Transportation Research Part B 58 (2013) 243–263

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

# Simulation based population synthesis

Bilal Farooq [a,*], Michel Bierlaire [b,1], Ricardo Hurtubia [c], Gunnar Flötteröd [d]

[a] Département des Génies Civil, Géologique et des Mines, École Polytechnique Montréal, 2500 Ch. Polytechnique Montréal, H3T 1J4 Montréal, Canada
[b] Transport and Mobility Laboratory, ENAC, École Polytechnique Fédérale de Lausanne, Station 18, CH-1015 Lausanne, Switzerland
[c] Departamento de Urbanismo, Facultad de Arquitectura y Urbanismo, Universidad de Chile, Santiago, Chile
[d] Division for Traffic and Logistics, Royal Institute of Technology, Teknikringen 72, 11428 Stockholm, Sweden

ARTICLE INFO

*Keywords:*
Markov chain Monte Carlo simulation
Population synthesis
Agent based model
Integrated urban systems planning

ABSTRACT

Microsimulation of urban systems evolution requires synthetic population as a key input. Currently, the focus is on treating synthesis as a fitting problem and thus various techniques have been developed, including Iterative Proportional Fitting (IPF) and Combinatorial Optimization based techniques. The key shortcomings of these procedures include: (a) fitting of one contingency table, while there may be other solutions matching the available data (b) due to cloning rather than true synthesis of the population, losing the heterogeneity that may not have been captured in the microdata (c) over reliance on the accuracy of the data to determine the cloning weights (d) poor scalability with respect to the increase in number of attributes of the synthesized agents. In order to overcome these shortcomings, we propose a Markov Chain Monte Carlo (MCMC) simulation based approach. Partial views of the joint distribution of agent's attributes that are available from various data sources can be used to simulate draws from the original distribution. The real population from Swiss census is used to compare the performance of simulation based synthesis with the standard IPF. The standard root mean square error statistics indicated that even the worst case simulation based synthesis (SRMSE = 0.35) outperformed the best case IPF synthesis (SRMSE = 0.64). We also used this methodology to generate the synthetic population for Brussels, Belgium where the data availability was highly limited.



## 1. Introduction

Large-scale activity based travel demand and land use evolution models that take into account the individual agent decisions and interactions, are actively been developed in research and practice (Waddell, 2002; Miller and Roorda, 2003; Arentze and Timmermans, 2004; Balmer et al., 2006; Miller et al., 2011). These behavioral and market oriented models are an active tool for detailed impact forecasting of transportation, land use, environmental, and energy related policies.

Among other data, these simulations require at least a base year population of agents (households, families, and/or persons) and their attributes. These attributes are then used in various behavioral models estimated on a sample and implemented in the simulations for forecasting. The attributes are needed for not only the base year population, but also for the future years population. Agent level data on the complete population in a study area is almost never available–not including the few exceptions like Switzerland where the complete census is available for research. Instead in most cases a microsample called public use micro sample (PUMS) with out any high resolution location information is available. It

* Corresponding author. Tel.: +1 514 340 5121x4802.
*E-mail addresses:* bilal.farooq@polymtl.ca (B. Farooq), michel.bierlaire@epfl.ch (M. Bierlaire), rhurtubia@uchilefau.cl (R. Hurtubia), gunnar.floetteroed@abe.kth.se (G. Flötteröd).
[1] Tel.: +41 21 693 9327.

0191-2615/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trb.2013.09.012

may not have information on associations of agents either. Travel surveys conducted by governmental bodies (e.g. municipality) provide a sample for use. In addition to that census and travel surveys also provide aggregate level data at various zonal systems. There might also be various other bits and pieces of information that are available to the researcher about the population. These various sources that are partial views of the population are used to reconstruct it, using synthesis techniques. Future year population can either be generated using the endogenous demographic update mechanism or by generating new population every year using a synthesis technique. Demographic update is not the topic of this paper, but is extensively covered in Farooq et al. (2009) and elsewhere.

The existing population synthesis techniques focus on fitting a single contingency table to the available data. Using that as the ground truth the microsimulations are run to produce the outputs. In existing literature there is no investigation or discussion on the error that may be propagated forward via this approach due to: (a) incompleteness of the data (b) systematic and deliberate tampering of the data at source to conserve privacy (c) differences in the definitions, aggregation levels, scale, etc. (d) assumptions and short coming of the fitting procedures. Moreover, due to unavailability of the data on real population (ground truth) in most of the cases, there rarely has been a complete and systematic analysis done on the performance of the proposed techniques. In this context we propose a new approach that uses all the partial views of the joint distribution of the real population, available through various data sources, in order to draw the synthetic populations from it. We have access to the census on the real population of Switzerland, which we used for the comprehensive performance assessment. The proposed approach is able to overcome major issues faced by the existing approaches, while maintaining at least the same level of accuracy as the leading approaches outlined in the existing literature.

The rest of the paper is organized as follow: we first describe the types of available datasets that can be used in the synthesis. Existing literature is outlined and key shortcomings are discussed. We then formally introduce the problem statement and present our methodology to address the problem. Various performance comparison experiments and a case study are presented. In the last section we discuss key features of the proposed approach and present conclusions.

## 2. Available data sources

Traditionally, primary data sources to construct a synthetic population have been census and travel surveys. Other sources include: household spending survey, labor force survey, statistics from revenue agency, real estate cadaster etc. Although, they are rarely used in the existing literature. The information from these sources is available in two different forms: sample of individual agents and cross-classification tables. These data are associated to one or more spatial zoning systems.

### *2.1. Zoning systems*

The data is available at certain aggregations of space that is defined by a zoning system. The aggregation may be based on certain maximum density levels, physical obstacles (river, street etc.), and political boundaries. There may also be hierarchy of aggregations within each zoning system. For instance, in case of Canadian census the lowest level of zone is called dissemination area where 400–700 persons are living/working. One level above is the census tract where the limit is 2500 to 8000 persons. Further aggregations are census sub-division, division, and municipality, respectively. The zoning system also changes with time i.e. a discrimination area in year 2001 census, may have been divided into two in 2010 census, so as to satisfy the constraint on number of persons.

The travel surveys are available at the lowest level of aggregation called Traffic Analysis Zone (TAZ). TAZs are usually defined based on road network; its size may vary depending on the agency conducting the survey; and may not overlap with any of the census zoning system. Another zoning system that may be used is the postal code system.

### *2.2. Sample of individuals*

Statistics bureau of a country among other surveys also conducts periodic census of the entire population. The periodicity of this census range from 5 (in case of Canada) to 10 years (USA, Switzerland etc.). While the whole dataset is almost never available for the research (with some exceptions, like that of Switzerland), bureaus do provide a representative sample for public use. In North America, this sample of individual agents is called Public Use Micro Sample (PUMS) and in the UK and few other countries, Sample of Anonymised Records (SARs). In this paper we use the term PUMS. This sample is only available at a large spatial area (for instance, City of Toronto, London etc.), so as to make sure that the privacy of the individuals is protected. The size of sample may range from 1% to 5% of the total population. The sample may contain range of demographic and socioeconomic information on households, families, and persons. The exact location, income details, and some other details may be missing due to privacy concerns. Furthermore, census bureau may hide information on certain individuals, if they deem it to be exposing individuals' identity.

Another source of the sample is the travel survey, usually conducted by the urban regions, municipalities, counties, etc. The focus of this survey is on the travel demand patterns of agents and mode shares, but it also has some information on socieconomics and demographics of agents. There is more fluctuation here in terms of the details, size, and periodicity of

travel survey among the regions/countries. The periodicity of travel surveys of the neighboring regions may not coincide and it may also not coincide with that of census.

