An integrated framework for forecasting travel behavior using Markov Chain Monte Carlo simulation and profile Hidden Markov Models

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

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

Abstract

Recent advances in agent-based micro-simulation modeling have further highlighted the importance of a thorough full synthetic population procedure for guaranteeing the correct characterization of real-world populations and underlying travel demands. In this regard, we propose an integrated approach including Markov Chain Monte Carlo (MCMC) simulation and profiling-based methods to capture the behavioral complexity and the great heterogeneity of agents of the true population through representative micro-samples. The population synthesis method is capable of building the joint distribution of a given population with its corresponding marginal distributions using either full or partial conditional probabilities or both of them simultaneously. In particular, the estimation of socio-demographic or transport-related variables and the characterization of daily activity-travel patterns are included within the framework. The fully probabilistic structure based on Markov Chains characterizing this framework makes it innovative compared to standard activity-based models. Moreover, data stemming from the 2010 Belgian Household Daily Travel Survey (BELDAM) are used to calibrate the modeling framework. We illustrate that this framework effectively captures the behavioral heterogeneity of travelers. Furthermore, we demonstrate that the proposed framework is adequately adapted to meeting the demand for large-scale micro-simulation scenarios of transportation and urban systems.

Keywords: Travel behavior analysis, travel demand modeling, activity sequences, daily activity-travel patterns, simulation-based population synthesis, profiling analysis

1. Background

Activity-based analyses of travel behavior within large-scale micro-simulation models are particularly adapted to understanding the dynamics and the transitional patterns of travel behavior. In this context, most activity-based models require a full (synthetic) population to obtain greater behavioral realism within such complex transport and urban systems. However, the lack of highly disaggregate data suggests the use of more efficient methods for synthesizing individual/household socio-demographic attributes as well as their daily activity information (Beckman et al., 1996). Obtaining accurate individual-level information for a large population is a great challenge especially in the context of restrictive data availability. Typically, such detailed data can be derived from national censuses. However, in practice, only aggregate information is available to researchers and practitioners (Anderson et al., 2014).

*Corresponding author
Email address: ismail.saadi@ulg.ac.be (Ismaïl Saadi)

Furthermore, micro-samples may also include incomplete observations, making the true population identification process through a virtual process more complex. To overcome this problem, several techniques have been developed to address multiple data sources and incomplete datasets. Generally, detailed individual-level micro-data and aggregate large-scale datasets are the two components used as input for population synthesizing purposes. In the literature, the classical techniques identified are Iterative Proportional Fitting (IPF) (Beckman et al., 1996; Mohammadian et al., 2010), Iterative Proportional Updating (IPU) (Ye et al., 2009), Combinatorial Optimization (CO) (Williamson et al., 1998; Voas and Williamson, 2000) and probabilistic models (Farooq et al., 2013; Sun and Erath, 2015; Saadi et al., 2016b).

In this paper, we opt for a simulation-based approach because of its ability to address incomplete datasets using partial conditional distributions (Farooq et al., 2013). More specifically, this operational approach based on a Gibbs sampler can generate agents using partial views of the true joint distribution as if the synthesized agents were drawn from the real population. In this context, the method enables the generation of populations of any given size. A comparative study established by Farooq et al. (2013) between IPF and simulation-based approaches revealed that the latest technique clearly outperforms IPF using, for both methods, almost the same amount of data. We build a representation of the true population by implicitly merging different information provided by multiple micro-data sources using the simulation-based approach. Then, from this representative/synthesized population, we can group the population into homogeny clusters characterized by similar attributes to understand their related travel behavior and to enable a comparative attribute-based study based on the activity sequences.

For several decades, characterizing people's activity-travel behavior has been an important issue to researchers (Pendyala and Goulias, 2002; Auld et al., 2015; Saadi et al., 2016a). To include the sequential dependencies of daily activities, some studies have suggested the use of Sequence Alignment Methods (SAMs) (Joh and Timmermans, 2011; Joh et al., 2002; Wilson, 1998), whose inputs representing the daily activity behavior in the form of activity sequences are extracted from activity-travel diaries (Bhat and Singh, 2000; Spissu et al., 2009). Typically, the application of SAMs includes pairwise sequence alignment algorithms for scoring and comparing activity sequences in-between them. In this way, the main activity patterns can be identified quantitatively and qualitatively within their respective subset of activity chains (Joh et al., 2006).

When a relationship can be established between the derived patterns and the variables of interest, general activity-travel behavior trends can be analyzed to investigate more deeply the determinants influencing those travel patterns (Wilson, 1998). Although SAMs have been extensively used for characterizing activity-travel patterns, such methods clearly suffer from serious limitations. Indeed, SAMs are unable to identify the complete nature of a cluster. Even if the information regarding the frequent activity patterns is extracted, irregular activities are not considered; thus, only parts of the clusters are characterized (Liu et al., 2015). In this context, a number of questions have been raised regarding the consistency of such sequence alignment techniques.

In an attempt to obtain improvements, Liu et al. (2015) proposed a profiling method called profile Hidden Markov Models (pHMMs) for characterizing the complete set of activity sequences including irregular activities. pHMMs belong to the family of sequence profile methods that are essentially used to characterize protein sequences. Furthermore, pHMMs may be described as position-specific scoring parameters emanating from a multiple alignment of a group of protein sequences (Durbin et al., 1998).

From a methodological point of view, the positions of alignment are categorized according to three possible states: match, insertion and deletion. Fundamentally, the configuration of the model is a successive layout of match states representing the conserved activities that have been identified within their respective cluster. This successive layout forms the reference activity sequences or a base template characterized by the most frequent activities. In parallel, insertion states model the introduction of new conserved residues, which can be interpreted as additional activities incorporated into the previously defined base template.

Conversely, deletion states represent the omission of particular conserved activities from the base template. In this regard, random differences between the activity sequences within the overall characterization process are implicitly included throughout these two states. Subsequently, estimated pHMMs can generate new activity sequences so that the relationships with their corresponding cluster are preserved (Liu et al., 2015).

Furthermore, pHMMs can be considered as quantitative descriptors that assign weights - probabilities of occurrence - to each activity at each corresponding match state characterizing both regular and irregular activities. For instance, considering daily activity-travel sequences, some travelers might see their behavior differing from the general behavior of a cluster (e.g., work cluster). This difference is captured by pHMMs by introducing new activities or, inversely, canceling or ignoring others. Subsequently, the scored activity sequences can be assessed to measure how similar or dissimilar these are from the cluster (Liu et al., 2015).

In this paper, we propose an integrated framework including both a population synthesis approach (Farooq et al., 2013) and a profiling method (Liu et al., 2015) capable of describing and assigning the activity sequences for each individual of the synthetic population. The model is capable of considering, in an efficient manner, the behavioral effects of different influencing factors, which might vary between clusters. In this regard, the main trends in terms of travel behavior can emerge from the characterization achieved using the pHMMs.

The remainder of this paper is organized as follows. First, the data, obtained from the Belgian National Household Travel Survey, are described. Subsequently, in Section 3, the details of the integrated framework are provided to enable implementation in different contexts. Then, the main results and the validation of the integrated framework are discussed (Section 4). Finally, the concluding remarks are formulated in Section 5.

2. Data

To investigate individuals' daily activity-travel behaviors, we use data from the Belgian National Household Travel Survey of 2010 (Cornelis et al., 2012). The data include 37,680 recorded trips with respect to 15,821 individuals distributed across the country. For each respondent, the data include a detailed sequencing of trips with their respective characteristics (e.g., time expenditure, start-end locations, trip purpose, and mode preferences). With respect to the variables of interest, age, gender, socio-professional status and working time expenditure are considered in the modeling framework (Table 1). Furthermore, public transport subscription and driving license ownership are also synthesized as transport-related variables.

Variable	**Basic Statistics**
Age	Mean: 46.54, Std. Dev.: 21.08
Gender	Male: 47.87%, Female: 52.13%
Socio-Professional status	Not schooled children: 0.08% - Student: 17.23% - Housewife (husband): 4.30%
	Job seeker: 5.54% - Pensioner: 28.14% - Disabled person: 2.23% Blue-collar worker: 7.72%
	White-collar worker (executive): 3.51% White-collar worker (non-executive): 21.72%
	Self-employed person: 3.96% - Liberal profession: 1.11% - Teacher: 3.76% Farmer: 0.23% - Other: 0.48%
Work time expenditure	Mean: 36.68 h/week, Std. Dev.: 11.36 h/week
Public transport subscription	No: 77.33%, Yes: 22.67%
Driving license ownership	In progress: 4.90%, No: 24.87%, Yes: 70.23%

Table 1: Data description of the arbitrary explanatory variables

Table 2 presents the percentage of the recorded trips where, the outcome is the trip purpose and the explanatory variables are age, gender, socio-professional status of the travelers, working time expenditure, public transport subscription and driving license ownership. The values of the different explanatory variables are expressed in terms of proportion (%) except for working time expenditure, which is expressed in

