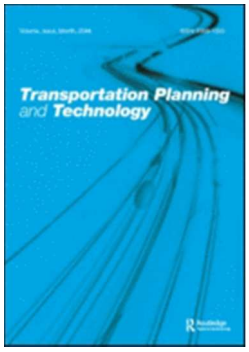

Investigating the scalability in population synthesis: A comparative approach

Journal:	*Transportation Planning and Technology*
Manuscript ID	GTPT-2016-0179
Manuscript Type:	Original Article
Date Submitted by the Author:	16-Nov-2016
Complete List of Authors:	Saadi, Ismaïl; Universite de Liege, ArGEnCo - Local Environment Management & Analysis (LEMA) Eftekhar, Hamed; Universite de Liege Teller, Jacques; Universite de Liege Cools, Mario; Universite de Liege
Keywords:	Iterative Proportional Fitting (IPF), simulation-based approach, population synthesis, scalability, agent-based micro-simulation

Université de Liège

Faculté des Sciences Appliquées
Département d'Architecture, Géologie, Environnement et Constructions
Ismaïl SAADI, PhD Candidate

ArGEnCo

Liège, 16 November 2016.

Object: Submission to Transportation Planning and Technology

Dear Editor-in-chief,
Dear Editor,
Dear Reviewers,

On behalf of myself and my co-authors, I would like to submit our article entitled "Investigating the scalability in population synthesis: A comparative approach" to the journal "Transportation Planning and Technology ". We consider that our paper corresponds perfectly to the scope of your journal.

I thank you in advance for considering our paper, and keeping me informed about the review process.

Yours sincerely,

Ismaïl Saadi
Ph.D Candidate

LEMA

Secteur Architecture & Urbanisme (A&U) - LEMA
Quartier Polytech 1, Allée de la Découverte 9
4000 Liège (Belgique)
Parking P52
Tél. +32 (04) 366 94 44 Fax. +32 (04) 366 29 09
Email : Ismail.Saadi@ulg.ac.be www.lema.ulg.ac.be

Investigating the scalability in population synthesis: A comparative approach

Ismaïl Saadi[a], Hamed Eftekhar[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*

Corresponding author: Ismaïl Saadi (Tel.: +32 4 366 96 44 - Email: ismail.saadi@ulg.ac.be)

Co-authors: Hamed Eftekhar (Tel.: +32 4 366 98 69 - Email: h.eftekhar@ulg.ac.be), Jacques Teller (Tel.: +32 4 366 94 99 - Email: jacques.teller@ulg.ac.be), Mario Cools (Tel.: +32 4 366 48 13 - Email: mario.cools@ulg.ac.be)

Word cound : 4145 words (excluding tables, figures, captions and references)

Acknowledgements: The research was funded by the ARC grant for Concerted Research Actions for project no. 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 no. 5128 entitled "Assessment of sampling variability and aggregation error in transport models", both financed by the French Community of Belgium (Wallonia-Brussels Federation).

Investigating the scalability in population synthesis: A comparative approach

In this paper, we investigate the influence of scalability on the accuracy of different synthetic populations using both fitting and generation-based approaches. Most activity-based models need a base-year synthetic population of agents with various attributes. However, when several attributes need to be synthesized, the accuracy of the synthetic population may decrease due the mixed effects of scalability and dimensionality. We analyze the two population synthesis methods for different level of scalability, i.e. two to five attributes and different sample sizes, i.e. 10%, 25% and 50%. The results reveal that the simulation-based approach is more stable than Iterative Proportional Fitting (IPF) when the number of attributes increases. However, IPF is less sensitive to changes in sample size when compared to the simulation-based approach. We also demonstrate the importance of choosing the correct metric to validate the synthetic populations as the trends in terms of RMSE/MAE are different from those of SRMSE.

Keywords: Iterative Proportional Fitting (IPF); simulation-based approach; population synthesis; scalability; agent-based micro-simulation modelling

1. Literature review

In general, agent-based micro-simulation models for transportation, e.g. activity-based models, and urban systems require highly disaggregated data, at individual level. Typically, such data consists of a series of attributes describing the individuals and their behavior. Collecting such type of data, while preserving the required level of disaggregated information for each agent, could be subjected to specific restrictions, i.e. confidentiality and important costs. In this regard, generating synthetic population data has been considered as an efficient alternative for providing agent-based micro-simulation models with reasonably accurate synthetic populations (Müller and Axhausen 2011; Ye et al. 2009; Zhu and Ferreira 2014).