Various fitting based procedures (e.g. IPF) use the sample to initialize the contingency table, which is a multidimensional table representing the frequency of each combination of attributes categories (Beckman et al., 1996).

### 2.3. Cross-classification tables

At various zoning level (for instance, dissemination area and census tracts in case of Canada. Or sectors and communes in case of Belgium and France), Statistics bureau also releases the cross-classification tables for socioeconomics and demographics (for instance, income by age at sector level). These tables are usually 1–3 dimensional tables. Marginal distribution/counts of an attribute can be directly constructed from 1 dimensional table or by aggregating the higher dimension tables for each category of an attribute. They are used in fitting based procedures as control totals. The higher dimensional cross-classification tables can also be used to generate conditional-marginal distribution/count of one attribute on the other attributes that are present in the table. In the proposed methodology we make use of both marginals and conditional-marginals distribution/counts.[2]

Higher dimensions tables are not released due to privacy concerns–the cell values may be small enough to reproduce the population. In the available tables, if the cell values were very small, usually they had been rounded off to zero by the bureau. There is also a random rounding of values done by the Statistics bureau, so as to make it further difficult to reproduce the baseline population.

Details on various techniques used to anonymize the census data can be found in Sweeney (2002),Dalenius and Reiss (1982), Brand (2002), and else where. The point to be noted here is that the available data has already been treated with various such anonymization techniques and thus a population synthesized by fitting based techniques may not be completely representative of the real population.

## 3. Literature review

Dealing with the incomplete/missing multivariate dataset has extensively been studied in statistics and applied probability literature. Schafer (1997) presented various approaches to create imputations on multivariate datasets with missing values. The imputation techniques can be divided into two groups: joint modeling and fully conditional specification (Buuren, 2007). In joint modeling, a parametric distribution $\pi(X|\theta)$ is assumed for the data $X$. With the appropriate prior for the parameter $\theta$, Bayesian framework is used to draw from the posterior predictive distribution $\pi(x^{mis}|x^{obs})$. Here the assumption is that the data gathering mechanism is ignorable. Common forms of parametric distributions used are multivariate normal, log-linear, and the general location model (Holford, 1980). One of the most commonly used specification, especially in transportation research is the log-linear model. In this case, the cell probability in a contingency table is assumed to be made up of two components (Schafer, 1997): (a) multiplicative effect of each variable (b) effect of association among the variables. Using specific assumptions and eliminating certain terms (the details of which can be found in Deming and Stephan (1940) and Schafer (1997)) bring in the equality constraint called odd-ratio on the contingency table. Using this constraint and the aggregate marginals an optimization based classical fitting method has been developed by Deming and Stephan (1940). This method is commonly known as Iterative Proportional Fitting (IPF) method and was introduced in transportation literature by Duguay et al. (1976) to synthesize households survey data. It was later used by Beckman et al. (1996) to create synthetic population for TRANSIMS (LaRon et al., 1996) using census cross tabulations and sample. Since then IPF has been the workhorse for synthesizing population for activity based travel demand and land use models.

The IPF based techniques involve two step process. In the fitting step, a Contingency Table (CT) is fitted to the available marginals. PUMS is used to initialize the CT. The underlying assumption here is that the sample represents the true correlation structure among the attributes. And the odd ratio property of IPF will ensure that the fitted CT will maintain this structure. Final fitted CT is generated by iteratively adjusting the cell values so as to minimize deviation from the marginals of the attributes. For a fitting to work the attribute should be present in the sample and its marginals should be available. IPF involves maintaining $\Pi_{i=1}^{I}(k_i)$ cells in memory, where $I$ is the number of attributes, $k$ represents categories. Note that all the attributes have to be discrete and with limited categories. Fitting for large number of attributes quickly becomes computationally and memory-wise expensive. Pritchard and Miller (2012) used sparse matrix manipulation techniques, but was limited to synthesizing 8 attributes. IPF fitting cannot differentiate between structural and sampling zero. In the literature, various methods have been suggested to avoid sampling zero issues, including Guo and Bhat (2007),Auld et al. (2009), and others. Recent literature in IPF based fitting (Arentze et al., 2007; Ye et al., 2009) has also concentrated on simultaneous fitting of CTs for different types of agents (for instance, household and persons together). Schafer (1997) proposed a Bayesian procedure for computing the fitted CT. Here the cell values are treated as random variables with constrained Dirichlet priors. A Markov Chain Monte Carlo (MCMC) process is designed to retrieve a realization of CT from the posterior Dirichlet. While this method is a step forward in terms that the cell values are treated as RVs, the method has limitations. First the assumption of the prior distribution is very restrictive–it has to be such that the posterior can be retrieved. Secondly, the synthesized

[2] From here on we will use the term *marginal* to represent the *marginal distribution/counts* and *conditional* to *conditional-marginal distribution/counts*.

population is still the result of one realization of the fitted cell values out of posterior. To our knowledge, this extension of IPF has not been explored in the transportation literature.

The second step of IPF based techniques involves creating the synthetic population using the fitted CT. This is done by cloning/replication of the sample, based on cell weights. The fractions are incorporated in the synthetic population using Inverse Transform Sampling in a Monte Carlo simulation (Beckman et al., 1996).

Another fitting based method that has been used to some extent in the transportation literature is based on Combinatorial Optimization (CO). In this method a weight ($w = \{0, 1\}$) is added to the sample. For each zone the population is synthesized by replicating the sample and optimizing these weights in order to minimize the difference from the zonal marginals (Voas and Williamson, 2000; Lu, 2011). Openshaw and Rao (1995), Williamson et al. (1998), and Voas and Williamson (2000) used Simulated Annealing (SA) as an optimization tool to produce the synthetic population. Ryan et al. (2009) compared the CO based methods with the IPF and concluded that CO produced lesser variance. Another advantage is that it has lower memory requirements, though the convergence time of CO based techniques is very high. In case of SA based optimization, the process can be very wasteful and time consuming. The resulting synthetic population from CO based methods is still cloning/replication based. Here too, more attributes mean more constraints that will result in complicated optimization. Like IPF, CO also requires the information on attributes both in the sample and marginal level. There is also, no guarantee that the optimal solution in terms of matching marginals can be achieved in reasonable amount of time.

Barthelemy and Toint (2013) developed a sample-free synthesis method that is based on various discrete and continuous optimization procedures. Their method is a step forward in fitting based methods, as it gives more flexibility in terms of data needs. The developed procedure however involves various complicated and hierarchical fitting steps (for each aggregation level where the data is available), entropy maximization, tabu search, and various ad hoc matching rules. The approach does not guarantee the simultaneous matching of the control totals for both households and persons. The method was applied to synthesize Belgian population with limited attributes. The generalization of the methodology, so as it can be used for other cases is not very clear.

One of the early examples of directly drawing from distribution can be found in TORUS (Miller et al., 1987), which microsimulated the households' location choice decision in Toronto area. In TORUS, to generate an agent's attributes, all the available distributions were sampled independently while making sure that there were no logical inconsistencies among the realized attributes (for instance, a 2 years old cannot have a university degree). It used the zonal marginals and partial conditionals available from the census. A comprehensive literature review on population synthesis methods can be found in Müller and Axhausen (2011), Lu (2011), and Pritchard (2008).

