Transportation Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Synthetic Population Generation Without a Sample

Johan Barthelemy, Philippe L. Toint,

To cite this article:
Johan Barthelemy, Philippe L. Toint, (2013) Synthetic Population Generation Without a Sample. Transportation Science 47(2):266-279. https://doi.org/10.1287/trsc.1120.0408

TRANSPORTATION SCIENCE
Vol. 47, No. 2, May 2013, pp. 266–279
ISSN 0041-1655 (print) | ISSN 1526-5447 (online)

http://dx.doi.org/10.1287/trsc.1120.0408

Synthetic Population Generation Without a Sample

Johan Barthelemy, Philippe L. Toint
Namur Research Center for Complex Systems (NAXYS), FUNDP-University of Namur, B-5000 Namur, Belgium
{johan.barthelemy@fundp.ac.be, philippe.toint@fundp.ac.be}

The advent of microsimulation in the transportation sector has created the need for extensive disaggregated data concerning the population whose behavior is modeled. Because of the cost of collecting this data and the existing privacy regulations, this need is often met by the creation of a synthetic population on the basis of aggregate data. Although several techniques for generating such a population are known, they suffer from a number of limitations. The first is the need for a sample of the population for which fully disaggregated data must be collected, although such samples may not exist or may not be financially feasible. The second limiting assumption is that the aggregate data used must be consistent, a situation that is most unusual because these data often come from different sources and are collected, possibly at different moments, using different protocols. The paper presents a new synthetic population generator in the class of the Synthetic Reconstruction methods, whose objective is to obviate these limitations. It proceeds in three main successive steps: generation of individuals, generation of household type's joint distributions, and generation of households by gathering individuals. The main idea in these generation steps is to use data at the most disaggregated level possible to define joint distributions, from which individuals and households are randomly drawn. The method also makes explicit use of both continuous and discrete optimization and uses the χ^2 metric to estimate distances between estimated and generated distributions. The new generator is applied for constructing a synthetic population of approximately 10,000,000 individuals and 4,350,000 households localized in the 589 municipalities of Belgium. The statistical quality of the generated population is discussed using criteria extracted from the literature, and it is shown that the new population generator produces excellent results.

Key words: synthetic population; microsimulation; limitations of iterative proportional fitting based procedures; sample-free generator; nonexisting sample
History: Received: December 2010; revision received: June 2011; accepted: October 2011. Published online in *Articles in Advance* April 5, 2012.

1. Introduction

Synthetic population generation has recently received considerable attention in the literature (Müller and Axhausen 2011 present a good overview of the techniques available in 2011). It is often motivated by the observation that microsimulations, such as activity-based travel demand models in transport, usually involve a large number of agents and that it may be impossible or too expensive to obtain a fully disaggregated data set describing the agents of interest. Moreover, if such a data set were available, its use may also be problematic in some countries because of stringent privacy laws. A way to address these issues is to construct an artificial population starting from known data about the true one. Because it is obvious that the representativeness of the synthetic population is critical for the simulation accuracy, a synthetic population generator should therefore produce a population approximating the correlation structure of the true population as accurately as possible.

Techniques for synthetic population generation typically belong to either the Synthetic Reconstruction (SR) techniques or the Combinatorial Optimization (CO) methods. The SR methods generate a synthetic population given joint distributions of the population's attributes, generally using a sample of the population and the iterative proportional fitting procedure (IPFP) to generate the desired joint distributions (see Wilson and Pownall 1976; Beckman, Baggerly, and McKay 1996). The CO category is far less common. The CO methods divide the area of interest in mutually exclusive zones for which a set of marginal distributions of the desired attributes is available. Then a subset of a sample taken over the whole population is fitted to the given set of margins for each zones. We refer the reader to Voas and Williamson (2001) and Huang and Williamson (2002) for a formal and complete description of the latter methods.

However, both SR and CO approaches usually make strong assumptions on the data used in the process, and it is not always possible to ensure that they can be satisfied in practice. In particular, this caused significant difficulties in the generation of a synthetic population for Belgium. These difficulties motivate the research presented here, where a new type of SR generator is developed, obviating these data-related issues.

The remainder of this paper is organized as follows. In §2, we first present the standard approach for building a synthetic population, from which the other Synthetic Reconstruction techniques are derived. Section 3 then describes an alternative method, belonging to the SR family, obviating the limitations of the conventional generation methods. We next present in §4 the results of this new methodology applied to the generation of a synthetic population for Belgium. Section 5 then compares the new generator with an IPFP-based methodology. Concluding remarks are discussed in §6.

2. The Standard Approach

To date, the standard approach for building synthetic populations is based on the method developed by Beckman, Baggerly, and McKay (1996), whose main idea consists in merging aggregate data from a source covering the whole population with disaggregated data from a sample in order to get a disaggregated data set for the population of interest. Typically the aggregate data set is extracted from an existing census and the disaggregated data set is drawn from a survey over a sample of the population. The aggregate data consist of a set of marginal distributions for the characteristics of interest of the true population: we refer to these distributions and variables as target and control variables. The disaggregated data provide full information about the attributes of interest, but only for a sample of agents, and are referred to as the seed.

The population synthesis procedure usually starts with identifying the relevant (categorical) socio-demographic variables of the agents. Assuming that there are n attributes of interest in the seed and denoting by $V = \{v_1, v_2, \ldots, v_n\}$ the vector of variables representing these attributes, each combination of values of v_i therefore defines a socio-demographic group. The synthetic population is then generated by a two-step procedure:

1. Starting from the seed, estimate the k-way joint distribution of the true population, where $k \leq n$ is the number of control variables, such that the resulting distribution is consistent with the marginal distributions (margins) of the target and preserves the correlation structure of the seed.

2. Select agents from the sample and copy them to the synthetic population in a proportion derived from the distribution computed in the previous step.

These steps are discussed in the next two subsections, followed by a description of the limitations of this first approach and the proposed improvements obviating these limitations.

2.1. Estimating the Attributes' Joint Distribution Using IPFP

The most popular way to estimate a k-way joint distribution table based on known marginal distributions and a sample is the well-known iterative proportional fitting procedure (IPFP) originally described by Deming and Stephan (1940). This procedure is detailed below for $k = 2$ but can easily be extended to higher dimensions.

Assume that a two-way contingency table is built from the seed with initial components $\pi_{ij} \in \mathbb{R}^+$ where i and j, respectively, correspond to the level of the first and the second variable. These π_{ij} correspond to the number of agents in the sample for each combination of levels. Assume also that desired marginal distributions $\{x_{i\bullet}, x_{\bullet j}\}$ (the target) are known $\forall\, i, j$. The IPFP then iteratively updates the cells' values depending on the marginal distributions of the target until the margins of the computed table match the target's ones; i.e., $\pi^*_{i\bullet} = x_{i\bullet}$ and $\pi^*_{\bullet j} = x_{\bullet j}$ where the π^*_{ij} are the component values at the last iteration. The adjustments at iteration l are computed by the equations

$$\pi^{l'}_{ij} = \pi^{l-1}_{ij} \cdot \frac{x_{\bullet j}}{\pi^{l-1}_{\bullet j}} \quad \forall\, i, j; \tag{1}$$

$$\pi^{l}_{ij} = \pi^{l'}_{ij} \cdot \frac{x_{i\bullet}}{\pi^{l'}_{i\bullet}} \quad \forall\, i, j. \tag{2}$$

To produce an accurate estimate of the true distribution, the procedure ideally requires an initial representative sample of the true population for building the initial multiway table (even if, technically, a multiway table of ones can be used as a starting point of the procedure). This requirement is important because Mosteller (1968) pointed out that the procedure preserves the interaction structure of the sample as defined by the odd ratios

$$\frac{\pi_{ij} \cdot \pi_{hk}}{\pi_{ik} \cdot \pi_{hj}} = \frac{\pi^l_{ij} \cdot \pi^l_{hk}}{\pi^l_{ik} \cdot \pi^l_{hj}} \tag{3}$$

at each iteration l, where $i \neq h$ and $j \neq k$ stand for different levels for each variable respectively. Moreover, according to Ireland and Kullback (1968), the IPFP also produces the π^*_{ij} minimizing the discrimination information, also known as the relative entropy or the Kullback-Leibler divergence, defined by

$$\sum_i \sum_j \pi^*_{ij} \ln\left(\frac{\pi^*_{ij}}{\pi_{ij}}\right). \tag{4}$$

Finally, Little and Wu (1991, p. 94) showed that IPFP results in a maximum likelihood estimator for the RAKE model, which was judged "the best overall choice . . . in the absence of knowledge of the form of the selection model [for fitting to known marginals]."