Trip purpose (*)	1	2	3	4	5	6	7	8	9	10	11	12
Age												
6-31 years	4.89	40.03	9.05	1.30	14.51	2.09	7.19	1.54	7.51	2.13	7.06	2.68
31-45 years	13.46	37.20	16.77	2.96	0.63	2.41	10.62	2.42	4.99	1.99	4.25	2.30
45-59 years	7.25	38.83	16.73	3.06	0.49	1.92	13.06	3.59	5.74	3.15	3.82	2.37
59+	5.36	40.36	3.10	0.72	0.32	1.88	20.27	5.94	7.94	5.30	4.71	4.10
Gender												
Male	6.64	39.32	12.79	3.10	4.09	2.32	11.14	2.95	6.31	3.05	5.44	2.85
Female	8.77	38.96	10.14	1.04	4.25	1.79	14.04	3.70	6.78	3.15	4.54	2.85
Socio-professional Status												
Student	3.51	41.72	1.49	0.19	23.78	1.75	5.72	1.04	7.21	2.08	8.58	2.92
Housewife (husband)	13.79	39.41	0.74	0.00	0.99	1.48	19.21	4.93	7.88	3.69	4.19	3.69
Job seeker	9.30	39.47	2.09	0.76	3.04	2.28	15.94	4.17	9.11	5.69	4.17	3.98
Pensioner	5.82	40.62	0.99	0.23	0.33	1.83	20.92	6.29	8.11	5.53	5.11	4.22
Disabled person	9.55	40.76	0.64	0.00	1.27	2.55	17.20	7.01	7.64	5.10	5.10	3.18
Blue-collar worker	7.08	40.05	24.52	4.63	0.27	1.09	8.72	1.91	5.45	1.77	3.00	1.50
White-collar worker (executive)	8.64	35.95	21.41	3.93	0.59	3.73	9.63	2.55	4.13	1.77	5.11	2.55
White-collar worker (non-executive)	9.43	37.45	20.48	2.91	0.39	2.26	10.72	2.51	5.52	2.08	4.23	2.01
Self-employed person	8.82	36.65	21.49	6.79	0.23	2.26	8.82	2.71	4.52	1.81	3.17	2.71
Liberal profession	7.78	34.73	20.36	9.58	0.60	2.99	9.58	2.99	4.79	1.80	2.99	1.80
Teacher	10.71	38.16	16.35	1.50	0.56	2.07	12.22	3.01	6.58	2.07	4.32	2.44
Farmer	10.53	42.11	10.53	5.26	0.00	0.00	10.53	0.00	10.53	5.26	5.26	0.00
Other	8.00	38.00	8.00	2.00	2.00	6.00	12.00	4.00	4.00	4.00	8.00	4.00
Working time expenditure												
Mean (h/week)				42.19								
Std. Dev. (h/week)				11.12								
Public transport subscription												
No	8.48	38.92	11.71	2.28	2.92	1.99	13.04	3.32	6.62	3.07	4.77	2.89
Yes	5.04	40.28	10.40	1.24	8.55	2.30	10.98	3.33	6.25	3.20	5.71	2.72
Driving license ownership												
In progress	5.69	40.66	6.95	0.63	11.52	2.67	10.18	2.95	7.23	2.74	5.62	3.16
No	3.49	42.39	4.05	0.31	15.87	1.60	10.36	2.69	6.84	3.14	6.61	2.66
Yes	8.81	38.39	13.39	2.53	1.05	2.14	13.23	3.49	6.43	3.11	4.56	2.88

(*) 1-bring/get, 2-home, 3-work, 4-for work, 5-education, 6-meal, 7-daily-shopping, 8-service, 9-visit, 10-tour, 11-entertainment, 12-other

Table 2: Cross-classification of recorded trips by purpose within each demographic segment (in %)

hours/week. Regarding the general distribution of the trips, one can clearly observe that the trips toward home are the most important in terms of proportions.

Furthermore, it should be emphasized that commuting patterns also account for a relatively significant share of trips and are mainly represented by professionally active people (see socio-professional status). Young people essentially undertake trips (14.51%) with the objective of attending a school or university to study. In parallel, they are active for visiting friends and family (7.51%) and also participating in extra-activities (7.06%) (e.g., sports and entertainment). Individuals belonging to the oldest age category spend a significant amount of time on daily shopping (20.27%) and visiting (7.94%).

Regarding working time expenditure, we can observe that the average number of hours is relatively high (42.19 hours/week). This variable is an indicator of activity time expenditure (work). As mentioned later in the paper, the pHMMs do not represent exact temporal information. Thus, it is technically possible to synthesize a variable that can be used afterward to classify the population and obtain some trends in terms of activity durations.

3. Methodology

3.1. Framework

To develop the integrated travel demand modeling framework presented in Figure 1, two types of datasets are used: [A] socio-demographic/transport-related variables derived from the individuals' file of the Belgian Household Travel Survey and [B] the trips file including the activity-travel diaries. As detailed before, we propose to select a set of pertinent socio-demographic explanatory variables as the basis for comparing the different clusters in terms of activity-travel behavior. Subsequently, activity sequences will be derived from the trip diaries. In this context, we will obtain a set of individuals with socio-demographic information (e.g., age, gender, socio-professional status and working time expenditure) and transport-related variables (i.e., public transport subscription and driving license ownership) associated with their respective activity sequences.

At this level, the data processing will be performed in parallel and in a completely independent manner. On one side, a synthetic population [2] is built using a simulation-based approach [1]. The population synthesis procedure plays a key role within the modeling framework because it provides a better estimate of the heterogeneity of the population in comparison to standard population synthesis techniques such as IPF (Farooq et al., 2013). In this regard, the synthesized population represents a better approximation and is true regardless of the selected attribute. A direct implication is that this approach enables the estimation of the precise proportion of clusters regarding the whole population.

In parallel with the population synthesis, we derive the activity sequences from the activity-travel diaries [B] so that every respondent is associated with a daily activity-travel pattern/plan. At the end of step [3], we will have a full synthetic population as well as a detailed list of individuals characterized by socio-demographic attributes and transport-related variables as well as their respective activity sequences.

In the following step [4], it is possible to establish completely homogeneous clusters to isolate the effects of the various explanatory variables to achieve more accurate analysis. Moreover, it is also important to measure the coupled effects of mixed factors for investigating potential interactions.

Regarding the characterization of the activity sequences with profile Hidden Markov Models, a procedurally less complicated version of Liu et al. (2015) is implemented to gain computational efficiency in the calibration phase. In the approach proposed by Liu et al. (2015), before calibrating the Markov Chain profiles, it is necessary to classify the activity sequences according to their longest activities. Then, an identification process of the most recurring activities, based on the definition of the regularity (Liu et al., 2015), is recommended along with the determination of the most frequent activity transitions. Finally, for every subdivision, templates are defined so that the activity sequences related to the clusters are aligned by employing multiple sequence alignment methods based on their respective template. It is only after these three steps that performed that the calibration of the Markov profiles becomes possible. We can clearly notice that the method of Liu et al. (2015) is relatively heavy to implement in its entirety. In this context, we suggest a less heavy methodology to implement while maintaining the key component of the modeling chain represented by the pHMM. This can be realized by estimating the effects of a combination of explanatory variables of the activity sequences.

