# Synthetic population generation without a sample

Barthelemy, Johan; Toint, Ph

*Published in:*
Transportation Science

*DOI:*
10.1287/trsc.1120.0408

*Publication date:*
2013

*Document Version*
Early version, also known as pre-print

FUNDP
NAMUR

# Synthetic population generation in presence of data inconsistencies

by J. Barthelemy and Ph. L. Toint

Report NAXYS-12-2010 23 December 2010

naXys
Namur Center for Complex SYStems

University of Namur
61, rue de Bruxelles, B5000 Namur (Belgium)

http://www.fundp.ac.be/sciences/naxys

# Synthetic population generation in presence of data inconsistencies

Johan Barthélemy*, Philippe L. Toint†

23 December 2010

**Abstract**

The advent of microsimulation in the transportation sector has created the need for extensive disaggregate data concerning the population whose behaviour is modelled. Due to 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. While 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 which is most unusual because this data often comes from different sources and is 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 proper. The main idea in these generation steps is to use data at the most disaggregate 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 used the $\chi^2$ 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 its is shown that the new population generator produces excellent results.

**Keywords**

Synthetic population, microsimulation, limitations of iterative proportionnal fitting based procedures, sample-free generator.

## 1 Introduction

Synthetic population generation has recently received considerable attention in the literature (see, for instance, Müller and Axhausen (2010)). It is often motivated by the observation that micro-simulations, 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 due to stringent privacy laws. A way to address these issues is to construct an artificial population starting from known data about the true one. As it is obvious that the representativeness of the synthetic population is critical for the simulations accuracy, a synthetic population generator should therefore produce a population approximating the correlation structure of the true population as accurately as possible.

*Namur Research Center for Complex Systems (NAXYS), FUNDP-University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium. Email: johan.barthelemy@fundp.ac.be

†Namur Research Center for Complex Systems (NAXYS), FUNDP-University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium. Email: philippe.toint@fundp.ac.be (corresponding author)

Techniques for synthetic population generation typically belong to either the Synthetic Reconstruction techniques (SR) 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 Willson and Pownall (1976) and 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 sub-set 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 motivates 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 Section 2, we first present the standard approach for building a synthetic population, from which the others 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 Section 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 finally discussed in Section 6.

# 2 The standard approach

To date, the standard approach for building synthetic populations is based on the method developed by Beckman et al. (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 consists in 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 provides full information about the attributes of interest, but only for a sample of agents and is 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 $V = \{v_1, v_2, ..., v_n\}$ the vector of variables representing these attributes, each combination of values of $v_i$'s therefore defines a socio-demographic group. The synthetic population is then generated by a two steps 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 in 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 on 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 2-way contingency table is built from the seed with initial components $\pi_{ij} \in \mathbb{R}^+$. Assume also that desired marginal distributions $\{x_{i\bullet}, x_{\bullet j}\}$ (the target) are known. 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 $\pi^*_{ij}$'s are the component values at the last iteration. The adjustments at iteration $l$ are computed by the equations

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

$$\pi^{l}_{ij} = \pi^{l'}_{ij} . \frac{x_{i\bullet}}{\pi^{l'}_{i\bullet}} \qquad \forall i, j.$$

In order to produce an accurate estimate of the true distribution, the procedure requires an initial representative sample of the true population for building the initial multiway table. This requirement is important since Mosteller (1968) pointed out that the procedure preserves the interaction structure of the sample as defined by the odd ratios

$$\frac{\pi_{ij} . \pi_{hk}}{\pi_{ik} . \pi_{hj}} = \frac{\pi^l_{ij} . \pi^l_{hk}}{\pi^l_{ik} . \pi^l_{hj}}$$

at each iteration $l$. Moreover according to Ireland and Kullback (1968), the IPFP also produces the constrained maximum relative entropy estimator of the true contingency table

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

Finally, Little and Wu (1991) have demonstrated that the IPFP results in a maximum likelihood estimator of the true contingency table.

## 2.2 Generating the synthetic population

Once the expected numbers of agents in every socio-demographic groups are estimated, each sampled agent is associated with a probability of being selected 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, agents are randomly drawn 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 in 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, 2005), MOBEL (Hubert and Toint (2002)) or NTS (Office of UK National Statistics, 2010) suggest that the travel behaviour 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 dataset detailing the joint-distribution of individuals' and households' attributes simultaneously. Since the estimation step of the algorithm described in Section 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 not both. Note that households' distributions accuracy has often been preferred Ye, Konduri, Pendyala, Sana and Waddell (2009).