2.2. Generating the Synthetic Population

Once the expected numbers of agents in every socio-demographic group are estimated, each sampled agent is associated with a probability of being included

in the synthetic population. This probability typically depends on the agent's sampling weight and the expected number of similar agents in the true population. Based on these probabilities, the approach of Beckman, Baggerly, and McKay (1996) randomly draws agents from the sample using a Monte Carlo procedure until the expected number of agents is reached for each socio-demographic group. When a sampled agent is drawn, then all its attributes, including the uncontrolled ones, are pasted to a new synthetic agent who is added to the synthetic population.

2.3. Limitations and Improvements of the Approach

Recent mobility surveys such as EGT (Direction Régionale de l'Équipement d'Île-de-France 2004), MOBEL (Hubert and Toint 2002), or NTS (Avery 2011) suggest that the travel behavior of an individual is significantly influenced by the type and composition of his/her household. This points to a first limitation of the conventional approach: it is very unlikely that analysts have access to a single data set detailing the joint distribution of individuals' and households' attributes simultaneously. Because the estimation step of the algorithm described in §2.1 is designed to deal with a single contingency table, the conventional approach can consequently account either for individual-level or for household-level control variables but not for both. In other words this process results in a synthetic population where either the households' or individuals' joint distributions match the desired ones, but both cannot. Note that households' distributions accuracy has often been preferred (see Ye et al. 2009).

This strong limitation led several authors to propose interesting improvements to this basic algorithm. Guo and Bhat (2007) designed a method to overcome this problem by simultaneously controlling the individual- and household-level variables. Their algorithm generates a population where the household-level distributions are close to those estimated using the IPFP, while simultaneously improving the fit of person-level distributions. Arentze, Timmermans, and Hofman (2007) propose another method using relation matrices to convert distributions of individuals to distributions of households, such that marginal distributions can be controlled at the person level as well. Ye et al. (2009) further build on these contributions and suggest a practical heuristic approach called iterative proportional updating (IPU), based on adjusting households' weights such that both household- and individual-level distributions can be matched as closely as possible. Control for households' and individuals' relationship, improvements, or alternatives to the standard approaches are also investigated in Auld, Mohammadian, and Wies (2010), Pritchard and Miller (2009), and Srinivansan, Ma, and Yathindra (2008).

However, these improved approaches remain based on the IPFP (or the IPU) and thus rely on the same assumptions on data quality, i.e., that the aggregate data of the target is consistent in the sense that margins extracted from available but different joint distributions are equal. This is critical for practical convergence of IPFP. They also assume that a significant sample of the population of interest is available at the desired level of disaggregation, from which synthetic agents can be extracted an duplicated. For example if a class of agents is not represented in the seed, then this particular class will remain unpopulated in the final synthetic population. This could also be cured by introducing small initial values in the unpopulated classes, but this approach remains unsatisfactory because it introduces unwanted bias.

These two strong requirements unfortunately limit the applicability of the IPFP in real situations, such as the generation of a synthetic population for Belgium at the municipality level. Indeed these requirements could not be met in this particular case. First, a representative sample at the municipality level (which is the desired spatial disaggregation level) is not available, and, even if it were, the privacy issue would remain because the IPFP repeats the observations of the sample as many times as necessary. A second problem is that all necessary information, i.e., distributions, is not available from a single source (which would hopefully guarantee consistency) but has to be extracted from different data sets, typically produced by different institutions and/or using different protocols or data cleaning mechanisms. This results in significant differences between margins, as illustrated in Table 1 (extracted from Cornélis, Legrain, and Toint 2005) for the Charleroi district. Data were provided by Groupe d'etrude de démographie appliquée (GéDAP) and the Directorate-general Statistics and Economic Information of the Belgain Federal Government (INS).

In this table, the total number of inhabitants in the district is compared among the data files used in the population synthesis. If one takes (for instance) the first data set as a reference, one immediately notices inconsistencies between the different estimations, with differences up to 12% irrespective of the

Table 1 Inconsistencies Between Margins Extracted from Different Sources

Joint distribution	Data source	Margins	Prop.
Municipality × gender × age	GéDAP, 2001	405.491	1.00
Municipality × household type	GéDAP, 2001	380.653	0.94
Municipality × education level	GéDAP, 2001	426.372	1.05
Municipality × activity status	GéDAP, 2001	396.594	0.97
District × household type × age	INS, 2001	357.884	0.88
District × education level	INS, 2001	398.582	0.98

data source (see last column of Table 1). These inconsistencies prevent the IPFP process from converging. This could possibly be cured by considering the frequencies rather than the number of agents themselves, but the issue of the missing sample nevertheless remains. These difficulties motivate our proposal for an alternative population synthesis tool that would not suffer from the lack of a representative sample at the most disaggregated level and/or from (moderate) inconsistencies between different data sources. This is the object of §3.

3. A New Population Synthesis Technique

We start the presentation of our proposal by outlining its main steps before the more formal description.

Our general philosophy is to construct individuals and households by *drawing their characteristics or members at random within the relevant distribution at the most disaggregated level available, while maintaining known correlations as well as possible*. The algorithm implementing this principle, and illustrated in Figure 1, consists of a three-step procedure for each spatial aggregation unit:

1. a pool of individuals is generated, which we denote by *Ind*;

2. the households' joint distribution is estimated and stored in the contingency table *Hh*; and

3. the synthetic households are constructed by randomly drawing individuals from the individuals' pool *Ind*. This is achieved while preserving the distribution computed in the second step. Once a household has been built, it is added to the synthetic population.

We now provide detailed information on each of these successive steps.