There are few major issues that can be pointed out in the existing approaches. First of all, it is assumed that these approaches are robust enough that they can reproduce one contingency table that is representative of the real population. Given the fact that the data is incomplete and has purposely been tampered with sophisticated anonymizing techniques, the assumption does not remain valid. Second, usually these techniques fit the marginals and controlling for the additional conditionals/joint distribution results in combinatorial issues. For instance, if *A*, *B*, and *C* are to be synthesized from marginal on *A*, *B*, *C* and joint distribution on *AB*. This will require adding another dimension in the contingency table so that we can fit for *AB* in addition to *A*, *B*, and *C*. Moreover, it has to be done for each additional constraint, which can become an issue when agents with large number of attributes have to be synthesized. Third, there are two distinct levels of error that are introduced by these methods: (a) matching of the marginals only, may result in distortion of the underlying joint distribution during the fitting step (b) based on the weights from the fitted CT, the actual realization of the synthetic population is achieved by replication and running Monte Carlo simulation on the PUMS. This may further add to the error. Effectively, the population is synthesized by increasing/decreasing the mass at known points in the attribute space (usually known from the PUMS), rather than reproducing a continuous distribution surface in that space. A consequence of which is that it is not possible to synthesize an agent with attribute values present in the real population, but missing in the sample (such can occur more frequently in continuous attributes). Fourth, these methods require a well defined form of raw data i.e. (a) big enough and well representative microsample (b) the zonal control totals on the attributes that are needed to be synthesized. Additionally, they require that the information on the attributes is present at both sample and marginal level. Lastly the convergence is very slow–especially if there is a zero value for a category in the marginal, but a non-zero value for any cell associated to it in the initialized contingency table (Bishop et al., 1975; Brown and Fuchs, 1983). Note that this situation may occur if the sample is not consistent.

## 4. Methodology

### 4.1. Problem definition

In a spatial region under consideration at any point in time there exists a true population. The individual agents in the population are characterized by a set of attributes $X = (X^1, X^2, \ldots, X^n)$.[3] These attributes may be discrete (e.g. marital status) or continuous (e.g. income). In the true population, they have a unique joint distribution, represented by $\pi_X(x)$. We do not have access to $\pi_X(x)$ and most likely it is hard to draw from. Instead, only partial views of $\pi_X(x)$ from various types of data sources are

[3] Where $n$ is the number of attributes we are interested to synthesize.

available. These partial views are in the form of marginals, conditionals, and samples. We want to develop a synthesis procedure that lets us use these views to draw a synthetic population as if we were drawing from $\pi_X(x)$. At the same time, it can be ensured that the empirical distribution $\pi_{\widehat{X}}(\hat{x})$ of $\widehat{X}$ in the realized synthetic population is as close to $\pi_X(x)$ as possible.

### 4.2. Simulation based approach

As mentioned in sub Section 4.1, we are interested in synthesizing independent populations by drawing agents from the joint distribution of the attributes in the real population, instead of fitting a single optimization based solution. This distribution is not known and is hard to directly draw from. Markov Chain Monte Carlo (MCMC) methods are computer based simulation techniques that can be used to simulate a dependent sequence of random draws from very complicated stochastic models/processes (Hastings, 1970). These methods provide flexibility in terms of using various data sources at various spatial scale; bring in prior knowledge in a systemic way; wherever the data is not available, implement assumptions in a coherent manner; and are computationally and memory-wise robust. These techniques have been extensively used in various other domains including physics, image processing, etc.

Here we propose using MCMC techniques to draw from the real population distribution $\pi_X(x)$ to obtain a synthetic population, instead of using the conventional fitting procedures. Using MCMC techniques can overcome the major shortcoming of fitting based techniques that are pointed out in Section 3, while maintaining at least the same quality of the synthetic population, with at most the same amount of data as the fitting based techniques require. In this section, we first describe how such a MCMC technique can be operationalized. We then design and implement comparative experiments to compare its performance with the IPF. The last part of the section discusses the results and implication.

#### 4.2.1. Using Gibbs sampling for synthesis

As the joint distribution $\pi_X(x)$ of attributes $X$ is unknown and is highly complex to directly draw from, we propose to use Gibbs sampling to generate the synthetic population. Gibbs sampler is a MCMC method that uses the conditionals $\pi(X^i|X^j = x^j$, for $j = 1, \ldots, n$ & $i \neq j) = \pi(X^i|X^{-i})$ for $i = 1, \ldots, n$ to simulate drawing from the joint distribution $\pi_X(x)$ (Geman and Geman, 1984). The key challenge here is to prepare the conditional distributions of the attributes using all the available data about the attributes of the population.

#### 4.2.2. Preparation of conditionals

In the straightforward case, these conditionals can be counts by category for each attribute. This may be available from the census zonal statistics table or can be directly constructed from the PUMS. In practice though, it is rarely the case where the full-conditionals in this form are available for all the attributes in $X$. Here, we can use parametric models to construct the conditional distributions. The flexibility of using such parametric models is that the data from various sources (PUMS, zonal marginals, etc.) can be combined to estimate the parameter values. In Section 5, we show how discrete choice models can be estimated, where the dependent variable is used from PUMS; some of the independent variables are from the sample as well, and other independent variables are from the zonal marginals.

#### 4.2.3. Dealing with incomplete conditionals

There may be cases where not enough data is available to construct the full conditional for an attribute over all the other attributes. Suppose that in $\pi(X^1|X^{-1}) = \pi(X^1|X^{(2\ldots k)}, X^{((k+1)\ldots n)})$ only the incomplete conditional $\pi(X^1|X^{(2\ldots k)})$ is available. In such a situation, we can assume the conditional independence of $X^1$ on $X^{((k+1)\ldots n)}$, given $\pi(X^1|X^{(2\ldots k)})$. Thus $\pi(X^1|X^{-1}) = \pi(X^1|X^{(2\ldots k)})$. In the worst case, where only marginals are available, we can use $\pi(X^1|X^{-1}) = \pi(X^1)$. Furthermore, we can also use the domain knowledge about the incomplete part of the conditional to construct full conditionals. This may result in a case where $\pi(X^1|X^{(2\ldots k)}, X^{((k+1)\ldots n)} = a) = \pi^a(X^1|X^{(2\ldots k)})$ and $\pi(X^1|X^{(2\ldots k)}, X^{((k+1)\ldots n)} = b) = \pi^b(X^1|X^{(2\ldots k)})$. To give a concrete example, for instance, if we do not have the data on the head of the household conditioned upon age, we can assume that the probability of a child to be head of the household as zero, while making a different assumption for adults.

#### 4.2.4. Realization of synthetic population

Using the full and consistent conditionals, if the Gibbs sampler is ran for an extended amount of iterations, it eventually reaches a stationary state. At that point any draw will be as if the draw was from $\pi(x)$ (Train, 2003). To avoid the correlation between the consecutive draws, certain number of draws between two recorded draws are skipped. Now using this mechanism, a synthetic population can be realized by simply drawing the number of individuals equaling the size of the required population. Note that we can realize any number of synthetic populations $\widehat{X}$. Depending on the quality of the data, the distribution $\pi_{\widehat{X}}(\hat{x})$ of $\widehat{X}$ in the resulting synthetic population will be as close to $\pi_X(x)$ as possible. Moreover, two independently generated populations will have similar statistics (with some simulation noise), but the agents in the two populations may not be the same. For computational efficiency, one can store the warmed up state of the Gibbs sampler. So every time a new synthetic population is needed, the sampler can directly start from the previously stored state. Another option can be that a large pool of agents are being drawn and stored using the warmed up Gibbs sampler. Later whenever a realization of synthetic population is needed it can be drawn from the pool.