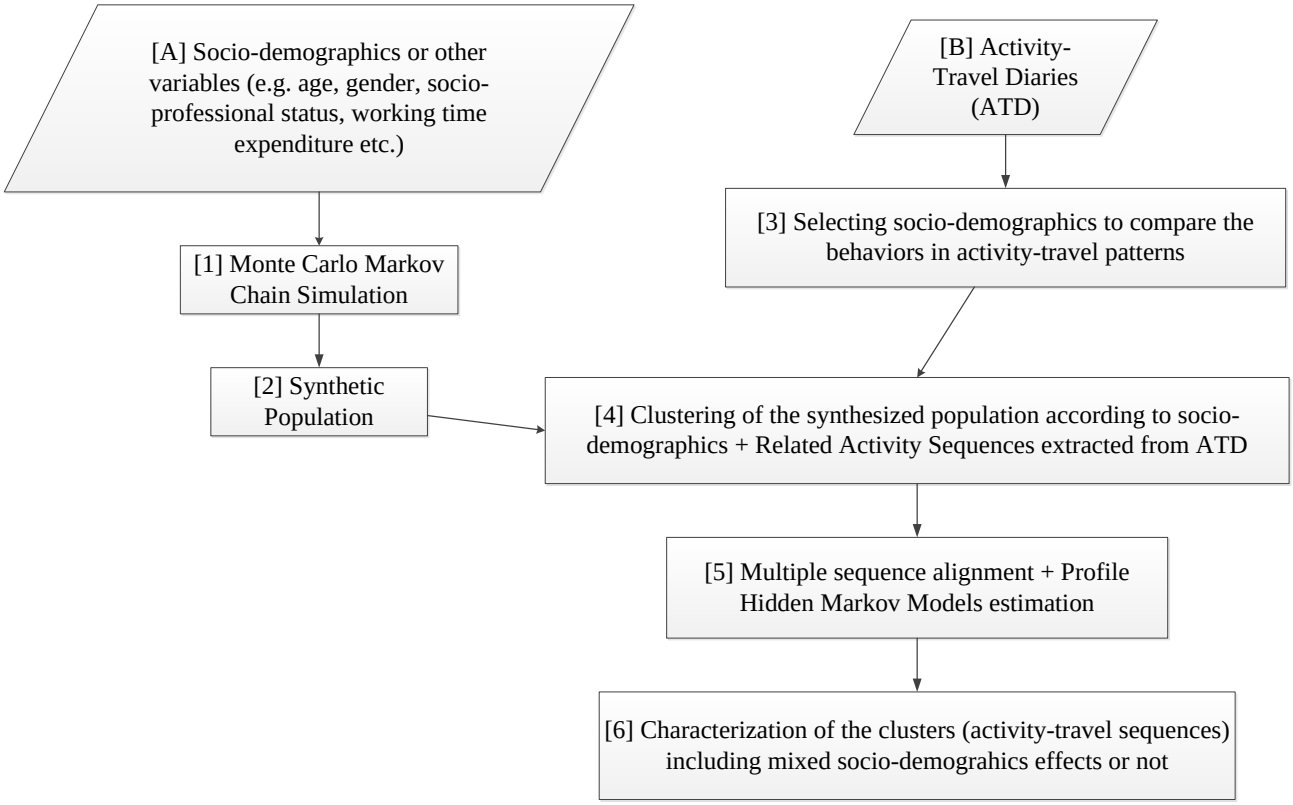


Figure 1: Overall modeling framework

With respect to the estimation of the profile Hidden Markov Models [5], the procedure consists of aligning the activity sequences within each cluster using an existing multiple sequence alignment approach (SAM). Subsequently, the transition and emission probabilities of the different pHMMs can be estimated so that every group is characterized by its dedicated model. In this context, the daily activities of the individuals are qualitatively and quantitatively simulated. In addition, both regular and irregular activities are implicitly included in the estimation conducted by the pHMM.

The strength of this framework lies in its capacity to characterize daily activity-travel sequences of very heterogeneous groups; in contrast, a purely descriptive statistical analysis would become quasi-incomprehensible and difficult to dress.

3.2. Synthetic population

Posterior samples are generated by a Gibbs sampling procedure using partial views of the true joint probability, i.e., the conditional probabilities. Theoretically, full conditional distributions should be defined and included within the algorithm; however, full conditionals are rarely available in practice. Thus, partial or even marginal distributions are used as substitutes (Farooq et al., 2013). Let $q(\mathbf{x})$ be an initial set of random attributes, $\mathbf{x} = (x_1, x_2, ..., x_i, x_n)$ the set of attributes and π the conditional probability. Then, the algorithm is structured as follows:

Step 1: initialize $\pi(\mathbf{x}) \sim q(\mathbf{x})$

Step 2: samples from CD q

$x_1 \sim \pi(x_1 \mid x_2, x_3, ..., x_i, ..., x_n)$

$x_2 \sim \pi(x_2 \mid x_1, x_2, ..., x_i, ..., x_n)$

$x_3 \sim \pi(x_3 \mid x_1, x_2, ..., x_i, ..., x_n)$

...

$x_i \sim \pi(x_i \mid x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n)$

...

$x_n \sim \pi(x_n \mid x_1, x_2, ..., x_i, ..., x_{n-1})$

Step 3: repeat until reaching the population size

As mentioned previously, four socio-demographic and two transport-related variables are included in the form of conditionals as input for the Gibbs sampler. In this regard, data preparation represents an advantage compared to the amount of data necessary for the IPF (Farooq et al., 2013). In particular, we attempt to draw agents from the partial views of the true joint probability using Gibbs sampling. This technique is particularly flexible for handling multiple data sources with different spatial scales (Farooq et al., 2013). A simulation-based approach typically needs a set of conditional probabilities $\pi(x_i \mid x_j)$, where i and j are respectively referring to the dependent and independent variables.

To reach steady state, it is advisable to warm the Gibbs sampler (i.e., using approximately 1000 iterations in our case). To reduce computation time, we save the generated population in a .csv file. In this way, it is possible to extend the size of the population by starting with this file. We simply have to extract the characteristics of the last agent to use it as the initial condition for the Gibbs sampler. In addition, in the context of this study, we intend to generate a population of 250,000 agents to serve as a basis. Then, from

Dependent variable	Independent variables
Socio-professional status (full)	Age - Gender - WT expenditure (**) - PT subscription (*) - Driving license ownership
Gender (full)	Age - Socio-professional status - WT expenditure - PT subscription - Driving license ownership
Age (full)	Gender - Socio-professional status - WT expenditure - PT subscription - Driving license ownership
WT expenditure (full)	Age - Gender - Socio-professional status - PT subscription - Driving license ownership
PT subscription (full)	Age - Socio-professional status - WT expenditure - Gender - Driving license ownership
Driving license ownership (full)	Age - Socio-professional status - WT expenditure - Gender - PT subscription

(*) Public Transport subscription
(**) Working Time expenditure

Table 3: List of Conditional Distributions

this basis, a sub-sample of 110,000 agents is randomly extracted. This approach allows for a reduction in the possible correlations between successive draws. The sub-sample represents approximately 1% of the population in Belgium. Because we are only interested in the proportions of specific groups with respect to the full population, it is not necessary to generate a larger number of agents. Indeed, the proportions will remain asymptotically identical. In addition, if we were addressing an agent-based micro-simulation problem, it would also have been necessary to approach the problem with a similar population size.

As outlined by Farooq et al. (2013), although the method provides a more accurate representation of the population, the simulation-based approach is not yet capable of synthesizing a full household synthetic population with multiple individuals. In this particular situation, an alternative technique, i.e., IPU, could be adopted. Moreover, the synthetic population generation procedure can be easily replaced by IPU. Indeed, the integrated framework presented in this paper is designed to ensure a sufficient level of modularity. Alternatively, when the dataset includes the associations in-between households and their related individuals, i.e., through a referencing system, it is possible to include mixed household and individual effects. First, the target cluster of households is isolated. Then, the activity sequences associated with the individuals within the selected households are processed. In this way, household effects can also be considered.

3.3. Activity sequences characterization

After sorting beforehand the activity sequences according to their main activity, the standard characterization method of activity sequences is structured according to three different steps (Liu et al., 2015). (a) The first step consists of measuring the regularity of the activities with respect to their related clusters. In the same way, the most probable sequential order of the activities is identified for every group. In this context, a complete template (reference activity chain) characterizing the most frequent activities as well as their sequential order is defined.

Subsequently, (b) the following step consists in aligning the activity sequences based on the templates and with respect to every cluster. This approach allows activity sequences that are perfectly aligned, with identical dimensions, to be obtained.

Finally, (c) the aligned activity sequences are characterized by calibrating the pHMMs. The characterization implies the estimation of the transition, emission, insertion and deletion probabilities.

Note that, within the framework of our study, we do not group the agents according to their main activities. Thus, it is not necessary to perform a cluster analysis to apply step (a) and allow the identification of the template. On the contrary, we only attempt to estimate the effects of selected explanatory variables. In this context, step (a) can be bypassed. Indeed, because the population is extremely heterogeneous, it is not possible to define a template if activity chains of the same group possess different types of main activities. One can refer to the research of Liu et al. (2015) for a more thorough description of the modeling framework.

Figure 2 describes the full parameters of a pHMM, including the emission and transition conditional probabilities, as well as the match m_i, insertion i_i and deletion d_i states. Note that both the match and

insertion states are capable of emitting an activity type A_i. m_0 is the beginning state, and m_N the final state. N is the length of the chain containing the largest sequence of activities within the cluster. The bold arrows represent an illustration of all the possible transition combinations between the states m_2 and m_3. Furthermore, an insertion state can evolve toward the same insertion state (symbolized by the loop); otherwise, the following match state is selected. The parameters of the pHMM and the SAM have been estimated using the Bioinformatics Toolbox of MATLAB.

