Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking

Author(s):
Casati, Daniele; Müller, Kirill; Fourie, Pieter J.; Erath, Alexander; Axhausen, Kay W.

Publication date:
2015

Permanent link:
https://doi.org/10.3929/ethz-b-000086853

Originally published in:
Transportation Research Record 2493, https://doi.org/10.3141/2493-12

Synthetic Population Generation by Combining a Hierarchical, Simulation-Based Approach with Reweighting by Generalized Raking

Daniele Casati, Kirill Müller, Pieter J. Fourie, Alexander Erath, and Kay W. Axhausen

A recent approach for generating populations of synthetic individuals through simulation is extended to produce households of grouped individuals. The contingency tables of the generated populations match external controls on the individual and household levels while exhibiting far greater variety in composition than existing approaches can offer. The method involves a two-step approach. The first consists of a procedure based on Gibbs sampling, which has only recently been applied to population generation in transportation modeling and is generically called Markov chain Monte Carlo (MCMC). For this work, the model was generalized, and an extension was developed, hierarchical MCMC, which was able to generate a hierarchical structure. The second step, a postprocessing step, uses generalized raking (GR), which reweights the output from hierarchical MCMC to perfectly satisfy known marginal control totals on the individual and household levels. The application input data—a demographic sample and some known marginals from Singapore—added further complexities to the problem, which had not yet been explored in the current literature. Despite data challenges, consecutively applying the methods above produced realistic synthetic populations. Results confirm their goodness of fit and their generated hierarchical structures.

Activity-based transport demand models based on individual agents are widely regarded in research and practice as the most suitable tools for evaluating transport infrastructure and policy. These models require a population of agents that is representative of the actual population and that features sociodemographic attributes to be used in various behavioral submodels, which define, for example, activity schedule or mode choice.

Because of privacy concerns but also monetary constraints, individual-level data on the complete population are normally not available. As an alternative, statistical offices usually provide microsamples with rather limited spatial resolution. For example, in the United States, public use microdata sample (PUMS) files contain records representing 5% or 1% samples of housing units and the people living in the occupied units. Alternatively, travel surveys conducted by governmental bodies usually also cover disaggregated information on a household and an individual level. However, other kinds of agents and hierarchies can be considered, such as employees and firms. To reconstruct statistically representative complete populations, various synthesis techniques have been proposed and applied in the literature.

The conventionally applied population synthesis techniques in transport modeling [see Müller and Axhausen for an overview (*1*)] apply variants of iterative proportional fitting (IPF), first introduced by Deming and Stephan (*2*). This technique operates only on categorical data, which can be organized in a contingency table that defines the control totals over multiple dimensions. The population is synthesized by replicating individuals and households as represented in the microsample. The desired contingency table is fitted by iterating over all controls in a round-robin fashion until convergence; in each iteration, the cell values are rescaled uniformly to perfectly satisfy the current control.

To address the shortcoming that the basic configuration of IPF can control only for either individual- or household-level attributes, iterative proportional updating (*3*) and hierarchical IPF (*4*), as well as entropy-based methods (*5, 6*), have recently been introduced as techniques that ensure that expansion factors are consistent on both levels. However, a major drawback of IPF is that combinations of attributes that are not part of the reference sample cannot be generated; this problem is known as the "zero-cell problem." Hence, the heterogeneity in regard to individual and household attributes in the resulting synthesized population is limited to the observations in the reference sample.

To overcome those issues, Farooq et al. proposed a Markov chain Monte Carlo (MCMC) simulation-based approach that uses partial views of the joint probability distribution (*7*). They successfully demonstrated that the resulting synthetic population not only outperformed an IPF-based method in relation to fit with the actual full population, but also featured a higher level of heterogeneity. However, the proposed method is restricted to nonhierarchical data, and the extension of the method to generate associations between households and individual people has been identified as a relevant strand for further research. Another unaddressed problem is that MCMC does not allow for directly imposing any control totals on the synthetic population.

In this work, an extension of the MCMC simulation-based population is proposed to combine individual and household attributes at the same time, in a process that is called hierarchical MCMC. Furthermore, generalized raking (GR) is introduced as a technique

D. Casati, K. Müller, P. J. Fourie, and A. Erath, IVT, ETH Zürich, Raemistrasse 101, CH-8093 Zürich, Switzerland. K. W. Axhausen, IVT, ETH Zürich, Wolfgang-Pauli-Strasse 15, ETH-Honggerberg, HIL F 32.3, CH-8093 Zürich, Switzerland. Corresponding author: A. Erath, alexander.erath@ivt.baug.ethz.ch.

Transportation Research Record: Journal of the Transportation Research Board, No. 2493, Transportation Research Board, Washington, D.C., 2015, pp. 107–116.
DOI: 10.3141/2493-12

to fit the simulated synthetic population to actual observed control totals (*8*). The proposed method is applied with travel survey data from Singapore, and the resulting synthetic populations are tested for representativeness and consistency on an individual and a household level.

METHOD

Two methods were combined: (*a*) an extension of the MCMC method that allows producing hierarchies of people grouped into households and (*b*) an optional postprocessing of the output from the first method to perfectly satisfy known control totals on the individual and household levels (*7*).

Markov Chain Monte Carlo

The technique used to generate populations belongs to the family of Markov chain Monte Carlo (MCMC) methods. Farooq et al. first explored this approach for population generation in the field of transportation modeling to synthesize a population of individuals (iMCMC) (*7*). An extension to their approach has been developed; it is capable of synthesizing a hierarchical structure, grouping individuals into households (hMCMC). In what follows, the original approach is first explored, and then it is compared with the extended method.

iMCMC: Individual Population Synthesis Through MCMC

Instead of directly approximating the joint probability distribution of the attributes that describe the agents, iMCMC (schematized in Figure 1) exploits the conditional distributions of each variable with respect to all other variables or a subset of them, as in Farooq et al. (*7*). These conditionals are fitted a priori from an available demographic sample through a model, such as multinomial linear logistic regression.