This strong limitation lead several authors to propose interesting improvements to this basic algorithm. Guo and Bhat (2007) propose 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 built on these contributions and propose 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.

However these improved approaches remain based on the IPFP 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 iterations. 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[(1)].

These three 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 the case of point. Firstly, a representative sample at the municipality level (which is the desired spatial disaggregation level) is not available. A second problem is that all necessary informations, *i.e.* distributions, are not available from a single source (which would hopefully guarantee consistency), but has to be extracted from different datasets, 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 2.1 (extracted from Cornélis, Legrain and Toint, 2005) for the Charleroi district.

| **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 |

Table 2.1: Inconsistencies between margins extracted from different sources

In this table, the total number of inhabitants in the district using the most reliable of the data sources (first row of the table) is compared with the same statistic extracted from other data files to be used in the population synthesis (row 2 to 5). One immediately notices inconsistencies between the different estimations with differences up to 10%, irrespective of the data source. These inconsistencies prevent the IPFP process to converge. This could possibly be cured by considering the frequencies rather that 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 which would not suffer from the lack of a representative sample at the most disagregated level and/or from (moderate) inconsistencies between different data sources. This is the object of Section 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 disaggregate level available, while maintaining known correlations as well as possible.* The algorithm implementing this principle consists in a 3-steps procedure:

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 $Hh$;

[(1)] Introducing small initials values in the unpopulated classes remains unsatisfactory as this procedure introduces unwanted bias.

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 in the synthetic population.

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

## 3.1 Step 1: Generating the pool of individuals

The first step aims at building the $Ind$ pool of synthetic individuals, 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. However, not every distribution for $V_i$ is known at the most disaggregate 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 (n.d.) and 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 disaggregate 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 disaggretated data available (at municipality level in our case). The missing attributes for each individual in this table are then determined by finding the most disaggregate level at which a joint distribution for the missing attribute and some already know 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.

Since 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. A correction is then made to $Ind$ to make it consistent with the margins at the level 0. This correction is computed by suitably shifting some of the attributes' value of certain individuals. Only shifts between two contiguous modality are allowed.

## 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 the of household-related attributes and by $w_j$ the value taken by a particular household for the $j^{\text{th}}$ 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 a number of particular household of a type specified by a combination of the $w_j$'s (which we call a household type). This problem is solved in two steps: a maximum entropy estimate of $Hh$ is first generated and is subsequently improved by using a tabu-search optimization process.

### 3.2.1 Entropy maximisation 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 amongst 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, we are however, interested in a more aggregate 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

$$E(Hh) = \frac{(\sum_i x_i)!}{\prod_i x_i!}. \tag{3.1}$$

The function $E(\cdot)$ is called the entropy function. As 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 (3.1). Using the natural logarithm and Stirling's short approximation (Dwight (1961) and Kreyszig (1972)), the corresponding objective function of this problem can then be approximated by

$$\min_x \sum_i x_i \ln(x_i) - x_i \tag{3.2}$$

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

Unfortunately, due to the inconsistent nature of the available data, as exposed in Section 2.3, the constraints of this optimization problem are also formally in inconsistent. Our approach is then to impose only a subset $\Omega$ of them corresponding to the data of highest quality as strict constraints, the others being then 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 $\sigma$ is a penalization parameter and $n_p$ is the number of households involved in $p$.

Denoting by $A$ and $b$ the matrix and the vector derived from the subset of the scaled inconsistent constraints, the new objective function can now be formulated as

$$(EN) \qquad \min_x \|Ax - b\|_2^2 + \sum_i x_i \ln(x_i) - x_i \tag{3.3}$$

and the minimization is then carried out under the constraints in the set $\Omega$ only. In general the solution of this optimization problem yields a non-integer solution, which is unsuitable for representing households numbers. The solution's components of this optimisation problem are then rounded and the value

$$f_{EN}(\hat{x}) = \sum_i w_i|\hat{c}_i - c_i| + \sum_i \hat{x}_i \ln(\hat{x}_i) - \hat{x}_i \tag{3.4}$$

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 $i^{\text{th}}$ constraint and the associated weight depending on the quality of the associated data source, respectively. This value 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 $\sigma$, 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 subsection.

### 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 neighbourhood 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 neighbourhood $\mathcal{N}(x)$ 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), Glover (1990) and Glover and Laguna (1997).

In this paper, the tabu list is a list $T$ of size $n$, which 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 follow:

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

where the notation $x_j^{i-1} \pm 1$ stands for two variations of the $j$-th 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 $f_{EN}(x^i) < f_{EN}(x^*)$ then $x^* = x^i$;
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 trade-off between computing cost and quality of the estimate.