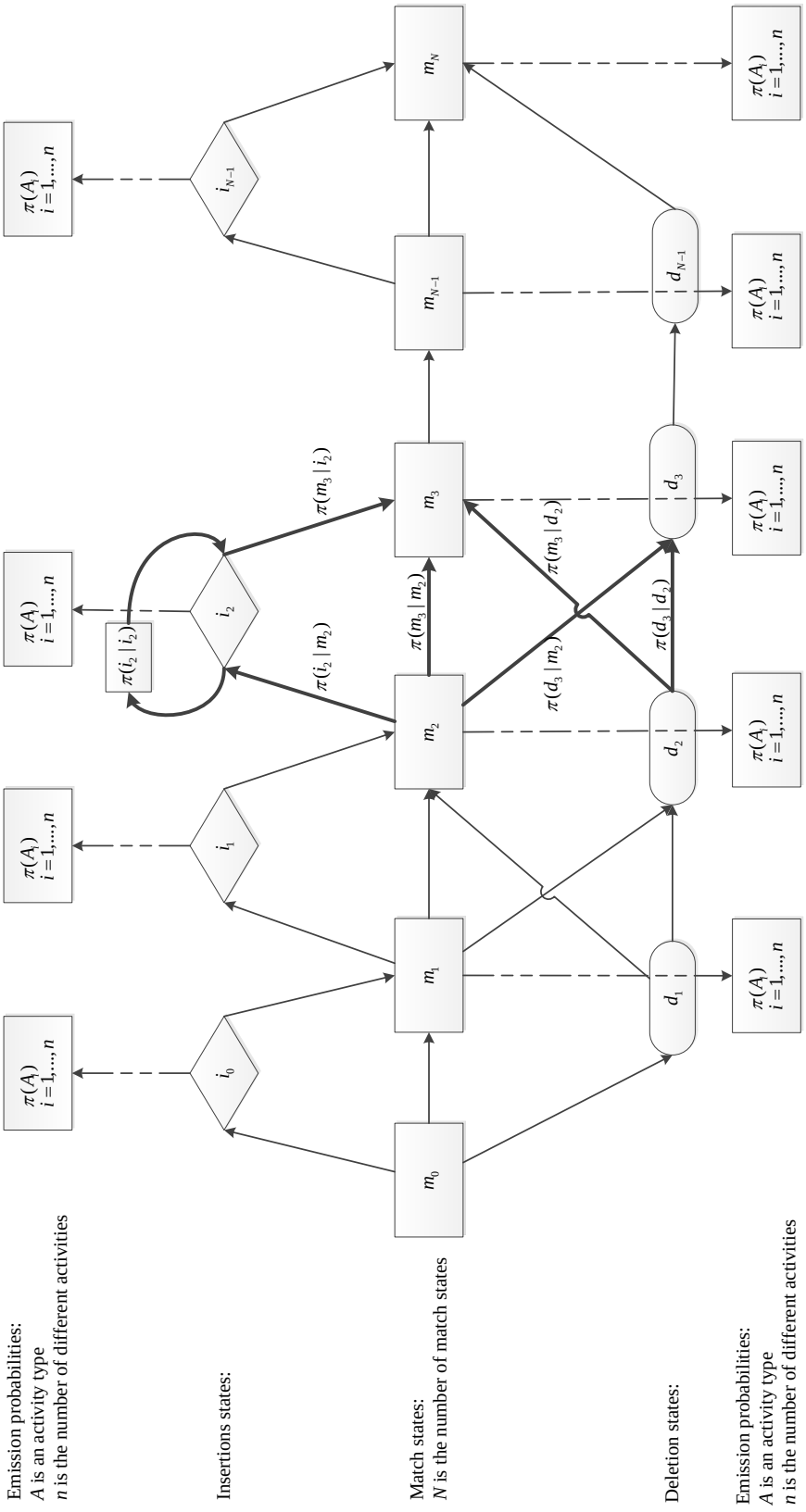


Figure 2: Parameters of the profile Hidden Markov Model

4. Results

4.1. Synthetic population

Figure 3 presents the comparison between the synthetic population, considering every explanatory variable, and the reference dataset. The results clearly indicate that the generated synthetic population is a suitable approximation. Furthermore, the simulation-based approach provides good estimates of the marginal distributions for the selected attributes. With respect to the joint distributions presented in Fig. 4, the results demonstrate a good fit between the synthetic population and the reference dataset. Indeed, the slope is close to 1, with an R^2 value of 0.87. Note that each point represents the proportion of the combination of six attributes. In this study, only full conditions have been implemented in the framework. Therefore, the spread of the data points cannot be explained by the use of partial conditions. Moreover, the combined effects of scalability and dimensionality can explain slight deviations in the joint distributions of the simulation-based approach.

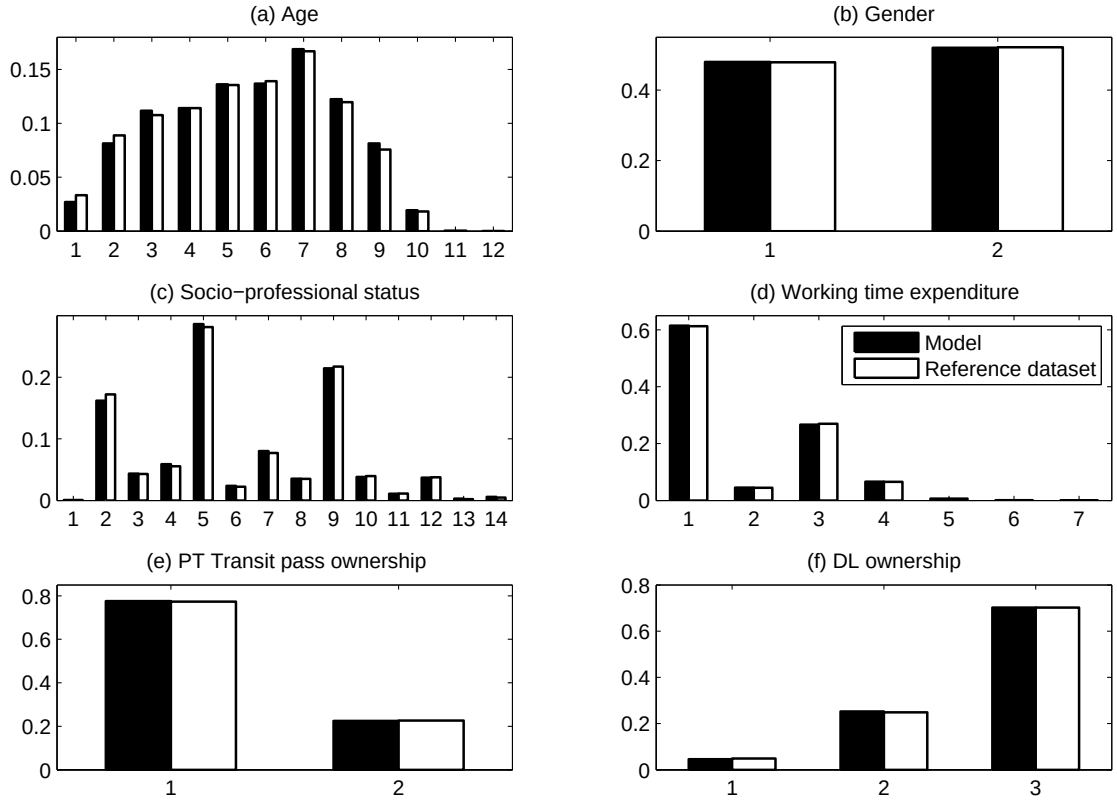


Figure 3: Comparison between the synthetic population and the reference dataset

In addition, some slight deviations may be observed within the marginal distributions (Fig. 3) as well as the joint conditional distributions (Fig. 4) due to the random effects included in the Gibbs sampler, which are mainly related to the stochastic nature of the model. In this regard, we assume that an increase in the size of the micro-data can play an effective role in the mitigation of the error.

Furthermore, studies have shown that an increase in the sample fraction is conducive to mitigating the input uncertainty (Rasouli and Timmermans, 2014). To enhance the stability of the forecasts, we propose to run the Gibbs sampler repeatedly such that the final outcome is the average of multiple model runs. As

outlined by Rasouli et al. (2012), this procedure also contributes to decreasing uncertainty but from the model perspective only.

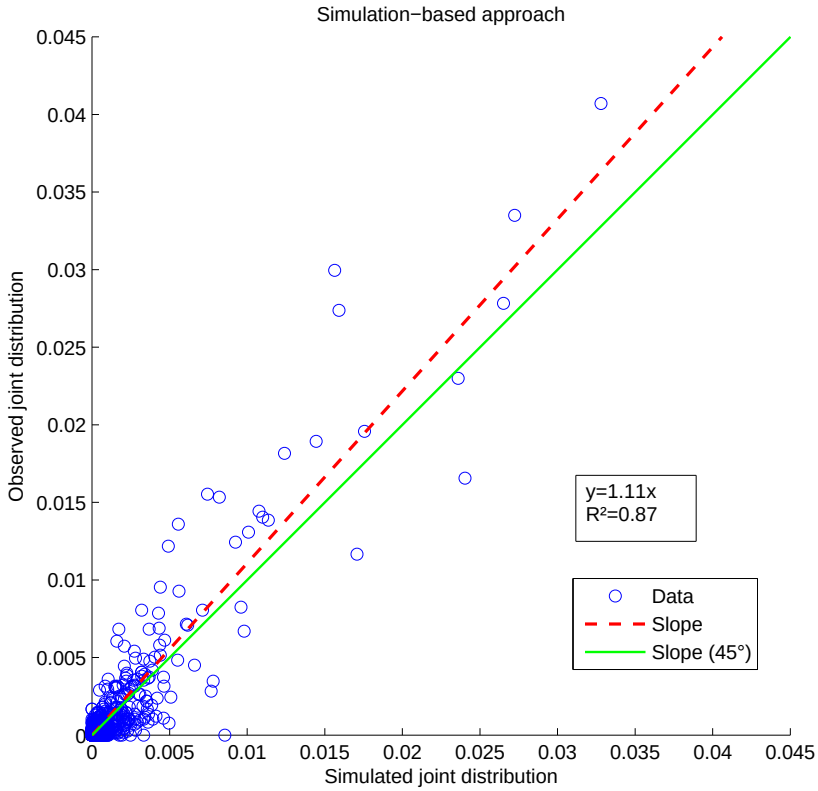


Figure 4: Comparison between the simulated and observed joint distributions

To assess the reliability of the population synthesis method, a more disaggregate approach consisting in a detailed clustering of the simulated and observed populations with respect to all the levels of the attributes defined in Table 3. Based on this clustering procedure, the deviations between the two populations from their related joint distributions are estimated. In this regard, the Root Mean Square Error (RMSE) is an indicator which allows to assess how close the synthesized population is to the observed one. The RMSE is defined as follows:

$$RMSE = \sqrt{E((\tilde{\theta} - \theta)^2)} = \sqrt{\frac{\sum_{i=1}^{n}(\tilde{y_i} - y_i)^2}{n}} \quad (1)$$

where $\tilde{\theta}$ is the estimator of the population, θ is the observed population, and n is the number of predicted values. Vovsha et al. (2015) used the RMSE to estimate the goodness-of-fit of the synthesized populations for different zonal systems. The RMSEs presented in Table 4 show that, even in the context of a finer analysis, the model is capable of maintaining a satisfactory level of accuracy. Overall, most of the errors are close to 0, except for the values combined with level 13 of the socio-professional variable (i.e., the farmers). However, this only represents a small portion of the full population (0.23% (Table 1)). In this regard, when the number of observations within a specific combination of variables is low, it is more probable to observe important deviations in terms of RMSE.

