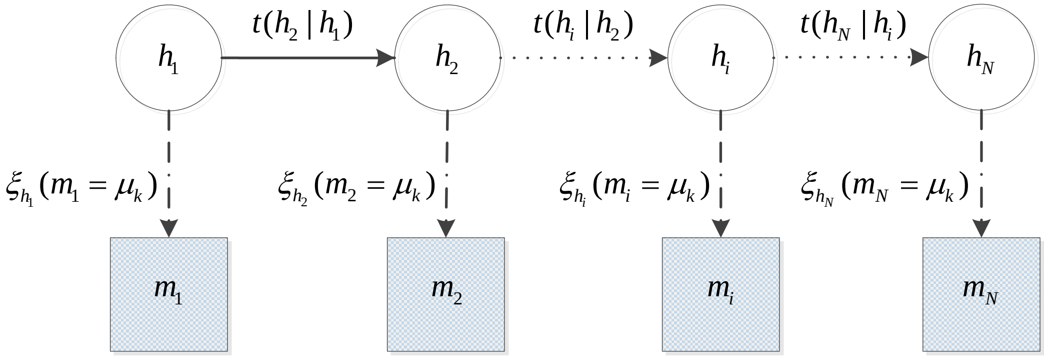


Figure 1: Probabilistic representation of the first-order hidden Markov Chain

Thus, based on Figure 1, the probability of observing the random sequence $\mathbf{m} = \{m_1, m_2, ..., m_N\}$

where N is the length of the hidden Markov chain is given by

$$P(\mathbf{h}, \mathbf{m}) = P(h_1)P(m_1 \mid h_1) \times \prod_{i=2}^{N} P(m_i \mid h_i)P(h_i \mid h_{i-1}) \tag{1}$$

where $P(h_1) \equiv \pi_i$, $P(m_i \mid h_i) \equiv \xi_{h_i}(\mu_k)$ and $P(h_i \mid h_{i-1}) \equiv t_{ij}$ are, respectively ,the initial, emission and transition probabilities.

The parameters of the HMM can be written in a compact form: $\theta = (\mathbf{T}, \mathbf{Z}, \mathbf{\Pi})$. Following this mathematical formalism, the objective is to determine the parameters $\theta = (\mathbf{T}, \mathbf{Z}, \mathbf{\Pi})$ of the HMM from the observed data sets such that the probability $P[\mathbf{h}, \mathbf{m}|\theta]$ of generating the sequence of hidden states $\mathbf{h}$ and the corresponding observation sequence $\mathbf{m}$ for the given parameters θ is maximized. With respect to the synthetic population framework, the ML problem can be translated into the following mathematical formulation:

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} P[\mathbf{m}, \mathbf{h} \mid \theta] \tag{2}$$

3.2. Extension for a higher-order HMM

Fig. 2 presents N hypothetic variables or attributes $\mathbf{h}_{i,1} = X_1, ..., \mathbf{h}_{i,N} = X_N$ organized in the form of columns where i is the i^{th} level within the attribute. One can also observe that the columns may differ in length. The model can be applied to any given number of attributes. Each attribute may have several categories/levels. For categorical variables, the number of levels is equal to the number of states.

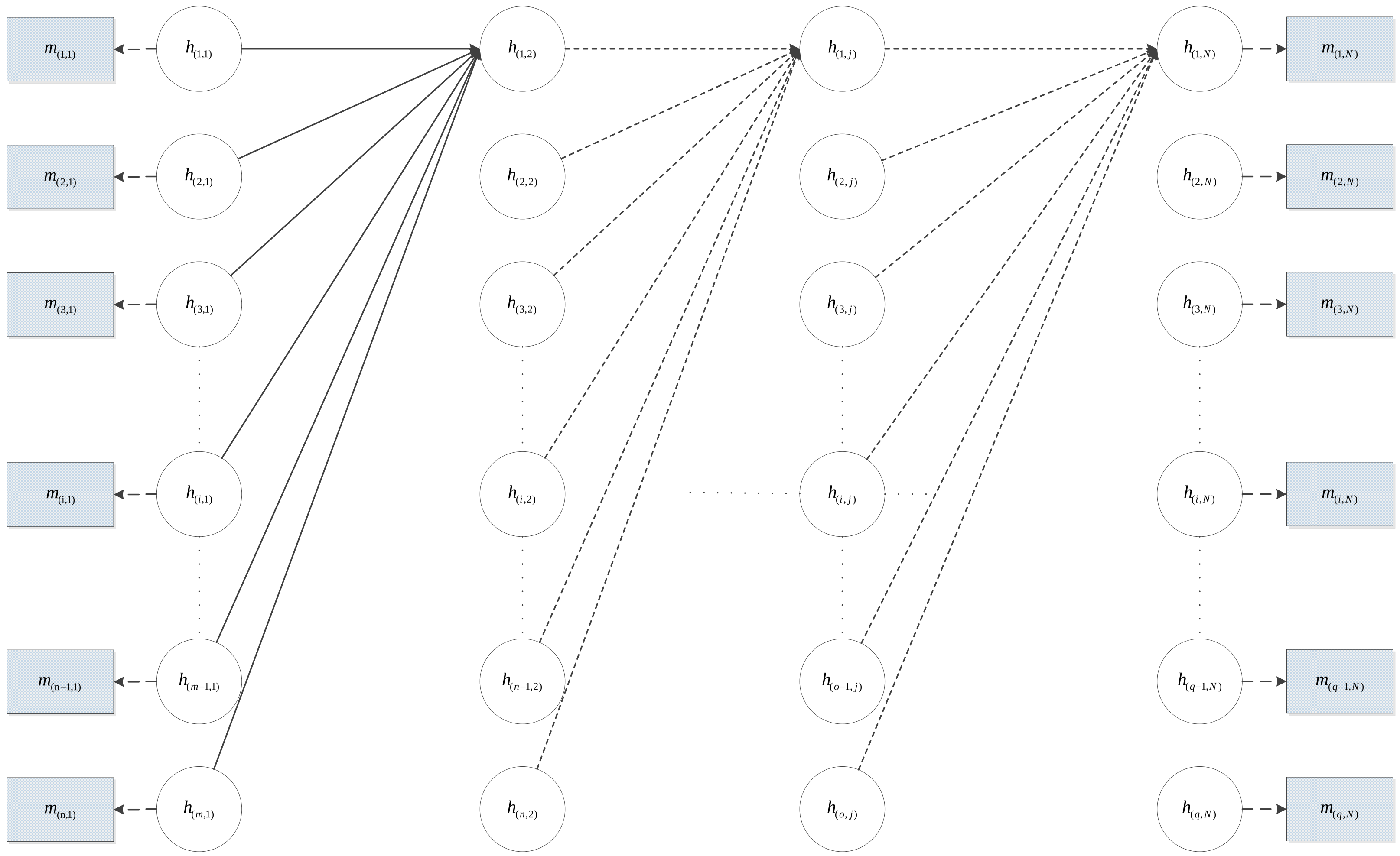


Figure 2: Graphical representation of the higher-order HMM structure

$$\begin{aligned} P(h_{i1}, h_{i2}, ..., h_{ij}, ..., h_{iN}; m_{i1}, m_{i2}, ..., m_{ij}, ..., m_{iN}) &= P(h_{1i})P(m_{1i} \mid h_{1i}) \\ &\times \prod_{k=2}^{N} P(m_{ik} \mid h_{ik})P(h_{ik} \mid h_{1,k-1}, h_{2,k-1}, ..., h_{i,k-1}..., h_{m,k-1}) \end{aligned} \tag{3}$$

$$P(\mathbf{h}; \mathbf{m}) = P(h_{1i})P(m_{1i} \mid h_{1i}) \times \prod_{k=2}^{N} P(m_{ik} \mid h_{ik})P(h_{ik} \mid \mathbf{h}_{k-1}) \tag{4}$$

However, for a continuous variable, the number of states is fixed by the user. Let us assume that one of the attributes is continuous. Including continuous variables in standard synthesizing processes (fitting and reweighing methods) is an important computational issue. As a result, continuous variables are often aggregated, resulting in an important loss of information (Farooq et al., 2013). The strength of a HMM lies in its ability to handle both continuous and discrete variables. Given the fact that a Markov process is characterized by discrete states, continuous variables need to be discretized to be included in the modeling process. As an example, age could vary from 1 to 100. In the most detailed representation, all 100 different states can be considered (Fig. 3c). However, if required, the marginal distribution could be aggregated using a fixed number of bins (Fig. 3a-b). In this case, within each bin, the emission probability matrix $\mathbf{M}_E$ could be used to indicate the sub-distribution, instead of randomly selecting values in between these intervals for synthesizing the considered attribute.

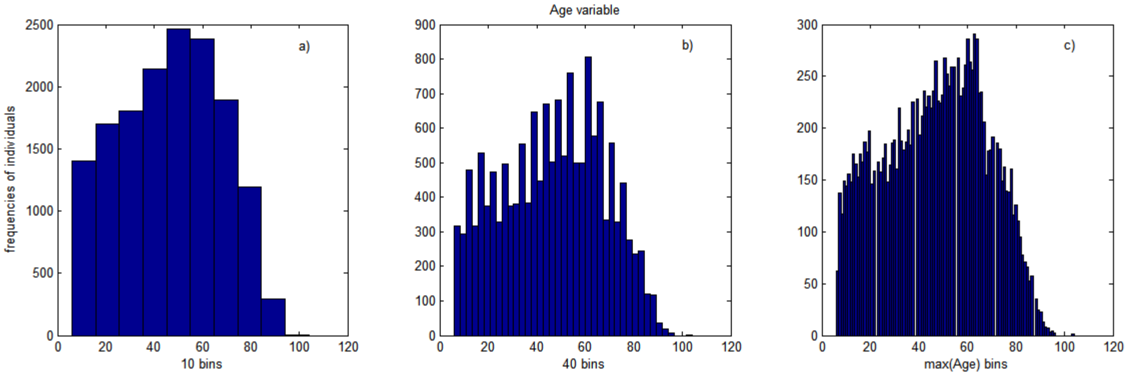


Figure 3: Discretization of a continuous variable

Links in between attributes represent the transitions occurring in the micro-sample. For the sake of clarity, Fig. 2 partially indicates the transition probabilities as well as the emission probabilities. Each state of attribute i is connected to all the states of the following attribute $i + 1$. In practice, the number of transition probabilities ϵ that should be defined is determined as follows:

$$\epsilon = \sum_{i=1}^{n-1} \phi(X_i).\phi(X_{i+1}) \tag{5}$$

where n is the number of attributes, X_i is the attribute i, and ϵ is the number of transition probabilities to be determined. ϕ is an operator that provides the number of categories of an attribute. The key point of the modeling process is the design of the transition probability $\mathbf{M}_T$ and the emission probability $\mathbf{M}_E$ matrices.

$$\mathbf{M}_T = \begin{pmatrix} t_{11} & & \cdots & & t_{1\epsilon} \\ & \ddots & & \iddots & \\ \vdots & & t_{ij} & & \vdots \\ & \iddots & & \ddots & \\ t_{\epsilon 1} & & \cdots & & t_{\epsilon\epsilon} \end{pmatrix} \tag{6}$$

Elements of the transition probability matrix $\mathbf{M}_T$ are defined using the following formula:

$$t_{ij} = \frac{p_{ij}}{\sum_{k=1}^{N} p_{ik}}, \forall i, j = 1, ..., \epsilon \tag{7}$$

where t_{ij} is the transition probability between the category i of the left-side attribute and the category j of the right-side attribute, e_{ij} is the emission probability[1] of symbol j within state i, ϵ is the number of states, γ is the number of levels of the attribute containing the highest number of levels, p_{ij} is the number of transitions occurring between state i of an attribute and state j of the directly following attribute, and N is the total number of transitions starting from state i.