### 4.3. Experiments on a real population: Swiss census

As most of the synthesis methods in recent literature are primarily based on IPF (for instance, Guo and Bhat, 2007; Ye et al., 2009; Auld et al., 2009; Pritchard and Miller, 2012), this paper is restricted to comparison of the simulation based methodology to IPF only. Comparison between IPF and combinatorial optimization based fitting can be found in Ryan et al. (2009). The experiments here are designed as such that both methods (IPF and simulation based) are provided with the same amount of data about the real population. The output of each method is then analyzed in terms of how good the marginals and joint distribution of the real population are reproduced. In the end, Standardized Root Mean Square Error (SRMSE) based goodness of fit test is performed on each case and results are compared.

#### 4.3.1. Data description and preparation

The Swiss census for year 2000 was used as the data source for the experiments, where we had access to the attributes of real population. We selected the spatial area associated with postal code CH1004 (western side of the city of Lausanne) as the testbed. The population of CH1004 in year 2000 was 28,533 persons. For synthesis experiments, we selected four attributes: age (8 categories), sex (2), household size (6), education level (4).

Based on what is commonly available to the researcher (see Section 2 for details), two different set of information were made available to both methods: (a) microsample (b) aggregate conditionals (for instance, $\pi(age|sex, hhld\_size, edu\_level)$, $\pi(sex|age, hhld\_size, edu\_level)$, ... ) in the form of cross-tabulations for the population in CH1004. The experiments were designed such that on one side the size of the sample was changed while on the other side the completeness of the conditionals was changed.

The sample sizes that were used for the experiments are 1%, 3%, 5%, 10%, and 20%. Depending on the country the PUMS is available at 1%, 3%, and 5%. However, we also wanted to test the effect of larger samples, so we included the 10%, and 20% sizes in the samples as well. We prepared these samples by drawing from the real population of CH1004 using a uniform selection probability and without replacement of the already sampled individuals.

For the conditionals, experiments were designed such that we started from full-conditionals (every attribute conditioned upon other). Then sequentiality removing the sex related information from the conditionals of other attributes (for instance, from $\pi(age|sex, hhld\_size, edu\_level)$ to $\pi(age|hhld\_size, edu\_level)$). It was assumed that for conditionals where the sex is missing the distribution is independent of sex given other attributes. All the conditionals were prepared directly from the real population of CH1004 by counting the agents for each category of the attributes.

A complete list of available information can be found in Table 1 and 2. For IPF, not all the 20 combinations of sample and conditionals were needed to be tested. The reason being that IPF in its basic form converts all the conditionals to marginals. So, for IPF we performed 5 experiments with varying sample sizes and marginals generated by collapsing full conditionals (*FullCond*). For simulation based methods, all the 20 combinations could have been tested, but here we tested for only 4 different conditionals without using any of the samples. Note that it means that lesser information was used in the simulation based method as compared to the IPF. Later in Section 5, we have illustrated how both sample and partial conditionals/marginals can be combined to generate input for the simulation based method. Next three sub-sections describe the individual details of these experiments, results, and comparison.

#### 4.3.2. IPF based synthesis

We used the standard two step process outlined in Beckman et al. (1996) to generate synthetic population using IPF procedure (i.e. fitting, and cloning process). For fitting process, the given sample was used to initialize the contingency table (CT). We knew the location of structural zero cells from the real population. The sample zero cells were thus initialized to a small value. This made sure that the sampling zero cells can evolve during the fitting process. The CT contained 384 cells that represents the combination of categories of the four attributes. As the standard fitting process uses marginals only, the conditionals were converted to marginals for all 4 attributes. Note that IPF can also use conditionals, but it very quickly becomes a combinatorial issue. The CT has to be fitted not only to the conditionals but also to the marginals for the attributes in the conditionals. Moreover, all the attributes-combinations have to be present in sample too. To get a close fit between generated and available marginals, a minimum absolute difference of 0.00001 and maximum iteration size of 10 billion was used.

In the second step, cloning of the sample using the fitted CT was performed. Fractions in the CT table were dealt with running a Monte Carlo simulation as suggested by Beckman et al. (1996). Note that, for IPF, the completion of conditionals

**Table 1**
List of available sample sizes.

| No. | Sample size (%) |
|---|---|
| 1 | 20 |
| 2 | 10 |
| 3 | 5 |
| 4 | 3 |
| 5 | 1 |

**Table 2**
List of available conditionals.

| No. | ID | Conditionals |
|---|---|---|
| 1 | *FullCond* | $\pi(age\|sex, hhld\_size, edu\_level)$<br>$\pi(sex\|age, hhld\_size, edu\_level)$<br>$\pi(hhld\_size\|age, sex, edu\_level)$<br>$\pi(edu\_level\|age, sex, hhld\_size)$ |
| 2 | *Partial_1* | $\pi(age\|,hhld\_size, edu\_level)$<br>$\pi(sex\|age, hhld\_size, edu\_level)$<br>$\pi(hhld\_size\|age, sex, edu\_level)$<br>$\pi(edu\_level\|age, sex, hhld\_size)$ |
| 3 | *Partial_2* | $\pi(age\|,hhld\_size, edu\_level)$<br>$\pi(sex\|age, hhld\_size, edu\_level)$<br>$\pi(hhld\_size\|age, edu\_level)$<br>$\pi(edu\_level\|age, sex, hhld\_size)$ |
| 4 | *Partial_3* | $\pi(age\|,hhld\_size, edu\_level)$<br>$\pi(sex\|age, hhld\_size, edu\_level)$<br>$\pi(hhld\_size\|age, edu\_level)$<br>$\pi(edu\_level\|age, hhld\_size)$ |