The behavioral realism of an agent-based micro-simulation framework depends highly on the quality of the generated synthetic population. In this regard, a crucial choice consists of applying the most appropriate population synthesis approach among the existing ones. Most of the population synthesis methods require either aggregate data, i.e. target marginal distributions, or disaggregate data, i.e. micro-samples. Moreover, both types of data can be used at the same time in the case of fitting-based approaches, e.g. Iterative Proportional Fitting (IPF). The Public Use Micro-Sample (PUMS), also called the "initial seed" in the case of IPF, is a disaggregated dataset that usually contains detailed-enough information regarding the target population, but the number of observations is generally limited, e.g. less than 10% of the full population. In contrast, the target marginal distributions refer to the total frequencies of a one-dimensional distribution of an attribute.

In literature, the most common techniques used for generating a synthetic populations are Iterative Proportional Fitting (Beckman, Baggerly, and McKay 1996; Mohammadian, Javanmardi, and Zhang 2010), Iterative Proportional Updating (Ye et al. 2009), Combinatorial Optimization (Voas and Williamson 2001; Williamson, Birkin, and Rees 1998) and probabilistic models using Markov Chains concepts (Farooq et al. 2013; Saadi, Mustafa, Teller, Farooq, et al. 2016; Sun and Erath 2015).

The Iterative Proportional Fitting (IPF) procedure involves the generation of the desired joint distribution for a given sample of the target population (Beckman, Baggerly, and McKay 1996; Deming and Stephan 1940). In this regard, the first step consists in the calibration of a k-way contingency-table based on the initial PUMS. Then, the table is fitted to the target marginal distributions, while preserving the weights present in the PUMS. As soon as the multi-dimensional contingency table of the target population is entirely fitted, a synthetic population is produced by sampling a fixed

number of households or individuals from the seed data. As IPF consists of fitting a k-way contingency table, an increase in the number of attributes would considerably enlarge the size of the multi-dimensional-table (Müller and Axhausen 2011). This aspect is particularly important in the context of scalability, i.e. the sensitivity of a population synthesis approach to the number of synthesized variables. Indeed, if each of the considered variables includes an important number of categories, the total number of cells can increase significantly, leading to a curse of dimensionality (Sun and Erath 2015).

The Combinatorial Optimization (CO) approach, which like IPF adopts an iterative algorithm, was firstly proposed by Williamson et al. (1998). The CO technique begins with a random subset of the households and iteratively replaces the households with a new set of households from a data source. Then, the replacements are checked using a goodness-of-fit indicator. If the replacement improves the fit of the subset, the new replaced household is retained. Otherwise, the replacement is reversed and a new household is selected from the source file. The quality of the fit is repeatedly checked until the algorithm converges towards the most accurate synthetic population (Voas and Williamson 2001).

Based on IPF, the Iterative Proportional Updating (IPU) includes an additional component in the form of a heuristic algorithm (Ye et al. 2009). The idea behind IPU consists in adjusting the sample households' weights such that both individual and household-level distributions are matched (Barthelemy and Toint 2013). Particularly, the constraints, i.e. base-year marginal distributions, for both individual and household-levels are estimated using an IPF procedure and then the sample households' weights are estimated using the IPU algorithm. After estimating the households' proportions

based on the determined weights, synthetic populations can be generated by drawing from the weighted k-way table.

With respect to recent population synthesis methods, the Bayesian Network (BN) is a data-driven approach that characterizes the inherent joint distribution of the true population under a probabilistic framework. The BN represents the probabilistic relations, e.g. causality or dependence, between a set of features within a graphical structure. Such a graphical representation enables inferring the true population's structure from a certain number of PUMS. Additional information regarding the application of BN for population synthesis can be found in Sun and Erath (2015).

In a similar way, Farooq et al. (2013) used a Markov Chain Monte Carlo (MCMC) algorithm to draw a synthetic joint distribution from partial views of the true population, i.e. conditional distributions. The simulation-based approach overcomes the weaknesses existing in previous methodologies, e.g. multiple solutions for matching contingency tables, loss of inherent heterogeneity in the micro-data, and scalability issues upon increasing the number of intended attributes.

Finally, the Hidden Markov Model (HMM)-based approach is a probabilistic representation of the true population, where connections between attributes are estimated in the form of transition probabilities. The HMM-based approach is characterized by an important flexibility and efficiency in terms of data preparation. The HMM-based approach is capable of inferring the structure of a given population from an unlimited number of micro-samples and only one marginal distribution (Saadi, Mustafa, Teller, Farooq, et al. 2016). Contrary to the BN approach, the HMM-based population synthesis procedure does not include model selection using AIC/BIC criterions. In this way, the HMM-based approach is more straightforward in

approximating accurate synthetic populations, while limiting the complexity with respect to the implementation.