The second step of iMCMC is to actually implement a Markov chain for the generation of agents through Gibbs sampling. Given an initial seed agent characterized by a vector of N attributes $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_N^{(0)})$, sample the category of attribute i of agent k according to the conditional distribution:

$$P\left(X_i \middle| x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_{i+1}^{(k-1)}, \ldots, x_N^{(k-1)}\right) \tag{1}$$

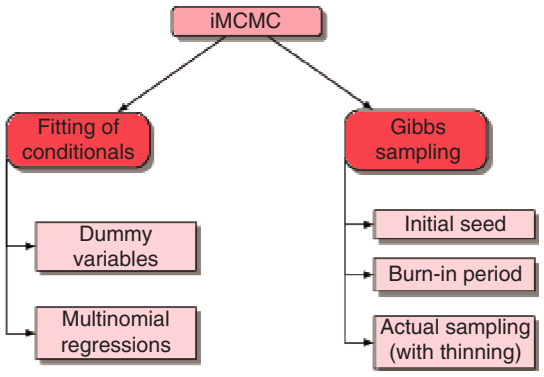


FIGURE 1 Diagram of iMCMC.

[Here and in the following, $P(Y|X = x)$ is abbreviated by $P(Y|x)$, and in turn $P(Y = y|x)$ is abbreviated by $P(y|x)$]. This conditional distribution is modeled under a multinomial logit regression with coefficients β by

$$\log \frac{P\left(x_{ij} \middle| x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N\right)}{P\left(x_{iM} \middle| x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N\right)} = \beta_{ij0} + \sum_{\substack{n=1 \\ n \neq i}}^{N} \beta_{ijn} \cdot x_n \tag{2}$$

for all $j = 1, 2, \ldots, M - 1$, where M is the number of categories for the ith attribute. Under the multinomial logit model, one category (the Mth in this equation) is chosen as the pivot category, that is, its conditional weight $P(x_{iM}|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N)$ can be obtained from the normalization condition.

Gibbs sampling then ensures that the stationary probability distribution of this Markov chain is the joint distribution $P(X_1, X_2, \ldots, X_N)$.

As is usual with simulated Markov chains, to reset the memory of the initial seed, it is necessary to discard some sampled agents before collecting them for the new population (burn-in period). The first samples should be discarded even when the initial seed is made of randomly chosen variables. Indeed, if the seed is an improbable agent (something very likely with demographic data, for which the number of feasible combinations of attributes is low compared with the total), the new population would be too dependent on it, and not a fair sample of the joint distribution.

Another characteristic of Gibbs sampling is that two subsequently generated agents tend to have rather similar sets of attributes. To reduce autocorrelation, the generated chains are thinned and all but every kth observation are discarded.

The characteristics of this method, compared with the more common IPF-based approaches in transportation modeling, can be summarized as follows:

- The joint distribution, which is difficult to model, does not need to be accessed directly.
- Not too many assumptions on the underlying structure of the data are imposed, apart from considering the conditional distributions as smooth functions. Thus continuous variables can be considered instead of categorical ones. However, iMCMC cannot impose control totals (although the marginals obtained are similar to those of the reference sample), which makes a postprocessing operation necessary.
- iMCMC does not suffer from the IPF zero-cell problem, that is, nonexistent combinations of attributes in the reference sample are possible in the generated populations. Indeed, fitted conditionals under a smooth model such as the multinomial logit can never be completely zero, and since they are the transition weights of the MCMC Markov chain, some very unlikely, but possible, steps can lead to these combinations. However, if a category of an attribute is missing from the reference data set, no model to fit conditionals can consider it.

Consequently, the results of iMCMC are not too dependent on the reference sample, allowing as many heterogeneous new populations as desired. However, iMCMC can have some flaws when one deals with multipolarized data sets, as will be shown in the section on the demographic data sample.

hMCMC: Hierarchical MCMC

This extension of iMCMC, aimed at synthesizing populations with a hierarchical structure, is based on ordering the agents living in the

same household according to their household roles. A "household role" variable can already be present in the reference sample as an individual variable [something usual in the data treated by the literature on transportation modeling, as in Pritchard and Miller (*9*)], or it has to be defined as in this work.

The general formulation of hMCMC (schematized in Figure 2) is based on the definition of three groups of agent types, each to be generated differently: (*a*) "owners," (*b*) "intermediate types," and (*c*) "others." "Owners" are synthesized under iMCMC and also characterized by variables on the household level. "Intermediate types," in the case discussed here, are spouses and children. The predictors of their conditionals take into consideration some variables of the already generated agent types, so that the model of their conditionals becomes larger the further the intermediate type is from the respective owner. The number of intermediate types to be defined can vary according to the complexity of the model one wants to implement. The conditionals of "others" do not change and are dependent only on their owners and intermediate types, regardless of the step in which they are generated.

The segmentation of the household into owners, intermediate types, and others is performed with a rule-based approach. For instance, the owner of a household can be identified by a sequential selection process; first the person(s) with the highest reported "inc" (income) is selected, and this subsample can then be further screened for other selection criteria until a single agent is identified as the owner. Similar strategies are then also applied to classify the remaining people of the household. Referring to a conventional nuclear family model, the intermediate types can be described as "spouse" (second agent) and "child" (third agent).

However, those descriptions should be interpreted with care as the proposed MCMC approach does not impose a certain household composition model. Segmenting a single-parent household, for example, will identify a child as the first intermediate type, which is described here as spouse. The accordingly fitted conditionals for the first intermediate types will then ensure that the MCMC process will generate the appropriate number of household types, which might not necessarily correspond to the family model used to describe the different agent types.

Hence, to generate a household, its owner is sampled first according to the iMCMC approach described in the previous subsection. This agent wholly represents its household since the variables characterizing the household are drawn together. (An example of these kinds of variables can be the number of agents living there, hereafter referred to as "numpax.")

Then, if numpax > 1 for the current owner, the other inhabitants of its household are also generated accordingly. In Gibbs sampling, the conditionals of these subsequent agents also depend on some attributes of the already generated agents: for example, when the age of the spouse of a household is drawn, the corresponding conditional is

$$P(\mathrm{AGE}_{\mathrm{SPOUSE}} | \mathrm{other\_variables}_{\mathrm{SPOUSE}}, \mathrm{some\_variables}_{\mathrm{OWNER}}) \quad (3)$$

and drawing attributes for a child conditions on some attributes of the spouse as well. In the case of a large numpax, the attributes of type "other" are not used to condition the generation of later agents of the same type. This decision was made to keep the algorithm simple and avoid overfitting since most of the agents of the reference sample belonged to the first three types.

Fitting to Known Control Totals

Gibbs sampling, based on a reasonably specified and estimated joint probability distribution, will ensure that the simulated populations will be representative for the full population in regard to the variables used. However, this MCMC procedure does not ensure meeting stringent control totals as published by statistical offices or predefined scenarios. For that purpose, the simulated sample needs to be reweighted.