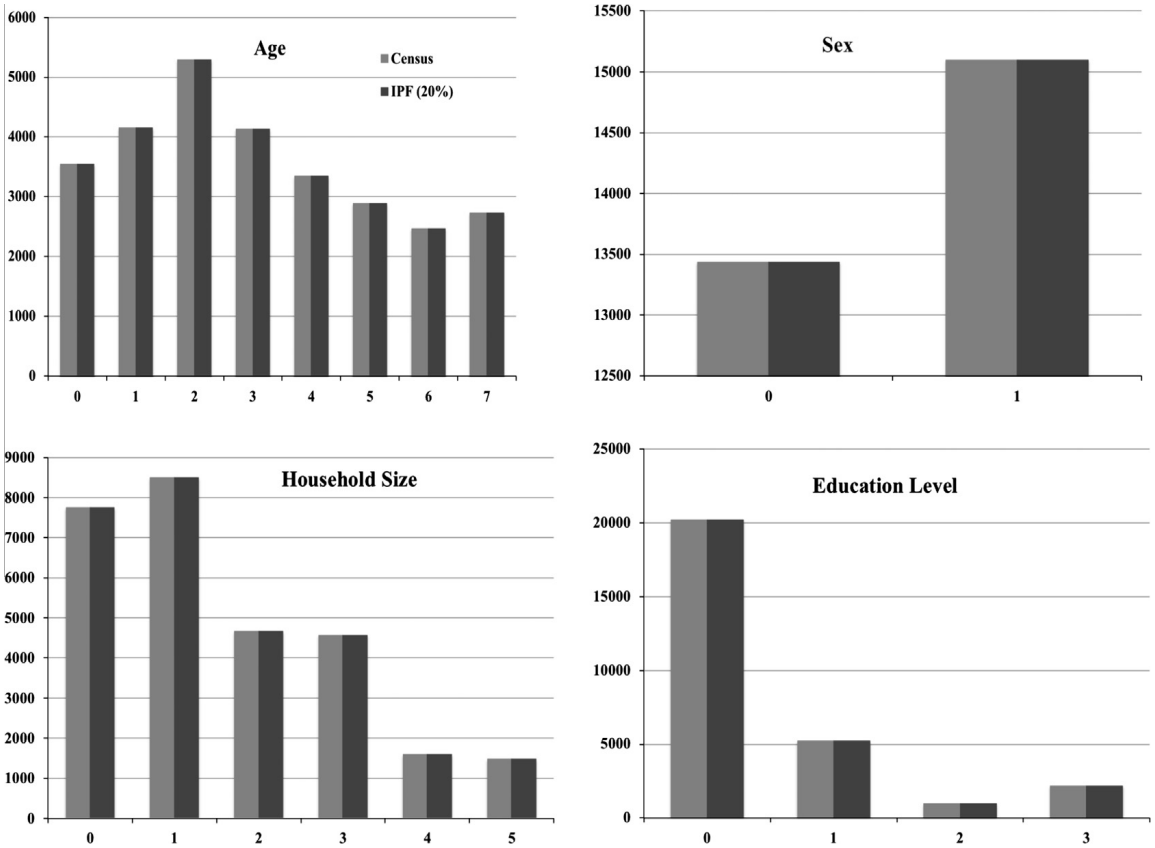


**Fig. 1.** Comparison of marginals for the four attributes (IPF).

does not matter, as they have to be converted to marginals anyway. Hence, we focused our analysis on size of the sample only.

Fig. 1 shows the comparison between the census and IPF with 20% sample and the marginals of the four attributes. The IPF produces a near perfect fit for the marginals. It is because the CT fitting algorithm is specifically designed to iteratively reduce the difference between available marginals and that of the CT. However, if we compare the joint distribution of the agents synthesized by IPF to the distribution of real population in the census (Fig. 2), we see more variation. With the value of 0.97 for the slope, IPF is under predicting the distribution. The $R^2$ value is 0.91, which indicates that even with as large as 20% sample, there is still a 9% variation in the population that is not reproduced in the synthetic population. By looking

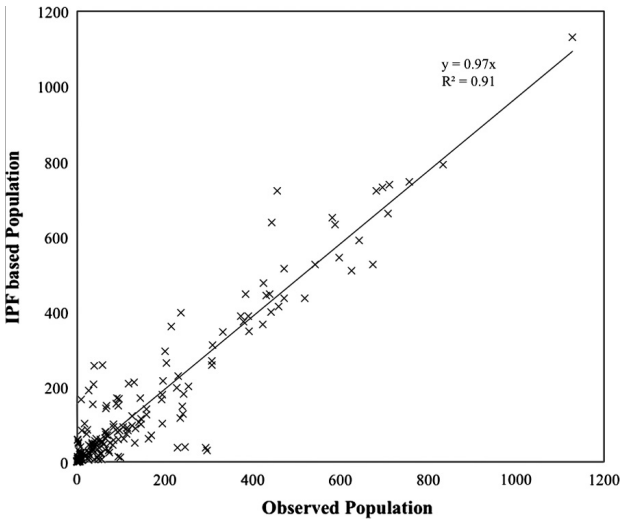


**Fig. 2.** Fit between real and IPF population (20% sample).

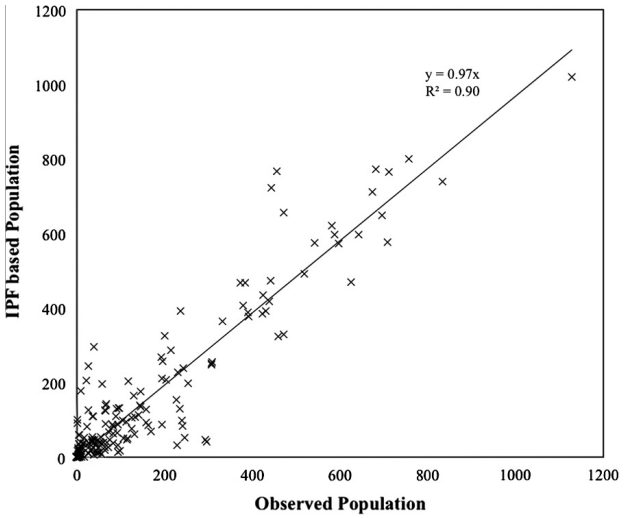


**Fig. 3.** Fit between real and IPF population (10% sample).

closely it can be observed that the fit is worst for the points in the four dimensional attributes space, where the probability values are low. It means that the type of agents who are very few in the real population (for instance, 100 years old male with a university degree) are not very well reproduced in the synthetic population.

By decreasing the sample size to 10% there is no effect on the slope of the fit (Fig. 3), but the $R^2$ value is decreased marginally. It can thus be concluded that there is not much of a gain in using a 20% sample than 10%, as both samples are well representative of the population. Here again, the fit is worst for the combination of attribute categories for which the probability of existence is low. Moreover, the fit for other combinations decreases as well.

Note that in most of the cases, the PUMS or any other microsample is only available for 5% or less of the population. Figs. 4–6 show the fits for the 5%, 3%, and 1% sample with the real population, respectively. For 5% sample the fit is still comparable to larger samples, but for 3% and 1% the scatter in the fit significantly increases. The most probable reason behind it is the fact that IPF is only focusing on fitting the marginals. So, it may be able to reproduce marginals, but if the sample is not representative enough, there is no guarantee of reproducing the actual joint distribution of the population. These results suggest that for IPF to reproduce the joint distribution reasonably, a microsample size of at least 5% is needed.

Another important point to note here is that in all the above experiments, the information on all the attributes was present at both sample and marginals level. In practical cases, usually this is not the case. It is thus expected that the quality of the synthesized joint distribution will decrease further. A similar experiment can easily be designed to prove this point, by