As a second input, a matrix for emission probabilities is required by the HMM to indicate the probability of emission of symbols at a given state. Note that in this case, we consider that each state emits only one symbol such that the probability emission is equal to 1. Additionally, the columns of $\mathbf{M}_E$ are subjected to the following constraint:

$$\sum_{j=1}^{\gamma} e_{ij} = 1, \forall i = 1, 2, ..., \epsilon \tag{8}$$

As a result, the emission probability matrix is the identity matrix.

$$\mathbf{M}_E = \begin{pmatrix} e_{11} & & \cdots & & e_{1\gamma} \\ & \ddots & & \iddots & \\ \vdots & & e_{ij} & & \vdots \\ & \iddots & & \ddots & \\ e_{\epsilon 1} & & \cdots & & e_{\epsilon\gamma} \end{pmatrix} \tag{9}$$

where ϵ is the total number of states and γ is the highest possible number of symbols emitted by one of the ϵ states. The sum of the rows of $\mathbf{M}_E$ is equal to one. In other terms, each row of $\mathbf{M}_E$ corresponds to a state of the HMM or a level of an attribute. Furthermore, within each row, it is possible to determine the

[1] Because we are mainly dealing with transitions in the current framework, the emission probability matrix is considered to be an identity matrix. In practice, this means that each state will systematically emit its related category. In the case of continuous variables, emission probabilities can be defined within each interval chosen for discretization. However, this practice is not advised, as part of the transition information is lost. Test results indicate that the more a continuous variable is aggregated, the less accurate the joint distribution is from a disaggregate perspective.

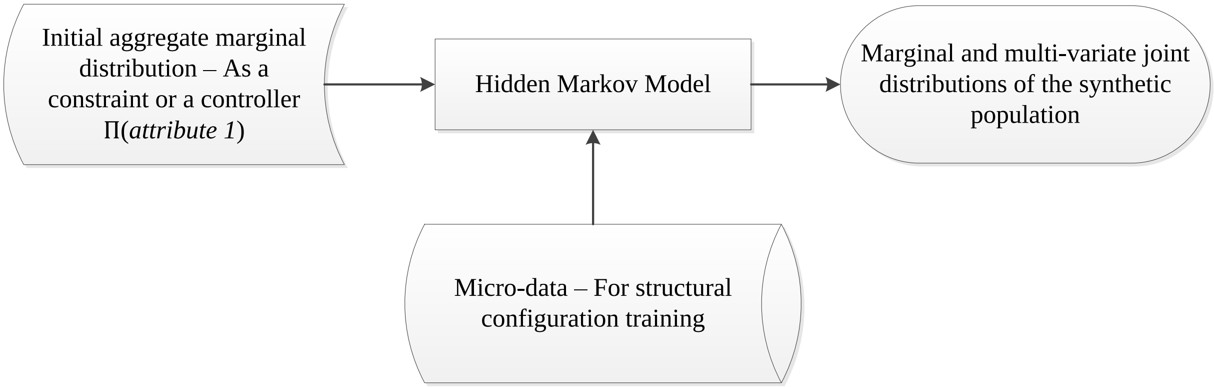

Initial aggregate marginal distribution – As a constraint or a controller Π(*attribute 1*)
Hidden Markov Model
Marginal and multi-variate joint distributions of the synthetic population
Micro-data – For structural configuration training

Figure 4: Proposed framework to include a controller

horizontal distribution of the symbols. Based on these two matrices, the HMM will be able to generate any individual sequence of attributes as generated from the synthesized joint distribution $\delta(\mathbf{X})$.

Note that we proposed to set $\mathbf{M}_E$ as an identity matrix to obtain the most disaggregated synthetic population. The goal was to show how to handle, as accurately as possible, the continuous variables so that all the information is preserved. However, if we consider the example of the variable age, depending on the nature of the problem, synthesizing it according to a few intervals (e.g., 4, 5 etc.) is largely sufficient. Here, the existence of $\mathbf{M}_E$ becomes very useful. Indeed, when the synthetic population is included in an agent-based micro-simulation model, one should specify the age of each agent. In this context, $\mathbf{M}_E$ can be used to set the "sub-distributions" within each interval such that the outputs are already defined. In addition, such a choice is judicious because it can reduce the computational time especially in the context of several continuous variables.

3.3. Initial distribution

If X_1 or $\mathbf{h}_{1i}$ is supposed to be the first attribute, generating attribute chains requires the definition of an initial distribution Π. If the micro-sample is sufficiently reasonable in terms of sample size and contains limited missing information such that it is a good representation of the population, it is possible to consider the initial distribution as a controller or a constraint from an aggregate data source (e.g., census). For example, every year, Statistics Belgium publishes the marginal distributions of the whole population according to different variables, such as age, gender, and spatial location at the municipality level. The additional use of these data is more suitable because the structural configuration of the population is captured by the HMM (micro-data) and initiated by the initial marginal distribution as presented in Fig. 4.

Then, the synthesized marginal distributions can be, in the best case, compared with the common available marginal distributions (age, gender and geographical locations) for validation. However, this option is relatively limited when the user plans to synthesize other kinds of attributes, especially in the context of data privacy. In this regard, we propose to use the marginal distribution directly from the training data set to illustrate the proposed methodology.

3.4. Framework

Fig. 5 presents the steps for performing a HMM-based population synthesis approach. The data preprocessing step or data preparation should be performed carefully. The training and test data sets cannot contain some missing values; otherwise, the HMM would not be able to determine the transition and emission probabilities. In this regard, some techniques related to machine learning (Yang et al., 2012) might be used to solve the problem of missing values in a data set. However, as the purpose of this paper is to present a new methodology and its application, we prefer to avoid adding data imputation issues. As shown in Fig. 5, the data set is split into two parts. The first, called the training data set, is used to set the parameters of the HMM. Generally, depending on the sample size and the nature of the problem, we chose to set p equal to a value between 70-80%. The second part is used to validate the synthesized population. One can refer to Section 4 to know more about the validation issues. One should ensure that the variable classification is performed according to the descending order of the number of categories. In this regard, highly disaggregate continuous variables are generally placed at the beginning of the chain. By referring to the notations in Figure 2, the constraint that should be respected to maintain the best approximation of the synthetic population is the following:

$$m \geq n \geq ... \geq o \geq ... \geq q \tag{10}$$

The following step consists of determining the values of $\mathbf{M}_T$ and $\mathbf{M}_E$ following the guidelines provided previously. Using Equ. 7, a routine can be implemented to determine the transition probabilities of $\mathbf{M}_T$. $\mathbf{M}_E$ should be an identity matrix in order to obtain the most representative and accurate population, as outlined in Section 3.

Once the HMM is calibrated, a sequence of agent attributes can be generated from the model depending on the size n of the population.

3.5. Model estimation

In the previous sections, guidelines have been proposed to generate $\mathbf{M}_T$ and $\mathbf{M}_E$ matrices. Here, we propose some directives to generate the agents. In practice, various packages are available depending on the programming language used. For the R statistical language, Visser and Speekenbrink (2010) implemented a framework to define and estimate standard Markov models as well as latent and hidden Markov models. Additionally, the MATLAB Toolbox also contains a package with a set of functions able to generate sequences from an estimated HMM. The latter toolbox has been used for the analysis presented in this paper.

One should pay attention to how the HMM is estimated. For example, the HHM models within the Statistical and Machine learning toolbox of MATLAB start, by default, from state 1. This option can cause problems for population synthesis because all the initial probabilities will be systematically located within the first level of the first attribute. Thus, only a small portion of the population will be synthesized. In this regard, a dummy state should be added before the first attribute, whose transition probabilities will be specified by the initial aggregate marginal distribution. When $\mathbf{M}_T$ and $\mathbf{M}_E$ are determined, a further step is necessary to ensure the generation of the full synthetic population. Indeed, from the $\mathbf{M}_T$ ($\epsilon \times \epsilon$ matrix) and $\Pi^\top$ ($1 \times N$ vector), where N is the number of states of the first attribute, an extended matrix

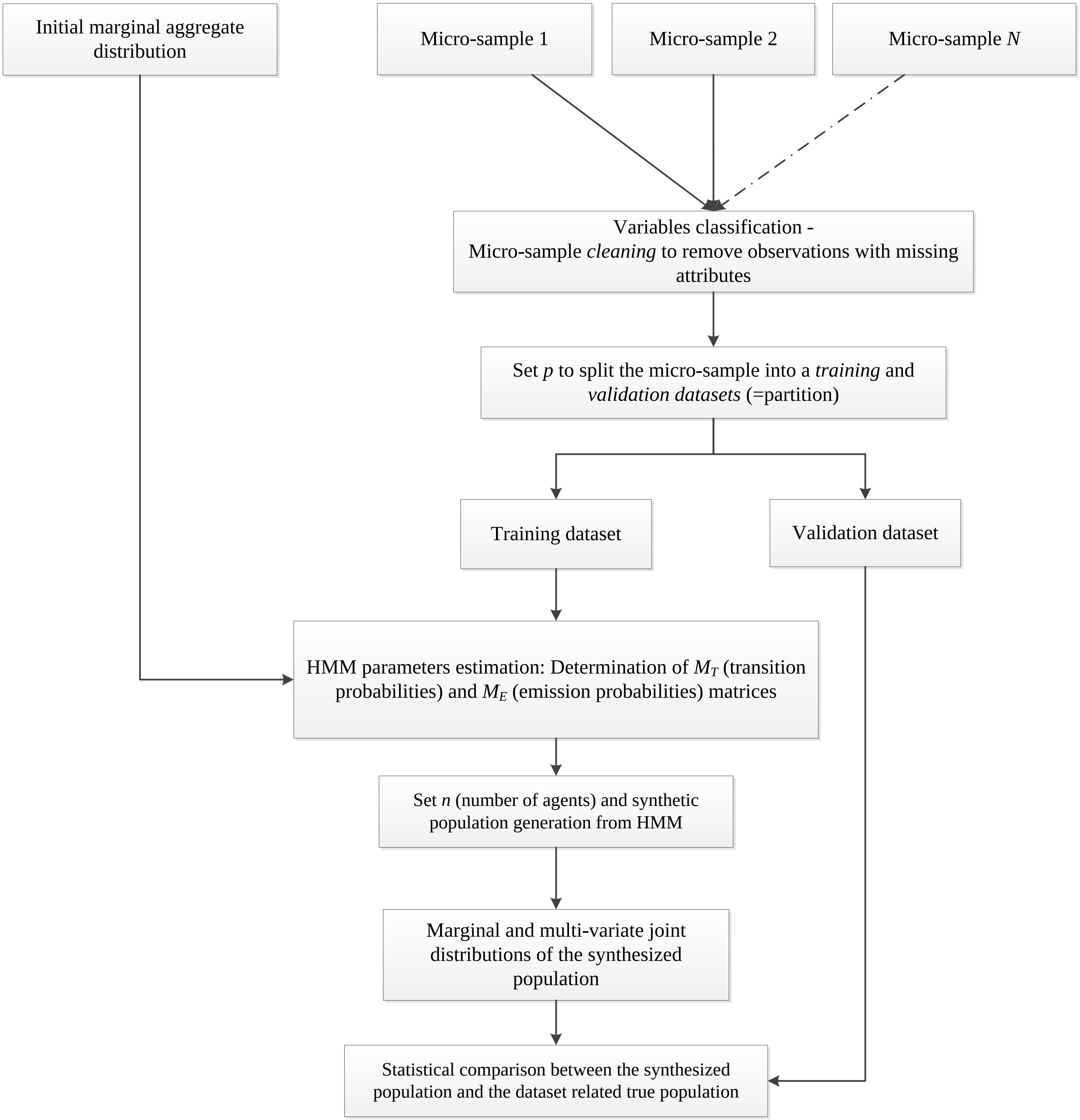

Initial marginal aggregate distribution
Micro-sample 1
Micro-sample 2
Micro-sample *N*
Variables classification - Micro-sample *cleaning* to remove observations with missing attributes
Set *p* to split the micro-sample into a *training* and *validation datasets* (=partition)
Training dataset
Validation dataset
HMM parameters estimation: Determination of M_T (transition probabilities) and M_E (emission probabilities) matrices
Set *n* (number of agents) and synthetic population generation from HMM
Marginal and multi-variate joint distributions of the synthesized population
Statistical comparison between the synthesized population and the dataset related true population