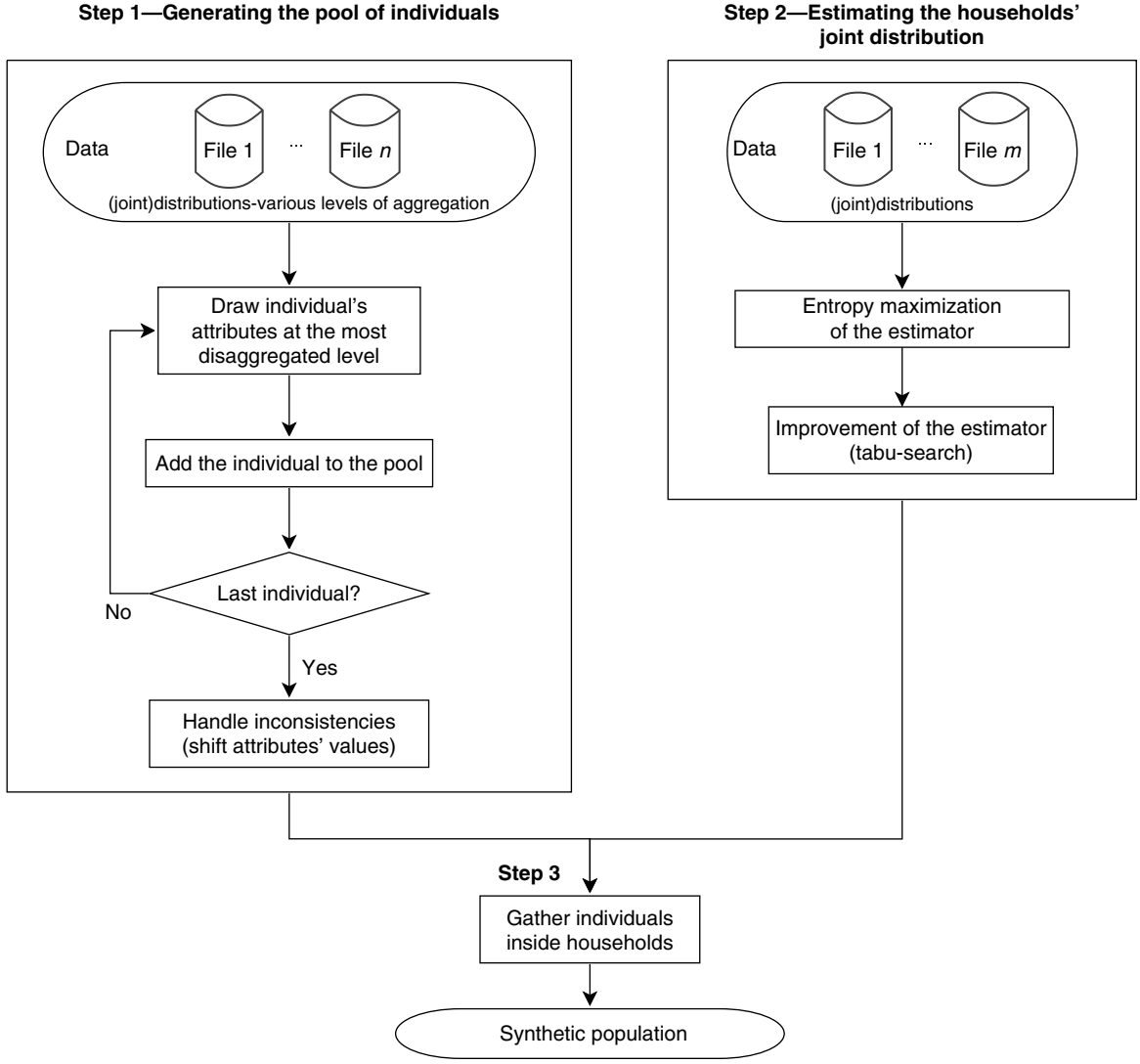

Figure 1 Synthetic Population Generator

3.1. Step 1: Generating the Pool of Individuals

The first step aims at building the *Ind* pool of synthetic individuals for the area of interest by generating them one by one. In our method, each individual is characterized by a vector of attributes $V = (V_1, \ldots, V_n)$ whose components may take a discrete set of values. We denote by v_i the value taken by the characteristic V_i for a particular individual. We would like to draw each v_i from known empirical distributions derived from available data. However, not every distribution for V_i is known at the most disaggregated level, and we thus face a hierarchy of levels. Our first step is then to merge the various distributions available at the same disaggregation levels using the IPFP technique (Frick and Axhausen 2004; Guo and Bhat 2007), possibly substituting less reliable values by their frequencies to handle inconsistent margins. This results in a set of distributions V^k, where k denotes the level of disaggregation (in our case, municipality, district, nation). In accordance to the general principle stated above, our idea is then to use, for each such characteristic, the most disaggregated level available.

Specifically, a table V^0 corresponding to the numbers of individuals with the attributes $(v_1^0, \ldots, v_{n_0}^0)$ is first constructed from the most disaggregated data available (at municipality level in our case). The missing attributes for each individual in this table are then determined by finding the most disaggregated level at which a joint distribution for the missing attribute and some already known characteristic of the considered individual is available. The first of these is then determined by a random draw in this (conditional) distribution. Once all characteristics of an individual are defined, the pool *Ind* is updated.

Because some of the individuals' characteristics are determined by draws from distributions at aggregate levels, the margins extracted from the pool *Ind* for these particular characteristics may be inconsistent with the known true ones. For each attribute, a correction is then made to the agents of *Ind* to make it consistent with their respective margins at the level zero. This correction is computed by suitably shifting the attribute's value of certain number of individuals, determined by the number of individuals with problematic attribute's values and the known distribution of this attribute. Only shifts between two contiguous modality are allowed. For instance consider an attribute U whose modalities are u_1, u_2, u_3, and u_4, where

$$u_1 \le u_2 \le u_3 \le u_4. \tag{5}$$

If an individual is initially characterized by u_2, then the shift is either u_1 or u_3. Note that these shifts can only be applied to numerical or ordinal variables.

3.2. Step 2: Estimating the Households' Joint Distribution

We now consider the second step of our population synthesis procedure. Denote by $W = (W_1, \ldots, W_m)$ the vector of household-related attributes and by w_j the value taken by a particular household for the jth such attribute. Now that a pool of individuals has been built, the next step is to find an estimator of the households' type contingency table, denoted by Hh, given data provided by several different sources. Each cell of Hh thus corresponds to the number of particular household of a type specified by a combination of the w_js (which we call a household type). This problem is solved in two steps: a maximum entropy estimate of Hh is first generated and subsequently improved by using a tabu-search optimization process.

3.2.1. Entropy Maximization of the Estimator. In our algorithm, the initial estimation of Hh is obtained as the solution of an optimization problem, where the entropy is maximized under the (linear) constraints implied by the known margins on households' types. This approach has the advantages of producing a more reasonably spread-out distribution among all types while keeping the constraints satisfied than would be produced by a least-squares formulation, say. The entropy maximization approach is introduced here in an intuitive way, inspired by Bierlaire (1991) and Ortúzar and Willumsen (2001). For a more formal description, see Wilson (1974).

Consider a system consisting of a large number of distinct elements. A full description of such a system requires the complete specification of each micro-state of the system, which involves in our case completely identifying each household. At this stage, however, we are interested in a more aggregated level called the meso-state, corresponding to the households' distribution Hh. Typically one meso-state can be associated with different micro-states. For instance if two household heads with similar attributes are exchanged, then the meso-state is unchanged but the associated micro-states are different. Finally, the last and highest level of aggregation, called the macro-state, is the available data on the system as a whole.

The basic idea of the method is to accept that unless we have information on the contrary, all micro-states consistent with the macro-state are equally likely. This consistency is enforced, in our approach, by imposing equality constraints given by the macro-state. If $x = (x_1, \ldots, x_p)$ is the vector of unknown cells of Hh, Wilson (1970) showed that the number of micro-states $E(Hh)$ associated with the meso-state Hh is given by

$$E(Hh) = \frac{(\sum_i x_i)!}{\prod_i x_i!}. \tag{6}$$

The function $E(\cdot)$ is called the entropy function. Because it is assumed that all micro-states are equally likely, the meso-state corresponding to the largest number of micro-states (and thus the most likely) is that maximizing (6). Using the natural logarithm and Stirling's short approximation (Dwight 1961; Kreyszig 1972), the corresponding objective function of this problem can then be approximated by

$$\min_x \sum_i x_i \ln(x_i) - x_i \tag{7}$$

under the constraints on households types given by the macro-state.

Unfortunately, because of the inconsistent nature of the available data, as exposed in §2.3, the constraints of this optimization problem are also formally inconsistent. Our approach is then to impose only a subset Ω of them, corresponding to the data of highest quality as strict constraints, the others then being incorporated in the objective function in a form penalizing their violation. Each of these latter constraints p is affected with a weight defined by $n_p\sigma$ where σ is a penalization parameter and n_p is the number of households involved in p.

Note that the problem constraints are all linear and can therefore be represented in matrix form by the system $Ax = b$, where A contains the coefficients of the variables and b the independent terms. Denoting by A_σ and b_σ the matrix and the vector derived from the subset of the scaled inconsistent constraints, the new objective function can now be formulated as

$$\text{(EN)} \quad \min_x \|A_\sigma x - b_\sigma\|_2^2 + \sum_i x_i \ln(x_i) - x_i, \tag{8}$$

and the minimization is then carried out under the constraints in the set Ω only. This optimization problem is solved using an augmented Lagrangian algorithm, as implemented in the (freely available) LANCELOT package (Conn, Gould, and Toint 1992; Gould, Orban, and Toint 2003). In general the solution of this optimization problem yields a noninteger solution, which is unsuitable for representing households' numbers. The solution's components of this optimization problem are then rounded and the value

$$f_{\text{EN}}(\hat{x}) = \sum_i w_i|\hat{c}_i - c_i| + \sum_i \hat{x}_i \ln(\hat{x}_i) - \hat{x}_i \tag{9}$$

is computed, where $\hat{x}$, $\hat{c}_i$, c_i, and w_i denote the rounded solution of (EN), the computed and the desired value of the ith constraint, and the associated weight depending on the quality of the associated data source, respectively. The latter are admittedly somewhat arbitrary: we have chosen to penalize violation of consistent constraints 10 times more than those associated with inconsistent ones, but the results seem relatively insensitive to this choice. The value of f_{EN} can be seen as a performance measure describing how well the rounded integer solution fits the whole set of initial constraints. We then loop over a set of values for the penalization parameter σ, and the best rounded solution x^* associated with the lowest value of f_{EN} is determined. This solution is finally used as the starting point of a combinatorial optimization problem using a tabu-search algorithm in order to get a final estimation of Hh. Details on this process are provided in the next section.