Socio-professional status × Age	1	2	3	4	5	6	7	8	9	10	11	12
1	-	0.027	0.058	-	-	-	-	-	-	-	-	-
2	0.013	0.026	0.017	0.027	0.049	0.074	0.080	-	-	-	-	-
3	0.036	0.059	0.018	0.034	0.055	0.024	0.025	0.032	-	-	-	-
4	0.027	-	0.023	0.018	0.021	0.017	0.011	0.055	-	-	-	-
5	-	-	-	0.056	0.031	0.023	0.016	0.007	0.015	0.041	0.070	0.154
6	0.118	-	0.056	0.031	0.023	0.016	0.014	0.061	-	-	-	-
7	0.019	-	0.018	0.030	0.027	0.032	0.037	0.078	-	-	-	-
8	0.036	-	0.036	0.032	0.035	0.029	0.033	0.071	-	-	-	-
9	0.020	-	0.007	0.007	0.012	0.012	0.011	0.037	-	-	-	-
10	0.038	-	0.043	0.043	0.040	0.039	0.047	0.026	-	-	-	-
11	0.051	-	0.069	0.040	0.052	0.035	0.038	0.094	-	-	-	-
12	0.021	-	0.038	0.023	0.015	0.018	0.023	0.072	-	-	-	-
13	**0.154**	-	**0.126**	0.045	0.062	0.065	0.066	**0.154**	-	-	-	-
14	0.081	-	0.063	0.046	0.043	0.058	0.056	-	-	-	-	-

Table 4: Root Mean Square Error (RMSE) of the Simulated Joint Distributions with respect to different combinations of age and status

4.2. Characterization of the activity sequences

After building the synthetic population, we define criteria of analysis (e.g., gender and socio-professional status) to extract the activity-travel patterns. The objective of the population synthesis is to estimate the proportion of the categories of the studied people within the overall population. Subsequently, the activity sequences are extracted from the activity-travel diaries describing the succession of the activities of the studied clusters.

Regarding the simulation tools, various programs in bioinformatics are available for aligning multiple activity sequences and also estimating the parameters of pHMMs. Thus, it is not necessary to develop a full code for sequence alignment and parameter estimation. This proves that this framework can be applied in a rather fast and effective manner.

As advised by Liu et al. (2015), some assumptions have been made regarding all the activity sequences. Indeed, we suppose that all the sequences begin and end with the same activity, in this particular case, home (H). Thus, the first and last positions possess a probability that is approximately equal to 1. In this regard, no deletion states occur in the first position, and no transition probabilities need to be defined in the last position because everything is shifting toward home.

Note that the number of positions is defined by the activity chain with the highest number of successively different activities, i.e., 13 activities in this paper. If a smaller size is taken, the profile-HMM will have to aggregate some information to be able to estimate the probabilities, thus reducing the quality of the information. On the other hand, if a greater profile size is fixed, more parameters will have to be estimated; however, this would not improve the accuracy and would make the analysis more complex. In this regard, one can understand that determining the number of positions depends exclusively on the longest chain. By referring to the work of Liu et al. (2015), the approach is similar except that the longest activity is called a template and is built from an identification of the regular and irregular activities as well as their sequential order. Note that in Liu et al. (2015), built clusters are only based on the main activities, i.e., the work cluster. By including socio-demographics in our paper, some additional travel patterns can be revealed, e.g., education for young people. In this regard, the activity sequencing is sensitive to changes in the socio-demographics.

To illustrate the methodology, we propose some case studies wherein the differences in behavior are highlighted through the parameter estimates. Table 5 presents the transition and emission probabilities resulting from an estimated pHMM, where the results for the full population of Belgium are included to highlight the main patterns of conduct.

For each column of the emission probabilities, the highest values are in bold. In this way, we can extract some key information. For example, throughout all the positions, the commuting patterns are quite significant in terms of importance compared to the remainder of the population and specifically in position 4.

The trends suggest that some activities, such as leisure, sports and visiting family and friends (see between positions 9-12), are preferred to be conducted at the end of the day. In contrast, the daily shopping activity is distributed throughout the day.

Position (k)	1	2	3	4	5	6	7	8	9	10	11	12	13
Transition probabilities													
$\pi(m_{k+1}, m_k)$	0.548	0.583	0.270	0.644	0.503	0.649	0.633	0.747	0.093	0.999	0.977	1.00	–
$\pi(i_k, m_k)$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.031	0.000	0.000	0.000	–
$\pi(d_{k+1}, m_k)$	0.452	0.417	0.730	0.356	0.497	0.351	0.367	0.253	0.876	0.000	0.023	0.000	–
$\pi(m_{k+1}, i_k)$	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.510	0.500	0.500	0.500	–
$\pi(i_k, i_k)$	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.490	0.500	0.500	0.500	–
$\pi(m_{k+1}, d_k)$	0.263	0.041	0.475	0.035	0.056	0.076	0.098	0.115	0.018	0.066	0.186	0.648	–
$\pi(d_{k+1}, d_k)$	0.737	0.959	0.525	0.965	0.943	0.924	0.902	0.885	0.982	0.934	0.814	0.352	–
Emission probabilities													
$\pi(H)$ (home)	**1**	0.002	**0.691**	0.046	**0.457**	0.282	0.154	**0.533**	0.062	**0.097**	**0.489**	0.004	**1**
$\pi(W)$ (work)	–	0.018	0.018	**0.651**	0.254	**0.287**	**0.237**	0.055	0.080	0.065	0.125	0.237	–
$\pi(G)$ (bring/get)	–	0.140	0.036	0.079	0.029	0.052	0.095	0.093	0.204	0.065	0.057	0.091	–
$\pi(E)$ (education)	–	0.206	0.009	0.041	0.004	0.017	0.041	0.005	0.009	0.032	0.011	0.008	–
$\pi(M)$ (meal)	–	0.020	0.021	0.002	0.043	0.012	0.018	0.017	0.049	0.064	0.023	0.008	–
$\pi(S)$ (daily shopping)	–	**0.337**	0.124	0.072	0.082	0.098	0.219	0.121	**0.227**	**0.097**	0.046	0.079	–
$\pi(A)$ (service)	–	0.080	0.033	0.017	0.029	0.046	0.047	0.028	0.009	0.065	0.011	0.012	–
$\pi(F)$ (visit)	–	0.062	0.018	0.022	0.029	0.069	0.095	0.028	0.191	0.065	0.046	**0.289**	–
$\pi(V)$ (tour)	–	0.066	0.009	0.017	0.014	0.029	0.012	0.022	0.009	**0.097**	0.023	0.174	–
$\pi(R)$ (leisure, sports etc.)	–	0.051	0.009	0.029	0.025	0.052	0.024	0.044	0.116	0.032	0.057	0.059	–

Table 5: Parameter Estimates for the Full Population of Belgium

It has been reported in various studies (Bhat and Singh, 2000) that most professionally active individuals go shopping after work on the road returning home or even later in the day. Because this category of people is important in term of size with respect to the full population, it is thus logical to obtain significant values (i.e., 21.9% in position 7 – 22.7% in position 9) later in the day.

Furthermore, the bring/get activity is significantly present throughout the day as well. However, it is necessary to note that various groups of the population are implicitly included in the results highlighted in Table 5, which makes the clusters' specific features more difficult to capture. In this context, a more detailed cluster analysis is necessary to allow one to distinguish which part of the population conducts shopping activities and during approximately which period of the day.

When characterizing the full population, it is indeed less obvious to consider what are the proportions as well as the categories of people who perform a given activity at a given moment of the day.

Table 6 presents the parameter estimates of the emission and transition probabilities for individuals less than 31 years of age. This category groups most of the students and also some young workers. After isolating the target sub-population, we can note the important increase in the education activity; this is synonymous with important trips toward schools and universities.

Note that young people also dedicate a portion of their time to conducting sports or entertainment activities. Furthermore, they organize, as a general rule, such activities after their courses (see position 5). If we need any reminder of this, we simply need to observe the sequencing of the activities within the positions. The great majority of the educational activities are grouped in position 2 (43.3%). Subsequently, it is only from position 4 that young people perform secondary activities such as shopping. Indeed, an important

portion of young people (students/novice workers) live alone during their studies or at the beginning of their professional lives; they also have to fulfill their vital needs by moving quasi-daily to shop.