Figure 5: Methodology to synthesize and validate populations

$\hat{\mathbf{M}}_T$ ($[\epsilon+1] \times [\epsilon+1]$ matrix) is built such that

$$\hat{\mathbf{M}}_T = \begin{pmatrix} 0 & \Pi^\top \mid \mathbf{0} \\ 0 & \mathbf{M}_T \end{pmatrix}. \tag{11}$$

where **0** is a $1 \times (\epsilon - N)$ vector of null values. In parallel, the matrix $\mathbf{M}_E$ is extended into the following form $(\epsilon+1) \times \gamma$, where the emission probability of the new dummy state is specified:

$$\hat{\mathbf{M}}_E = \begin{pmatrix} \mathbf{0} \\ \mathbf{M}_E \end{pmatrix}. \tag{12}$$

Subsequently, the function `hmmgenerate(arg)` can be used to generate the full synthetic population. The inputs are defined as follows: arg$\leftarrow$(s,$\hat{\mathbf{M}}_T$,$\hat{\mathbf{M}}_E$). Note that s indicates the length of the full sequence. Let us suppose that we synthesize N attributes and a population size of M agents. In this context, s will be equal to $(1+N) \times M$, as the dummy state should also be taken into account. Of course, when the list of agents is defined, the first column (vector of ones) can be removed as it is of no more interest.

4. Numerical analysis (case study)

4.1. Data

In contrast with fitting and reweighting methods, the proposed methodology can incorporate one or multiple micro-samplings, including the variables of interest as input and an initial marginal distribution (information related to the full population). In this paper, we used the 2013 Survey on the Workforce of 2013 (EFT) and the Belgian National Household Travel Survey (BELDAM) to ensure a sufficiently large sample size and data quality and to investigate the scalability of the proposed methodology. The main objective of the first survey is the classification of the active population according to three distinct categories: professional active, unemployed and inactive people. The survey provides annual information related to the activity status of 95,940 Belgian inhabitants 15 years of age and older. The variables used from this data set are municipality location (spatial information), travelled distances, age, education level, profession and gender. Table 1 briefly describes the main statistical indicators of these variables as well as the data accuracy and sampling size.

In this paper, we illustrate the HMM-based approach using various simulations. The first simulation tests the combined effects of scalability and dimensionality. The second simulation compares the HMM-based approach with the most common technique for population synthesis, i.e., IPF. Finally, a third simulation is carried out to demonstrate the advantage of the HMM approach over IPF using multiple samples.

In the first numerical simulation, six attributes were extracted from the EFT data set in order to assess the scalability effects of the HMM. In this regard, incomplete observations can be removed from the training and validation data set. We propose synthesizing three (SP3), four (SP4), five (SP5) and six (SP6) attributes. Table 1 presents the characteristics related to each prepared data set. Note that the relative sample size with respect to the original data set decreases when the number of synthesized attributes increases. Indeed, when a new attribute is introduced, the observations related to the missing or incorrect values (0 or NaN) are removed. Therefore, the proportions of the categorical variables as well as the means and standard deviations of the continuous variables are relatively changing from one sample to the other. We also discuss the dimensionality and its contribution within the overall error rate.

Furthermore, the HMM approach is compared to IPF to show how the error rate can be improved, especially in the context of small sample sizes. To test the stability of these approaches, the BELDAM data are used. The goal of this second simulation is to demonstrate that for various sampling rates, the HMM is capable of outperforming IPF, even when less data are used. In addition, IFP-based synthesis may be affected by an eventual bias that is caused by the zero-cell problem in the case of a larger number of attributes. In this regard, a set of 4 attributes are synthesized to ensure a fair comparison between both methods.

Finally, the third simulation investigates the contribution of a population synthesis using different data sources. In contrast to IPF, the HMM approach is able to merge information provided by an unlimited number of micro-samples with varying sample sizes. In this paper, data from the EFT data set are used to illustrate how data fusion is performed through the transition probability matrix $\mathbf{M}_T$.

Attributes	Level	Label	OD*	SP3*	SP4*	SP5*	SP6*
Municipalities	-		547/589	-	547/589	547/589	547/589
Travelled distances	-		Mean: 21.39 km	-	-	-	Mean: 21.57 km
	-		Std. Dev.: 82.96 km	-	-	-	Std. Dev.: 48.22 km
	NaN		58862(-)	-	-	-	-
Age**	-		Mean: 37.67	Mean: 44.30	Mean: 44.29	Mean: 41.29	Mean: 40.65
	-		Std. Dev.: 21.26	Std. Dev.: 16.82	Std. Dev.: 16.81	Std. Dev.: 11.32	Std. Dev.: 11.05
Education	1	Primary school	8506(10.90%)	8506(11.50%)	8506(11.50%)	1819(4.51%)	1314(4.28%)
	2	Secondary school	6624(8.49%)	6624(8.96%)	6624(8.96%)	1742(4.32%)	1327(4.32%)
	3	Higher education	4202(5.38%)	4202(5.68%)	4202(5.68%)	1855(4.60%)	1357(4.42%)
	4	Technical education	4911(6.29%)	4911(6.64%)	4911(6.64%)	2095(5.19%)	1605(5.22%)
	5	Vocational education	9665(12.39%)	9665(13.07%)	9665(13.07%)	4431(10.98%)	3475(11.32%)
	6	Further education (not university/universty college)	9123(11.69%)	9123(12.34%)	9123(12.34%)	5945(14.73%)	4583(14.93%)
	7	Bachelor (Professional)	6587(8.44%)	6587(8.91%)	6587(8.91%)	4280(10.61%)	3249(10.58%)
	8	Bachelor (High School)	2008(2.57%)	2008(2.71%)	2008(2.71%)	1551(3.84%)	1157(3.77%)
	9	Bachelor (Univ.)	10955(14.04%)	10955(14.82%)	10955(14.82%)	8070(20.00%)	6174(20.11%)
	10	Continuing education	1604(2.06%)	1604(2.17%)	1604(2.17%)	1193(2.96%)	902(2.94%)
	11	Transition studies (Bac. to Ms)	648(0.83%)	648(0.88%)	648(0.88%)	280(0.69%)	204(0.66%)
	12	Master (High school)	271(0.35%)	271(0.37%)	271(0.37%)	217(0.54%)	159(0.52%)
	13	Higher non-university education	1474(1.89%)	1474(2.00%)	1474(2.00%)	1172(2.90%)	904(2.94%)
	14	Master (University)	6202(7.95%)	6202(8.39%)	6202(8.39%)	4735(11.73%)	3538(11.53%)
	15	University studies	788(1.01%)	788(1.07%)	788(1.07%)	677(1.68%)	529(1.72%)
	16	Second master	376(0.48%)	376(0.51%)	376(0.51%)	289(0.72%)	221(0.72%)
	0	No answer	4089(5.24%)	-	-	-	-
	NaN		17907(-)	-	-	-	-
Profession	1	Executive	10510(25.50%)	-	-	9871(24.46%)	8006(26.08%)
	2	Employee	15116(36.67%)	-	-	15043(37.28%)	12569(40.94%)
	3	Civil servant	6518(15.81%)	-	-	6496(16.10%)	5077(16.54%)
	4	Contractual	3015(7.31%)	-	-	2970(7.36%)	2281(7.43%)
	5	Self-employed worker without staff	3811(9.25%)	-	-	3753(9.30%)	1627(5,30%)
	6	Self-employed worker with staff	1738(4.22%)	-	-	1718(4.26%)	1040(3.39%)
	7	Volunteer	511(1.24%)	-	-	500(1.24%)	98(0.32%)
	NaN		41219(-)	-	-	-	-
Gender	1	Male	47536(49.55%)	36481(49.34%)	36481(49.34%)	21556(53.42%)	16567(53.97%)
	2	Female	48404(50.45%)	37463(50.66%)	37463(50.66%)	18795(46.58%)	14131(46.03%)
Sample size	-	-	95940	73944	73944	40351	30698
Relative sample size with respect to original dataset	-	-	-	77.07%	77.07%	42.06%	32.00%

*OD: original dataset, SPx: Synthetic Population x

**Later in this paper, Age is divided into 10 bins ([0,10[,[10,20[,[20,30[,...) for model validation

Table 1: Description of the inputs (EFT of 2013)

4.2. HMM-based synthesis

In order to test the population synthesis algorithm, we first select the variables of interest from the reference data set. In this regard, a training data set representing around 70% of the initial micro-sample is extracted. The rest of the data set (30%) is used for validation. Using Equ. 5 and taking into account that age varies from 15 to 101 in the reference data set, education contains 16 categories and gender is binary, it can be derived that the total number of states ϵ considered by the HMM is 105 (=101-15+1+16+2). Based on the training data set, the M_T matrix was built step by step (as an image of the structure of the population) using Eqs. 6 and 7. As soon as M_T and M_E are defined, the calibrated HMM can generate any number of attribute sequences.

In this simulation, we propose to generate a set of 100,000 agents from SP3, SP4, SP5 and SP6 using the HMM. In this way, comparisons between the same attributes from different synthetic populations will be possible. Similar to IPF or MCMC, HMM-based population synthesis can reproduce the marginal distributions related to the synthesized attributes (Fig. 6). Moreover, Fig. 7 indicates a quasi-perfect fit between the synthesized and the true population in terms of the slopes and R².

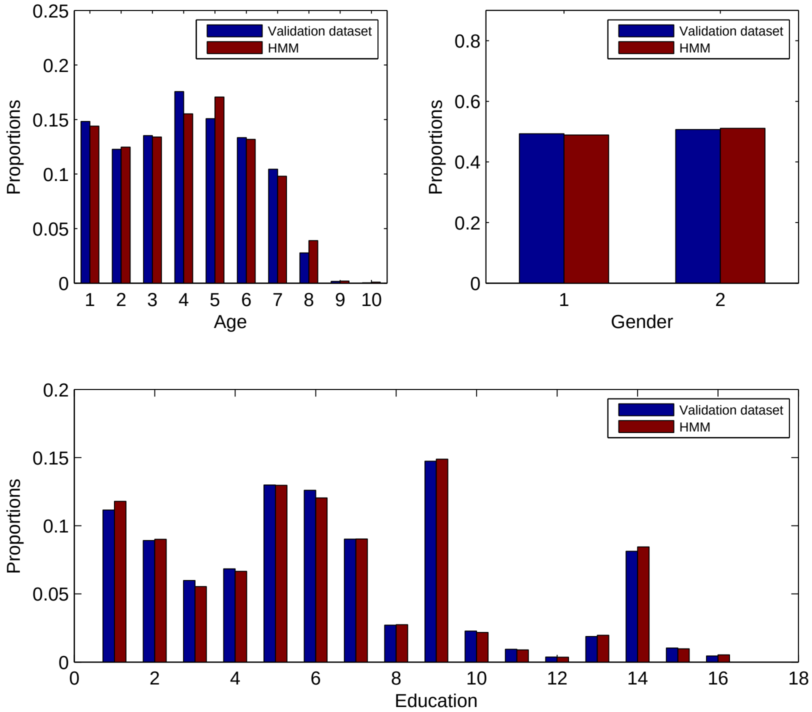


Figure 6: Comparison of marginal distributions between the attributes using an HMM (SP3)

In addition to a univariate marginal distribution, we also compare the HMM approach and the validation data set in terms of the use of multivariate joint distributions (Age × Status, Status × Gender, Age × Gender, Age × Status × Gender). It can be concluded that the fits between the real population and the synthesized population are highly acceptable especially in the context of a small micro-sample size (1%). Note that the synthesized population is compared with the validation data set. In other terms, we are performing a comparison with a sample size representing approximately 0.2% of the population.