3.2.2. Improvement of the Estimator Using Tabu-Search. Tabu-search is a local-search meta-heuristic originally proposed by Glover (1986), which can be used for solving combinatorial optimization problems. This procedure iteratively moves from one solution x to a solution $x' \in \mathcal{N}(x)$, a neighborhood of x containing a list of candidate solutions, until a stopping criterion (such as a given number of iterations N) has been reached. In order to avoid cycling, the neighborhood $\mathcal{N}(x)$ is modified to exclude some solutions encountered in previous iterations (these solutions constitute the "tabu list"). For a complete description of this optimization technique, we refer the reader to Glover (1989, 1990) and Glover and Laguna (1997).

In this paper, the tabu list is a list T of size n that contains the solutions visited in the last n iterations. If we denote by x^i the candidate solution at iteration $i > 0$, x^0 being x^*, i.e., the rounded solution computed above, $\mathcal{N}(x^i)$ is then defined as follows:

$$\mathcal{N}(x^i) = \left\{ x^i_{j\pm} = (x_1^{i-1}, \ldots, x_j^{i-1} \pm 1, \ldots, x_p^{i-1}) \,|\, j = 1, \ldots, p \right\},$$

where the notation $x_j^{i-1} \pm 1$ stands for two variations of the jth component around its value x_j^{i-1}. The following steps are then executed iteratively N times:

1. define a new candidate by randomly drawing $x^i \in \mathcal{N}(x^{i-1})$ such that $x^i \notin T$;
2. if $f_{\text{EN}}(x^i) < f_{\text{EN}}(x^*)$, then $x^* = x^i$; and
3. replace the oldest component of T by x^i and go back to step 1.

This procedure results in an updated and improved estimate x^* of Hh. Note that the quality of the improvement depends on the size of the tabu list and the number of iterations allowed. These parameters must therefore be chosen to obtain a reasonable tradeoff between computing cost and quality of the estimate. However, the impact of varying these parameters appears to be small in our application.

3.3. Step 3: Households' Generation

Individuals' and households' distributions having been estimated, the last step of our generator consists of gathering individuals into households by randomly drawing households' constituent members. We proceed in two successive stages: the first is to select a household type and the second is to draw the individuals to form a household of this type.

The selection of the household type is performed in order to keep the distribution of already completed

households statistically close to the estimated one. The goal is achieved by choosing the type of the next household to assemble such that the distribution Hh' of the already generated households (including the household being built) minimizes the observed χ^2 distance between the Hh and Hh', which is given by

$$d_{\chi^2}(Hh', Hh) = \sum_i^p \frac{(x_i' - x_i)^2}{x_i^2}.$$

This minimization is extremely simple because the number of household types is very limited. Once the household type is selected, households' members are generated as follows: a household head is first drawn from the pool of individuals *Ind*, and then, depending on the household's type, additional individuals are also drawn from the pool if relevant. All these draws from *Ind* are made without replacement.

We now provide some detail on this last drawing process. If we assume that a household is made up of a head and possibly a mate, children, and additional adults, the construction starts with the selection of its head. Depending on the household type, the head's attributes are either obtained directly (for instance for an isolated man) or randomly drawn according to known joint distributions, e.g.,

- household type × head's gender × head's age;
- household type × municipality type × head's age × head's activity status; and
- household type × municipality type × head's age × head's education level.

More formally, this selection procedure is organized in three steps:

1. Determine the desired attributes' values (i.e., the v_i's) for the household head:
   - some can be derived directly from the current household type; and
   - the remaining missing attributes are either randomly drawn according to known distributions or, if different values are feasible and equally likely for V_i, determined in order to minimize the χ^2 distance between the generated and estimated distributions.
2. Add the head to the household being generated:
   - if the corresponding individual's class is still populated in the individuals' pool, extract an individual from this class and make it the household's head; and
   - else find a suitable household head by random search in the members of the previously generated households. This last individual is then replaced with an appropriate one randomly drawn in the pool of the remaining individuals. If the generator fails to find a head, then the generation is ended.
3. The estimated and generated contingency tables are updated according to the actions performed in step 2.

Depending on the household type, the generator may pursue the construction of the current household by selecting a head's partner, children, and additional adults. The corresponding selection procedures are similar to the head's, with the only exception that individuals' characteristics may no longer be determined by the household type only but are randomly drawn according to known distributions on couple formation such as

- household type × head's gender × head's age × mate's age;
- household type × head's gender × head's education level × mate's education level; or by predefined rules (a child must be younger than his/her parents).

The household generation for the current municipality terminates if all households have been constructed or the generator fails to find a household member, e.g., if the pool of individuals is empty or if it is impossible to find a suitable individual in the previously generated households.

When the procedure stops after exhausting either the pool of individuals or the pool of households, inconsistencies of two types may remain in the generated population: in the first case the final number of households is smaller than anticipated, whereas in the second case the final number of individuals is smaller than estimated.

4. Generating a Belgian Synthetic Population

4.1. The Application and Data Sources

The procedure outlined in the previous section has been used to generate a synthetic population of 10,637,107 individuals gathered in 4,334,281 households for the 589 municipalities of Belgium in 2001. The municipalities (NUTS-5 level) themselves belong to 43 districts (NUTS-3 level) containing between 2 and 35 municipalities each. Table 2 presents basic statistics on these municipalities. The individuals' and households' attributes are respectively described in Tables 3 and 4. Data available at the municipality or

Table 2 Basic Statistics for Municipalities

	Min	Max	Mean
Individuals	85	461,115	18,059.6
Households	35	212,707	7,358.9

Table 3 Individuals' Characteristics

Attribute	Values
Gender	Male; female
Age class	0–5; 6–17; 19–39; 40–59; 60+
Activity	Student; active; inactive
Education level	Primary; high school; higher education; none
Driving license ownership	Yes; no

Table 4 Households' Characteristics

Attribute	Values
Type	Single man alone
	Single woman alone
	Single man with children (and other adults)
	Single woman with children (and other adults)
	Couple without children (and other adults)
	Couple with children (and other adults)
Number of children	0 to 5
Number of other adults	0 to 2 (mate not included)

district aggregation levels is provided from the following sources:

- *Directorate-general statistics and economic information* of the Belgian Federal Government (2001);
- *Service public fédéral Mobilité et Transports* of the Belgian Federal Government (2000);
- *Groupe d'étude de démographie appliquée* (GéDAP) centre of the University of Louvain-la-Neuve (Belgium) (2001); and
- the *MOBEL* mobility survey (Hubert and Toint 2002).

The synthetic population generator has been implemented in a single threaded program written in the Perl 5.10 programming language and executed on a desktop computer running with a 3 GHz CPU and 3 GB of RAM under a 32-bit Linux environment. The generation process uses a TABU list of length 1,000 and takes approximately 16 hours and 30 minutes to handle all the 589 municipalities.

4.2. Verification of the Household Generation Procedure

4.2.1. Absolute Percentage Difference. Having generated a synthetic population, one is then faced with the question of estimating its quality. As in Guo and Bhat (2007), one possible performance measure to assess the generator accuracy is the absolute percentage difference (*APD*) between the estimated contingency tables computed in the first steps (steps 1 and 2) of the generator and the corresponding ones resulting from the household generation step (step 3). This measure is calculated for a particular cell $(u_1, \ldots, u_p)$ as follows:

$$APD_{T,T'}(u_1, \ldots, u_p) = \left| \frac{T'[u_1] \cdots [u_p] - T[u_1] \cdots [u_p]}{T[u_1] \cdots [u_p]} \right|$$

where T and T' denote, respectively, the estimated (steps 1 and 2) and the generated (step 3) tables. The lower the *APD*, the better the generated table fits the estimated one. Results are reported in Table 5.

Table 5 Generated Agents

	Estimated	Generated	Difference	*APD*
Individuals	10,637,107	10,635,691	1,416	<0.001
Households	4,334,281	4,333,448	833	<0.001

Table 6 *AAPD* Statistics

Distribution	Min	Max	Std dev	Mean
Ind′	0.000	0.005	<0.001	<0.001
Hh′	0.000	0.082	0.003	<0.001

First note that the procedure was able to generate 10,635,695 individuals gathered in 4,333,425 households, meaning that it could build a synthetic population where the numbers of households and individuals are very close to the estimated ones and differ less than 0.1% for the number of agents. This is highly encouraging.