Earlier approaches to population synthesis [see Müller and Axhausen for a review (*1*)] are based on reweighting a reference sample to known control totals (fitting) and then using those weights as a probability distribution to sample from so as to produce synthetic populations that satisfy the controls (generation). These control totals specify the number of individuals or households with given characteristics in a geographic area.

Here such a reweighting technique is briefly introduced: generalized raking (GR), which is applied to the results of the hMCMC process (*8*).

Generalized Raking

Several extensions of the IPF algorithm to handle hierarchical structures have been proposed in the literature on transportation

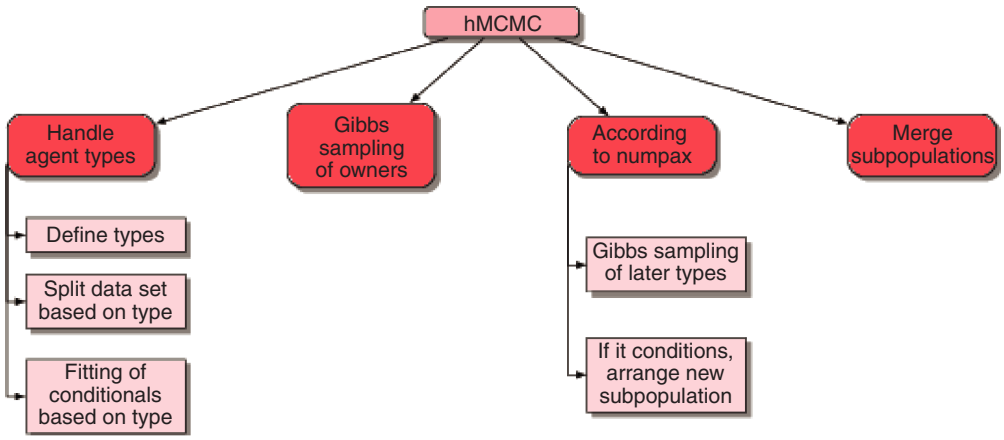


FIGURE 2 Diagram of hMCMC.

modeling (*3–6, 10, 11*). As shown in Müller and Axhausen, most of these techniques are very similar or in fact equivalent to (a special case of) the much more powerful GR, which is used in the domain of survey statistics to estimate weights for a survey to reflect known exogenous totals and is also capable of processing continuous variables (*12*).

For hierarchical structures, each category on the individual level is mapped to a count variable on the group level, which in this case corresponds to households, and the sum of each such count variable is controlled. Households are identified by their signature, that is, the number of individuals for each category; households with identical signatures can be pooled for this algorithm to reduce run times.

Example: When the overall number of persons of each sex is controlled for, two count variables sex = M and sex = F will be controlled for. The externally imposed distribution of the categorical sex variable on the individual level is converted to a vector of desired sums over the sex = M and sex = F variables for the signatures. For a family household with husband, wife, and one daughter, the signature has a value of 1 for sex = M and a value of 2 for sex = F. A single household where a mother lives with a son and a daughter would have the same signature with respect to the sex variables.

The estimation of weights is performed by solving a constrained optimization problem with standard numerical methods. These weights, now on the household level, can be again treated as a probability distribution; sampling from this distribution will lead to populations which, in expectation, satisfy the controls.

Here GR is used to postprocess an hMCMC synthetic population so that marginal controls on the individual and household level are perfectly satisfied. In this way, the hMCMC population can be used as a large reference sample for the usual fitting-generation approach described above.

As hMCMC is able to create as many households as desired, GR can also be applied to an oversampled population (i.e., where the sample size is greater or equal to the population size), which leads to improved heterogeneity of the final synthetic population.

EXPERIMENT

This section will describe the experiment to validate the discussed method in two steps:

- Synthesizing a population using hMCMC and
- Optional postprocessing by GR to fix its marginals.

In both cases, for one reference sample, one population about 10 times the size of the sample is created by using hMCMC, some assumptions having been made on the reference data.

With GR, the hMCMC population is calibrated to the sample by using its marginals as control totals as a proof of concept. Validation is then performed by comparing interactions between attributes in the synthetic population with the reference sample, allowing assessment of the error introduced by the method.

In addition, a reweighting against actual marginal totals has been performed; here, the distribution of the resulting weights will be analyzed briefly.

Demographic Data Sample

The demographic sample used in this work was derived from the household records of the 2008 Household Interview Travel Survey of Singapore, commissioned by Singapore's Land Transport Authority. It is unusual in population synthesis to use travel surveys as a source for reference samples, as the PUMS offered by many census authorities makes possible larger, richer data sets that are coded consistently with any marginal totals those authorities might release. However, in the case of Singapore, the stewards of the census, Statistics Singapore, release only summary information on the population, and no PUMS exists.

The sample consisted of 35,448 agents living in 10,640 different households, making up approximately 1% of the resident population of the island. From all the demographic variables surveyed, a selection of pertinent attributes was made; these are listed in Table 1 with their corresponding possible categories. The "ethnicity" attribute is a household variable as all the households of the sample were ethnically homogeneous. For the "dwell" attribute, the various Housing and Development Board (HDB) classes correspond to public housing flats of various sizes and standards.

An additional household index uniquely identifies each household. This index, shared by agents living together, encoded the hierarchical structure of the sample.

By multiplying the number of possible categories per attribute from Table 1, there are 155,232 obtainable combinations, which is the number of considered possible distinct agents. It is much larger than the possible combinations explored in Farooq et al., 384 (*7*). However, the number of distinct agents actually present in the sample was just 8,687, and that fact led to problems with hMCMC.

While MCMC does not suffer from the IPF zero-cell problem, multipolarization of the reference data makes Gibbs sampling flawed. Consider a distribution in which nonzero probabilities are concentrated in "islands" surrounded by an "ocean" of zero or very low probability. Here, the Markov chain would need to perform a large number of very unlikely (or even impossible) steps to "pass the ocean between these islands" and sample from the whole joint distribution.