In summary, two important streams of population synthesis can be distinguished: fitting versus generation-based approaches. Studies related to the generation-based approaches suggest that probabilistic or Markov Chains-based models outperform IPF. However, it is difficult to generalize this statement when no studies have rigorously investigated the effects of scalability or changes in sampling rates on the synthetic populations' accuracy. In this regard, the current paper contributes to the state-of-the-art by comparing the effect of scalability on the quality of the synthetic populations generated by the standard IPF procedure and the simulation-based approach. Furthermore, we also discuss the effects and eventual interactions between changes in sampling rates and scalability. To this end, we use multiple statistical metrics to highlight the importance of choosing reliable indicators. Finally, we extend the findings of Farooq et al. (2013), who compared IPF with MCMC on the basis of a four attributes-based comparison, by confirming that the simulation-based approach outperforms IPF for additional levels of scalability and different sampling rates.

2. Methodology

In this study, we investigate the effects of scalability by comparing a fitting-based approach (Beckman, Baggerly, and McKay 1996) with a generation-based approach (Farooq et al. 2013). As mentioned in the literature review, the fitting-based approach has been extensively used in the past for synthesizing populations in the context of activity-based and agent-based micro-simulation models. Recently, different generation-based approaches that outperform standard fitting-based techniques have been introduced in the literature. Recent studies suggest that the synthetic populations produced from fitting-based approaches are less accurate than the recently introduced

methods (Farooq et al. 2013; Saadi, Mustafa, Teller, and Cools 2016; Saadi, Mustafa, Teller, Farooq, et al. 2016; Sun and Erath 2015). To our knowledge, no studies really investigated the impact of scalability in the form of a comparative study apart from what has been discussed about the HMM-based approach of Saadi et al. (2016).

We propose different statistical indicators to assess the performance of the population synthesis approaches for different parameter settings. The results will be discussed on the basis of three metrics: the Root Mean square Error (RMSE), the Standardized Root Mean Square Error (SRMSE) and the Mean Absolute Error (MAE). The RMSE has been used in various studies to validate the accuracy of the simulated joint distribution with respect to the reference dataset (Lee and Fu 2011; Saadi, Mustafa, Teller, and Cools 2016; Vovsha et al. 2015). The mathematical formulation can be defined as follows:

$$RMSE = \sqrt{E((\tilde{\theta} - \theta)^2)} = \sqrt{\frac{\sum_{t=1}^{n} (\tilde{y}_t - y_t)^2}{n}}$$

where $E()$ represents the mean, $\tilde{\theta}$ and $\tilde{y}_t$ the simulated population, θ and y_t the observed population and n the total number of cells of the k-way contingency table.

Similarly, other studies adopted the SRMSE (Farooq et al. 2013; Pritchard and Miller 2012; Saadi, Mustafa, Teller, Farooq, et al. 2016; Sun and Erath 2015), especially when tables of different dimensions were tested. The related mathematical formulation can be defined as follows:

$$SRMSE = \frac{\sqrt{\frac{1}{n}\sum_i \sum_j \sum_k \ldots (\tilde{y}_{ijk} - y_{ijk})^2}}{\frac{1}{n}\sum_i \sum_j \sum_k \ldots {y_{ijk}}^2}$$

where $i, j, k, \ldots$ are respectively the subscripts of the first, second and third dimensions. Thus, the number of necessary subscripts depends on the number of attributes involved within population synthesis. And y_{ijk} is the number of agents combining attributes i, j and k within a cell.

Regarding the MAE, the mathematical formulation that has been used is the following:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} e_t = \frac{1}{n}\sum_{t=1}^{n} | \tilde{y}_t - y_t |$$

The accuracy of the population synthesis methods will be assessed for the synthesis of respectively two, three, four and five attributes. In addition, the procedure will be applied using samples of 10%, 25% and 50% with respect to the full population. In this regard, it will be possible to investigate the effects of sample size and scalability separately as well as their interactions.

As mentioned by Saadi et al. (2016), the concepts of scalability and dimensionality are related as they share the same induced effects that generally increase the error. In this paper, we select variables with a reasonable (not too high) number of levels to avoid the phenomenon of curse of dimensionality. No matter if the IPF or the simulation-based approach is followed, a high number of levels, e.g. more than 50, within a single variable can lead to stability problems. For example, conditional probabilities may not be calibrated correctly before being incorporated into the Gibbs sampler with respect to the simulation-based approach. In this regard, we have encountered such type of problems when it came to calibrate the MNL models.

With respect to IPF, the multi-dimensional contingency tables are fitted by using the package of Barthelemy and Suesse (2014). The IPF procedure is very popular in the literature to allow its implementation. However, regarding the simulation-based

approach, we will provide some additional details in order to facilitate a quick and efficient implementation of the approach.