### 3.3 Step 3: Households' generation

Individuals' and households' distributions being estimated, the last step of our generator consists in 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 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) minimize the observed $\chi^2$ 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, household's members are generated as follows: a household head is first drawn from the pool of individual $Ind$, 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 of a head and possibly a mate, children and additional adults, the construction starts with the selection of its head. Depending on the type, the head's attributes are either randomly drawn according to known joint-distributions on couple formation, or obtained directly (for instance for an isolated man). More formally, this selection procedure is organized in 3 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;
   - 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 $\chi^2$ 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;
   - else find a suitable household head by random search in the constituents 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 one, with the only exception that individuals' characteristics may no longer be determined by the household type only.

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, while the final number of individuals is smaller than estimated in the second case.

# 4 Generating a Belgian synthetic population

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 household 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 4.2 presents basics statistics on these municipalities. The individuals and households attributes are respectively described in Tables 4.3 and 4.4. Data available at the municipality or 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);
- *GéDAP*[(2)] centre of the University of Louvain-la-Neuve (Belgium) (2001);
- the *MOBEL* mobility survey (Hubert and Toint (2002)).

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

Table 4.2: Basic statistics for municipalities

| **Attribute** | **Values** |
|---|---|
| Gender | male ; female |
| Age class | 0-5; 6-17; 40-59; 60+ |
| Activity | student; active; inactive |
| Education level | primary; high school; higher education; none |
| Driving license ownership | yes; no |

Table 4.3: Individuals' characteristics

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 an 3Ghz CPU and 3Go of RAM under a 32 bits Linux environment. The generation process took around 16 hours and 30 minutes to treat all the 589 municipalities.

[(2)] Groupe d'étude de démographie appliquée

| Attribute | Values |
|---|---|
| Type | single man alone<br>single woman alone<br>single man with children (and other adults)<br>single woman with children (and other adults)<br>couple without children (and other adults)<br>couple with children (and other adults) |
| Number of children | 0 to 5 |
| Number of other adults | 0 à 2 (mate not included) |

Table 4.4: Households' characteristics

## 4.1 Verification of the household generation procedure

### 4.1.1 Absolute percentage difference

Having generated a synthetic population, one is then faced to 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 follow:

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

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 fits the estimated one. Results are reported in Table 4.5.

| | **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 4.5: Generated agents

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 number of households and individuals are very close to the estimated ones and differs less than 0.1% for the number of agents. This is highly encouraging.

Table 4.6 presents some basic statistics (maximum, minimum, standard deviation and mean) on the average $APD$ values ($AAPD$) 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. Due to 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 4.7 presents the same statistics of Table 4.6 where we have neglected this problematic municipality, showing it can be considered as a statistical outlier.

| **Distribution** | **Min** | **Max** | **Std dev** | **Mean** |
|---|---|---|---|---|
| $Ind'$ | 0.000 | 0.005 | $< 0{,}001$ | $< 0{,}001$ |
| $Hh'$ | 0.000 | 0.082 | 0,003 | $< 0{,}001$ |

Table 4.6: $AAPD$ statistics

At a more disaggregate level, Figures 4.1 and 4.2 illustrates 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 term of $AAPD$ and spatial coherence. Figures 4.3 and 4.4 gives a representation