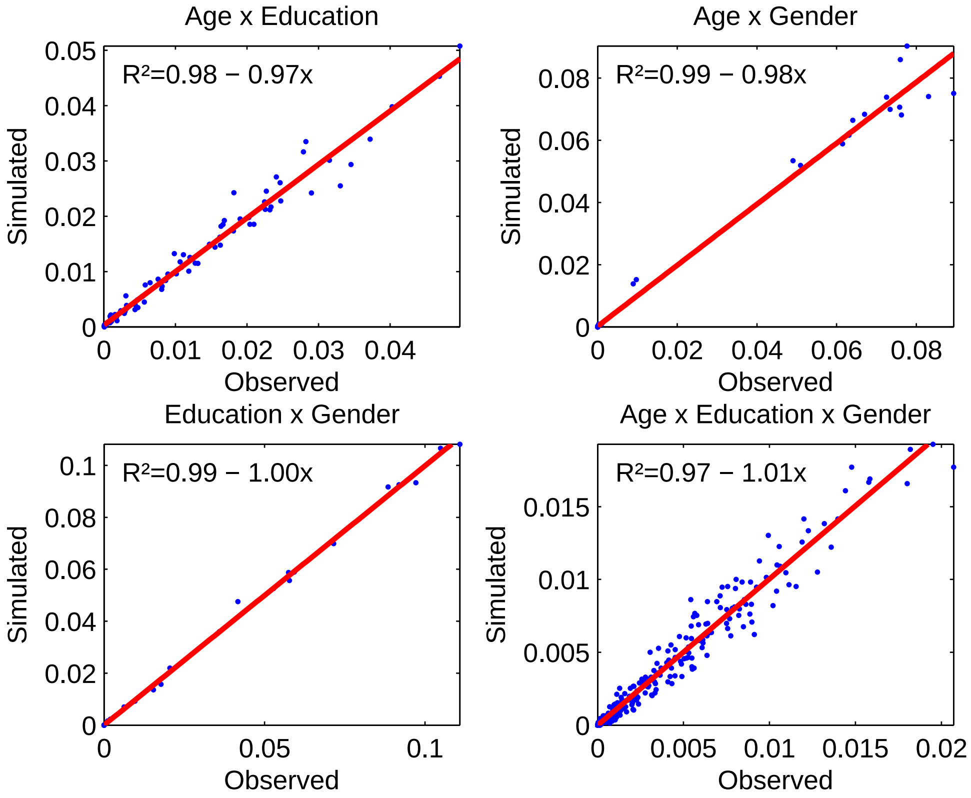


Figure 7: Fit between the real population and HMM-based approach synthesis (SP3)

The spread of the synthetic population is supposed to be more visible when a higher number of attributes is considered. To allow for a comparison with the SP3 results, HMMs with a higher number of attributes are constructed. Corresponding to the SP3 results, the marginal distributions for four (Fig. 8), five (Fig. A.13) and six (Fig. A.15) attributes are presented. Similarly, the joint distributions are respectively presented in Fig. 9 (SP4), Fig. A.14 (SP5) and Fig. A.16 (SP6).

Regarding the analysis of the scalability, the joint distributions indicate a very good fit between the synthesized and the true population. However, more substantial deviations appear between the joint distributions for a higher number of attributes. The observed R^2 values are generally a bit lower than those for the case where only three attributes are synthesized because of the coupled phenomenon of distances in between attributes and the number of levels within each attribute. Agents were synthesized according to the following scheme: Location (at municipality level) $\rightarrow$ Age $\rightarrow$ Education $\rightarrow$ Gender (based on SP4). Related joint distributions with attributes that are directly associated in the scheme such as Location × Age, Age × Education, Education × Gender have a near-perfect fit. However, if we consider the worst performance, i.e., Location × Gender, the slope is the lowest (=0.82).

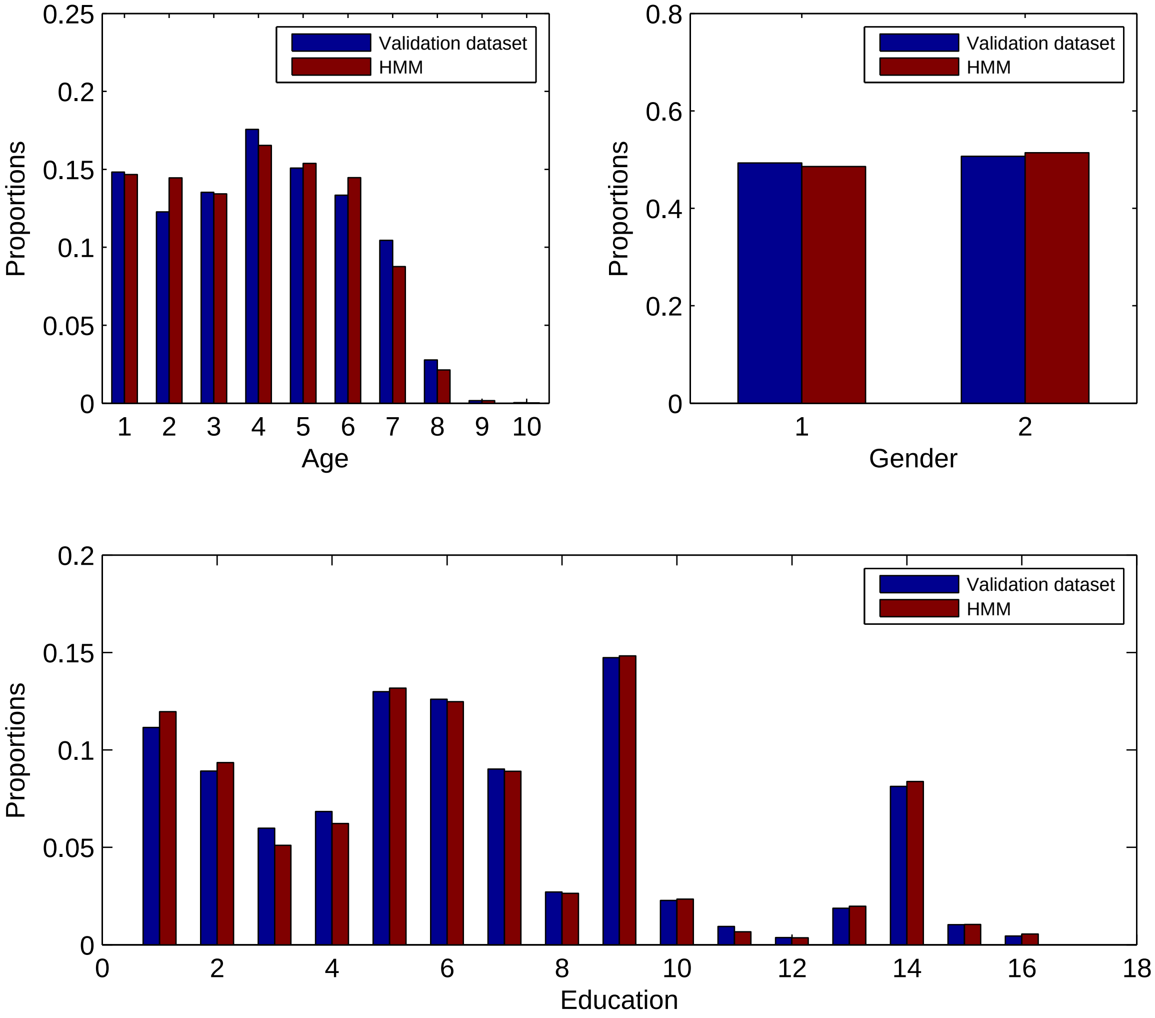


Figure 8: Comparison of the marginal distributions for different attributes (SP4)

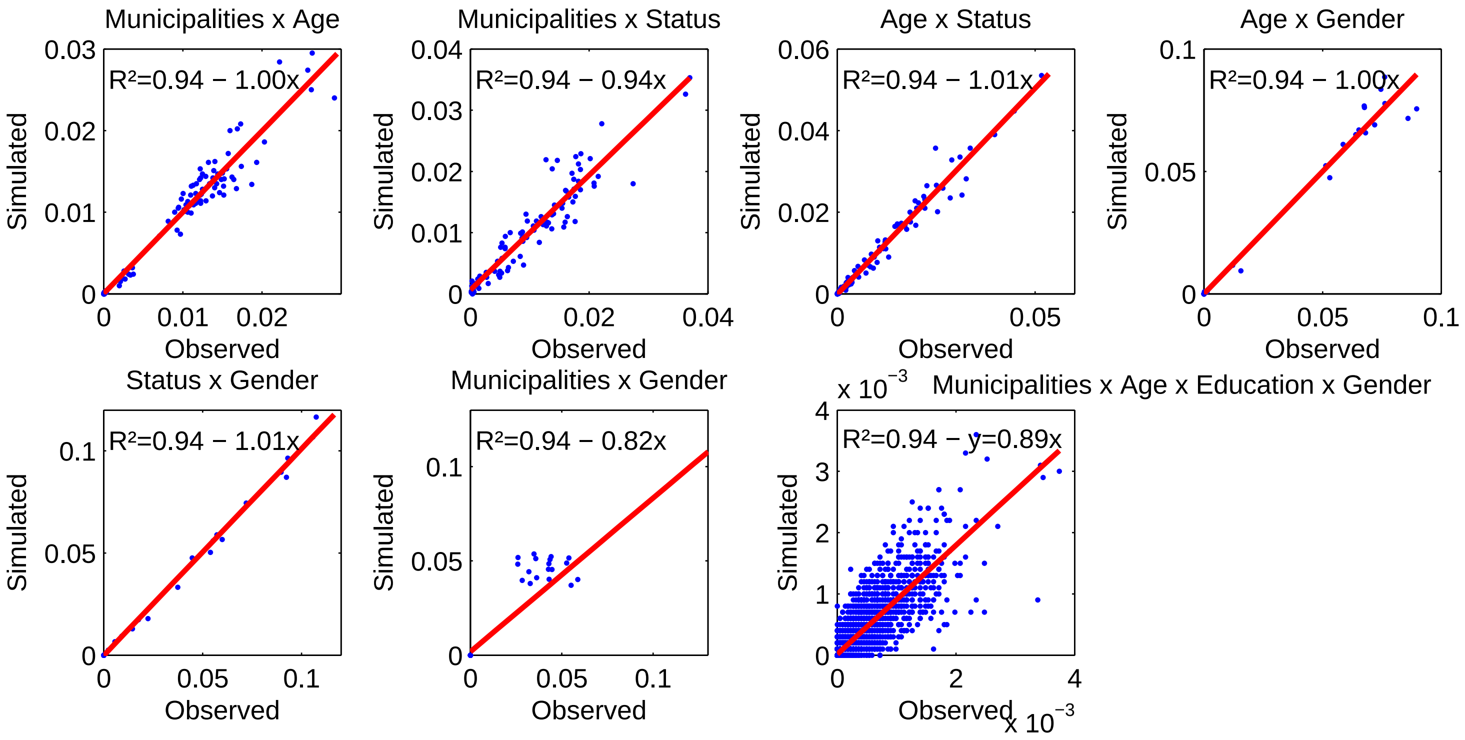


Figure 9: Fit between the real population and HMM-based approach synthesis (SP4)

Note that when the number of synthesized attributes increases, the slope of the joint distribution formed by the most distant attributes (Municipalities × Status and Municipalities × Gender) decreases (Fig. A.14).