Table 6 presents some basic statistics (minimum, maximum, standard deviation, and mean) on the average *APD* values (*AAPD*s) of the cells of the generated distributions computed across all the municipalities. As one can easily see, all these statistics also seem to indicate that the generator produces an accurate synthetic population. The maximum *AAPD* value for *Hh′* is associated with the municipality of Herstappe, which contains only 85 inhabitants gathered in 35 households. Because of its small size, a small deviation from the desired *Hh* can easily, in this case, result in a relatively large *AAPD* of 8.2%. Table 7 presents the same statistics as Table 6, but without this problematic municipality, showing it can be considered a statistical outlier.

At a more disaggregated level, Figures 2 and 3 illustrate the *AAPD*s' repartition for the individuals' and households' types across the Belgian municipalities and give some evidence of the synthetic population's accuracy in terms of *AAPD* and spatial coherence. Figures 4 and 5 give a representation of the *APD*'s mean and the standard deviation of each individual and household type over the 589 municipalities. Again, these figures suggest that the generator produces relatively small *APD* on average. Moreover, these *APD*s are associated with small standard deviations, meaning that *APD* values are relatively stable across the municipalities.

The synthetic population associated with the worst *AAPD* value for the individuals and the households are, respectively, Erezée and Herstappe. The details of these municipalities are described in Table 8. They clearly indicate that, even if these entities are the less accurate ones, the generated distributions are still reasonably close to the estimated ones: in average, the *APD* between the estimated and generated distributions for a given individual class is less than 0.5%,

Table 7 *AAPD* Statistics Without Herstappe

Distribution	Min	Max	Std dev	Mean
Ind′	0.000	0.005	<0.001	<0.001
Hh′	0.000	0.003	<0.001	<0.001