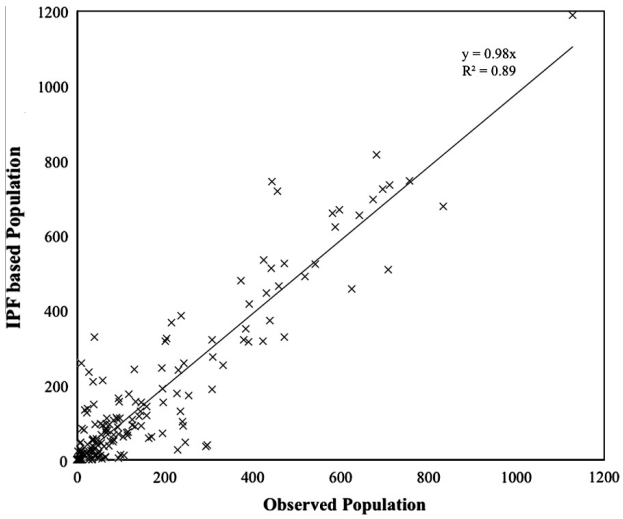


**Fig. 4.** Fit between real and IPF population (5% sample).

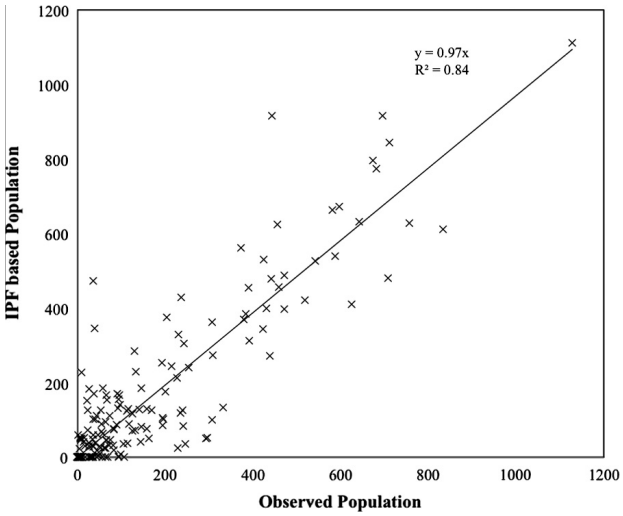


**Fig. 5.** Fit between real and IPF population (3% sample).

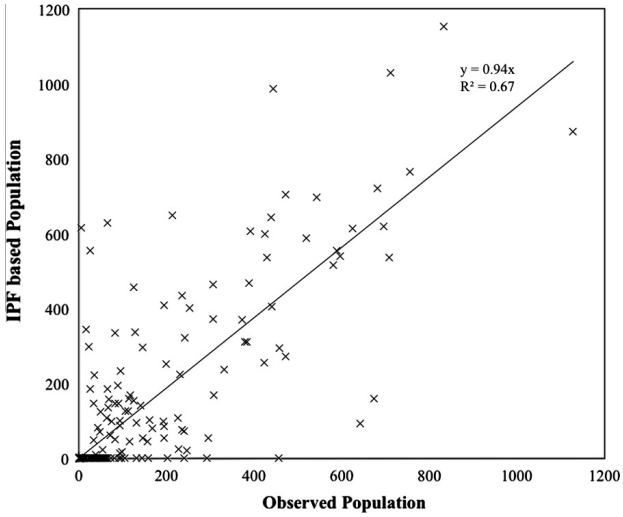


**Fig. 6.** Fit between real and IPF population (1% sample).

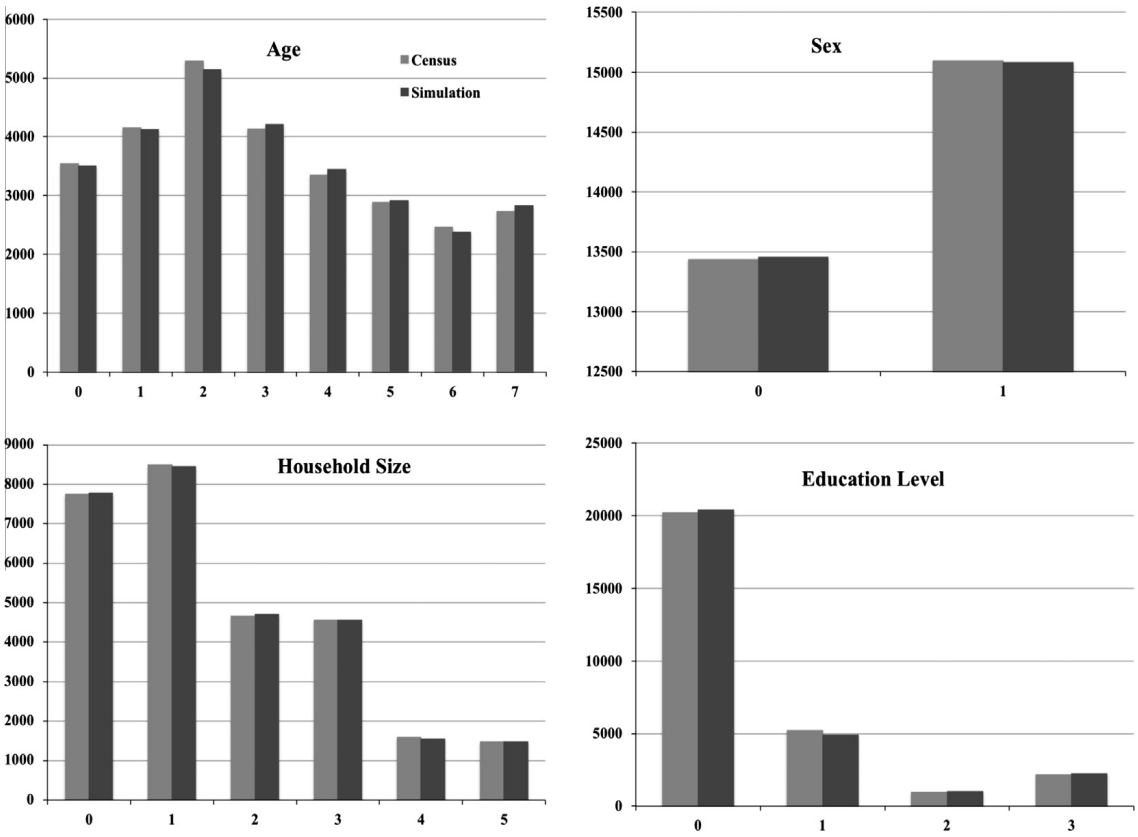


**Fig. 7.** Comparison of marginals for the four attributes (simulation).

using for instance only 3 marginals (assuming that 4th one is missing) and then running IPF fitting on the 3-dimensional CT.[4] In these experiments we made sure that there are no sampling zeros in the initial CT. In practice, again this is not the case. This is another source of further degradation of the synthesized population.

*4.3.3. Simulation based synthesis*

Exactly the same amount of data (samples and conditionals) was provided to the Gibbs sampler, as been previously provided to the IPF procedure. In the following results, the Gibbs sampler ignored the information in the sample and used the conditionals only. It thus used lesser information than IPF to generate the synthetic population. Later in the Section 5, we have illustrated that sample and aggregate information (for instance, partial conditionals or marginals) can also be fused together in various ways to produce better conditionals. Thus making the best use of all the information available to us from various sources.

In order to generate the results, we first warmed up the Gibbs sampler to a steady state by running it for 20,000 iterations. To avoid warming up every time a realization is needed, we drew a pool of 1 million agents. The agents were drawn from every 20th iteration after warmup, so as to avoid any correlation between the successive draws. This pool was then used to extract 20 populations for CH1004 (28,533 agents) and the results were generated based on the averages from them. Fig. 7 shows the marginals generated using full conditionals (*FullCond*):

$$\pi(age \mid sex, hhld\_size, edu\_level),\ \ \pi(sex \mid age, hhld\_size, edu\_level),$$
$$\pi(hhld\_size \mid age, sex, edu\_level), \text{ and } \pi(edu\_level \mid age, sex, hhld\_size).$$

Like IPF, simulation based synthesis is also able to reproduce the actual marginals for the four attributes. The fit between real population and synthesized population for full conditionals (Fig. 8) is practically perfect with slope of 1 and $R^2$ value of 0.999. The minor deviations in the values are due to the randomness in the simulation process. Unlike IPF, simulation is able to minimize deviations for both large and small probability values.

We also investigated the variance in the joint distributions from various realized populations. Fig. 9 shows the plot for the joint distribution of real population (black color), and 20 realizations of the synthesized populations superimposed upon it.

[4] In such an experiment performed but not reported here, it was observed that the fit decreased by 30%.