4.3. Influence of the scalability on SRSME

Introduced by Knudsen and Fotheringham (1986), the standardized root mean square error (SRMSE) is an interesting indicator for assessing the goodness of fit between two joint distributions (e.g., in a M-by-N matrix form). This indicator has also been used to assess the performance of other synthetic population methods (Farooq et al., 2013; Müller and Axhausen, 2010; Pritchard and Miller, 2012). This is a distance-based metric that yields 0 when the fit is perfect, similar to other related standard statistical indicators. Note that the absolute value has no significance by itself but is only a relative comparison of the source of information. The formula for comparing two joint distributions $\tilde{J}_{ijk...}$, $J_{ijk...}$ for any number of attributes is the following:

$$SRMSE = \frac{\sqrt{\frac{1}{N}\sum\sum\sum...(\tilde{J}_{ijk...} - J_{ijk...})^2}}{\frac{1}{N}\sum\sum\sum...J_{ijk..}^2} \tag{13}$$

where $\tilde{J}_{ijk...}$ and $J_{ijk...}$ represent the number of agents characterized by the combination of attributes i,j,k,... of the synthesized and observed population, respectively, and N is the total number of cells within the matrix. In other terms, it is the total product of the dimensions of $J_{ijk...}$.

To our knowledge, none of the existing studies has statistically investigated the effects of scalability in the context of population synthesis. In this regard, we propose to investigate scalability by successively increasing the number of synthesized attributes (from 3 to 6) for a constant number of generated agents. First, experiments show as expected that every time a new variable is introduced in the population synthesis, a relative deviation appear in the SRMSE. We can observe from Table 2 that a significant relative change compared to SP3 ($\frac{(0.1261-0.0145)}{0.0145} \times 100 = +769.66\%$) occurs because of the spatial variable "Municipalities". Indeed, this variable contains 547 sectors, which means that the same number of levels is, in fact, included in the HMM. As presented in Table 2, keeping such a level of disaggregation introduces a relative increase in the error in the modeling process of +769.66%. This clearly illustrates that the error contribution of a variable is highly related to its number of levels.

Synthetic population		**SRMSE**	**Relative change**
3 attributes (SP3)	Age x Education x Gender	0.0145	-
4 attributes (SP4)	**Municipalities** x Age x Education x Gender	0.1261	+769.66%
5 attributes (SP5)	Municipalities x Age x Education x **Profession** x Gender	0.4856	+285.09 %
6 attributes (SP6)	Municipalities x **Travelled distances** x Age x Education x Profession x Gender	0.5364	+10.46%

Table 2: Scalability effects on the SRMSE

The increase in the number of synthesized attributes also significantly affects the dimensionality: the number of cells is affected both by the number of attributes as well as the number of levels within each attribute. In this regard, the combined effect of scalability and dimensionality warrants particular attention in the context of continuous variables, as the number of bins used to discretize the continuous variables directly impacts the dimensionality. In practice, most variables used in activity-based models are categorical and do not exceed 6 categories (Yasmin et al., 2015). In the example of our paper, we considered a complex case to illustrate that the HMM is capable of maintaining an acceptable error rate, i.e., SMRSE = 0.5364 for 6 attributes.

4.4. Comparison with IPF

To assess the performance of the HMM-based approach, a comparison with the standard IPF procedure is made. The IPF approach is still widely used by both researchers and practitioners for synthesizing populations (Vovsha et al., 2015). Socio-demographic information from the BELDAM data set, which included information on 15,822 individuals grouped into 8,533 households, is used for the comparison. The 'mipfp' package (Barthelemy et al., 2015), written in the statistical language R, is used to synthesize the populations using multilevel IPF.

In practice, a comparison of methods on the same basis is quite difficult. IPF requires the definition of specific parameter settings, such as the convergence criterion, which influences both the run time and the quality of the solution. In addition, the same input data should be used. To ensure a fair comparison, we made the assumption that the travel survey represented the full population. In this context, all the aggregate marginal distributions can be exactly determined. In addition, various micro-samples with different sampling rates ψ={1%, 2%, 3%, 4%, 5%, 10%, 15%, 20% and 25%} have been extracted from the supposed full population. Only sampling rates ranging between 1 and 25% have been considered, as they are the most relevant for large-scale population synthesis. The first major advantage of the HMM approach should be highlighted at this stage: the amount of data used by both methods. Whereas the HMM needs a micro-sample (PUMS) and an aggregate marginal distribution related to the first attribute in the modeling procedure, IPF needs the PUMS, but with the full set of marginal distributions.

For the comparison, four attributes were synthesized, namely, housing location (20 categories/zones), age (14 categories), socio-professional status (14 categories) and gender. The housing location attribute has been extensively aggregated to reduce the effects of the zero-cell problem inherent to the IPF procedure. This problem negatively affects the accuracy of the results of IPF and may generate problems in terms of convergence. In contrast, modeling the selected attributes using the original number of categories does not generate any particular problems with the HMM-based approach, highlighting once again the robustness and flexibility of the approach in terms of the type of variables that are considered.

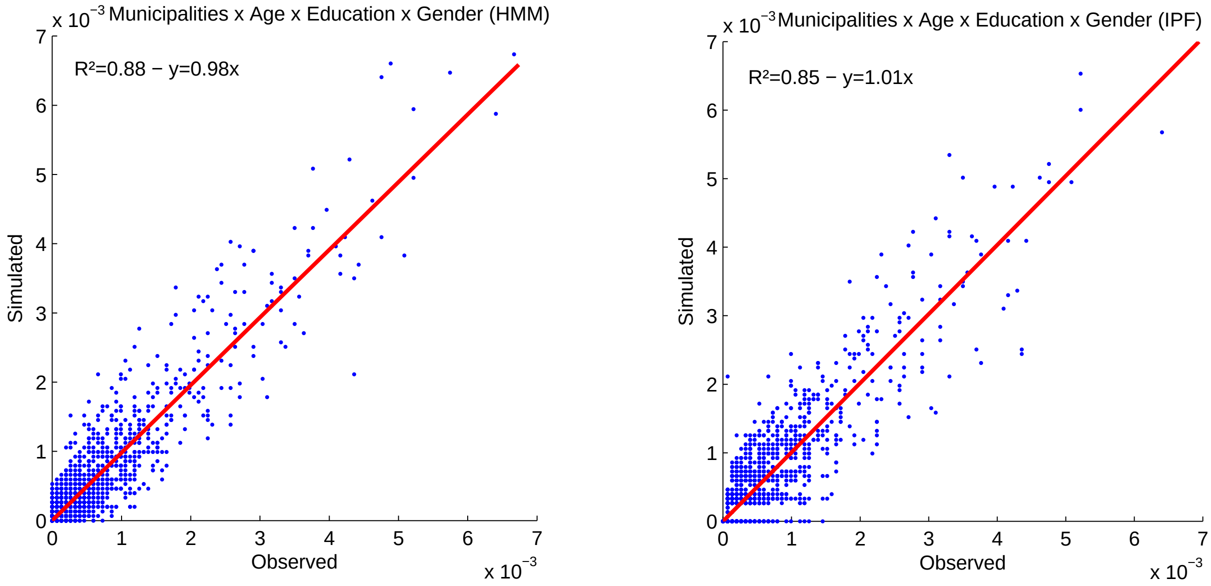
From comparing the multivariate joint distributions of the HMM approach (Fig. 10a) and the IPF approach (Fig. 10b), it appears that IPF is less capable of reproducing the complete heterogeneity present in the true population. This is especially the case for small proportions. Moreover, R^2 is slightly better for the HMM approach. A further analysis based on the SRSME, displayed in Table 3, indicates that the HMM is able to improve the quality of the synthetic population for small sampling rates when compared to IPF. This underlines the need to shift towards probabilistic approaches instead of the standard IPF when the sampling rates are (relatively) small.

4.5. HMM population synthesis using multiple data sources

Very often, not all variables of interest are included in a single data set. Depending on the nature of the problem, variables that need to be synthesized could be extracted from multiple and independent data sets. In such situations, IPF fails in synthesizing these attributes, unless more elaborate approaches consisting of multiple sub-models that incorporate IPF are considered. In this section, we illustrate that the HMM approach is particularly useful for synthesizing variables stemming from different micro-samples. Suppose that the EFT data set represents the full population; 2 sub-samples are drawn from this full population, for which the variables are indicated in Table 4. In total, 6 attributes are included in the two sub-samples, each sub-sample containing 2 variables that are not available in the others. Moreover, the sub-samples are drawn from the full population with different sampling rates (i.e., 15% and 10%). Note

Sampling rate(%)	IPF	HMM	Deviation with respect to IPF
1	0.558	0.303	0.255 (-45.70%)
2	0.370	0.266	0.104 (-28.11%)
3	0.329	0.226	0.103 (-31.31%)
4	0.263	0.183	0.080 (-30.42%)
5	0.199	0.125	0.074 (-37.19%)
10	0.137	0.098	0.039 (-28.47%)
15	0.104	0.082	0.022 (-21.15%)
20	0.091	0.077	0.014 (-15.38%)
25	0.076	0.073	0.003 (-3.95%)

Table 3: Difference between IPF and HMM in terms of SRSME



(a) HMM

(b) IPF

Figure 10: Multivariate joint distributions (ψ=20%)

that there is no limitation in the number of data sources. In this case, it is possible that the transition probabilities between two variables can be derived from multiple micro-samples. The transition probabilities can then be determined based on the most reliable micro-sample or by averaging the transition probabilities over the different micro-samples. The results of the synthetic population are compared to the marginal and joint distributions of the full population stemming from the reference data set.

Figure 11 presents the marginal distributions of the variables of interest. The HMM is capable of reproducing a correct estimation of the true population, although the information stems from three different sources: the aggregate initial distribution related to the municipalities and the two micro-samples (PUMS).