These results are logical and compatible with the descriptive analysis presented in Table 2. The main activity-travel patterns (e.g., education, visiting friends and family, entertainment and sports) have been characterized by the pHMM. Furthermore, this proves that the calibration of the pHMM was correct. This mode of comparison clearly reveals the added value of the pHMM compared to a classic analysis of descriptive statistics. Not only is the establishment of activity sharing possible throughout the day but the result also indicates that the sequencing of the activities can be obtained thanks to the positioning system.

Position (k)	1	2	3	4	5	6	7	8	9	10	11	12	13
Transition probabilities													
$\pi(m_{k+1}, m_k)$	0.876	0.529	0.333	0.362	0.462	0.090	0.998	0.352	0.411	0.757	0.988	1.000	–
$\pi(i_k, m_k)$	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	–
$\pi(d_{k+1}, m_k)$	0.124	0.471	0.667	0.638	0.538	0.910	0.001	0.648	0.589	0.243	0.012	0.000	–
$\pi(m_{k+1}, i_k)$	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	–
$\pi(i_k, i_k)$	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	–
$\pi(m_{k+1}, d_k)$	0.256	0.321	0.035	0.066	0.029	0.034	0.070	0.065	0.042	0.083	0.174	0.668	–
$\pi(d_{k+1}, d_k)$	0.744	0.679	0.965	0.934	0.971	0.997	0.930	0.935	0.958	0.917	0.826	0.332	–
Emission probabilities													
$\pi(H)$ (home)	**1**	0.000	**0.598**	**0.297**	0.180	**0.395**	0.046	0.099	**0.287**	**0.158**	**0.474**	0.007	**1**
$\pi(W)$ (work)	–	0.255	0.070	0.143	0.132	0.156	0.046	0.077	0.283	0.131	0.099	0.106	–
$\pi(G)$ (bring/get)	–	0.030	0.068	0.078	0.083	0.070	0.046	0.047	0.038	0.113	0.064	0.146	–
$\pi(E)$ (education)	–	**0.433**	0.033	0.064	0.132	0.091	0.046	0.077	0.007	0.018	0.020	0.024	–
$\pi(M)$ (meal)	–	0.011	0.036	0.023	0.019	0.049	0.046	0.017	0.026	0.045	0.035	0.061	–
$\pi(S)$ (daily shopping)	–	0.082	0.074	0.146	0.145	0.029	**0.136**	0.090	0.091	0.153	0.062	0.109	–
$\pi(A)$ (service)	–	0.020	0.015	0.026	0.035	0.017	0.046	0.022	0.019	0.018	0.027	0.026	–
$\pi(F)$ (visit)	–	0.070	0.060	0.092	0.038	0.033	0.046	0.249	0.113	0.131	0.072	**0.255**	–
$\pi(V)$ (tour)	–	0.030	0.008	0.064	0.024	0.008	0.046	0.009	0.030	0.063	0.030	0.042	–
$\pi(R)$ (leisure, sports etc.)	–	0.066	0.033	0.049	**0.186**	0.111	0.046	**0.270**	0.068	0.126	0.094	0.213	–

Table 6: Parameter Estimates for the Population of Belgium below 31 Years of Age

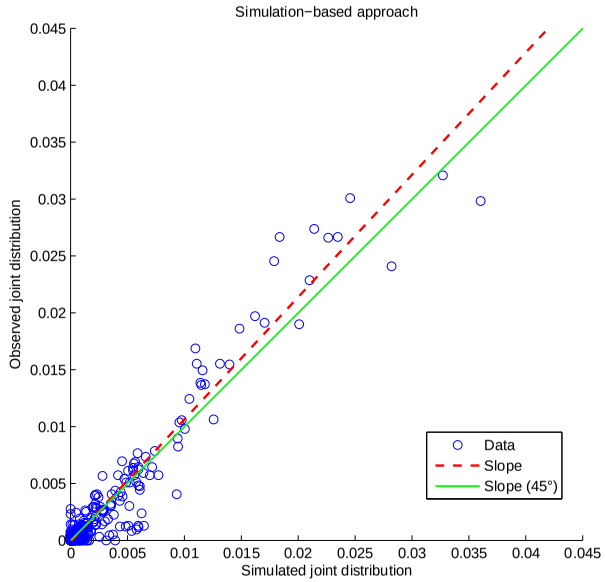
4.3. Comparison between the simulation-based approach and the IPU-based approach

To demonstrate the efficiency of the integrated framework with respect to standard approaches, we propose to compare, at each level of the framework, a sub-module with an existing technique. First, the simulation-based approach is compared with the iterative proportional updating (IPU) algorithm. Then, we show the added value of the pHMMs over a standard SAM. In this paper, it is important to distinguish the sequence alignment process (SAM) necessary for aligning all the activity sequences and the pHMM characterization. This step is fundamental because the activity sequences are structured such that deletion states are added to ensure equal lengths for each activity sequence. In this context, we can describe this integrated framework as a step above other previous approaches. Regarding the comparison between the IPU algorithm and the simulation-based approach, it is quite difficult to determine which method performs better from an absolute perspective. We must qualify the comparison in light of different aspects involved within each approach. For example, the IPU algorithm requires a huge amount of data (PUMS and all the total controls) to determine the weights associated with the corresponding agents (Ye et al., 2009). In contrast, the simulation-based approach is only based on the calibration of the conditional distributions (Farooq et al., 2013). In this context, only PUMS are used. Thus, a completely fair comparison is very complex. In addition, a throughout data preparation needs to be realized to classify all the variables in the correct format (table of frequencies) so that it can be included within IPU (Ye et al., 2009). In addition, all the total controls corresponding to all the levels of the set of attributes are needed, and the matching between

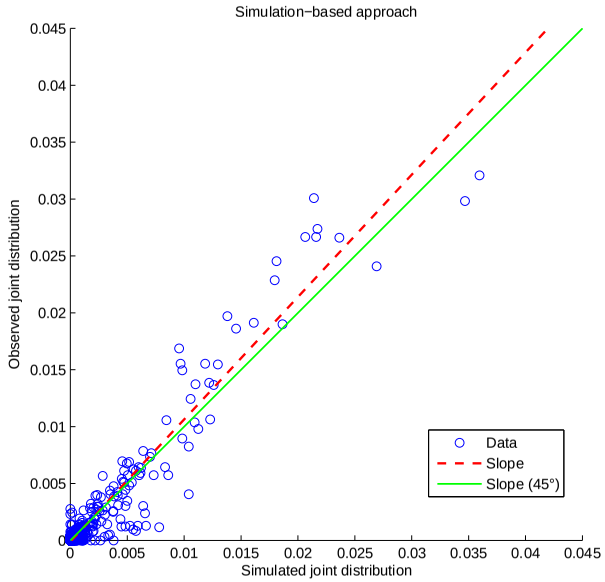
the format variables of the total controls and the PUMS should be the same. In practice, the implementation is quite heavy and computational intensive. However, the simulation-based approach is characterized by its great flexibility (Farooq et al., 2013). In addition, the PUMS can be used under their original format, i.e., as a set of observations with respect to different attributes. In practice, it is important to propose approaches that are capable of using a minimum amount of data to mitigate the phenomenon of data dependency in such a way that the quality of the results is preserved. To our knowledge, no studies have investigated the comparison between the simulation-based approach and IPU. Note that in Farooq et al. (2013), one can find the comparison with IPF. To ensure a fair comparison between both approaches, we will compare the methods for the synthesis of 4 attributes as advised by Farooq et al. (2013) so that zero-cell and zero-marginal problems, which can lead to non-convergence, can be avoided. In addition, let us consider the BELDAM travel survey because it describes the full population. In this context, the total controls can be derived for IPU as well as the seed. We suppose that the seed that is extracted represents 50% of the survey. In parallel, the same seed is included into the simulation-based approach. We will also consider a full seed for the simulation-based approach to provide it with full information and establish a relatively equal amount of inputs with respect to both methods. Table 7 presents the comparison between the different methods and configurations. Note that the intercepts are not included because they are all approximately zero.

Method	Simulation-based (full PUMS)	Simulation-based (50% PUMS)	IPU-based (50% PUMS + total controls)
R-square	94.6%	93.7%	92.4%
Slope	1.075	1.075	0.932
RMSE	0.00112	0.00121	0.00139

Table 7: Comparison between the simulation-based approach and IPU



(a) 50% PUMS



(b) Full PUMS

Figure 5: Comparison of the joint distributions with respect to the simulation-based approach

The results presented in Table 7 clearly show the advantage of the simulation-based approach over IPU. In this regard, the conclusions are in accordance with those presented by Farooq et al. (2013). We can see that using the full PUMS for the simulation-based approach does not improve the R-squared value significantly. A 50% PUMS is largely sufficient to ensure accurate results. Furthermore, although the amount of data is low compared to IPU, the simulation-based approach is capable of providing a synthetic population with an RMSE that is reduced by -14.88% and an R-squared that is improved by 1.4%.