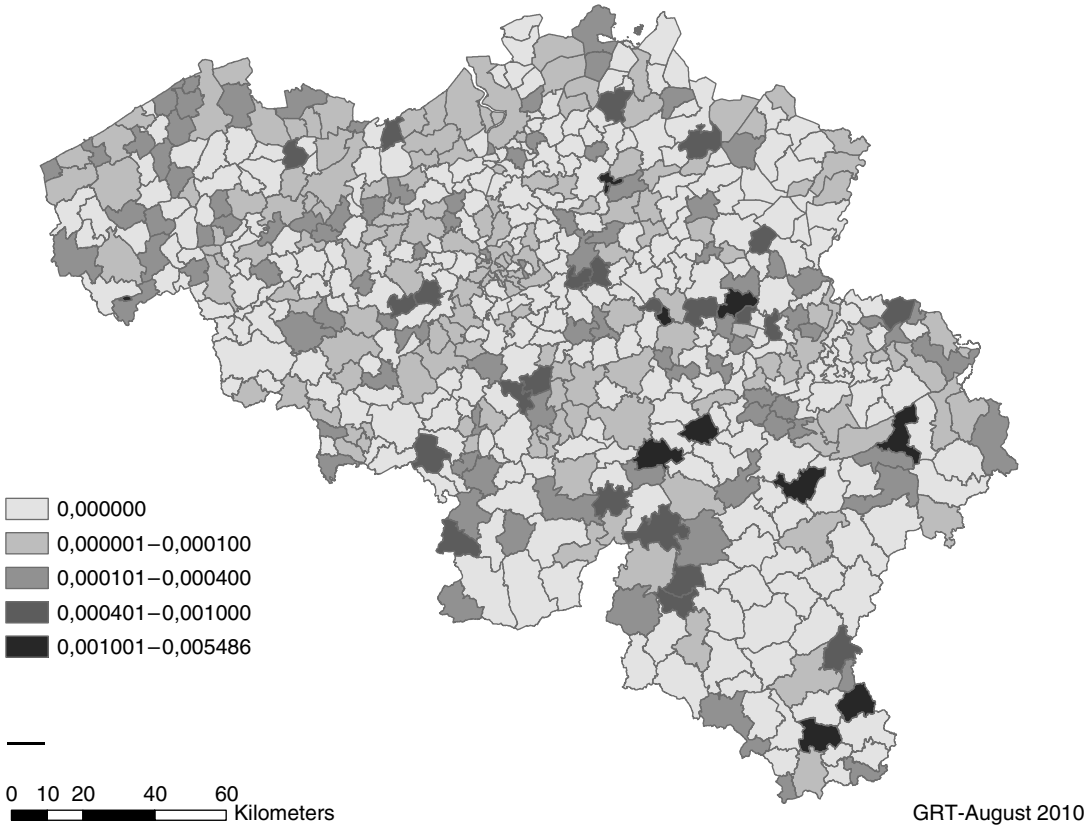

Figure 2 ***AAPD*'s Repartition for the Individual's Types**

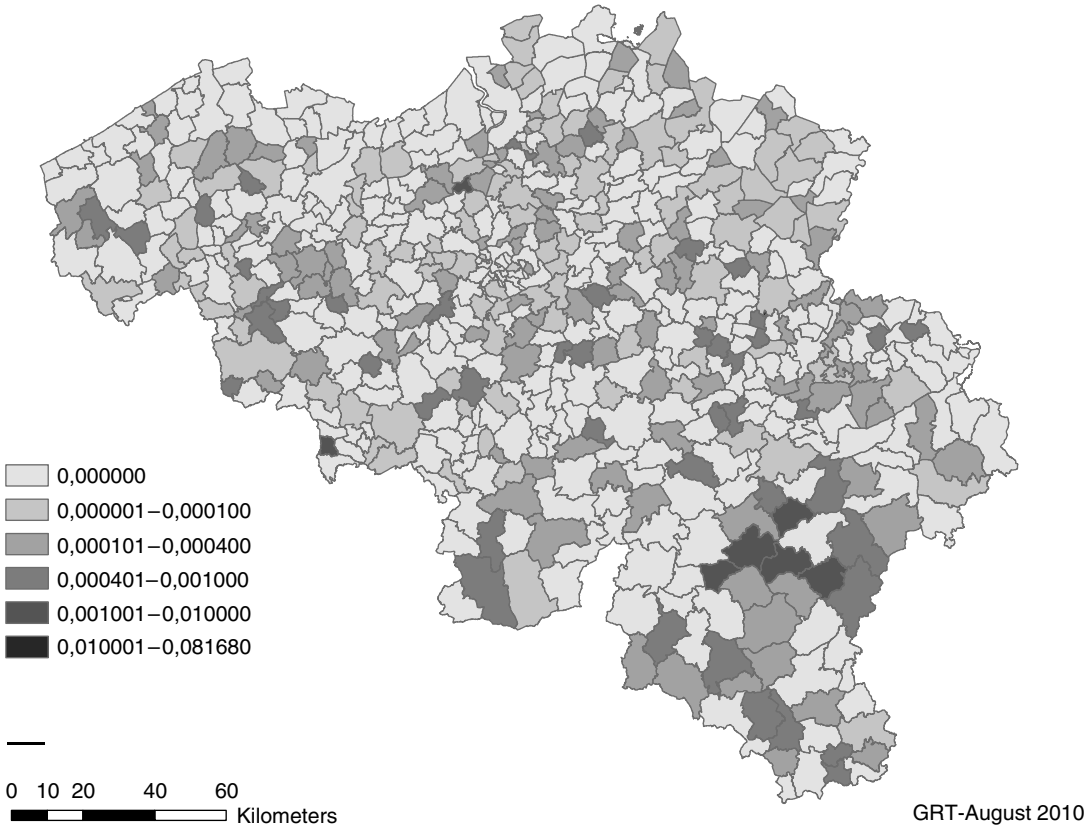

Figure 3 ***AAPD*'s Repartition for the Household's Types**

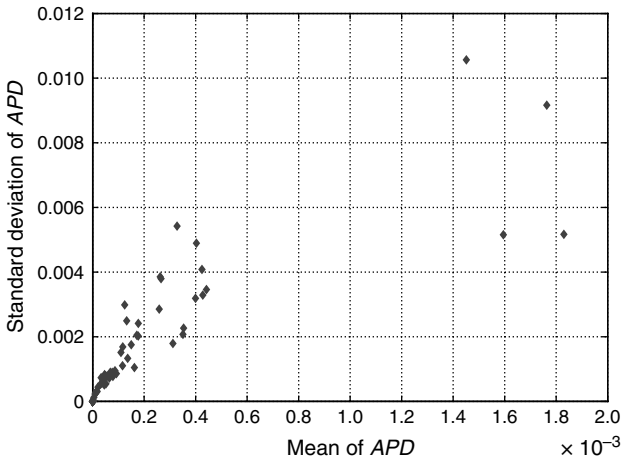

Figure 4 ***APD*'s Mean and Standard Deviation for Each Individual Type**

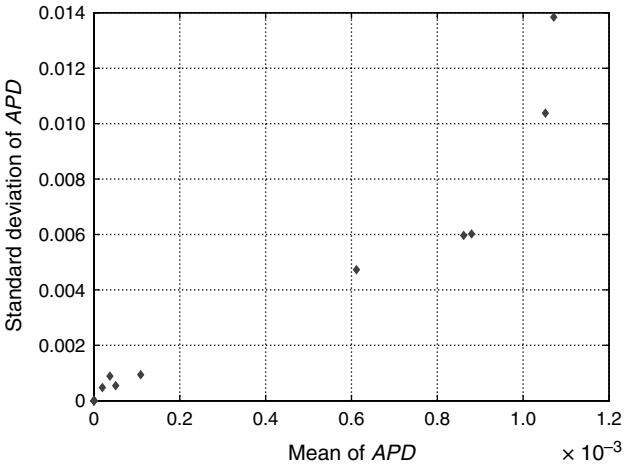

Figure 5 ***APD*'s Mean and Standard Deviation for Each Household Type**

and it less than 8.2% for a given household type. Moreover the generator produces a population having <0.1% fewer individuals and 7.9% fewer households than the estimated one. These results are illustrated in Figures 6 and 7 representing the number of agents generated against the number of estimated ones for each class of agents. As one can easily see, the contingency tables produced by the generator fit the initial ones quite accurately, given the initial level of data inconsistencies.

Table 8 **Erezée and Herstappe**

	Erezée	Herstappe
Distribution (D)	*Ind*	*Hh*
Agents estimated (E)	2,885	38
Agents generated (G)	2,869	35
Difference	16	3
$APD(E, G)$	<0.001	0.079
$AAPD(D, D')$	0.005	0.082

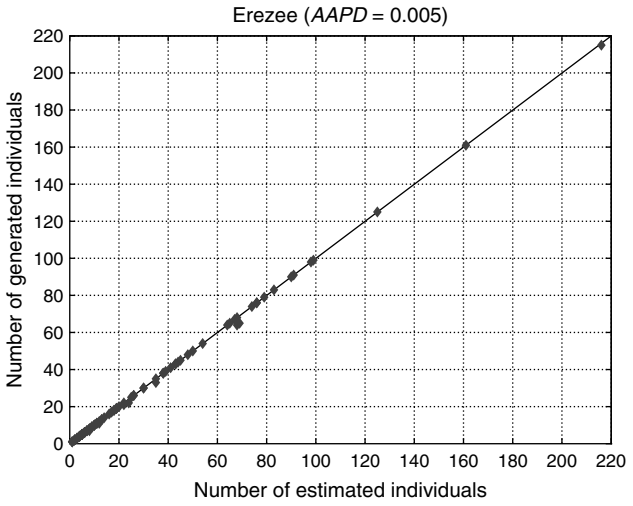

Figure 6 **Estimated vs. Generated Individuals for Erezée (the Worst Municipality for This Type of Agent)**

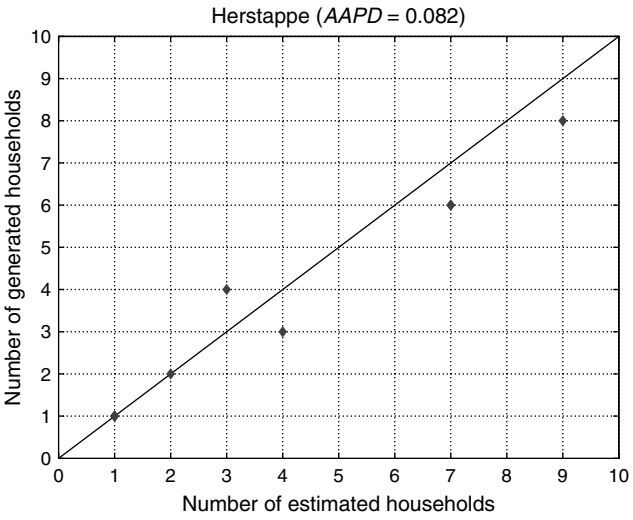

Figure 7 **Estimated vs. Generated Households for Herstappe (the Worst Municipality for This Type of Agent)**

4.2.2. Freeman-Tukey Goodness-of-Fit Test. Finally, we evaluated the goodness-of-fit of the distributions produced by the households generation procedure to the estimated ones at steps 1 and 2. This comparison is achieved by using the Freeman-Tukey (FT) statistic, defined by

$$\mathrm{FT}(T, T') = 4\sum_i \left(\sqrt{T_i} - \sqrt{T_i'}\right)^2,$$

where T and T' are, respectively, the estimated and generated (household or individual) distributions. This test, suggested by Voas and Williamson (2001) has the advantage over the classic Pearson χ^2 test in that it allows the presence of zeros in the cells of the

distributions. The FT statistic follows a χ^2 distribution with a number of degrees of freedom equal to one less than the number of cells in the compared distributions. The results of this goodness-of-fit test are highly promising because all generated distributions were statistically similar to the estimated ones both at a 95% and 90% level of confidence.

5. Comparison with an IPFP-Based Generator

5.1. Data and Parameters

In order to further assess the reliability and accuracy of the generator detailed in this paper, we now compare it with an IPFP-based synthetic population generator, namely, the extended IPFP generator described by Guo and Bhat (2007). The results obtained by this generator are strongly influenced by a parameter denoted by percentage deviation from target size (PDTS), whose value has been set to 0.10 for the comparison experiments. This value is recommended by Guo and Bhat (2007), and our experience also shows that it worked best with our data.

Assuming that the population generated in the previous subsection is a real population, we generate two synthetic populations by using the two generators. Because the true population is known, the required data necessary for running the two generators can easily be extracted; i.e., on one hand, a significant sample of households for each municipality and a set of margins for Guo and Bhat's generator and, on the other hand, various joint distributions (as specified earlier) at the municipality and district level for the new generator. Because the entire true population is known, the extracted data used are clearly consistent and the IPFP-related assumptions are met. The comparison can then be done without loss of accuracy. All the individuals' and households' attributes are used as control variables for the IPFP process. Because of its small size, the municipality of Herstappe is not considered in this test.

The minimal size of the sample for each municipality is given by Levy and Lemeshow (1999):

$$n \geq \frac{z_{1-\alpha/2}N}{z^2 + (N-1)l^2}, \tag{10}$$

where N is the total number of individuals of a municipality; l the margin of error; and $z_{1-\alpha/2}$ is the reliability coefficient associated with a confidence interval of $(1-\alpha)$%, that is, the $1-\alpha/2$-quantile of a standard normal distribution. In our experiments, $l = 0.025$, $\alpha = 0.05$, and $z_{0.975} = 1.96$. The samples are drawn using the simple random methodology.

Table 9 Generated Agents by Generator

		New generator		Guo and Bhat	
	True	Generated	*APD*	Generated	*APD*
Individuals	10,635,691	10,634,902	<0.001	9,731,686	0.085
Households	4,333,448	4,420,209	0.020	4,126,054	0.048

Table 10 *AAPD* Statistics

	Hh		*Ind*	
	New generator	Guo and Bhat	New generator	Guo and Bhat
Min	0.283	0.441	0.130	0.402
Max	6.807	13.188	0.603	2.511
Mean	0.596	0.665	0.322	0.658
Std dev	0.417	0.614	0.079	0.155
Median	0.533	0.578	0.314	0.630

5.2. Results

Table 9 shows that both procedures are able to produce a synthetic population having a number of agents (both households and individuals) close to the real one. However, we observe that the new generator's figures are closer to their correct values.

We pursue the comparison by reporting, in Table 10 statistics on the *AAPD* between the true and generated populations.