This multipolarization naturally arises when a large number of attributes are considered (the curse of dimensionality), but it can also occur with particularities of the data. In the present case, it manifested itself particularly for two variables of the reference sample that were included in initial experiments. As these variables reported the type

TABLE 1 Attributes of Demographic Data Sample

Individual Level			Household Level		
Sex	Age (years)	Income (SGD)	Ethnicity	Dwell	Numpax
F	4	No income	Chinese	Condo	1
M	9	Max. 1,000	Indian	HDB 1/2	2
NA	14	Max. 1,500	Malay	HDB 3	3
	19	Max. 2,000	Other	HDB 4	4
	24	Max. 2,500	na	HDB 5	5
	29	Max. 3,000	na	Landed property	6
	34	Max. 4,000	na		7
	39	Max. 5,000	na	Other	8
	44	Max. 6,000	na	na	9
	49	Max. 7,000	na	na	10
	54	Max. 8,000	na	na	11
	59	Over 8,000	na	na	na
	64	na	na	na	na
	65+	na	na	na	na

NOTE: SGD = Singapore dollars (1 SGD = US$0.677096 in December 2008); max = maximum; NA = not available.

of occupation and ongoing education, they applied only to subsets of the agents, namely, to people who were either economically active or students, respectively.

The problem was revealed through two exploratory data analysis techniques, multiple correspondence analysis (*13*) and self-organizing map analysis (*14*). These techniques showed that the data were clustered mainly into three groups: (*a*) infants with unspecified gender (XXX) and lowest age level (age 4), (*b*) students with "no inc," and (*c*) the economically active population, with their occupation specified.

Therefore, considering occupation and education (which were mainly responsible for the clusters) heavily polarized the reference sample into more than one cluster, and the fitted conditionals had some very low transition weights. Consequently, the populations produced by including these variables, while being considered acceptable by error measurement standards, showed many outliers. That outcome is why these attributes were not considered for the population synthesis. However, exactly because these variables are very polarized, they are ideally suited to be added in a postprocess using the set of attributes included in the synthesis as predictors, for example, by applying statistical matching as in Müller and Axhausen (*15*).

Model

The proposed hMCMC requires agent types, and these had to be defined on the demographic sample used. As discussed previously, their role is to simply order the inhabitants of a household, which are then synthesized; however, instead of names such as Type 1, Type 2, and so on, specific names were chosen for clarity, resembling the procedure through which the types were defined:

- Owner. First agent of its household, with the highest income. If there is a conflict, the one with the highest age is chosen. If they are still multiple, the owner is identified randomly.
- Intermediate types:
  - Second agent of the household, with the minimum age distance from the owner among those of the opposite sex. If no agent of the opposite sex is available, the agent with the minimum age distance is chosen. If there are multiple agents, the spouse is identified randomly.
  - Third agent of the household, with the maximum age distance from the owner. If there is a conflict, the agent with the maximum inc distance is chosen. If there are still multiple agents, the child is identified randomly.
- Other. Includes the remaining agents of the household.

In hMCMC these types were treated differently when their conditionals were being fitted. To be as general as possible, these were dependent on all the other synthetic attributes and on as many variables as possible from the already generated types. Hence, while owners were generated under iMCMC with all the attributes considered in the section on the demographic data sample (even the household ones), the synthetic variables of the intermediate types were only sex, age, and inc, conditioned by all the attributes of the already generated types living in the same household, as outlined in Table 2. However, incomplete models may also be considered.

For reasons of simplicity, in a father–son household the father will be classified as owner and the son as spouse. Using different models per household type is outside the scope of this paper.

Out of the 35,448 individuals in the reference sample, 7,927 were identified as other. However, only in 2,092 households were more than two individuals identified as other, as these households were composed of five or more individuals. As a result of this comparably small number and to avoid overfitting, the decision was made to use the same conditionals for any agent of type other, that is, for numpax ≥ 4.

The GR postprocessing applied after hMCMC used all attributes of the reference sample as marginal controls: sex, age, and inc for individuals and ethnicity, dwell, and numpax for households. Consequently, all one-dimensional marginal distributions, on the individual and household level, were identical for the reference sample and the final synthetic population after GR was applied.

Implementation

The implementation of hMCMC used in this work fitted the conditionals in R with multinomial linear logistic models and passed their parameters to Java, in which the probability weights were computed and used at each step of Gibbs sampling. The multinomial logit was chosen since the R package nnet through its function multinom proved very efficient.

It also managed to solve a possible problem of hMCMC: the number of categories of a variable characterizing a certain agent type can be different when this variable is involved in the generation of its agent type and when it conditions the generations of subsequent types. The developed implementation succeeds in dealing with this issue by discarding the combinations of variables that cannot be interpreted by models of later agent types (i.e., through an acceptance–rejection sampling).

Another relevant detail is that since MCMC can be performed as long as necessary and stopped at any time, the number of considered steps of an iMCMC Markov chain is equal to the size of the generated population, while this approximatively holds for hMCMC. The uncertainty arises because the controllable size of the synthetic population in hMCMC is actually the number of owners, and the total of agents is then controlled by the owners' numpax and by the number of owners that had to be removed because of the flaw of hMCMC discussed in the previous paragraph.

TABLE 2 Attributes Involved in Conditionals Fitted per Agent

Role of Attribute	Owner	Spouse	Children	Other
Both predictor and response (synthesized)	Sex, age, income, ethnicity, dwell, numpax	Sex, age, income	Sex, age, income	Sex, age, income
Only predictor (conditioning)	na	ETHNICITY, DWELL, NUMPAX, SEX_{OWNER}, AGE_{OWNER}, INC_{OWNER}	ETHNICITY, DWELL, NUMPAX, SEX_{OWNER}, AGE_{OWNER}, INC_{OWNER}, SEX_{SPOUSE}, AGE_{SPOUSE}, INC_{SPOUSE}	ETHNICITY, DWELL, NUMPAX, SEX_{OWNER}, AGE_{OWNER}, INC_{OWNER}, SEX_{SPOUSE}, AGE_{SPOUSE}, INC_{SPOUSE}, SEX_{CHILD}, AGE_{CHILD}, INC_{CHILD}

However, the fact that the size of the analyzed synthetic population from hMCMC (354,178 agents) was actually very close to 10 times the reference one (35,448) is evidence of the validity of the developed approach.

A number of owners 10 times larger was chosen because the results of hMCMC seemed to improve the larger the synthetic population was; its Markov chains had more time to reach the more separate combinations in the attributes' space and to properly explore the underlying joint distribution. Thus, it is possible to correctly obtain a small population from the hMCMC result only by subsampling a larger one (e.g., by increasing the thinning or through decimation).

However, the GR postprocessing assigns natural weights that sum up to the required totals, which can be used for weighted sampling; in particular, the total number of individuals (as well as of households) is fixed. Then, the oversampling rate becomes a compromise between heterogeneity and computational efficiency, as the run time for GR depends on the heterogeneity of the underlying population.