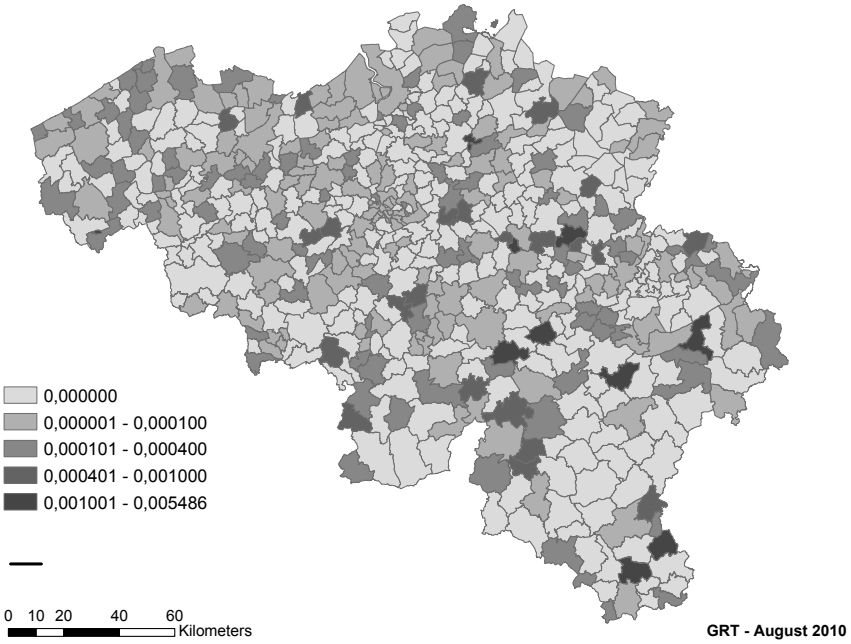


Figure 4.1: $AAPD$s' repartition for the individual's types

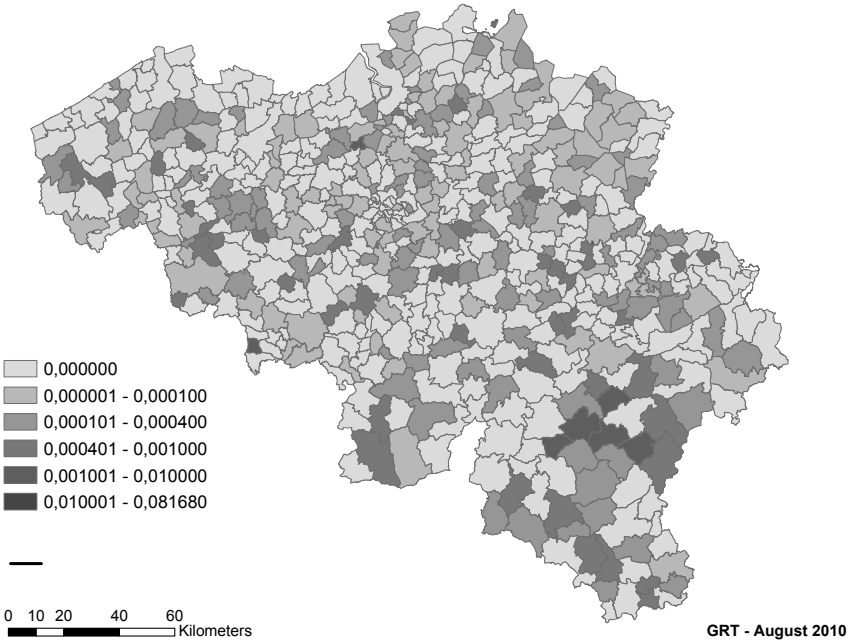


Figure 4.2: $AAPD$s' repartition for the household's types

| Distribution | Min | Max | Std dev | Mean |
|---|---|---|---|---|
| $Ind'$ | 0.000 | 0.005 | $< 0{,}001$ | $< 0{,}001$ |
| $Hh'$ | 0.000 | 0.003 | $< 0{,}001$ | $< 0{,}001$ |

Table 4.7: $AAPD$ statistics without Herstappe

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 Herezée and Herstappe. The details of these municipalities are described in Table 4.8. 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% while it less than 8.2% for a given household type. Moreover the generator produces a population having $< 0.1\%$ less individuals and 7.9 less households than the estimated one. These results are illustrated on Figures 4.5 and 4.6 representing the number of agents generated against the number of desired of estimated one for each class of agents. As one can easily see, the contingency tables produced by the generator fits the initial ones quite accurately, given the initial level of data inconsistencies.

| | Herezée | 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 |
| $AAPD(D,D')$ | 0.006 | 0.082 |

Table 4.8: Herezée and Herstappe

### 4.1.2 Freeman-Tukey goodness-of-fit test

Finally, we evaluated the goodness-of-fit of the distributions produced by households generation procedure to the estimated ones at Steps 1 and 2. This comparison is achieved by using the Freeman-Tuckey statistic defined by

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

where $T$ and $T'$ are respectively the estimated and generated (household or individual) distribution. This test, suggested by Voas and Williamson (2001), has the advantage over the classic Pearson $\chi^2$ test that it allows the presence of zeros in the cells of the distributions. The $FT$ statistic follows a $\chi^2$ 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 as all generated distributions were statistically similar to the estimated ones at a 95% level of confidence.