Fundamentally, the simulation-based approach is based on a Monte Carlo Markov Chain (MCMC) algorithm, better known as the Gibbs Sampler (Farooq et al. 2013). The principle consists of building the multi-variate joint distribution as well as the marginal distributions from a set of full or partial conditional distributions of the true population. The conditional distributions are generally estimated on the basis of travel or socio-demographic surveys. For example, in our analysis, all the variables contain multiple levels. In this way, all the conditional distributions are in the form of MNL models. The structure of a Gibbs Sampler can be defined as follows:

- Random initialization of the variables $x_1, x_2, ..., x_N$
- For iteration $k = 1, ..., n_{pop}$
- Sample $x_1^{k+1} \leftarrow p(x_1 \mid x_2^k, x_3^k, ..., x_N^k)$
- Sample $x_2^{k+1} \leftarrow p(x_2 \mid x_1^{k+1}, x_3^k, ..., x_N^k)$
- …
- Sample $x_N^{k+1} \leftarrow p(x_N \mid x_1^{k+1}, x_2^{k+1}, ..., x_{N-1}^{k+1})$
- End

where $n_{pop} = n_a + n_b + n_c$. Indeed, n_{pop} is defined by n_a, the size of the target population, in addition to n_b, the number of runs for warming the Gibbs Sampler and n_c, the sum of all the non-selected sequences.

In practice it may happen that one or more explanatory variables are not available. In this context, an alternative could be adopted by setting up partial conditional distributions. As some information is missing, the accuracy of the synthetic

population is generally smaller than the case where only full conditionals are used. When the conditional distributions are correctly estimated, the generation of sequences can be realized by running the Gibbs sampler. After each loop, an observation corresponding to a set of attributes is designed such that it corresponds to one agent of the full synthetic population. Note that a certain number of runs need to be done at the beginning before taking the observations into account. In the literature, this phenomenon is known as the warming process. Then, as specified by Farooq et al. (2013) and Saadi et al. (2016), the observations are selected step by step according to a fixed number of observations. This procedure mitigates eventual correlations in-between successive sequences.

3. Data

The data used in this study stem from the workforce survey of 2013 that has been carried out in Belgium. After cleaning the data, a dataset of 30,700 observations was retained, consisting of the following 5 variables: age, education level, gender, profession and province. The variables have respectively 7, 16, 2, 7 and 11 levels. Note that age and the spatial variable have been aggregated for the simulation purpose. In this way, the spatial variable that initially contains 547 municipalities is now aggregated into 11 provinces, which corresponds to the number of provinces in Belgium. We suppose that the dataset represents a real population. In this context, we can easily extract the marginal distributions (aggregate information) for IPF and their related micro-samples that are needed for the simulation-based approach.

Table 1 presents the descriptive statistics related to all the variables. From this table, one could depict minor variations between the proportions, means and standard deviations between the different samples (of different sample size). This is due to the random selection of the observations for the different samples. However, a sample size

of 10% is sufficiently acceptable to avoid any problem of heterogeneity. In this regard, as the smallest sample size is at least 10%, there is no risk of bad representation of the true population. Three sample sizes have been selected such that we can focus one three aspects: the sample size, the scalability, and the eventual interaction between both of them. In order to be consistent, we have also synthesized two and three variables, although in practice, the scalability is more important for a larger number of synthesized variables. In this way, we can better appreciate the trends in terms of error rate with the increase of the scalability.