These results clearly favor the new generator, an observation confirmed by a one-way ANOVA analysis (with p-value smaller than 0.001 at a level $\alpha = 0.05$), whose notched box-plots are shown in Figures 8 and 9. We also report in Figures 10 and 11 the maximum value of the *APD* per agent type. Again the conclusion favors the new approach, especially for the individuals.

At a more disaggregate level, Figures 12 and 13 give the same maxima but now computed for each agent type (sorted by class size) on the worst municipality in term of *APD* (separately for households and individuals) for each generator. Note that these figures are plotted using a logarithmic scale. As expected, the larger errors correspond to the smaller agent classes.

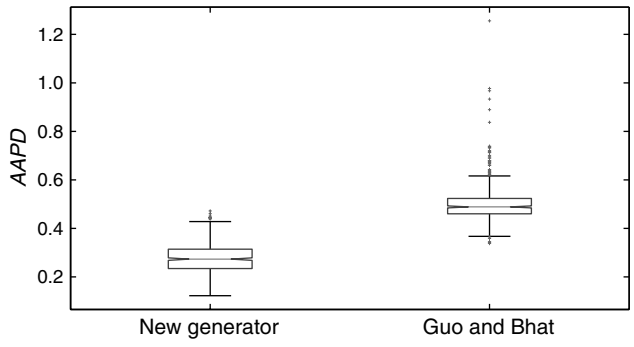

Figure 8 Notched Box-Plot of the *AAPD* for the Individual's Types

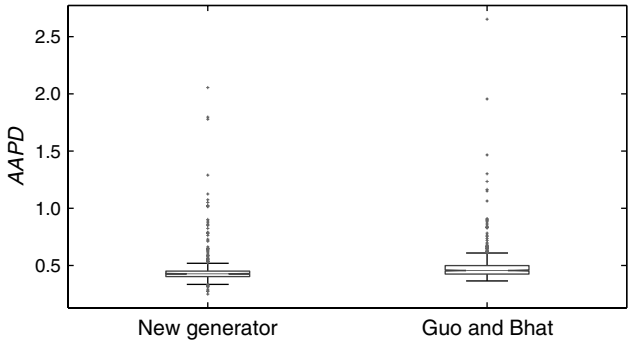

Figure 9 **Notched Box-Plot of the *AAPD* for the Household's Types**

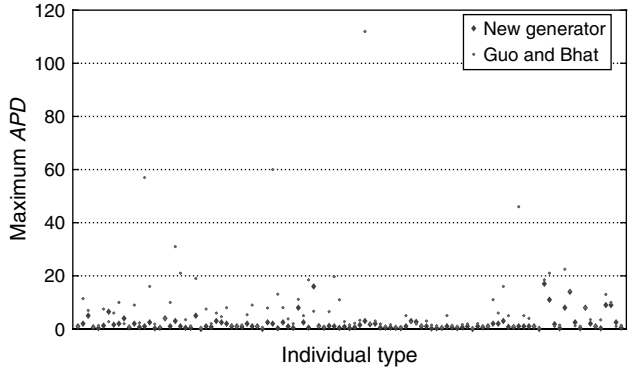

Figure 10 **Maximum *APD*'s Repartition for Disaggregated Individual Types**

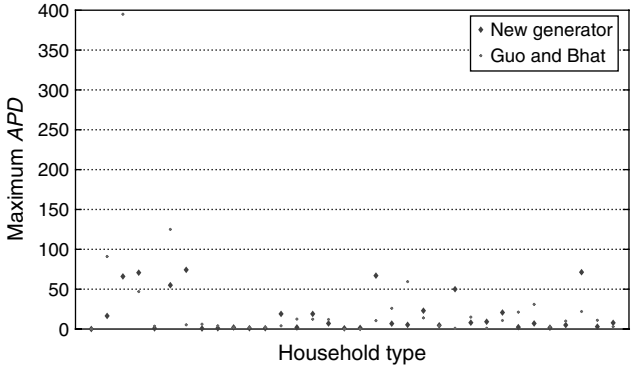

Figure 11 **Maximum *APD*'s Repartition for Disaggretated Household Types**

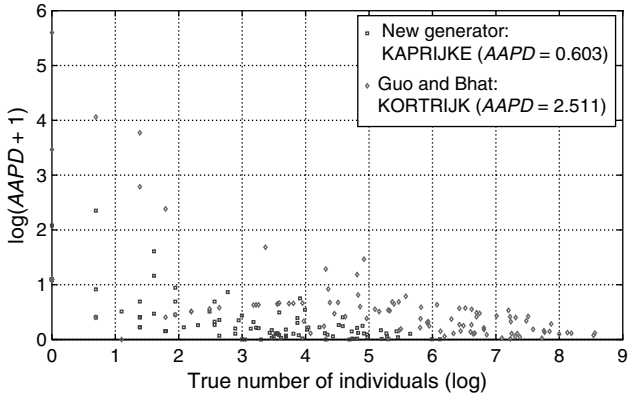

Figure 12 ***AAPD* vs. Generated Individuals by Generator for the Worst Municipality (Log Scale)**

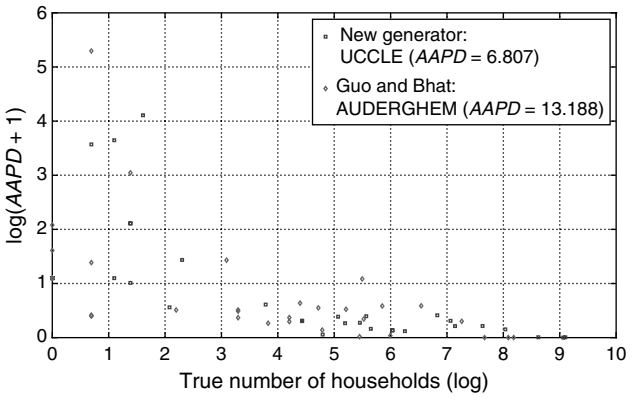

Figure 13 ***AAPD* vs. Generated Households by Generator for the Worst Municipality (Log Scale)**

The goodness-of-fit of the distributions produced by both households' generation procedures with respect to the estimated ones is considered in Table 11. This table gives the proportion of municipalities for which the generated distribution of the agents' attributes is statistically similar to the estimated one at a 95% level of confidence. The Freeman-Tukey statistic has been used to test the similarity. As one can see, both generators produce individuals' attributes' joint distributions for each municipality fitting accurately the estimed ones. This observation unfortunately no longer holds for the households' attributes' joint distributions. Indeed, whereas the distributions generated by the new generator still match the estimated ones, the IPFP-based approach performs poorly: less than 25% of the generated distributions adequately fit the estimated ones.

Considering the agent's dissagregation level, Figures 14 and 15 illustrate the distribution of *APD*'s means and standard deviations for each individual and household type by generator. These provide some evidence that the new generator outperforms that of Guo and Bhat (2007) in terms of *APD* between the estimated and the generated agents' attributes' joint distributions.

5.3. Sensitivity Analysis

The sensitivity of the proposed method with respect to data inconsistencies is also investigated to further assess the performance of the new generator. Different level of noise have been added to the data used for generating a synthetic population for Antwerpen (the largest Belgian municipality): the margins for all

Table 11 **Proportions of Municipalities Statistically Similar to the Estimation** ($\alpha = 0.05$)