Figure 12 presents the interconnections between the variables. One can observe that the HMM manages to capture the transition probabilities between all the variables with respect to the full population. R-squared values are close to 1 for most of the joint distributions, except for the combination Municipal-

Variable	Sample 1	Sample 2
Municipalities	×	
Travelled distances	×	×
Age		×
Profession	×	×
Gender	×	
Education		×
Sample size	15%	10%

Table 4: Variables distribution within both samples

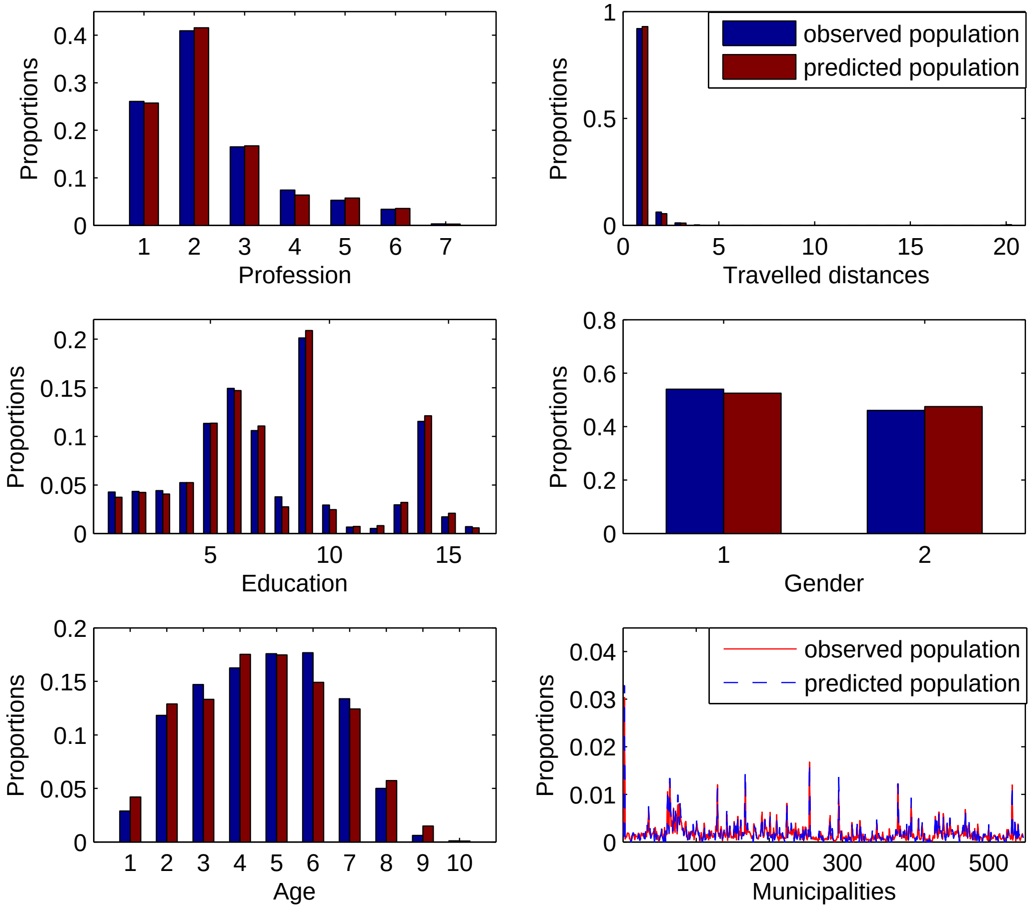


Figure 11: Marginal distributions

ities × Education with $R^2 = 0.83$. This is due to the combined effects of the propagation of the error within the transition probabilities and the significant number of categories within three variables of the data sets. From Equ. 4, the full joint probability is given by the product of the initial probability and the successive transition probabilities: P(Municipalities) × P(Travelled_distances | Municipalities) × P(Age | Travelled_distances) × P(Education | Age) × P(Profession | Education) × P(Gender | Profession). In this regard, the lowest R^2 is a result of the cumulation of the slight deviations appearing in both of the previous transition probabilities P(Age | Travelled_distances) and P(Education | Age). Furthermore, the high level of disaggregation related to the variables Municipalities (547 levels), Travelled distances (187 levels) and Age (62 levels) also increases their contribution in the overall error.

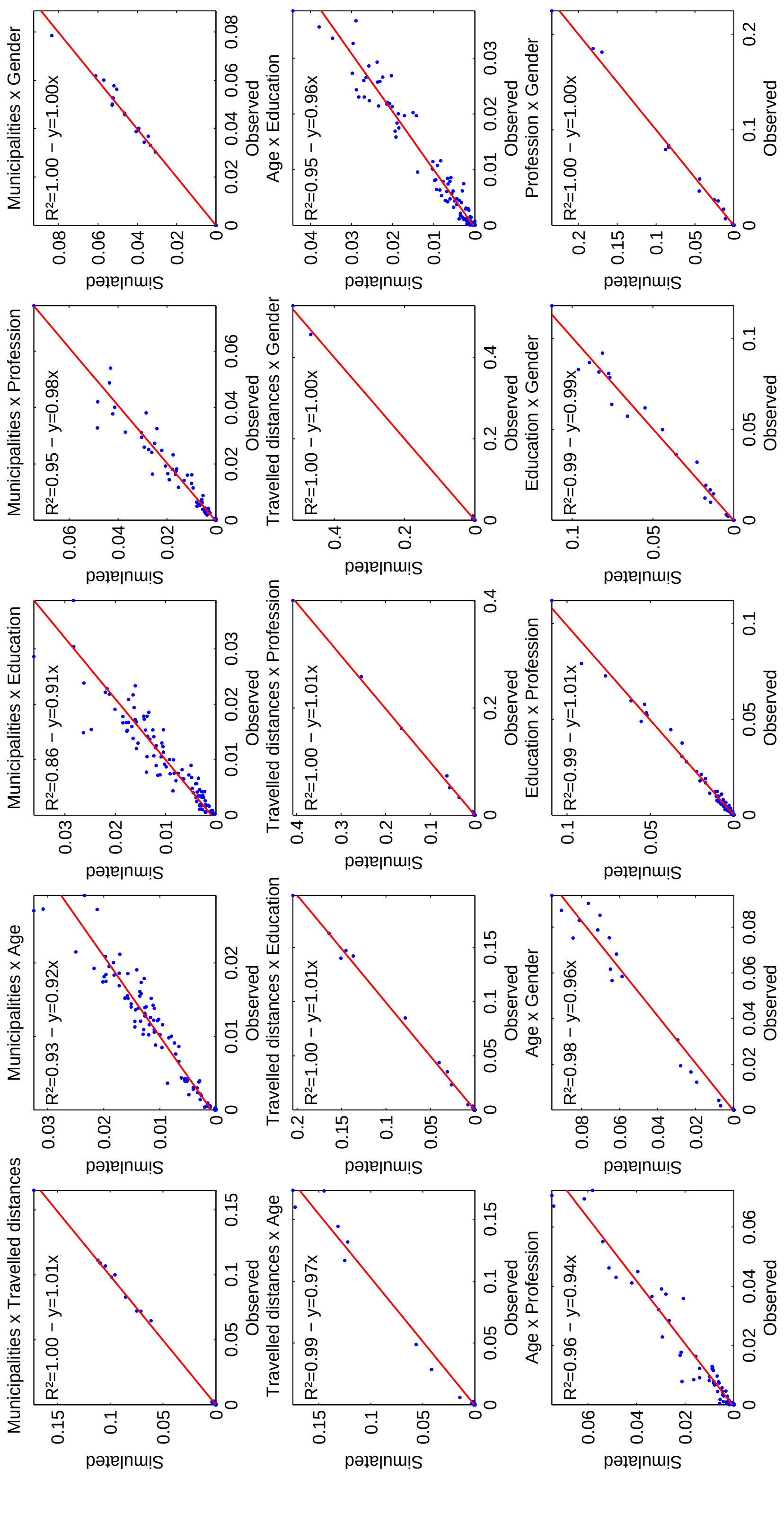


Figure 12: Joint distributions

5. Discussion and future directions

In this paper, we proposed an efficient alternative methodology to the standard approaches (fitting and reweighting methods) belonging to the Markov Process-based Methods (MPBM) to synthesize populations for micro-simulations of urban and transportation systems. This technique is able to replicate the configuration of a given population using at least one micro-sample and an initial marginal distribution. The HMM model can be controlled by an initial distribution extracted from either the micro-sample, in the worst case, or a census, if the data are available. To be more meaningful, the data sources should have been collected around the same time period. Any synthetic population size can be generated by the model because it has no influence on the computational complexity as the HMM belongs to the generation-based family. The positioning of the variables of interest should be arranged in the descending order of their number of categories to maintain a good approximation of the true population. Apart from the data cleaning procedure, data preparation requires specific attention because of the dependency of both the transition and (eventually) the emission probabilities.

As demonstrated, the HMM-based approach, as well as the standard approaches, are able to reproduce the marginal distributions of the proposed set of attributes, depending on the case study. The choices of transitioning from one variable to another are realized based on the transition probabilities. Thus, slight differences might appear because of the randomness included in the model when transitions are selected. A comparison of the joint distributions indicates small differences between the synthesized and observed populations. However, the analysis shows us a critical limitation of the model stemming from one of the following main assumptions of the Markov chain process:

$$P(\mathbf{h}_{t+1} = \mathbf{i} | \mathbf{h}_t = \mathbf{j}, \mathbf{h}_{t-1} = \mathbf{k}, ..., \mathbf{h}_0 = \mathbf{l}) = P(\mathbf{h}_{t+1} = \mathbf{i} | \mathbf{h}_t = \mathbf{j}) \tag{14}$$

where $\mathbf{h}_t$ is the vector of states at time t and P is the conditional probability. According to Equ. 14, it can be concluded that an attribute depends exclusively on the related previous attribute. In this regard, synthesizing a significant number of attributes could reduce the SRMSE as well as the R² . However, this limitation could be considered as an advantage insofar as agent attribute synthesis depends only on the previous attribute in such a way that the attribute chain is built for this agent. Note that no additional marginal distributions were necessary within the intermediary attributes, mitigating, in this manner, dependence on the data.

Regarding the scalability, there is theoretically no limitation in terms of the number of attributes to be synthesized. However, in practice, as demonstrated through the numerical examples, the more distant the attributes are in the chain, the larger the deviation between these attributes. In this context, specific attention should be paid to this phenomenon. In our paper, the number of attributes synthesized is 6. Even with a high level of disaggregation, a SRMSE of 0.54 was obtained. If a good trade-off is found between the study requirements and the number of levels per attribute, it is possible to extend the synthesis to a higher number of attributes while limiting the overall error rate.

As one of its major advantages, a HMM confers the ability to integrate an unlimited number of data sources. In multidisciplinary studies (e.g., transport and health), it often happens that the variables of interest are not recorded in a single data set. Thus, the HMM approach provides an ideal framework to obtain good estimates of populations for which the entire series of variables is available. Moreover, differences in sample sizes of the individual data sources do not affect the results.

A comparison of the HMM with the most common population synthesis technique, i.e., IPF, illustrates the advantages of HMM over IPF. The comparison reveals that for realistic sampling rates (< 25%), the HMM provides better results in terms of SRMSE. Moreover, the amount of data required by the HMM (1 micro-sample and 1 marginal distribution) is less than that for IPF (1 micro-sample and all marginal distributions).

Finally, matching between households and individuals is not investigated because it is not within the scope of this paper. Extending the HMM-based approach for grouping individuals into households according to the standard procedure can be realized. Existing matching methods could be applied to create a synthetic population consisting of both households and individuals simultaneously. Most of the association techniques operate according to the above described procedure (Anderson et al., 2014; Barthelemy and Toint, 2013; Pritchard and Miller, 2012; Ye et al., 2009). Given the importance of household characteristics on daily travel patterns Barthelemy and Toint (2013), further research is needed to include both individual and household information in the HMM-based population synthesis.

6. Acknowledgments

The research was funded by the ARC grant for Concerted Research Actions and by the Special Fund for Research, both financed by the Wallonia-Brussels Federation.