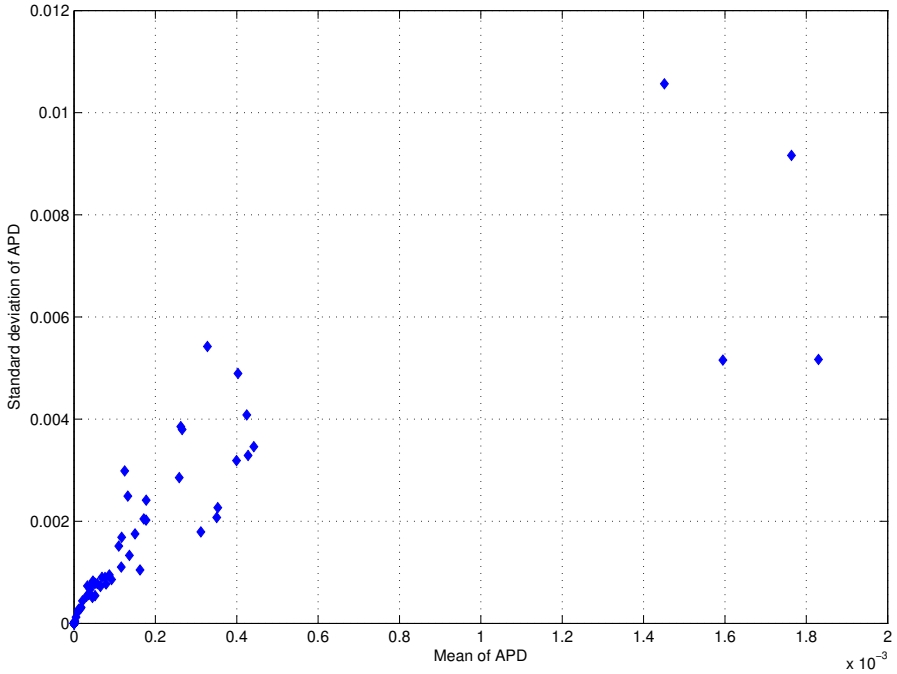


Figure 4.3: $AAPD$s' mean and standard deviation for each individual type

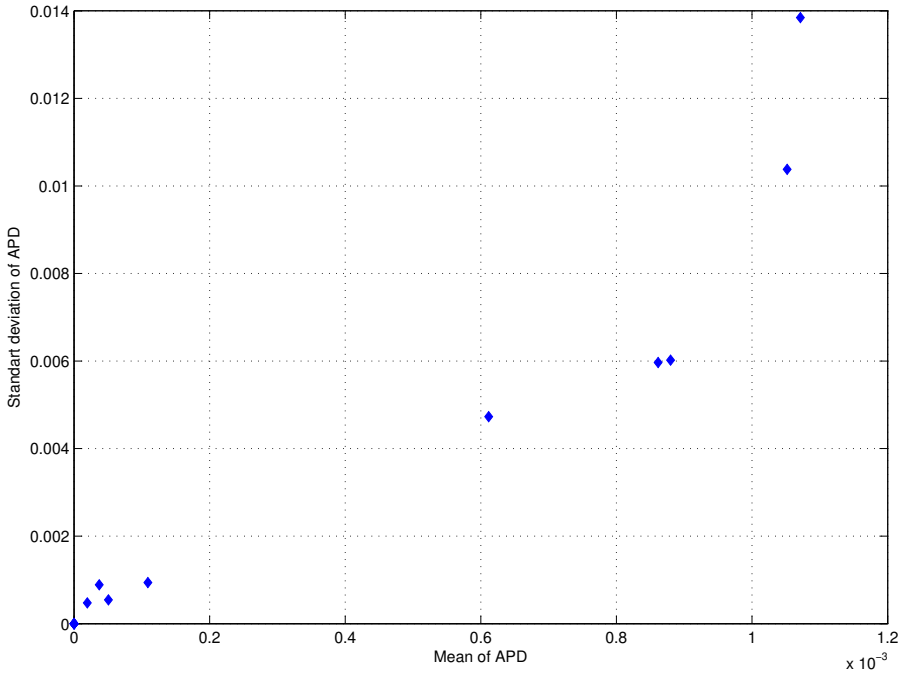


Figure 4.4: $AAPD$s' mean and standard deviation for each household type

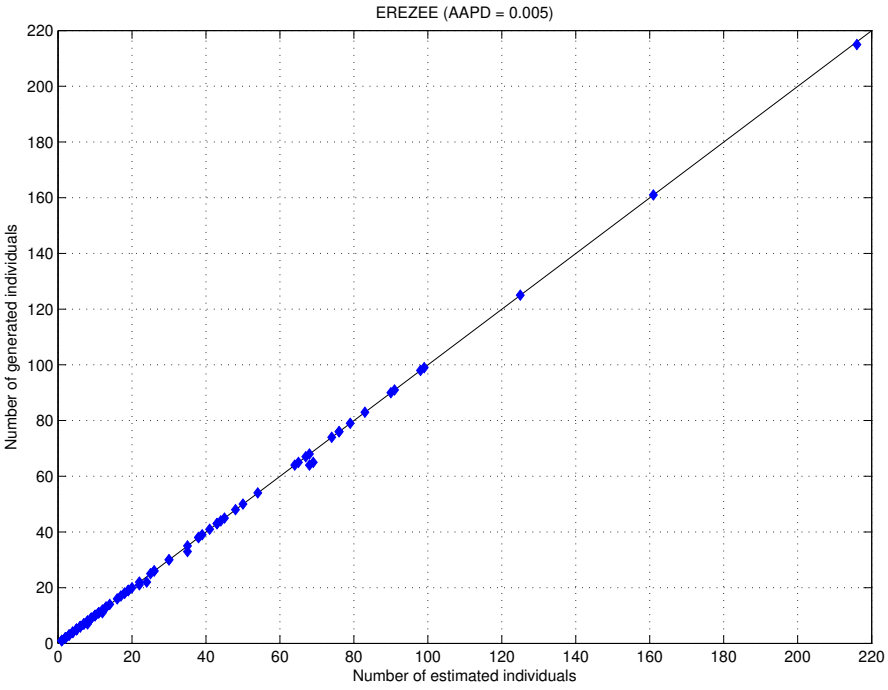


Figure 4.5: Estimated × generated individuals for Herezée

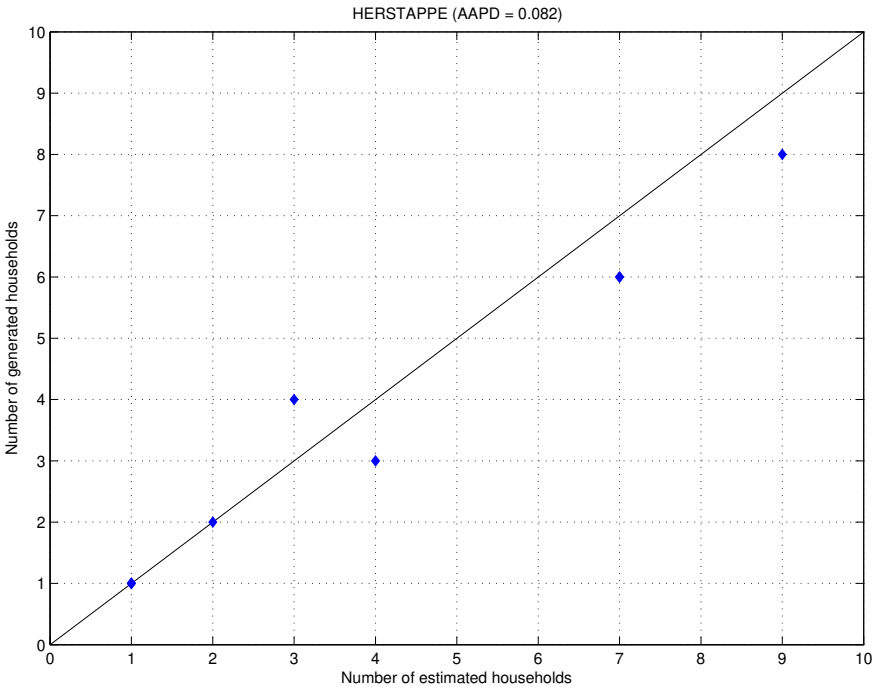


Figure 4.6: Estimated × generated households for Herstappe

# 5 Comparison with a IPFP-based generator

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 $PDTS$, whose value has been set to 0.10 for the comparison experiments[(3)].

Assuming that the population generated in the previous subsection is a real population, we generate two synthetic populations by using the two generators. As the true population is known, the required data by 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 at the muncipality and district level for the new generator. Since the entire true population is known, the extracted data used is clearly consistent and the IPFP-related assumptions are met. The comparison can then be done without loss of accuracy. Due to its small size, the municipality of Herstappe is not considered in this test.