Figures 5a, 5b and 6 present the comparison between the joint distributions of the simulated populations and the reference dataset. Each circle is a proportion of a combination of 4 attributes. Using these figures, we can see that the simulation-based approach is capable of preserving good estimates for the important proportions, whereas IPU provides a poorer performance. However, IPU maintains very good estimates for small proportions, which is not the case for the simulation-based approach. However, given that the absolute differences are more important for high proportions and less important for small proportions, the RMSE is thus higher under the simulation-based approach than under IPU.

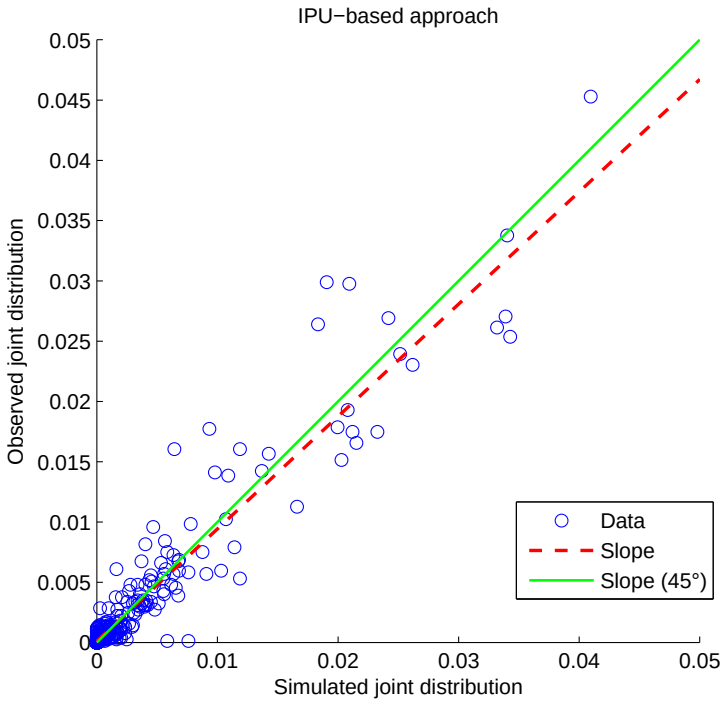


Figure 6: Comparison of the joint distributions with respect to IPU

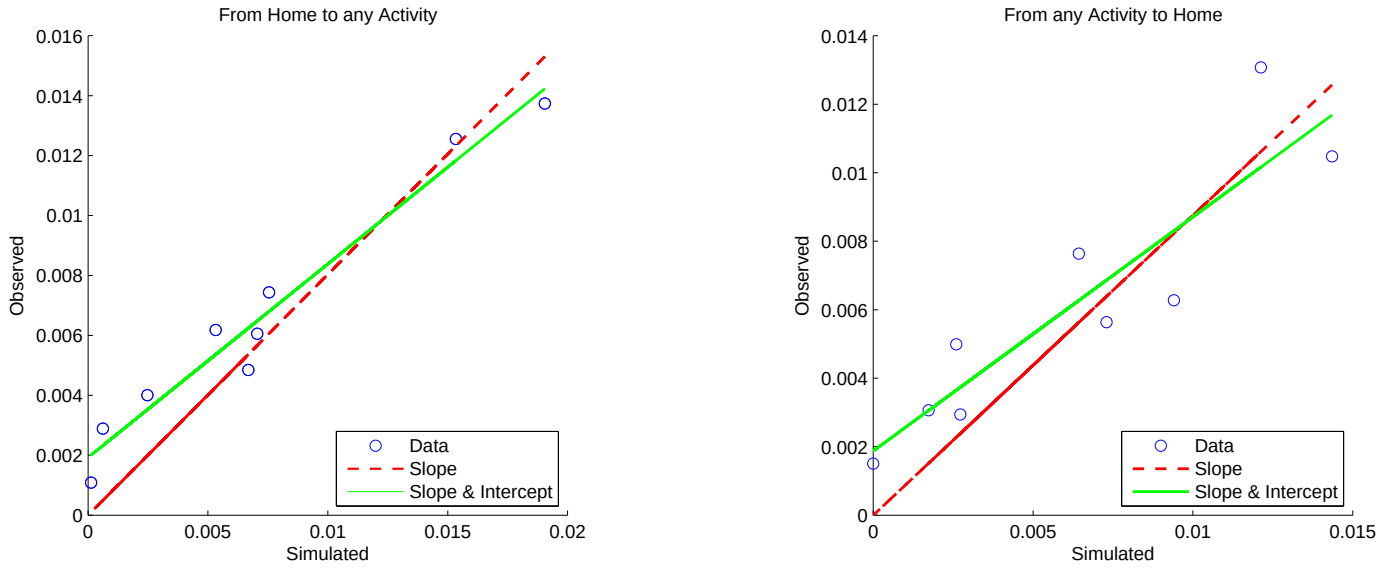
4.4. Profile-HMM validation results

To ensure that the estimated transition and emission probabilities are accurate, we propose two important indicators: (i) the proportions of the trip patterns and (ii) the occurrences of the different types of activities with respect to the full population. First, we generate a set of activity sequences from the estimated profile-HMM. Then, if the comparison between the indicators of the synthesized and observed activity sequences demonstrates that they are equivalent, then the parameter estimates have been estimated properly. As mentioned previously, we focus on the main trip patterns. In this regard, we compare all the trip patterns starting from home toward any other activity location and vice versa. The results presented in Figures **??** and 8 reveal that the main trip patterns have been correctly captured by the calibrated profile-HMM. This result is particularly important in the context of agent-based modeling, i.e., MATSim. Indeed,

the estimation of the traffic flows is especially affected by the quantity of trips and their related patterns. An over- or under-estimation may have a significant impact on the predictions in terms of traffic jams and/or traffic flows. To complete the comparison, we have also presented the fit of the simulated and observed proportions of trip patterns. Table 8 presents the main statistical metrics. Although some minor deviations might be depicted, we can consider that the model is able to produce good estimates of the trip patterns and their proportions within the overall modeling framework. The R-square values between brackets, presented in Table 8, correspond to the regression models that include the intercept. Note that these R-square values are better in comparison to those of the models without intercept. This can be explained by the under-estimation in the simulated values for uncommon trip patterns (see Fig. 8).

Model	Slope	R-squared	RMSE (in %)
Trip patterns 1 (H to A)	0.80 (0.65)	0.86 (0.96)	0.23
Trip patterns 2 (A to H)	0.87 (0.68)	0.72 (0.82)	0.21

Table 8: Comparison between the synthesized activity sequences and the observed sequences for different statistical metrics



(a) (b)

Figure 7: Comparison between the joint distributions with respect to the trip patterns

With respect to the comparison of the proportion of activities for the full population, the results presented in Figure 9 and Figure 10 also indicate a good match between the simulated and observed activity sequences (R-squared=0.99). In this regard, these results prove that the previously presented emission probabilities are sufficiently accurate and reliable.

5. Conclusion

In this paper, we present an integrated framework including a synthetic population approach (Farooq et al., 2013), together with a profiling method (Liu et al., 2015), for characterizing activity-travel patterns from both qualitative and quantitative perspectives.

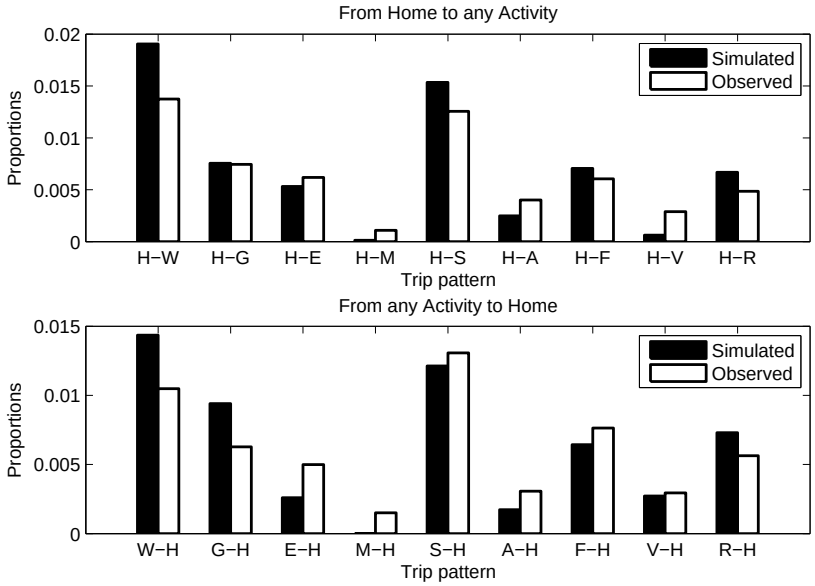


Figure 8: Comparison between the marginal distributions with respect to the trip patterns