The R package MultiLevelIPF was used for implementing GR (*16*). In the presence of very infrequent categories (such as numpax = 11), convergence could be achieved with only limited precision.

RESULTS

The synthetic population is analyzed at each of the two steps of the procedure described in the section on the experiment and compared with the reference sample. In the following, the corresponding populations are labeled as follows:

- RS—reference sample,
- MCMC—at the end of hMCMC (in table headers: MC), and
- GR—after GR is used to impose the controls defined by the sample on the hMCMC population.

MCMC and GR populations both have 354,178 observations, about 10 times larger than RS. The GR population is defined by the weighting; no sampling has been performed.

Individual Level

Results refer only to characteristics on an individual level. However, the related new population was still synthesized with hMCMC, and the generated hierarchical structure was simply ignored.

Figure 3 shows plots of the relative frequencies (normalized counts) of combinations of variables, thus allowing results of multidimensional attributes to be displayed in two-dimensional pictures. The left-hand part of the figure shows all feasible combinations of categories for the individual variables age, sex, and inc, ordered by their frequency in RS. Since the results of MCMC and GR are very similar (with one notable exception discussed below), only GR results are displayed. The top part shows the relative frequency for both populations; the bottom part shows the absolute error of GR compared with RS. As expected, since more agents are involved, the error increases with increasing frequency in RS but remains mostly below 0.1%. The right-hand part shows the same analysis for the household variables ethnicity, dwell, and numpax; no links between the individuals in a household are analyzed here.

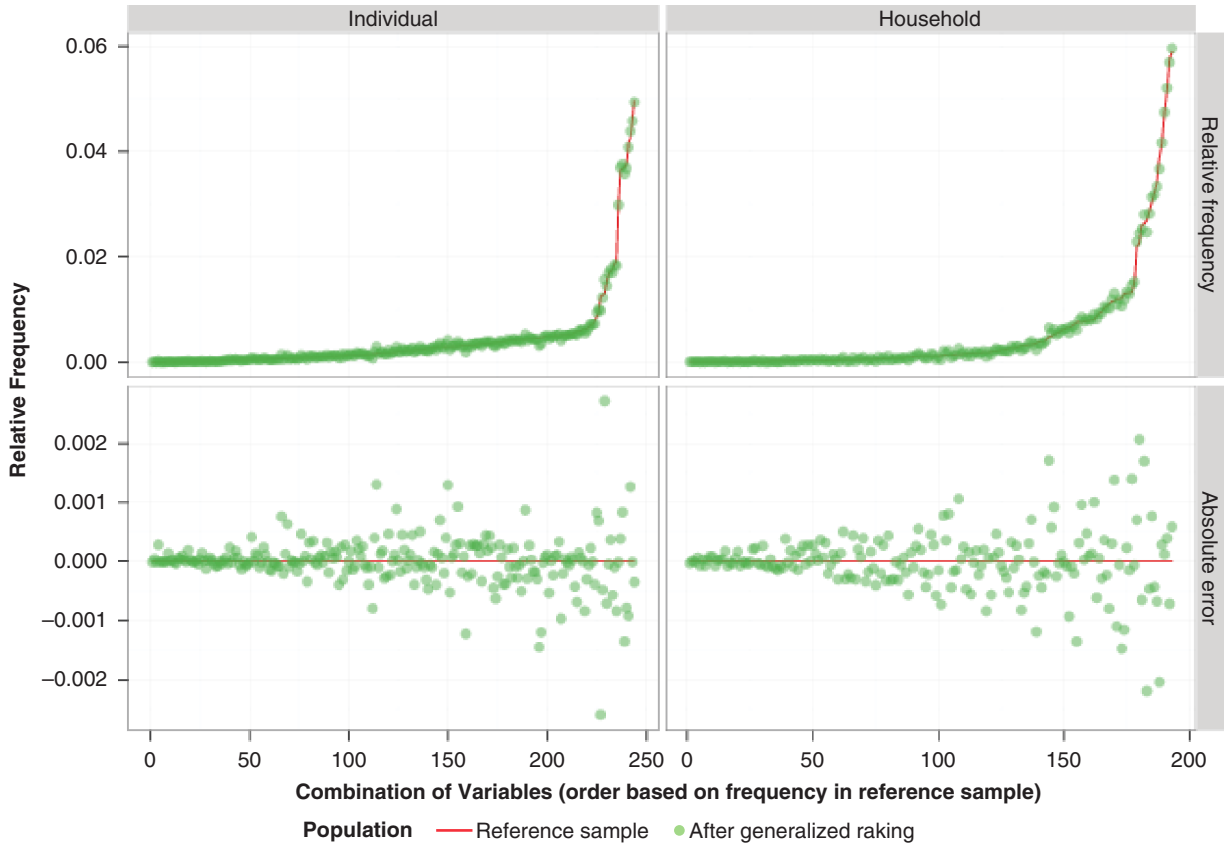


FIGURE 3 Frequency of combinations of variables.

TABLE 3 Most Recurrent Individuals and Households in RS

Attribute			Frequency (%)		
Sex	Age (years)	Income	RS	MC	GR
NA	4	No income	4.97	2.92	4.93
F	65+	No income	4.58	4.88	4.57
M	9	No income	4.26	4.94	4.38
M	14	No income	4.17	4.25	4.07
M	19	No income	3.76	3.52	3.68

Tables 3 and 4 report the five most recurrent individuals and households in RS, with their corresponding frequencies in the MCMC and GR synthetic populations. The failure of hMCMC to reasonably generate infants because of the flawed performance of Gibbs sampling with multipolarized distributions is obvious; in fact, these agents almost constitute a separate subsample of RS, given their (lack of) characteristics. However, the results are almost perfect after the application of GR, which fixed the marginals of sex and age.

To summarize Figure 3, standard root mean square errors will be provided. This error measure is typically used in the literature on transportation modeling and is defined by

$$\mathrm{SRMSE} = \sqrt{\sum_{m_1=1}^{M_1} \cdots \sum_{m_N=1}^{M_N} \left(f_{m_{1,\ldots,N}} - \hat{f}_{m_{1,\ldots,N}} \right)^2 (M_1 \ldots M_N)} \tag{4}$$

In this equation, $f_{m_{1,\ldots,N}}$ and $\hat{f}_{m_{1,\ldots,N}}$ are the relative frequencies of a combination of attributes that appear in the reference and synthetic population, with N the number of attributes. $M_1, \ldots, M_N$ is the product of the number of categories each attribute can take, that is, the total of possible distinct agents.