515 # AppendixA. Marginal and joint distributions related to SP5 and SP6

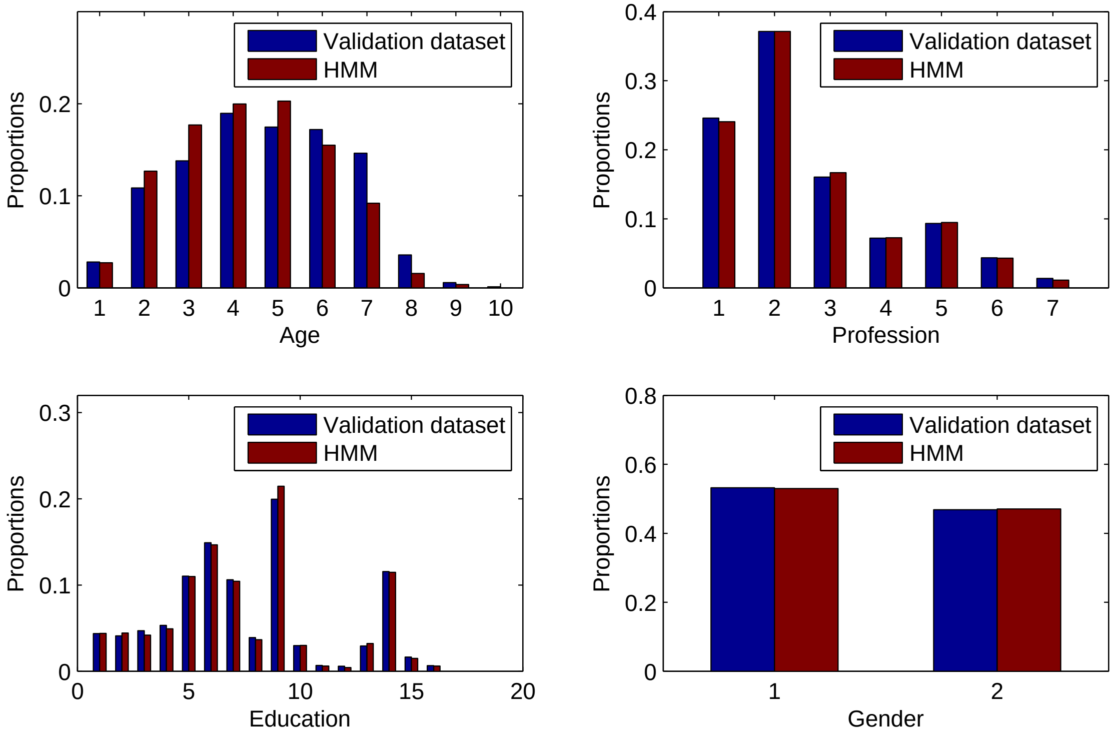


Figure A.13: Comparison of the marginal distributions for different attributes (SP5)

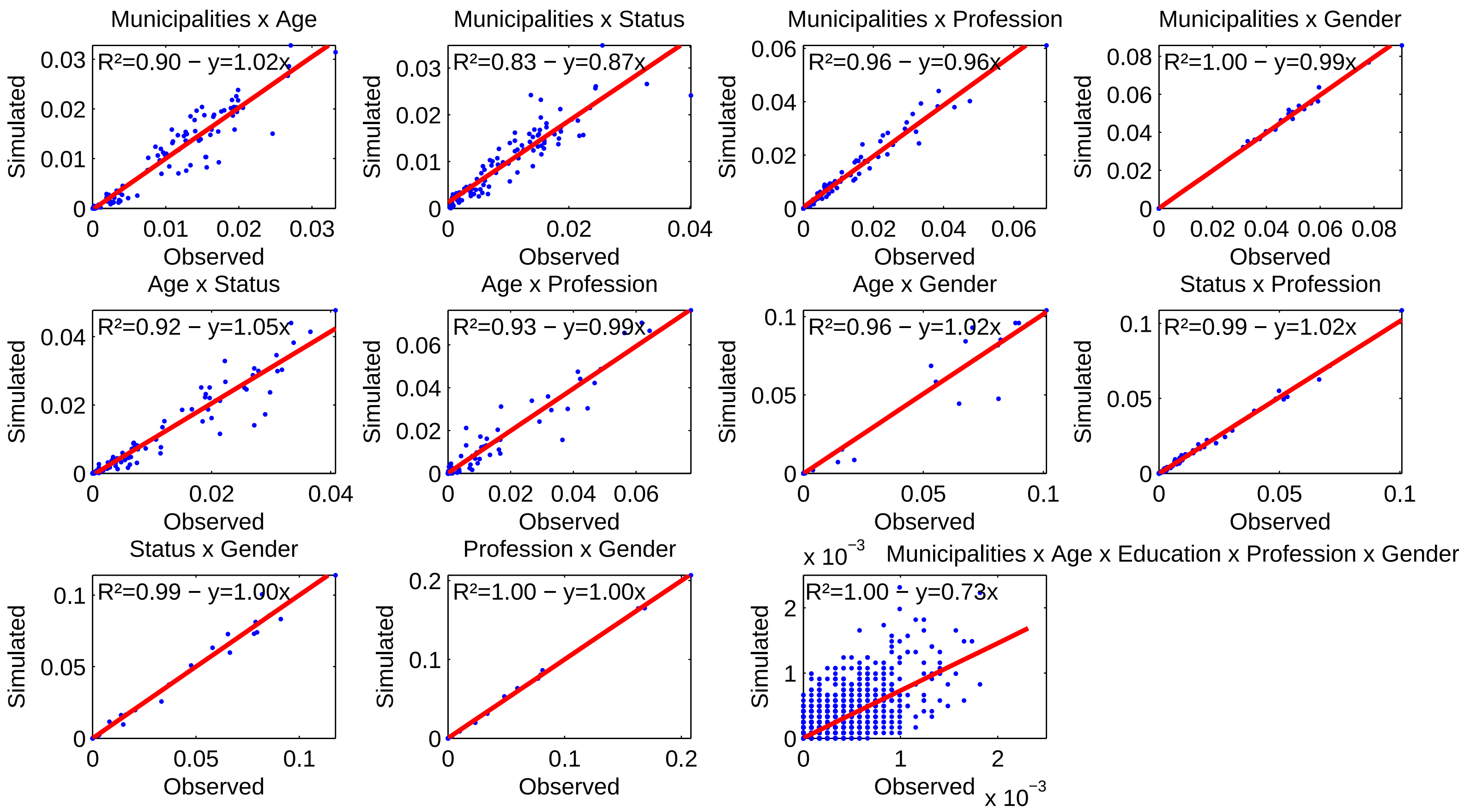


Figure A.14: Fit between the real population and HMM-based approach synthesis (SP5)

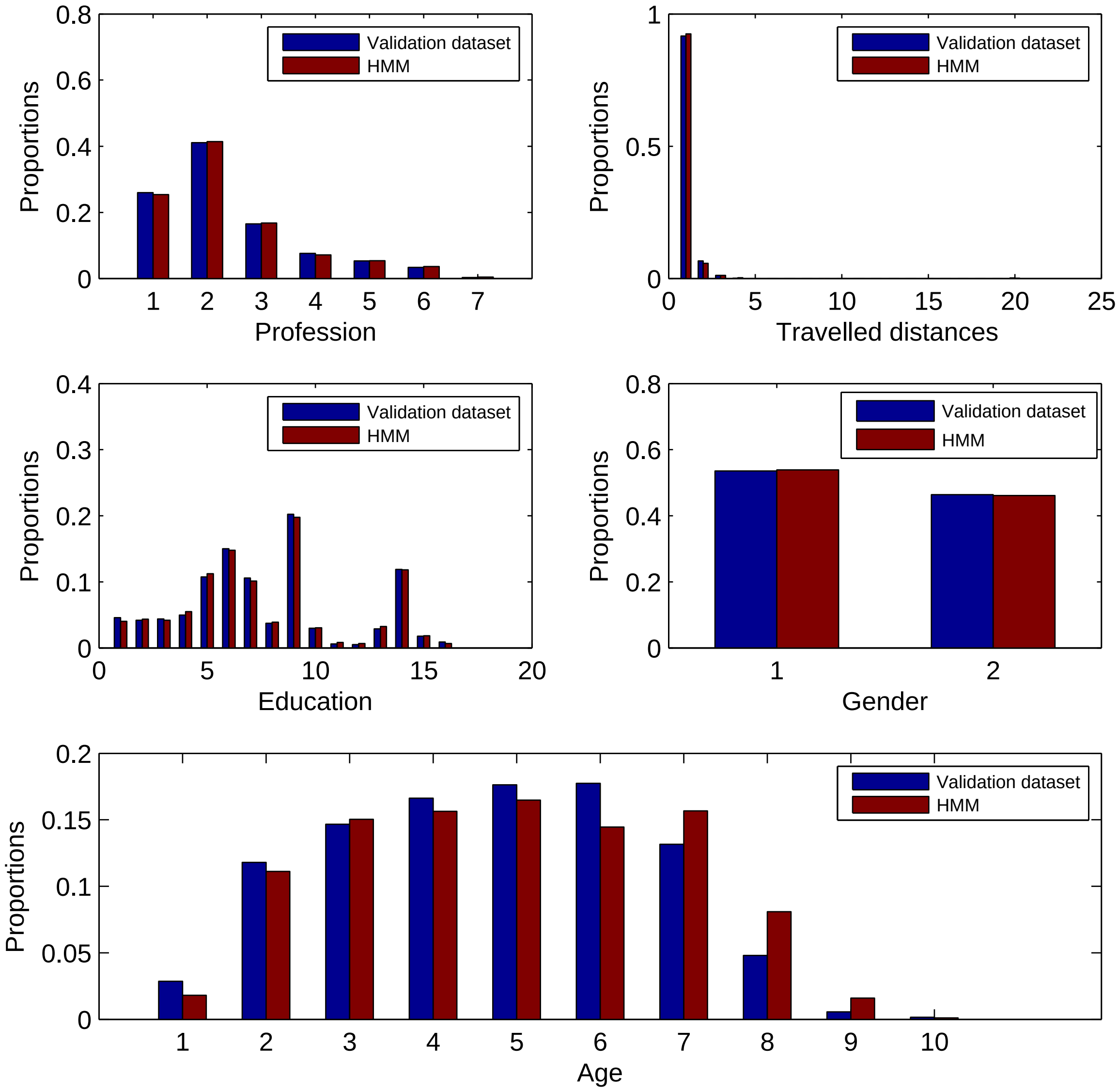


Figure A.15: Comparison of the marginal distributions for different attributes (SP6)