TABLE 1 Data description of the selected variables for different sampling rates

Rate	10%				25%				50%				100%			
	Levels	Pr.	Mean	S.D.	Levels	Pr.	Mean	S.D.	Levels	Pr.	Mean	S.D.	Levels	Pr.	Mean	S.D.
Age			40.3	10.9			40.6	11.0			40.7	11.1			37.7	21.3
Gender	1 / 2	54.1 / 45.9			1 / 2	53.7 / 46.3			1 / 2	54.2 / 45.8			1 / 2	49.5 / 50.5		
Status	1	4.1	-	-	1	4.1	-	-	1	4.2	-	-	1	11.5	-	-
	2	4.9	-	-	2	4.4	-	-	2	4.4	-	-	2	9.0	-	-
	3	3.9	-	-	3	4.2	-	-	3	4.6	-	-	3	5.7	-	-
	4	5.2	-	-	4	5.4	-	-	4	5.2	-	-	4	6.6	-	-
	5	10.7	-	-	5	11.2	-	-	5	11.3	-	-	5	13.1	-	-
	6	15.4	-	-	6	15.2	-	-	6	14.7	-	-	6	12.3	-	-
	7	11.3	-	-	7	11.0	-	-	7	10.5	-	-	7	8.9	-	-
	8	4.0	-	-	8	3.6	-	-	8	3.7	-	-	8	2.7	-	-
	9	19.7	-	-	9	19.9	-	-	9	20.3	-	-	9	14.8	-	-
	10	2.7	-	-	10	3.1	-	-	10	2.9	-	-	10	2.2	-	-
	11	0.8	-	-	11	0.8	-	-	11	0.6	-	-	11	0.9	-	-
	12	0.4	-	-	12	0.3	-	-	12	0.6	-	-	12	0.4	-	-
	13	3.0	-	-	13	3.1	-	-	13	2.9	-	-	13	2.0	-	-
	14	11.3	-	-	14	11.5	-	-	14	11.7	-	-	14	8.4	-	-
	15	2.0	-	-	15	1.7	-	-	15	1.7	-	-	15	1.1	-	-
	16	0.7	-	-	16	0.5	-	-	16	0.7	-	-	16	0.5	-	-
Profession	1	27.4	-	-	1	26.2	-	-	1	26.2	-	-	1	49.2	-	-
	2	39.6	-	-	2	41.1	-	-	2	41.2	-	-	2	21.2	-	-
	3	16.4	-	-	3	16.4	-	-	3	16.4	-	-	3	9.8	-	-
	4	7.8	-	-	4	7.5	-	-	4	7.3	-	-	4	12.4	-	-
	5	5.3	-	-	5	5.3	-	-	5	5.3	-	-	5	5.7	-	-
	6	3.1	-	-	6	3.2	-	-	6	3.3	-	-	6	1.7	-	-
	7	0.3	-	-	7	0.3	-	-	7	0.3	-	-	7	NA	-	-
Province	11(507)	-	-	-	11(538)	-	-	-	11(545)	-	-	-	11(547)	-	-	

4. Results and discussion

Tables 2-3 present the errors in terms of RMSE, SRMSE and MAE based on the comparison between the synthetic dataset and the reference dataset, i.e. the full population. The sampling rate is 50%. It means that in terms of data consumption, IPF includes all the marginal distributions as well as the initial micro-sample of 50%. In contrast, the conditional distributions of the simulation-based approach are only calibrated with the 50% PUMS. Besides, additional settings need to be defined regarding the convergence tolerance of the IPF algorithm, i.e. 10e-5, and the replacement of the zero-cells by very small values. The effects induced by the zero-cell problems are very low as we are ensuring that the number of levels per variable is reasonable. In this way, the number of cells of the k-way table will not be excessively important. One could depict that, in the case of the synthesis of 5 variables, the total number of cells is equal to 17,240.

Based on the RMSE, we can see that, for both methods, the errors are decreasing with higher levels of scalability (Figure 1), whereas one intuitively would expect that the error increases when the scalability increases. This counter-intuitive result is rooted in the mathematical definition of the RMSE. As the number of attributes decreases, the total number of cells n of the k-way contingency table will force the RMSE to decrease, despite the fact that the sum of the deviations is increasing. In other terms, the denominator takes precedence over the numerator. Regarding the MAE, similar trends can be observed. In general, both methods see their RMSE and MAE decreasing with an ascending scalability. In addition, we can clearly observe that the simulation-based approach provides better estimates. In contrast, if we observe the SRMSE (Figure 2), the simulation-based approach outperforms IPF by reducing the error by more than around 50%. In this way, the results confirm the findings of Farooq et al. (2013).

Although the simulation-based approach (Table 2) has a lower error rate than the IPF-based approach (Table 3), it is interesting to note a similar pattern in the growth of the SRMSE as the number of attributes increases. Indeed, as we go from three to five levels, the relative increase in error, e.g. from two to three, three to four and four to five attributes, is around 2, 5 and 8 (Table 2) and 3, 5 and 9 (Table 3), the errors seem to be increasing at roughly the same rate.

TABLE 2 Error rates for different level of scalability (IPF - sample=50%)

Levels	Scalability	RMSE	Tolerance	Nb. of cells	MAE	SRMSE
16						
7	2	0.0186133	10e-06	112	0.008928571	53.72502
2	3	0.009636428	10e-06	224	0.004464286	103.7729
7	4	0.001838705	10e-06	1,568	0.000637755	543.8611
11	5	0.000214848	10e-06	17,240	5.79777e-05	4654.455

TABLE 3 Error rates for different level of scalability (simulation-based - sample=50%)

Levels	Scalability	RMSE	MAE	SRMSE
16				
7	2	0.005483892	0.002369707	16.05
2	3	0.003787981	0.001708934	40.19
7	4	0.000755994	0.000281983	224.30
11	5	0.000107891	3.79631e-05	2337.34