On the individual level, the SRMSE of MCMC was 0.411, while with GR it became 0.122. The decrease is evident, in particular because GR fixed the frequencies of the infants. On the household level the improvement offered by GR is only minor: 0.117 after hMCMC and 0.116 after GR.

Finally, marginals are reported in Table 5. MCMC does not guarantee to respect the marginals of the reference sample, and although the MCMC marginals are still quite close to them, it is GR that offers an almost perfect match (recall that the marginals of the reference sample are used as control totals for GR).

TABLE 4 Most Recurrent Chinese Individuals and Households in RS

Attribute			Frequency (%)		
Ethnicity	Dwell	Numpax	RS	MC	GR
Chinese	HDB 4	4	5.89	5.92	5.95
Chinese	HDB 5	4	5.76	5.66	5.69
Chinese	HDB 4	3	5.16	5.19	5.20
Chinese	HDB 5	3	4.73	4.68	4.74
Chinese	HDB 3	2	4.14	4.14	4.16

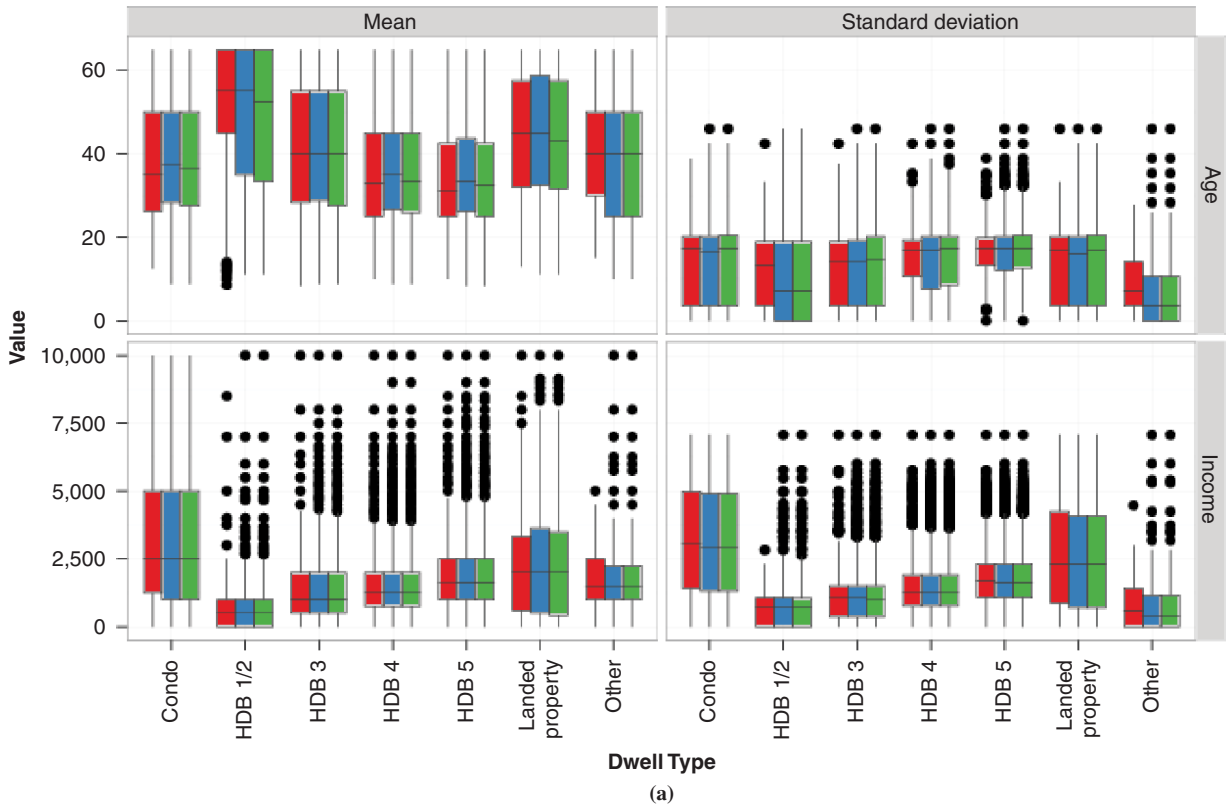
TABLE 5 Table of Marginals

		Frequency (%)		
Attribute	Category	RS	MC	GR
Sex	F	48.68	49.49	48.68
	M	46.36	47.57	46.36
	NA	4.97	2.94	4.97
Age (years)	Age 4	4.97	2.95	4.96
	Age 9	7.96	8.87	7.96
	Age 14	7.83	8.12	7.83
	Age 19	7.85	7.49	7.85
	Age 24	5.48	5.38	5.48
	Age 29	6.53	6.61	6.53
	Age 34	7.45	7.63	7.45
	Age 39	8.01	8.14	8.01
	Age 44	9.02	8.98	9.02
	Age 49	8.17	8.25	8.17
	Age 54	7.69	7.73	7.69
	Age 59	5.61	5.76	5.61
	Age 64	4.70	4.75	4.70
	Age 65+	8.73	9.33	8.73
Income	No income (SGD)	53.37	52.79	53.37
	Max. 1,000	4.99	4.96	4.99
	Max. 1,500	6.16	6.29	6.16
	Max. 2,000	6.82	6.96	6.82
	Max. 2,500	7.62	7.64	7.62
	Max. 3,000	4.40	4.50	4.40
	Max. 4,000	6.53	6.55	6.53
	Max. 5,000	3.54	3.62	3.54
	Max. 6,000	2.05	2.10	2.05
	Max. 7,000	0.90	0.94	0.90
	Max. 8,000	0.61	0.61	0.61
	Over 8,000	3.02	3.02	3.02
Ethnicity	Chinese	71.68	71.70	71.68
	Indian	12.66	12.60	12.66
	Malay	12.34	12.42	12.34
	Other	3.32	3.27	3.32
Dwell	Condo	13.28	13.31	13.28
	HDB 1/2	4.45	4.58	4.46
	HDB 3	19.47	19.39	19.47
	HDB 4	29.31	29.35	29.32
	HDB 5	25.84	25.76	25.84
	Landed property	6.67	6.62	6.67
	Other	0.97	1.00	0.97
Numpax	1	10.55	10.74	10.60
	2	20.23	20.30	20.23
	3	23.52	23.31	23.52
	4	26.02	25.98	26.02
	5	13.22	13.42	13.21
	6	4.53	4.28	4.51
	7	1.32	1.26	1.29
	8	0.44	0.42	0.41
	9	0.10	0.14	0.11
	10	0.03	0.08	0.05
	11	0.02	0.07	0.04