	Hh (%)	*Ind* (%)
New generator	100.0	100.0
Guo and Bhat	23.8	100.0

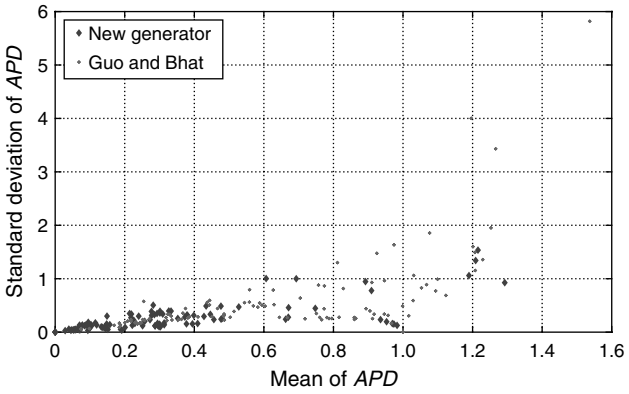

Figure 14 ***AAPD*'s Means and Standard Deviations for Each Individual Type**

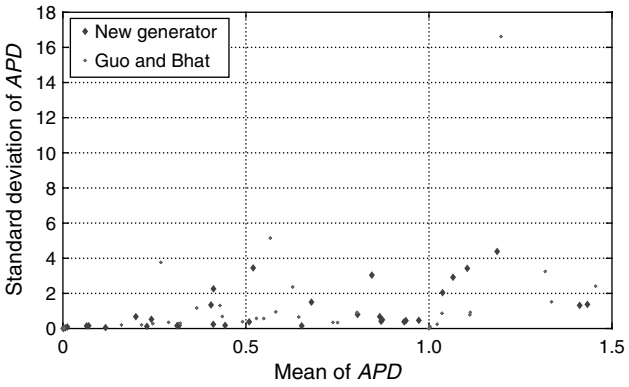

Figure 15 ***AAPD*'s Means and Standard Deviations for Each Household Type**

distributions have been corrupted by uniform relative random noise at 5%, 10%, and 15% levels. Given the level of inconsistency reported in §2.3 for the population of interest, generating higher levels of inconsistency seemed excessive. The results are reported in Table 12, where the difference in *AAPD* values compared to values quoted in §5.2 is explained by the more disaggregated level considered. We note that the generator produces similar *AAPD* at the individual and household level when data inconsistencies remain of the order of 10%, which is similar to the level observed for the Belgian data. When noise increases, the individual's *AAPD*s seem more stable than the household's ones.

Table 12 ***AAPD*'s Evolution for Antwerpen as a Function of the Noise Level**

	AAPD	
Noise (%)	Individual	Household
0	0.193	0.402
5	0.235	0.371
10	0.226	0.305
15	0.246	0.906

6. Conclusions

We have described a new synthetic population generation technique, belonging to the class of SR methods, which is designed to overcome some limitations of IPFP-based methods. In particular, the generator is sample-free and can handle (moderate) data inconsistency, which is common when data are extracted from several sources. Furthermore, its sample-free nature implies that it does not require an expensive survey to obtain the data needed for the generation of data protected by stringent privacy laws, which may apply in some countries such as Belgium.

Micro-simulations using synthetic data as input are obviously influenced by the quality of the population generated because the correlation structure of the resulting synthetic population reflects that given by the available data sets, but not necessarily the true one. Although it remains crystal clear that every effort should be made to invest in quality data, this is not always possible (with the desired standards) and the question then arises whether one should simply give up analysis or try to accomodate the available data sources. If the second path is chosen, which is our option in this paper, proportional care should be taken in interpreting the results.

The generator has been used to produce a synthetic population for Belgium at the municipality level. The results of the validation tests conducted on the households' generation procedure and the comparison with a more conventional approach indicate that the methodology has real potential to produce reliable synthetic populations. Coping with evolution of the database is also being investigated and will undoubtedly test the stability and practicality of the new algorithm further.

Acknowledgments

The authors thank the *Groupe d'étude de démographie appliquée* (GéDAP, University of Louvain-la-Neuve, Belgium) for providing the data, derived from the most recent available data sets collected for the 2001 Belgian census. Helpful corrections from Xavier Pauly and Fabien Walle are also gratefully acknowledged. The work of the first author has been funded by the DIDAM project within the Concerted Research Actions (ARC) research program of the Communauté Française de Belgique. Numerical simulations were made on the local computing resources at the Namur Center for Complex Systems (NAXYS).

References

Arentze T, Timmermans H, Hofman F (2007) Creating synthetic household populations: Problems and approach. Presentation, 86th Transportation Research Board Conference, Washington, DC.

Avery L (2011). National Travel Survey: 2010. *National Travel Survey* (Department for Transport, London, UK).

Auld J, Mohammadian AK, Wies K (2010) An efficient methodology for generating synthetic populations with multiple control levels. Presentation, 89th Transportation Research Board Annual Meeting, Washington, DC.

Beckman RJ, Baggerly KA, McKay MD (1996) Creating synthetic baseline populations. *Transportation Research A*. 30(6):415–429.

Bierlaire M (1991) Evaluation de la demande en trafic: quelques méthodes de distribution. *Annales de la Société Scientifique de Bruxelles* 105(1–2):17–66.

Conn AR, Gould NIM, Toint PhL (1992) *LANCELOT*: *A Fortran Package for Large-Scale Nonlinear Optimization* (*Release A*), No. 17. Springer Series in *Computational Mathematics* (Springer Verlag, Heidelberg, Berlin, New York).

Cornélis E, Legrain L, Toint PhL (2005) Synthetic populations: A tool for estimating travel demand. Jourquin B, ed., *BIVEC-GIBET Transport Research Day 2005*, Vol. 1 (VUBPRESS, Brussels University Press, Belgium), 217–235.

Deming WE, Stephan FF (1940) A least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* 11:428–444.

Direction Régionale de l'Équipement d'Île-de-France (2004) *Les déplacement des Franciliens en 2001-2002*, Enquê Globale de Transport. Documentation Française, Paris, France.

Dwight HB (1961) *Tables of Integrals and Other Mathematical Data*, 4th edition (The Macmillan Company).

Frick M, Axhausen K (2004) Generating synthetic populations using IPF and Monte-Carlo techniques: Some new results. *4th Swiss Transport Research Conference, Monte-Verita*.

Glover F (1986) Future paths for integer programming and links to artificial intelligence. *Comput. Oper. Res.* 13:533–549.

Glover F (1989) Tabu search—Part I. *ORSA J. Comput.* 1:190–206.

Glover F (1990) Tabu search—Part II. *ORSA J. Comput.* 2:4–32.

Glover F, Laguna M (1997) *Tabu Search* (Kluwer Academic Publishers, Boston).

Gould NIM, Orban D, Toint PhL (2003) *GALAHAD*—A library of thread-safe Fortran 90 packages for large-scale nonlinear optimization. *ACM Trans. Math. Software* 29(4):353–372.

Guo Y, Bhat CR (2007) Population synthesis for microsimulating travel behavior. *Transportation Res. Record* 2014:92–101.

Huang Z, Williamson P (2002) A comparison of synthetic reconstruction and combinatorial optimization approaches to the creation of small-area microdata. Working paper, Department of Geography, University of Liverpool, UK.

Hubert J-P, Toint PhL (2002) *La mobilité Quotidienne Des Belges*, No. 1. *Mobilité et Transports* (Presses Universitaires de Namur, Namur, Belgium).

Ireland CT, Kullback S (1968) Contingency tables with given marginals. *Biometrika* 55(1):179–199.

Kreyszig E (1972) *Advanced Engineering Mathematics*. 3rd edition (J. Wiley and Sons, Chichester, England).

Levy PS, Lemeshow S (1999) *Sampling of Populations—Methods and Applications*, 3rd edition (Wiley-Interscience Publication, USA).

Little RJA, Wu M-M (1991) Models for contingency tables with known margins when target and sampled population differ. *J. Amer. Statist. Assoc.* 86(413):87–95.

Mosteller F (1968) Association and estimation in contingency tables. *J. Amer. Statist. Assoc.* 63:1–28.

Müller K, Axhausen KW (2011) Population synthesis for microsimulation: State of the art. Presentation, 90th Annual Meeting of the Transportation Research Board.

Ortúzar JD, Willumsen LG (2001) *Modelling Transport*, 3rd edition (J. Wiley and Sons, Chichester, England).

Pritchard DR, Miller EJ (2009) Advances in agent population synthesis and application in an integrated land use and transportation model. Presentation, 88th Transportation Research Board Annual Meeting, Washington, DC.

Srinivasan S, Ma L, Yathindra K (2008) Procedure for forecasting households characteristics for input to travel-demand models. Final Report TRC-FDOT-64011-2009, Transportation Research Center, University of Florida, Gainsville.

Voas D, Williamson P (2001) An evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modeling* 5(2):177–200.

Wilson AG (1970) *Entropy in Urban and Regional Modelling* (Pion, London).

Wilson AG (1974) *Urban and Regional Models in Geography and Planning* (J. Wiley and Sons, Chichester, England).

Wilson AG, Pownall CE (1976) A new representation of the urban system for modeling and for the study of microlevel interdependence. *Area* 8:246–254.

Ye X, Konduri K, Pendyala RM, Sana B, Waddell P (2009) A methodology to match distributions of both household and person attributes in the generation of synthetic populations. Presentation, 88th Annual Meeting, Transportation Research Board, Washington, DC.