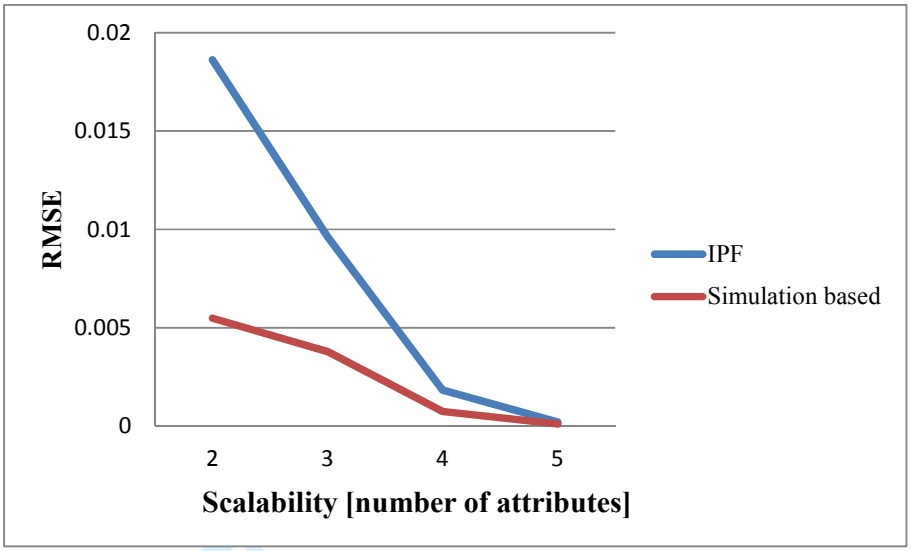

FIGURE 1 Comparison between IPF and SB in terms of RMSE for an increasing scalability

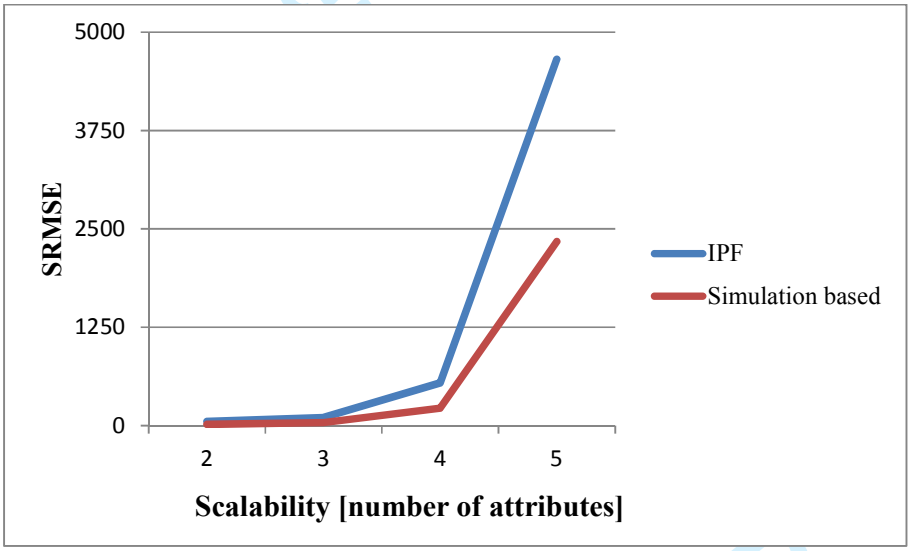

FIGURE 2 Comparison between IPF and SB in terms of SRMSE for an increasing scalability

Table 4 presents the results for a sample size of 25%. Note that the results related to IPF are not represented. Indeed, the changes in terms of error associated to IPF are so small that the values are similar through the different sampling rates. In contrast, some minor variations occurred in the case of the simulation-based approach.

In this regard, the decrease of the sampling rate has some minor influence on the error rates, i.e. variation in RMSE is +1.03% when shifting from a sampling rate of 50% to 25%. This can be explained by the fact that the conditional probabilities of the Gibbs Sampler are calibrated by using the PUMS. As a lower quantity of information is captured by a smaller sample size, the error generally increases. In this regard, the observed trends in terms of accuracy correspond to the findings of Saadi et al. (2016) where two sampling rates have been tested, i.e. 50% and 100%.

TABLE 4 Error rates for different level of scalability (simulation-based - sample=25%)

Levels	Scalability	RMSE	MAE	SRMSE
16				
7	2	0.005541	0.002501	15.99
2	3	0.003666	0.001663	39.48
7	4	0.000761	0.000287	225.14
11	5	0.000109	0.000038	2363.78

From Table 5, one can observe that the errors are still slightly increasing compared to the previous results. An important remark should be pointed out at this stage. One can see that for a sampling rate of 10%, the simulation-based approach provides better estimates than that of IPF in the case of 50% sample. In addition, the amount of input data for calibrating an IPF is more important. Globally, although the errors in terms of RMSE, MAE and SRMSE are varying negatively, the changes remain quite stable for both methods. Note that the sampling rate has been divided by 5, from 50% to 10%, while both methods preserve good estimates. In this regard, it is not necessary to establish travel surveys which size exceed 10%.