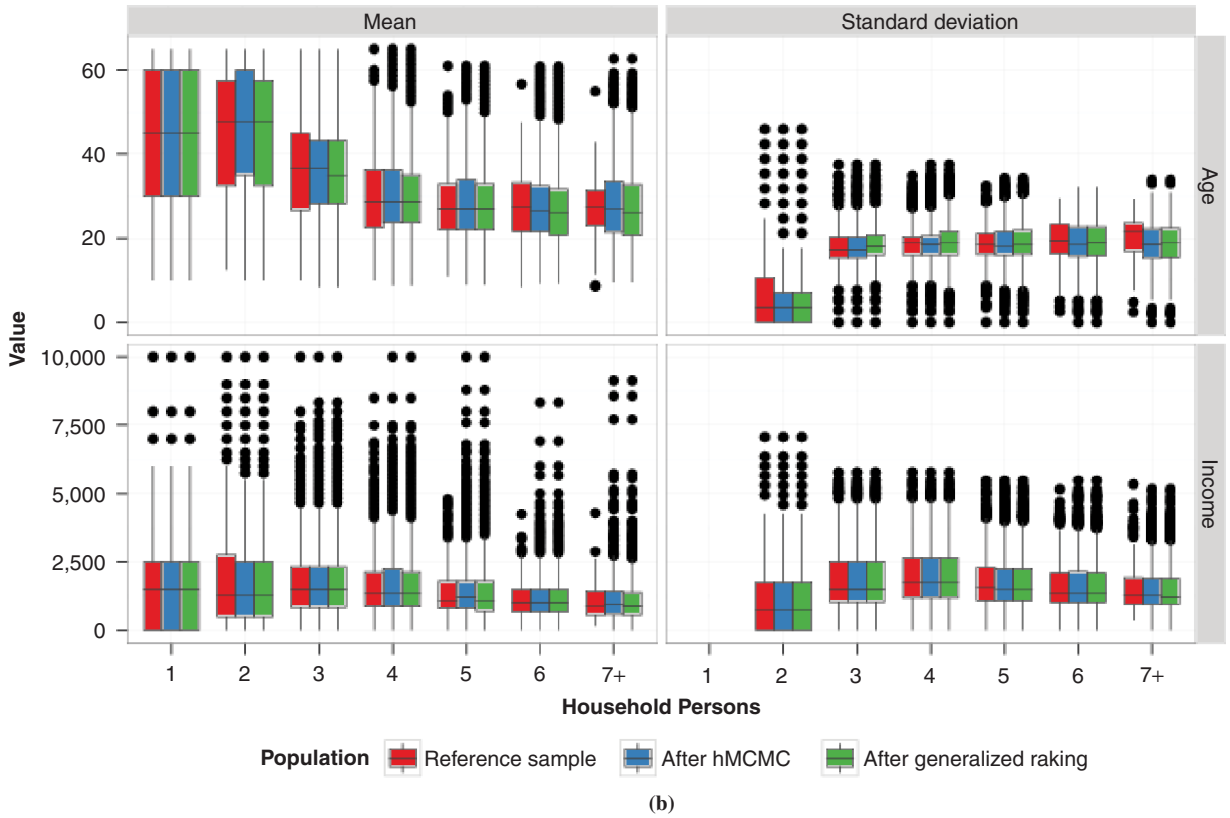
Household Level

The plots and tables of this section actually consider the household composition, thus analyzing the goodness of the generated hierarchical structure.

Figure 4 shows box plots of the distributions of per-household means and standard deviations of age and inc. For computing these values, the categories were transformed to their numeric equivalents. The distributions are analyzed for households with different dwell (top) and numpax (bottom).



(a)



(b)

FIGURE 4 Distribution of intrahousehold indicators.

TABLE 6 Most Recurrent Couples of Owners and Spouses in RS

Attributes of Owner and Spouse						Frequency (%)		
Sex	Sex of Spouse	Age	Age of Spouse	Income	Income of Spouse	RS	MC	GR
M	F	Age 65+	Age 65+	No income	No income	2.44	1.68	1.62
M	F	Age 65+	Age 64	No income	No income	0.66	0.54	0.53
M	F	Age 44	Age 44	Max. 2,500	No income	0.57	0.44	0.45
M	F	Age 44	Age 44	Max. 4,000	No income	0.45	0.36	0.37
M	F	Age 49	Age 49	Over 8,000	No income	0.41	0.37	0.37
M	M	Age 44	Age 44	Over 8,000	No income	0.40	0.35	0.35
M	F	Age 49	Age 49	Max. 2,000	No income	0.39	0.32	0.32
M	F	Age 54	Age 49	Max. 2,500	No income	0.39	0.32	0.33
M	F	Age 54	Age 54	Max. 4,000	No income	0.38	0.29	0.30
M	F	Age 49	Age 49	Max. 4,000	No income	0.37	0.32	0.33

It can be seen that outliers in the age and inc distributions tend to be more present in the synthetic populations. This outcome is expected because of their larger sizes and the heterogeneity of MCMC, which permits the generation of combinations absent in RS. At the same time, there are no remarkable differences between MCMC and GR because the latter fixed mainly the marginals of variables such as sex, which does not modify the age and inc distributions per household. Besides, the box plots with numpax along the *x*-axis display a better concurrence with RS at lower values, which can be explained by the definition of the conditionals of others and, more generally, the implicit assumptions made about the standard household composition through the definition of agent types.

The 10 most recurrent couples of owners and spouses in RS are illustrated in Table 6 to prove the validity of the generated matches; their frequencies in MCMC and GR are in fact quite close to those in RS. However, it seems that GR is not able to make the MCMC frequencies closer to RS, at least not by using only univariate marginals as control totals.

Distribution of Weights

To assess the performance of the method in a more realistic scenario, the hMCMC population has been calibrated also against countrywide controls obtained from the 2010 Singapore Census of Population (*17*). Here, the following controls were used:

- Ethnicity (person level),
- Age × sex × dwell (person level), and
- Dwell (household level).