Table 5.9 shows that both procedures are able to produce a synthetic population having a number of agents (both households and individuals) close to the real ones. However, we observe that the new generator's figures are closer to their correct values.

| | **True** | **New Generator** | | **Guo and Bhat** | |
|---|---|---|---|---|---|
| | | Generated | *APD* | Generated | *APD* |
| **Individuals** | 10,637,022 | 10,634,902 | $< 0.001$ | 9,731,686 | 0.085 |
| **Households** | 4,334,246 | 4,420,209 | 0.020 | 4,126,054 | 0.048 |

Table 5.9: Generated agents by generator

The same statististics as that described in Table 4.7 computed across the 588 municipalities are presented for the two generators in Table 5.10. These results indicate that our new method compares very favourably with Guo and Bhat's.

| | **Hh** | | **Ind** | |
|---|---|---|---|---|
| | New generator | Guo and Bhat | New generator | Guo and Bhat |
| Min | 0.000 | 0.006 | 0.000 | 0.072 |
| Max | 0.005 | 1.575 | 0.006 | 1.298 |
| Mean | $< 0.001$ | 0.175 | $< 0.001$ | 0.262 |
| Std dev | $< 0.001$ | 0.106 | $< 0.001$ | 0.074 |

Table 5.10: AAPD statistics

The goodness-of-fit of the distributions produced by both households generation procedures with respect to the estimated ones is considered in Table 5.11. 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-Tuckey statistic has been used to test the similarity. As one can see, both generator 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, while the distributions generated 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.

| | **Hh** | **Ind** |
|---|---|---|
| **New generator** | 100.0% | 100.0% |
| **Guo and Bhat** | 23.8% | 100.0% |

Table 5.11: Proportions of municipalities statistically similar to the estimation ($\alpha = 0.05$)

[(3)] This value is recommended by Guo and Bhat (2007) and our experience also shows that it worked best with our data.

Considering a more disaggrate level, figures 5.7 and 5.8 illustrate the distribution of $APD$s' means and standard deviations for each individual and household type by generator. Note that the scale of these figures is logarithmic. These provide some evidence that the new generator outperforms that of Guo and Bhat by several order of magnitude in terms of $APD$ between the estimated and the generated agents' attribute joint-distributions.