Besides, if we analyze the scalability of the methods separately, we can see that the SRMSE increase significantly, i.e. from 16.44 to 2342.55 while the number of cells is multiplied by around 70. In this regard, one should pay attention about the data preparation step. The levels within each variable should be limited as much as possible.

If a continuous variable need to be synthesized, the variable should be aggregated with the lowest number of categories necessary for the application. The number of cells of the k-way contingency table has an important influence on the calculation of the metrics.

TABLE 5 Error rates for different level of scalability (simulation-based - sample=10%)

Levels	Scalability	RMSE	MAE	SRMSE
16				
7	2	0.005697	0.002549	16.44
2	3	0.003631	0.001650	39.10
7	4	0.000791	0.000296	234.11
11	5	0.000110	0.000039	2382.55

In terms of scalability, the errors, related to the IPF approach and the simulation-based approach, are more or less constant, despite the fact that the sample size is decreasing. Based on the SRMSE, the increase of the error with respect to the simulation-based approach is lower than that for IPF.

Regarding the interpretation of the results using different metrics, we can learn that the information may be completely contradictory. For example, the RMSE are decreasing with an increased level of scalability. In contrast, the SRMSE are increasing. In this regard, it is necessary to select the most adapted metric to check the accuracy of synthetic populations. RMSE is more adapted when it comes to compare methods with the same k-way contingency table size.

Besides, based on the SRMSE, the difference between IPF and HMM increases significantly with the increase of the level of scalability. For example, in the case of a sample of 10%, the shift towards higher level of scalability leads to an increase of +73.46% (from 2 to 3), +378.98% (from 3 to 4) and +633.47 (from 4 to 5). Indeed, the relative differences are increasing because of the fact that the error inherent to IPF is increasing faster than HMM.

Globally, one could depict from Tables 3-5 that MAE provides almost similar trends than those stemming from RMSE. In this regard, the remarks formulated for RMSE can be similarly applied to the MAE.

5. Concluding remarks

In this paper, we investigated the effects of scalability on the accuracy for different synthetic populations by comparing results from a standard IPF algorithm (Beckman, Baggerly, and McKay 1996) with the ones of a simulation-based method (Farooq et al. 2013). Besides, we took into account the effects of sampling rates and checked the eventual interactions with scalability.

First, the findings reveal that for all the level of scalability and for all the sampling rates, the simulation-based approach outperforms IPF. In this context, the study extends the findings of Farooq el al. (2013) for additional scalability levels. Different reasons could explain these findings. Based on the random process present in the generation and the selection process of attribute' sequences, the MCMC algorithm is capable of building the joint distribution while incorporating some heterogeneity into the synthetic population. In this way, the simulated population may contain some combination of attributes that where not present in the training PUMS.

With respect to the reliability of the statistical metrics, we have highlighted the need of choosing the most adapted indicator based on the nature of the problem we are considering. In this regard, one can notice that from 5 synthesized attributes, the accuracy of an IPF gets close to that of the simulation-based approach according to the RMSE. In this regard, it would mean that from 5 attributes, both methods are equivalent. In contrast, the SRMSE reveal that the simulation-based approach is less sensitive to an increase of the level of scalability. In this regard, it should be emphasized that, based on their mathematical formulations, the SRMSE is a more appropriate

indicator for measuring the scalability. Given the fact that the RMSE is too sensitive to the number of cells, its standardized form can provide a better appreciation of synthetic populations when the level of scalability increases.

In conclusion, this study fills a serious gap in the literature regarding the effects of scalability on population synthesis accuracy. To our knowledge, there is no study that proposes a comparison between methods stemming from different population synthesis philosophies. Saadi et al. (2016) discussed the effects of scalability, but only in the context of a Hidden Markov Model-based approach. This paper highlights important aspects that need to be taken into account, and identifies additional issues associated to scalability which require further analysis. For example, more efficient statistical metrics could be used to better capture the effects of scalability. Also, depending on the complexity of a variable, i.e. number of levels, a more explicit link could be established between the added attribute and the loss in accuracy of the synthetic populations. Tests could be realized for smaller sample sizes, while scalability is increasing. In this regard, datasets containing a higher number of observations should be used. The important dependency on the micro-sample can play a negative role on the calibration of the simulation-based approach, especially when very small sampling rates (<5%) are considered. Thus, it is strongly recommended to preserve high sampling rates when travel or socio-demographic surveys are realized. While synthetic populations stemming from an IPF will be maintained by the aggregate source of information, i.e. marginal distributions, the simulation-based approach will depend essentially on the initial micro-sample. In such conditions, the effects of scalability could be analyzed to extend the conclusions of the study.