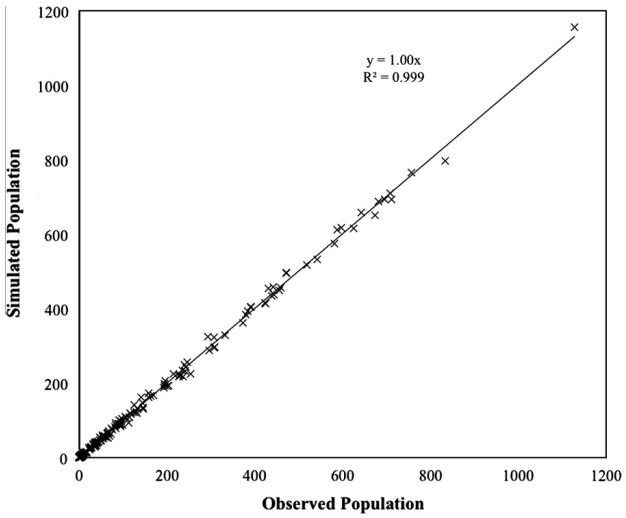


**Fig. 8.** Fit between real and simulation based population (*FullCond*).

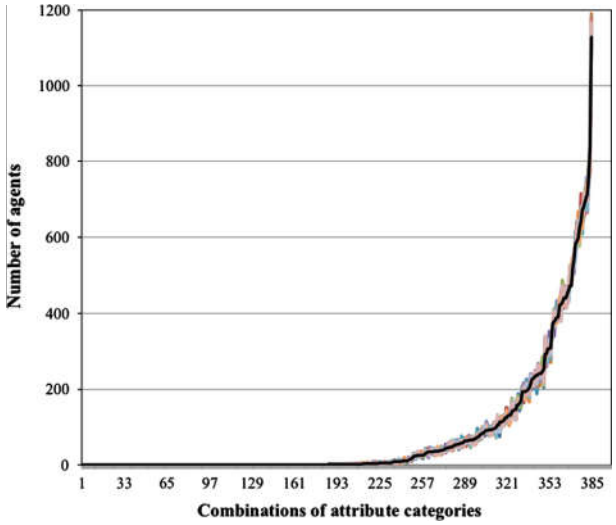


**Fig. 9.** Joint distribution of four attributes (20 simulation runs).

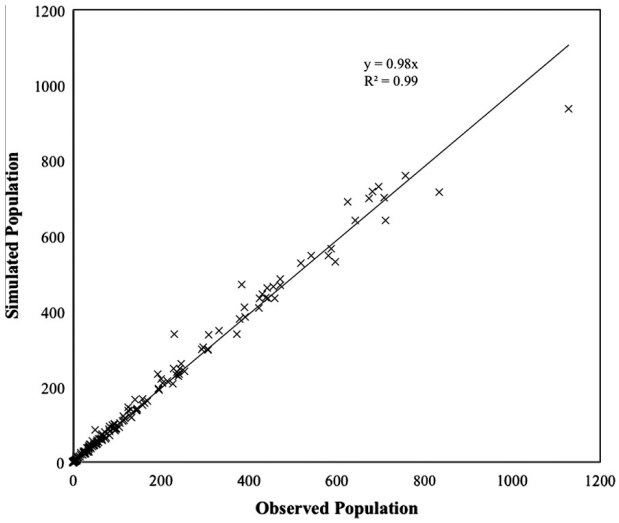


**Fig. 10.** Fit between real and simulation based population (*Partial_1*).

Each point on *x*-axis represents the combination of categories from 4 attributes, while *y*-axis represents the count. The values are sorted in ascending order, based on the real population. It can be observed that all the realizations have a close match among them and also with the real population.

In practice though, it is hardly the case that the full conditionals are available. In fact, there is a deliberate attempt to keep the conditionals incomplete in order to avoid any privacy concerns. Therefore, we also experimented with synthesizing the population using incomplete conditionals. Fig. 10 shows the fit of the synthetic population to the real population for partial conditionals (*Partial*_1):

$$\pi(age \mid hhld\_size, edu\_level),\ \pi(sex \mid age, hhld\_size, edu\_level),$$
$$\pi(hhld\_size \mid age, sex, edu\_level), \text{ and } \pi(edu\_level \mid age, sex, hhld\_size).$$

The simulation under predicts the slope by 2%, and the fit remains good ($R^2$=0.99). The smaller probabilities combinations were also predicted well. In the third experiment, we used further depleted conditionals (*Partial*_2):

$$\pi(age \mid hhld\_size, edu\_level),\ \pi(sex \mid age, hhld\_size, edu\_level),$$
$$\pi(hhld\_size \mid age, edu\_level), \text{ and } \pi(edu\_level \mid age, sex, hhld\_size).$$

Fig. 11 shows the fit. Again, the slope is very close to 1 and the fit is very high. In the fourth experiment, we further depleted the conditionals by not having any attribute conditioned upon the sex (*Partial*_3):

$$\pi(age \mid hhld\_size, edu\_level),\ \pi(sex \mid age, hhld\_size, edu\_level),$$
$$\pi(hhld\_size \mid age, edu\_level), \text{ and } \pi(edu\_level \mid age, hhld\_size).$$

In Fig. 12 one can observe that there is no further depletion in the fit. It is because of the fact that our assumption that education level's partial conditional is uniform along age is consistent with the actual situation in the population. This shows the flexibility of the approach. Especially, in cases where even if the information is incomplete, with the proper assumptions and domain knowledge, we can still synthesize a good population.

We also analyzed the ability of simulation base procedure in terms of reproducing the correlation structure (including higher order correlations). Here too, a good fit was observed between real and simulated population. Appendix A reports the correlation, 1st and 2nd order partial correlation while in Appendix B the first seven raw standardized moments are reported.

#### 4.3.4. Statistical comparison of the results

In literature the performance of synthesis procedures has been assessed using the Standardized Root Mean Square Error (SRMSE) (Müller and Axhausen, 2011; Pritchard and Miller, 2012). For this purpose, in most of the cases only marginals or conditionals are used to assess the fit between real and synthetic population. Here we have access to the joint distribution so we evaluated the fit to that. SRMSE is defined in terms of the distance from the actual distribution (Pitfield, 1978). It can be computed as:

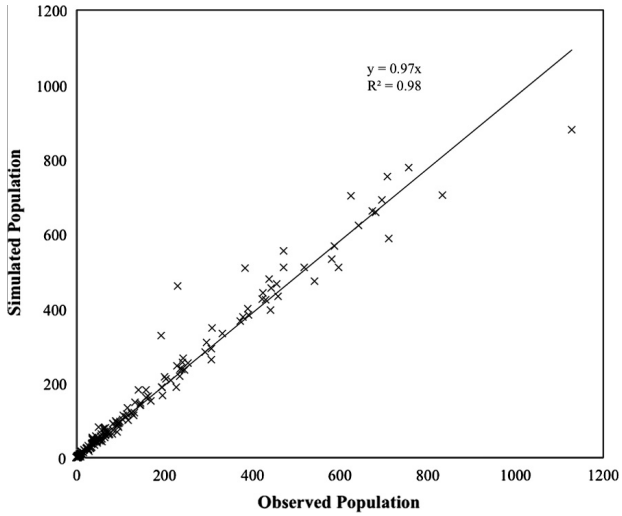


**Fig. 11.** Fit between real and simulation based population (*Partial*_2).

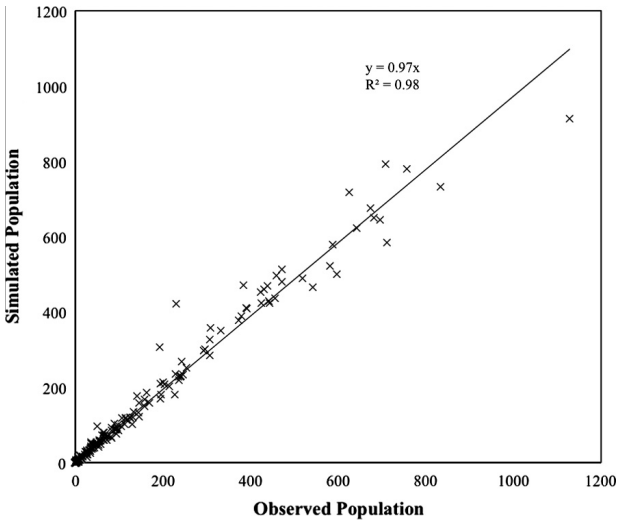


**Fig. 12.** Fit between real and simulation based population (*Partial_3*).

**Table 3**
Goodness of fit statistics (standardized root mean square error).

| Input | IPF | Simulation |
|---|---|---|
| 20% *Sample* | 0.637 | – |
| 10% *Sample* | 0.708 | – |
| 5% *Sample* | 0.750 | – |
| 3% *Sample* | 0.910 | – |
| 1% *Sample* | 1.420 | – |
| *FullCond* | – | 0.130 |
| *Partial_1* | – | 0.240 |
| *Partial_2* | – | 0.340 |
| *Partial_3* | – | 0.350 |