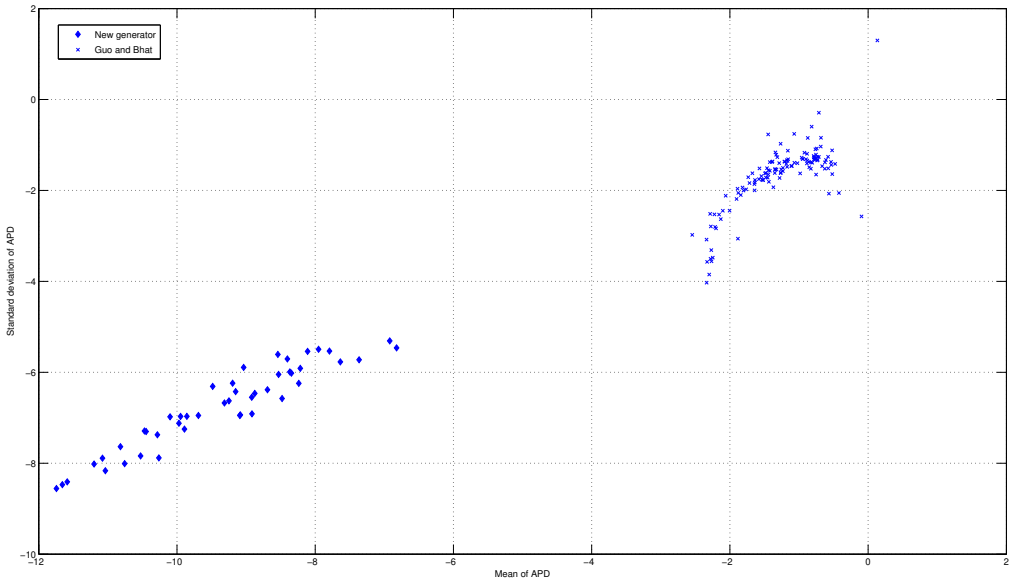


Figure 5.7: $AAPD$s' means and standard deviations for each individual type

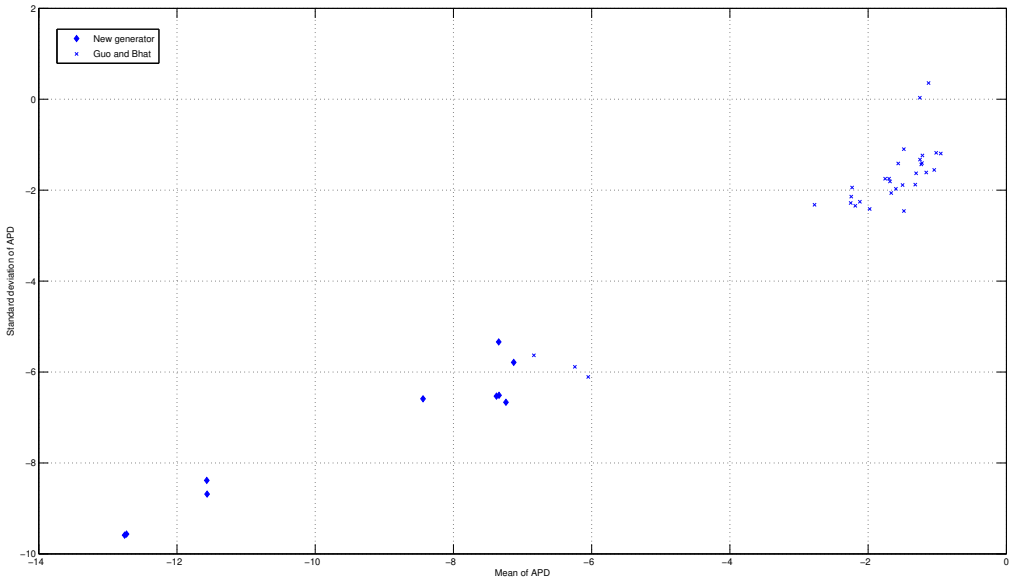


Figure 5.8: $AAPD$s' means and standard deviations for each household type

# 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 is 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.

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. Ongoing research covers a more detailed comparison of this generator with IPFP techniques in the (restrictive) case where the latter apply, but this is beyond the scope of this paper. 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 wish to 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 datasets 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.

# References

T. Arentze, H. Timmermans, and F. Hofman. Creating synthetic household populations: Problems and approach. Paper presented at the 86th Transportation Research Board conference, Washington DC, US, 2007.

R. J. Beckman, K. A. Baggerly, and M. D. McKay. Creating synthetic baseline populations. *Transportation Research A*, **30**(6), 415–429, 1996.

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

E. Cornélis, L. Legrain, and Ph. L. Toint. Synthetic populations: a tool for estimating travel demand. *in* B. Jourquin, ed., 'BIVEC-GIBET Transport Research Day 2005', Vol. 1, pp. 217–235. VUBPRESS Brussels University Press, 2005.

W.E. Deming and F.F. Stephan. A least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, **11**, 428–444, 1940.

H. B. Dwight. *Tables of integrals and other mathematical data*. The Macmillan Company, fourth edn, 1961.

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

F. Glover. Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research*, **13**, 533–549, 1986.

F. Glover. Tabu search - Part I. *ORSA Journal on Computing*, **1**, 190–206, 1989.

F. Glover. Tabu search - Part II. *ORSA Journal on Computing*, **2**, 4–32, 1990.

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

Y. Guo and C. R. Bhat. Population synthesis for the microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board*, **2014**, 92–101, 2007.

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

J.-P. Hubert and Ph. L. Toint. *La mobilité quotidienne des Belges*. Number 1 *in* 'Mobilité et Transports'. Presses Universitaires de Namur, Namur, Belgium, 2002.

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

E. Kreyszig. *Advanced engineering mathematics*. J. Wiley and Sons, Chichester, England, third edn, 1972.

R.J.A. Little and M.-M. Wu. Models for contingency tables with known margins when target and sampled population differ. *Journal of the American Statistical Association*, **86**(413), 87–95, 1991.

F. Mosteller. Association and estimation in contingency tables. *Journal of the American Statistical Association*, **63**, 1–28, 1968.

K. Müller and K. W. Axhausen. Population synthesis for microsimulation: State of the art. 10th Swiss Transport Research Conference, 2010.

J. D. Ortúzar and L.G. Willumsen. *Modelling Transport*. J. Wiley and Sons, Chichester (England), 3rd edn, 2001.

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

A. G. Willson and C.E. Pownall. A new reprensentation of the urban system for modeling and for the study of micro-level interdependence. *Area*, **8**, 246–254, 1976.

A. G. Wilson. *Entropy in urban and regional modelling*. Pion, London, 1970.

A. G. Wilson. *Urban and regional models in geography and planning*. J. Wiley and Sons, Chichester, England, 1974.

X. Ye, K. Konduri, R.M. Pendyala, B. Sana, and P. Waddell. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. *in* 'TRB 88th Annual Meeting Compendium of Papers DVD', Washington, U.S.A., 2009. Transportation Research Board - 88th Annual Meeting.