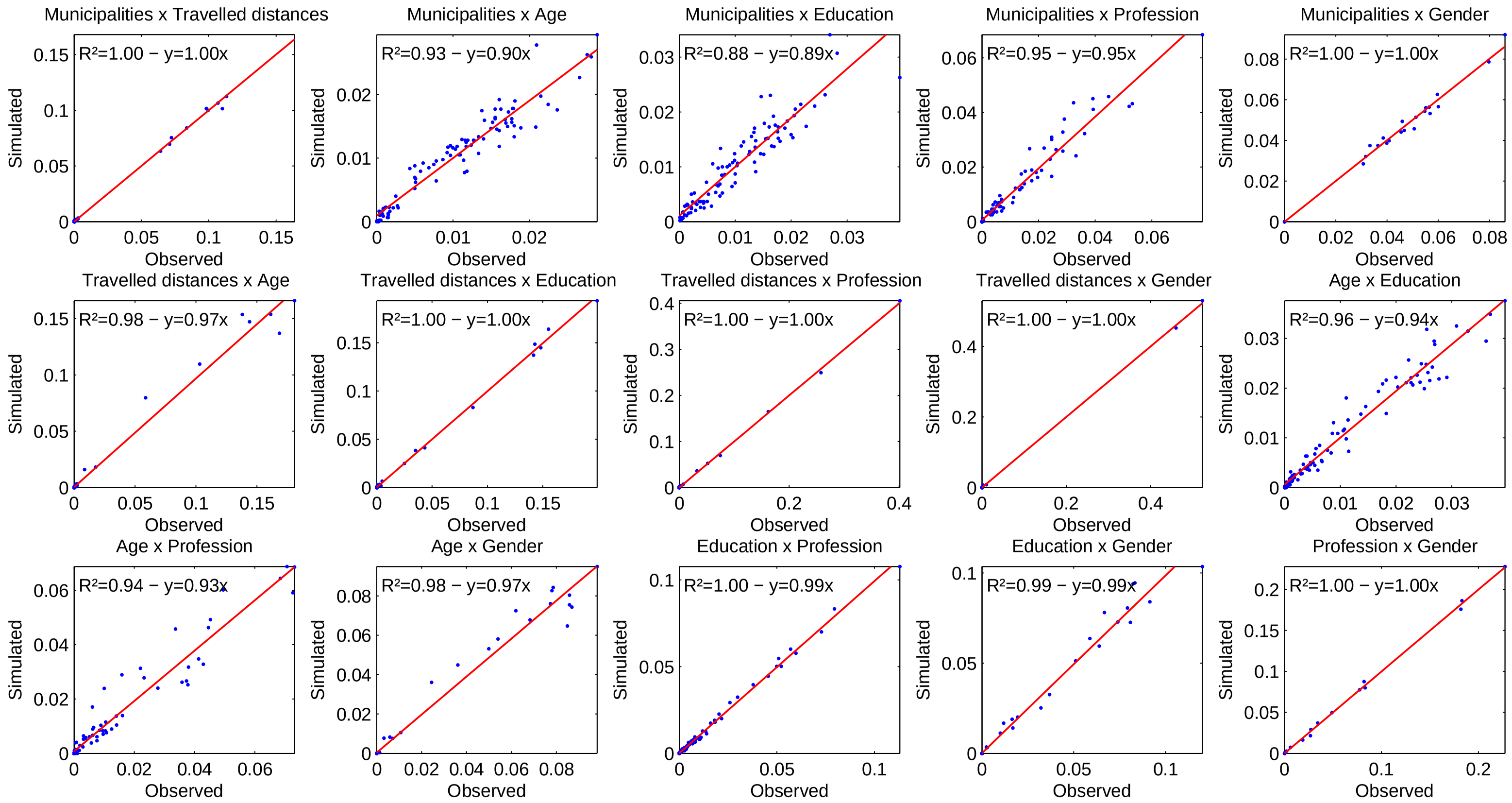


Figure A.16: Fit between the real population and HMM-based approach synthesis (SP6)

References

Anderson, P., Farooq, B., Efthymiou, D., Bierlaire, M., 2014. Associations generation in synthetic population for transportation applications. Transportation Research Record: Journal of the Transportation Research Board 2429, 38–50. doi:`http://dx.doi.org/10.3141/2429-05`.

Badsberg, J.H., Malvestuto, F.M., 2001. An implementation of the iterative proportional fitting procedure by propagation trees. Computational Statistics & Data Analysis 37, 297–322. doi:`http://dx.doi.org/10.1016/S0167-9473(01)00013-5`.

Balmer, M., Axhausen, K., Nagel, K., 2006. Agent-based demand-modeling framework for large-scale microsimulations. Transportation Research Record: Journal of the Transportation Research Board 1985, 125–134. doi:`http://dx.doi.org/10.3141/1985-14`.

Barthelemy, J., Suesse, T., Namazi-Rad, M., 2015. Multidimensional iterative proportional fitting and alternative models URL: `https://github.com/jojo-/mipfp`.

Barthelemy, J., Toint, P., 2015. A stochastic and flexible activity based model for large population. application to belgium. Journal of Artificial Societies and Social Simulation 18, 15. doi:`http://dx.doi.org/10.18564/jasss.2819`.

Barthelemy, J., Toint, P.L., 2013. Synthetic population generation without a sample. Transportation Science 47, 266–279. doi:`http://dx.doi.org/10.1287/trsc.1120.0408`.

Beckman, R.J., Baggerly, K.A., McKay, M.D., 1996. Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice 30, 415–429. doi:`http://dx.doi.org/10.1016/0965-8564(96)00004-3`.

Bekhor, S., Dobler, C., Axhausen, K., 2011. Integration of activity-based and agent-based models. Transportation Research Record: Journal of the Transportation Research Board 2255, 38–47. doi:`http://dx.doi.org/10.3141/2255-05`.

Caiola, G., Reiter, J.P., 2010. Random forests for generating partially synthetic, categorical data. Transactions on Data Privacy 3, 27–42. URL: `http://www.tdp.cat/issues/tdp.a033a09.pdf`.

Denteneer, D., Verbeek, A., 1985. A fast algorithm for iterative proportional fitting in log-linear models. Computational Statistics & Data Analysis 3, 251–264. doi:`http://dx.doi.org/10.1016/0167-9473(85)90088-X`.

Duguay, G., Jung, W., McFadden, D., 1976. SYNSAM: A Methodology for Synthesizing Household Transportation Survey Data. Urban Travel Demand Forecasting Project, Institute of Transportation Studies.

Endo, Y., Takemura, A., 2009. Iterative proportional scaling via decomposable submodels for contingency tables. Computational Statistics & Data Analysis 53, 966–978. doi:`http://dx.doi.org/10.1016/j.csda.2008.11.013`.

Farooq, B., Bierlaire, M., Hurtubia, R., Flötteröd, G., 2013. Simulation based population synthesis. Transportation Research Part B: Methodological 58, 243–263. doi:`http://dx.doi.org/10.1016/j.trb.2013.09.012`.

Gargiulo, F., Ternes, S., Huet, S., Deffuant, G., 2010. An iterative approach for generating statistically realistic populations of households. PloS one 5, 1–9. doi:http://dx.doi.org/10.1371/journal.pone.0008828.

Geard, N., McCaw, J.M., Dorin, A., Korb, K.B., McVernon, J., 2013. Synthetic population dynamics: A model of household demography. Journal of Artificial Societies and Social Simulation 16. doi:http://dx.doi.org/10.18564/jasss.2098.

Hermes, K., Poulsen, M., 2012. A review of current methods to generate synthetic spatial microdata using reweighting and future directions. Computers, Environment and Urban Systems 36, 281–290. doi:http://dx.doi.org/10.1016/j.compenvurbsys.2012.03.005.

Ibe, O.C., 2013. 14 - hidden markov models, in: Ibe, O.C. (Ed.), Markov Processes for Stochastic Modeling (Second Edition). Elsevier, Oxford, pp. 417–451. doi:http://dx.doi.org/10.1016/B978-0-12-407795-9.00014-1.

Jiroušek, R., Přeučil, S., 1995. On the effective implementation of the iterative proportional fitting procedure. Computational Statistics & Data Analysis 19, 177–189. doi:http://dx.doi.org/10.1016/0167-9473(93)E0055-9.

Knudsen, D.C., Fotheringham, A.S., 1986. Matrix comparison, goodness-of-fit, and spatial interaction modeling. International Regional Science Review 10, 127–147. doi:http://dx.doi.org/10.1177/016001768601000203.

Lenormand, M., Deffuant, G., 2012. Generating a synthetic population of individuals in households: Sample-free vs sample-based methods. Journal of Artificial Societies and Social Simulation 16. doi:http://dx.doi.org/10.18564/jasss.2319.

Müller, K., Axhausen, K.W., 2010. Population synthesis for microsimulation: State of the art. Eth zürich, institut für verkehrsplanung, transporttechnik, strassen-und eisenbahnbau (ivt) ed.

Namazi-Rad, M.R., Mokhtarian, P., Perez, P., 2014. Generating a dynamic synthetic population-using an age-structured two-sex model for household dynamics. PloS one 9, 1–16. doi:http://dx.doi.org/10.1371/journal.pone.0094761.

Pritchard, D.R., Miller, E.J., 2012. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. Transportation 39, 685–704. doi:http://dx.doi.org/10.1007/s11116-011-9367-4.

Rich, J., Mulalic, I., 2012. Generating synthetic baseline populations from register data. Transportation Research Part A: Policy and Practice 46, 467–479. doi:http://dx.doi.org/10.1016/j.tra.2011.11.002.

Rieser, M., Nagel, K., Beuck, U., Balmer, M., Rümenapp, J., 2007. Agent-oriented coupling of activity-based demand generation with multiagent traffic simulation. Transportation Research Record: Journal of the Transportation Research Board 2021, 10–17. doi:http://dx.doi.org/10.3141/2021-02.

Saadi, I., Mustafa, A., Teller, J., Cools, M., 2016. An integrated framework for forecasting travel behavior using markov chain monte carlo simulation and profile hidden markov models, in: Proceedings of the 95th Annual Meeting of the Transportation Research Board, Transportation Research Board of the National Academies, Washington, D.C.

Sun, L., Erath, A., 2015. A bayesian network approach for population synthesis. Transportation Research Part C: Emerging Technologies 61, 49–62. doi:http://dx.doi.org/10.1016/j.trc.2015.10.010.

Tirumalachetty, S., Kockelman, K.M., Nichols, B.G., 2013. Forecasting greenhouse gas emissions from urban regions: microsimulation of land use and transport patterns in austin, texas. Journal of Transport Geography 33, 220–229. doi:http://dx.doi.org/10.1016/j.jtrangeo.2013.08.002.

Visser, I., Speekenbrink, M., 2010. depmixs4: An r package for hidden markov models. Journal of Statistical Software 36, 1–21.

Voas, D., Williamson, P., 2000. An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. International Journal of Population Geography 6, 349–366. doi:http://dx.doi.org/10.1002/1099-1220(200009/10)6:5<349::AID-IJPG196>3.0.CO;2-5.

Vovsha, P., Hicks, J.E., Paul, B.M., Livshits, V., Maneva, P., Jeon, K., 2015. New features of population synthesis, in: Proceedings of the 94th Annual Meeting of the Transportation Research Board, Transportation Research Board of the National Academies, Washington, D.C.

Waddell, P., 2002. Urbansim: Modeling urban development for land use, transportation, and environmental planning. Journal of the American Planning Association 68, 297–314. doi:http://dx.doi.org/10.1080/01944360208976274.

Williamson, P., 2013. An evaluation of two synthetic small-area microdata simulation methodologies: Synthetic reconstruction and combinatorial optimisation, in: Spatial Microsimulation: A Reference Guide for Users. Springer Netherlands. Understanding Population Trends and Processes, pp. 19–47. doi:http://dx.doi.org/10.1007/978-94-007-4623-7_3.

Williamson, P., Birkin, M., Rees, P.H., 1998. The estimation of population microdata by using data from small area statistics and samples of anonymised records. Environment and Planning A 30, 785–816. doi:http://dx.doi.org/10.1068/a300785.

Yang, B., Janssens, D., Ruan, D., Cools, M., Bellemans, T., Wets, G., 2012. A data imputation method with support vector machines for activity-based transportation models, in: Foundations of Intelligent Systems. number 122 in Advances in Intelligent and Soft Computing, pp. 249–257. doi:http://dx.doi.org/10.1007/978-3-642-25664-6_29.

Yasmin, F., Morency, C., Roorda, M.J., 2015. Assessment of spatial transferability of an activity-based model, tasha. Transportation Research Part A: Policy and Practice 78, 200–213. doi:http://dx.doi.org/10.1016/j.tra.2015.05.008.

Ye, X., Konduri, K.C., Pendyala, R.M., Sana, B., Waddell, P., 2009. Methodology to match distributions of both household and person attributes in generation of synthetic populations, in: Proceedings of the 88th Annual Meeting of the Transportation Research Board, Transportation Research Board of the National Academies, Washington, D.C.