The GR weights specify the expected number of copies of an item so that the resulting synthetic population satisfies the exogenous controls. An advantage of the proposed method is that the generated samples can be made arbitrarily large, thus leading to lower average weights and more heterogeneous populations. Table 7 shows the weights for the different populations; here, GR-marginals correspond to calibration against the aforementioned control totals. No weighting has been performed for the RS and MCMC population, and the reported values represent the uniform expansion factors that need to be applied to each observation to produce a full population. The ratio between the MCMC and RS expansion factors corresponds to the oversampling ratio. For GR and GR-marginals, most of the weights are reasonably close to the median; however the weights are more extreme for GR-marginals because of the more restrictive control totals. In fact, very few observations are assigned a weight of zero, but only about 0.2% of observations have less than unit weight. The maximum weight for GR-marginals is less than the RS expansion factor, confirming that the population is more diverse, that is, any given household will be repeated only up to an expected 92.5 times, compared with the best-case scenario of uniform weights for RS, in which households have to be repeated more than a hundred times.

CONCLUSIONS

The results demonstrate that hMCMC is able to solve the problem of generating new populations with a hierarchical structure and GR can be applied for postprocessing these hMCMC populations to impose selected control totals on them. In this way, it complements the inability of MCMC to set control totals, and it can also help to overcome the flaws of MCMC when one is dealing with multipolarized data sets, as it was seen with the infants of the demographic sample explored here.

Exactly because of its lack of constraints, MCMC allows heterogeneous new populations not suffering from the IPF zero-cell problem. However, this method should not be applied blindly as illustrated by the problems experienced with occupation and education (compare with the section on data description). It is still possible to later impute additional attributes or categories that needed to be dropped because they contributed to data multipolarization in a postprocessing operation. Nevertheless, most likely such imputation would be performed individually for each attribute, and hence one could not control for potential correlation structures between these additional attributes.

TABLE 7 Quantiles of Distribution of Weights

Step	0%	25%	50%	75%	100%
RS	na	na	106.2	na	na
MC	na	na	10.6	na	na
GR	3.0	9.2	10.0	10.7	54.9
GR-marginals	0.0	7.6	10.2	12.7	92.5

The presented case study is somewhat simplistic as it was fit only against control totals from an independent sample, in this case the population census, and the structure of the generated microsample was not verified against it. Ideally, results generated with the proposed method would also be assessed for heterogeneity and correlation structure with an independent microsample. For the case of Switzerland, the census even features a full microsample, which would be ideal for such analysis, especially to test the behavior of postimputed attributes.

Further lines of research can deal with alternative extensions of MCMC handling hierarchies. A simple idea may be to implement iMCMC with type-based variables, such as age_{owner}, so as to generate households with their fully characterized populations at once. However, the curse of dimensionality can easily break this algorithm, and one may ultimately be handling only a few possible combinations of attributes to obtain acceptable results.

Another possible development stems directly from the developed hMCMC method, in which the generation of agent types living together always follows a certain order: households with an owner and a child but no spouse are not considered in this work since the agent types have to be defined a posteriori. Hence, if one would have predefined types in the reference sample, it would be possible to make the generation of the other agents after owner not simply dependent on numpax but on some other variables, such as "has spouse" or "no. of children," which could easily be deduced from this kind of data set. In that case, the first agent to be generated per household could actually be characterized only by household variables.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from the Swiss National Science Fund, as well as the Singapore Land Transport Authority for making the data available.

REFERENCES

1. Müller, K., and K.W. Axhausen. Population Synthesis for Microsimulation: State of the Art. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., 2011.
2. Deming, W.E., and F.F. Stephan. On the Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known. *Annals of Mathematical Statistics,* Vol. 11, No. 4, 1940, pp. 427–444.
3. Ye, X., K.C. Konduri, R.M. Pendyala, B. Sana, and P. Waddell. Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations. Presented at 88th Annual Meeting of the Transportation Research Board, Washington, D.C., 2009.
4. Müller, K., and K.W. Axhausen. Hierarchical IPF: Generating a Synthetic Population for Switzerland. Presented at 51st Congress of the European Regional Science Association, Barcelona, Spain, Sept. 2011.
5. Bar-Gera, H., K.C. Konduri, B. Sana, X. Ye, and R.M. Pendyala. Estimating Survey Weights with Multiple Constraints Using Entropy Optimization Methods. Presented at 88th Annual Meeting of the Transportation Research Board, Washington, D.C., 2009.
6. Lee, D.-H., and Y. Fu. Cross-Entropy Optimization Model for Population Synthesis in Activity-Based Microsimulation Models. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2255,* Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 20–27.
7. Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flötteröd. Simulation Based Population Synthesis. *Transportation Research Part B: Methodological,* Vol. 58, 2013, pp. 243–263.
8. Deville, J.-C., C.-E. Särndal, and O. Sautory. Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association,* Vol. 88, No. 423, 1993, pp. 1013–1020.
9. Pritchard, D.R., and E.J. Miller. Advances in Population Synthesis: Fitting Many Attributes per Agent and Fitting to Household and Person Margins Simultaneously. *Transportation,* Vol. 39, No. 3, 2012, pp. 685–704.
10. Auld, J., and A.K. Mohammadian. An Efficient Methodology for Generating Synthetic Populations with Multiple Control Levels. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2175*, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 138–147.
11. Guo, J.Y., and C.R. Bhat. Population Synthesis for Microsimulating Travel Behavior. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2014,* Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 92–101.
12. Müller, K., and K.W. Axhausen. Multi-Level Fitting Algorithms for Population Synthesis. Presented at 1st European Symposium on Quantitative Methods in Transportation Systems, Lausanne, Switzerland, Sept. 2012.
13. Greenacre, M.J., and J. Blasius (eds.). *Multiple Correspondence Analysis and Related Methods.* Statistics in the Social and Behavioral Sciences Series, Chapman & Hall/CRC, Boca Raton, Fla., 2006.
14. Kohonen, T. Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics,* Vol. 43, No. 1, 1982, pp. 59–69.
15. Müller, K., and K.W. Axhausen. Using Survey Calibration and Statistical Matching to Reweight and Distribute Activity Schedules. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2429*, Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 157–167.
16. Müller, K. MultiLevelIPF. 2014. https://www.github.com/krlmlr/MultiLevelIPF.
17. Statistics Singapore. *Census of Population.* 2014. http://www.singstat.gov.sg/publications/population.html#census_of_population.

The Standing Committee on Transportation Demand Forecasting peer-reviewed this paper.