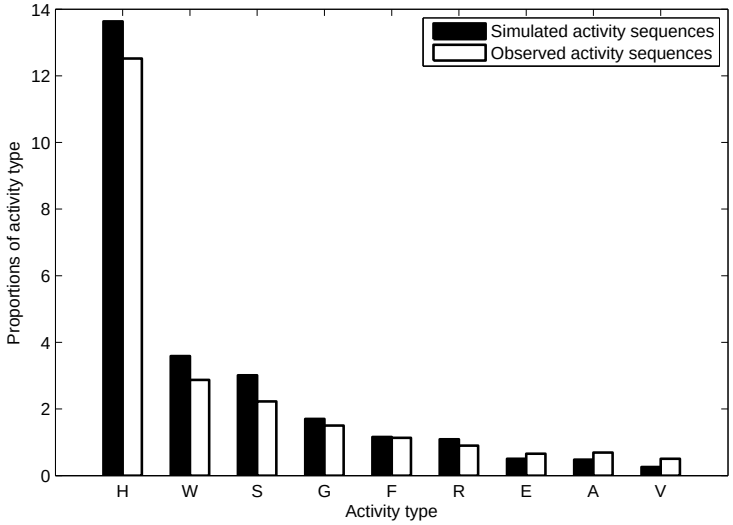


Figure 9: Comparison between the marginal distributions with respect to activity occurrence

The synthetic population sets up the connection between multiple micro-datasets. Indeed, the conditional probabilities are built so that they include the information resulting from all the available sources. In this regard, their correct determination represents a highly important result. Given the high performances of the simulation-based approach compared to standard methods (e.g., IPF), we opted for this technique to be introduced within the global modeling chain. Furthermore, the flexibility of this technique is particularly adapted to address partial micro-datasets. The results presented in Figure 4 clearly indicate that there is a scope that is able to fit the true population by implicitly merging different micro-datasets while ensuring high accuracy. Thus, these findings are in complete agreement with the conclusions of Farooq et al. (2013). A limitation of the simulation-based population synthesis is that the synthesis of households with multiple individuals is not yet possible. Nonetheless, as discussed in the methodological section, one could take into

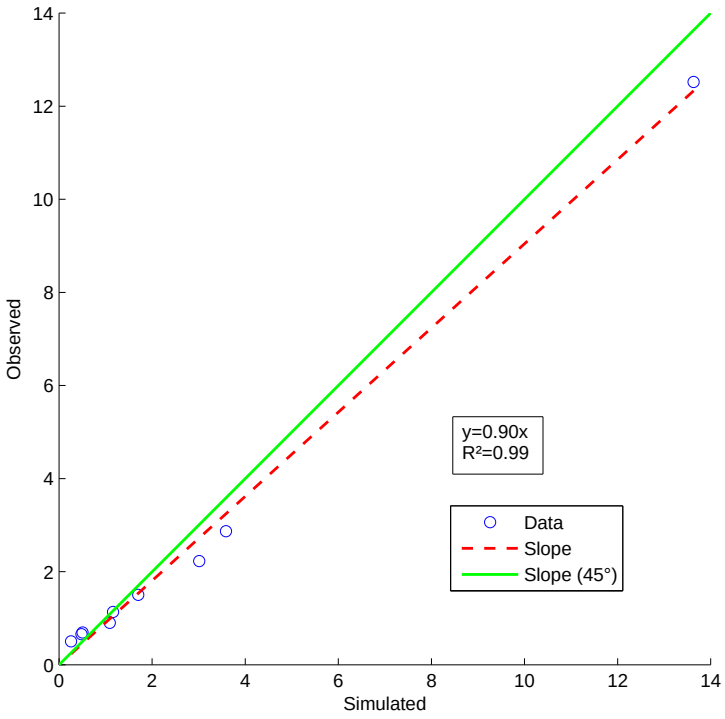


Figure 10: Comparison between the joint distributions with respect to activity occurrence

account household effects by using IPU in a similar way or by clustering the target households with their associated individuals. In this regard, we can conclude that the consecutive steps of our framework, i.e., the multiple-sequence alignment and the profile-HMM characterization steps, are not significantly affected. In the future and to preserve the presented framework, an extended population synthesis exclusively based on the simulation-based approach could be adopted at both levels (households and individuals) by considering an additional module able to generate associations between households and individuals. An interesting prelude in this regard can be found in Anderson et al. (2014).

The profiling approach enables the characterization of multiple activity sequences (activity regularity and sequential information) from only one model without neglecting any irregular activities. As a result, performing a comparison between clusters in terms of activity-travel patterns is much easier or investigating to what extent the activity-travel patterns of a specific group can be distinguished from the general behavior of a population.

Furthermore, when the pHMM is calibrated using a training dataset, a non-limited number of activity sequences can be regenerated from the estimated pHMM according to the size of the studied cluster. This application is particularly interesting in the context of agent-based micro-simulation models. Indeed, most of them require a synthetic population describing the attributes and the activity sequences of every individual. In this regard, we assume that such a modeling framework can be adapted to handle problems of multi-agent model generation and, as a result, provide new insights for further research.

Regarding the results, we indicated in Section 4.2 that the positions within a profile-HMM give the general trends of the activity sequencing from a temporal perspective. However, we consider that the approach presents a limitation at this level. Further developments of the framework should aim at the inclusion of the activity time dimension in a more explicit way.

Furthermore, by isolating the target populations, the model allowed one to characterize the proportion of

these sub-populations of the total population as well as the main travel behaviors. With respect to the young population, we have clearly shown that education appears to play an essential role in the need for travel. Note that, in the context of more elaborate analysis, it is possible to estimate quantitatively the disparity between different combinations of explanatory variables in terms of activity characteristics choice.

6. Acknowledgments

The research was funded by the ARC grant for Concerted Research Actions for project n°13/17-01 entitled 'Land-use change and future flood risk: influence of micro-scale spatial patterns (FloodLand)' and by the Special Fund for Research for project n°5128 entitled 'Assessment of sampling variability and aggregation error in transport models', both financed by the French Community of Belgium (Wallonia-Brussels Federation).

References

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

Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B., Zhang, K., 2015. Polaris: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. Transportation Research Part C: Emerging Technologies doi:http://dx.doi.org/10.1016/j.trc.2015.07.017.

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.

Bhat, C.R., Singh, S.K., 2000. A comprehensive daily activity-travel generation model system for workers. Transportation Research Part A: Policy and Practice 34, 1–22. doi:http://dx.doi.org/10.1016/S0965-8564(98)00037-8.

Cornelis, E., Hubert, M., Hunyen, P., Lebrun, K., Patriarche, G., De Witte, A., Creemers, L., Declercq, K., Janssens, D., Castaigne, M., Hollaert, L., Walle, F., 2012. Belgian Daily Mobility (BELDAM): Enquête sur la mobilité quotidienne des belges. Technical Report. SPF Mobilité & Transports. Brussels.

Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G., 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press.

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.

Joh, C.H., Arentze, T., Hofman, F., Timmermans, H., 2002. Activity pattern similarity: a multidimensional sequence alignment method. Transportation Research Part B: Methodological 36, 385–403. doi:http://dx.doi.org/10.1016/S0191-2615(01)00009-1.

Joh, C.H., Timmermans, H., 2011. Applying sequence alignment methods to large activity-travel data sets: Heuristic approach. Transportation Research Record: Journal of the Transportation Research Board , 10–17doi:http://dx.doi.org/10.3141/2231-02.

Joh, C.H., Timmermans, H., Arentze, T., 2006. Measuring and predicting adaptation behavior in multi-dimensional activity-travel patterns. Transportmetrica 2, 153–173. doi:http://dx.doi.org/10.1080/18128600608685659.

Liu, F., Janssens, D., Cui, J., Wets, G., Cools, M., 2015. Characterizing activity sequences using profile hidden markov models. Expert Systems with Applications 42, 5705–5722. doi:http://dx.doi.org/10.1016/j.eswa.2015.02.057.

Mohammadian, A.K., Javanmardi, M., Zhang, Y., 2010. Synthetic household travel survey data simulation. Transportation Research Part C: Emerging Technologies 18, 869–878. doi:http://dx.doi.org/10.1016/j.trc.2010.02.007.

Pendyala, R., Goulias, K., 2002. Time use and activity perspectives in travel behavior research. Transportation 29, 1–4. doi:http://dx.doi.org/10.1023/A:1012909228433.

Rasouli, S., Cools, M., Kochan, B., Arentze, T., Bellemans, T., Janssens, D., Timmermans, H., 2012. Uncertainty in forecasts of complex rule-based systems of travel demand: Comparative analysis of the albatross/feathers model system, in: 13th International Conference on Travel Behaviour Research, International Association for Travel Behaviour Research (IATBR).

Rasouli, S., Timmermans, H., 2014. Activity-based models of travel demand: promises, progress and prospects. International Journal of Urban Sciences 18, 31–60. doi:http://dx.doi.org/10.1080/12265934.2013.835118.

Saadi, I., Mustafa, A., Teller, J., Cools, M., 2016a. 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.

Saadi, I., Mustafa, A., Teller, J., Farooq, B., Cools, M., 2016b. Hidden markov model-based population synthesis. Transportation Research Part B: Methodological 90. doi:http://dx.doi.org/10.1016/j.trb.2016.04.007.

Spissu, E., Pinjari, A.R., Bhat, C.R., Pendyala, R.M., Axhausen, K.W., 2009. An analysis of weekly out-of-home discretionary activity participation and time-use behavior. Transportation 36, 483–510. doi:http://dx.doi.org/10.1007/s11116-009-9200-5.

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.

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.

Williamson, P., Birkin, M., Rees, P.H., et al., 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.

Wilson, W.C., 1998. Activity pattern analysis by means of sequence-alignment methods. Environment and Planning A 30, 1017–1038. doi:http://dx.doi.org/10.1068/a301017.

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.