6. References

Barthelemy, Johan, and Thomas Suesse. 2014. "Package 'mipfp.'" http://143.107.212.50/web/packages/mipfp/mipfp.pdf.

Barthelemy, Johan, and Philippe L. Toint. 2013. "Synthetic Population Generation Without a Sample." *Transportation Science* 47 (2): 266–279. doi:10.1287/trsc.1120.0408.

Beckman, Richard J., Keith A. Baggerly, and Michael D. McKay. 1996. "Creating Synthetic Baseline Populations." *Transportation Research Part A: Policy and Practice* 30 (6): 415–429. doi:10.1016/0965-8564(96)00004-3.

Deming, W. Edwards, and Frederick F. Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known." *The Annals of Mathematical Statistics* 11 (4): 427–444.

Farooq, Bilal, Michel Bierlaire, Ricardo Hurtubia, and Gunnar Flötteröd. 2013. "Simulation Based Population Synthesis." *Transportation Research Part B: Methodological* 58 (December): 243–263. doi:10.1016/j.trb.2013.09.012.

Lee, Der-Horng, and Yingfei Fu. 2011. "Cross-Entropy Optimization Model for Population Synthesis in Activity-Based Microsimulation Models." *Transportation Research Record: Journal of the Transportation Research Board* 2255 (December): 20–27. doi:10.3141/2255-03.

Mohammadian, Abolfazl (Kouros), Mahmoud Javanmardi, and Yongping Zhang. 2010. "Synthetic Household Travel Survey Data Simulation." *Transportation Research Part C: Emerging Technologies*, Special issue on Transportation Simulation Advances in Air Transportation Research, 18 (6): 869–878. doi:10.1016/j.trc.2010.02.007.

Müller, Kirill, and Kay W. Axhausen. 2011. "Population Synthesis for Microsimulation: State of the Art." In . http://trid.trb.org/view.aspx?id=1092120.

Pritchard, David R., and Eric J. Miller. 2012. "Advances in Population Synthesis: Fitting Many Attributes per Agent and Fitting to Household and Person Margins Simultaneously." *Transportation* 39 (3): 685–704. doi:10.1007/s11116-011-9367-4.

Saadi, Ismaïl, Ahmed Mustafa, Jacques Teller, and Mario Cools. 2016. "Forecasting Travel Behavior Using Markov Chains-Based Approaches." *Transportation*

Research Part C: Emerging Technologies 69 (August): 402–417. doi:10.1016/j.trc.2016.06.020.

Saadi, Ismaïl, Ahmed Mustafa, Jacques Teller, Bilal Farooq, and Mario Cools. 2016. "Hidden Markov Model-Based Population Synthesis." *Transportation Research Part B: Methodological* 90 (August): 1–21. doi:10.1016/j.trb.2016.04.007.

Sun, Lijun, and Alexander Erath. 2015. "A Bayesian Network Approach for Population Synthesis." *Transportation Research Part C: Emerging Technologies* 61 (December): 49–62. doi:10.1016/j.trc.2015.10.010.

Voas, David, and Paul Williamson. 2001. "Evaluating Goodness-of-Fit Measures for Synthetic Microdata." *Geographical and Environmental Modelling* 5 (2): 177–200. doi:10.1080/13615930120086078.

Vovsha, Peter, James E. Hicks, Binny M. Paul, Vladimir Livshits, Petya Maneva, and Kyunghwi Jeon. 2015. "New Features of Population Synthesis." In . http://trid.trb.org/view/2015/C/1339180.

Williamson, P., M. Birkin, and P. H. Rees. 1998. "The Estimation of Population Microdata by Using Data from Small Area Statistics and Samples of Anonymised Records." *Environment and Planning A* 30 (5): 785–816. doi:10.1068/a300785.

Ye, Xin, Karthik Charan Konduri, Ram M. Pendyala, Bhargava Sana, and Paul Waddell. 2009. "Methodology to Match Distributions of Both Household and Person Attributes in Generation of Synthetic Populations." In . http://trid.trb.org/view.aspx?id=881554.

Zhu, Yi, and Joseph Ferreira. 2014. "Synthetic Population Generation at Disaggregated Spatial Scales for Land Use and Transportation Microsimulation." *Transportation Research Record: Journal of the Transportation Research Board* 2429 (December): 168–177. doi:10.3141/2429-18.