$$SRSME = \frac{\left[\sum_{i=1}^{m}\cdots\sum_{j=1}^{n}(R_{i\ldots j}-T_{i\ldots j})^2/N\right]^{1/2}}{\sum_{i=1}^{m}\cdots\sum_{j=1}^{n}(T_{i\ldots j})/N}$$

where $N$ is the total number of agents; $R_{i\ldots j}$ is the number of agents with attribute values $i\ldots j$ in the population synthesized; and $T_{i\ldots j}$ is the number of agents with attribute values $i\ldots j$ in the actual population. A value of zero means perfect match, while the higher values represent a bad fit. From Table 3, one can clearly observe that even the 20% sample based IPF population (SRMSE = 0.637) is outperformed by the population generated by simulation for 3 out of 4 conditionals being incomplete (SRMSE = 0.350). We also tested, but have not reported the extreme case where only four marginals and no sample is available. In that case, both methods give the same statistical fit.

While the same amount of data was available to both methods, IPF is first of all not able to capitalize on the information in the conditionals (it converts them to marginals). Secondly, it is overly dependent upon the sample to keep the correlation, while fitting marginals only. Simulation based procedure however, in these experiments is using lesser information (no information from the sample) and even then is able to outperform the IPF in terms of reproducing the joint distribution.

Furthermore, the simulation based procedure is more flexible in terms that information coming from sample or any other sources can be fused together using a model or sampling procedure (e.g. Metropolis Hasting sampling) to draw from the full conditionals. This has a very strong positive implications in the cases where the data is very limited and all the available data has to be utilized in the most efficient way possible. In the next section we present such case and show that simulation based procedures are able to synthesize population even with very limited data availability.

## 5. Application: greater Brussels area

The proposed methodology is implemented for a real case study, where a synthetic population is generated for the base year of an integrated land use and transport model for the region of Brussels. The agents to synthesize in this case were households ($h$) and for this the 2001 Census of Belgium and a household survey was available from Hubert and Toint (2002). The area of study consists of 151 communes that are further divided in 4945 sectors ($i$). The Census statistics contained aggregate information for the 1.2 million households of the area of study. The household survey provided detailed

**Table 4**
Household attributes.

| Attribute | Levels |
|---|---|
| Income level of the household ($inc_h$) | 1 (0–1859 Euros)<br>2 (745–1859 Euros)<br>3 (1860–3099 Euros)<br>4 (3100–4958 Euros)<br>5 (>4959 Euros) |
| Household size ($size_h$) | 1, 2, 3, 4, 5+ |
| Number of children ($children_h$) | 0, 1, 2+ |
| Number of workers ($workers_h$) | 0, 1, 2+ |
| Number of cars ($cars_h$) | 0, 1, 2, 3+ |
| Number of people with university degree ($univ_h$) | 0, 1, 2+ |
| Dwelling type ($v_h$) | House (3 types), apartment |
| Sector (i) | 4945 |

information for a sample of 1367 households (approximately 0.1% of the total household population). For the land use model, the synthetic households needed to be described in terms of size, number of workers, number of children, car ownership, education level and income level. Table 4 describes the discrete levels for each of the required household attributes.

All of the attributes were available as marginals at the sector level from the Census and as variables in the travel survey. Additionally, the travel survey indicated the sector ($i$) in which a household's residence is located.

Because of the relatively small size of the detailed sample, complete conditionals cannot be generated directly from counts. Therefore, a set of models describing the conditional probabilities of some attributes was generated. These models allowed to explain the value of an attribute as a function of the rest of the attributes of the household (in the travel survey), while simultaneously relating them with marginal distributions associated with the location ($i$) of the household. Discrete choice models were estimated for the dwelling-type, car-ownership, education-level and income-level attributes. The conditionals for the three remaining attributes (household size, number of children and number of workers) were calculated directly from counts in the travel survey. The choice models explained the level of a particular household attribute ($x_h$) as a function of other household attributes and variables describing spatial attributes ($x_i$).

Estimation results for the four discrete choice models are shown in Appendix C. Table 5 shows the relationships between attributes that were modeled and the household and sector attributes that were used as explanatory variables. The introduction of spatial information produced a richer set of conditional distributions, which was equivalent to having a different conditional for each sector.

A pool of approximately 100 million households was generated. The synthetic population of approximately 1.2 million households was then realized by sampling out of it. Because the land use model required a single household per dwelling unit, the sample was performed by sector and by dwelling type. Therefore the synthetic population matched perfectly to the number of households by sector and dwelling type.

One of the most important attributes to consider in land use modeling is the income level. Because of this, we focused our analysis here only on this variable and compared the simulated average income per commune with the observed values, as shown in Fig. 13. The simulated income is computed from an average weighted by the number of households by income level group. The simulated population reproduces the observed income distribution with a reasonable fit ($R^2 = 0.799$). Fig. 14 shows the spatial distribution of the error in average income by commune. It can be observed that the maximum error is less than Euro 375. The remaining variables also reproduce observed distributions with an adequate fit.

## 6. Discussion and conclusions

Here we presented a novel and operational approach to synthesize the agent population required in the microsimulation of urban systems. This approach instead of fitting a single solution on the available information uses the partial views of the joint distribution for the agent attributes to draw agents as if they were drawn from the actual joint distribution. This enables us to draw any number of possible synthetic populations. The input to the approach can be prepared using models specified from the data (both cross-classification tables and discrete response models) and where no data is available, from assumptions based on domain knowledge. Simulation based approach was able to reproduce the marginals as good as the fitting

**Table 5**
Modeled conditionals for great Brussels area.

| Attribute | Household variables | Spatial variables |
|---|---|---|
| $inc_h$ | $cars_h$, $workers_h$, $univ_h$, $v_h$ | $income_i$ |
| $cars_h$ | $inc_h$, $workers_h$, $univ_h$, $v_h$, $children_h$ | $income_i$, $car\_ownership_i$ |
| $univ_h$ | $inc_h$, $workers_h$ | $univ_i$ |
| $v_h$ | $cars_h$, $size_h$ | $surface_{vi}$ |

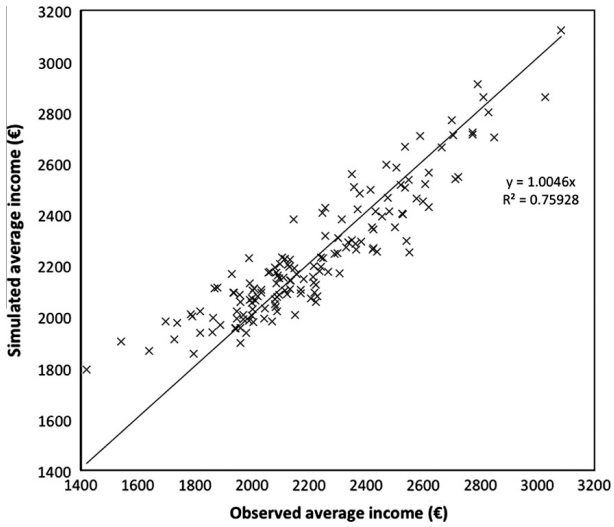


**Fig. 13.** Fit between observed and simulation based income.

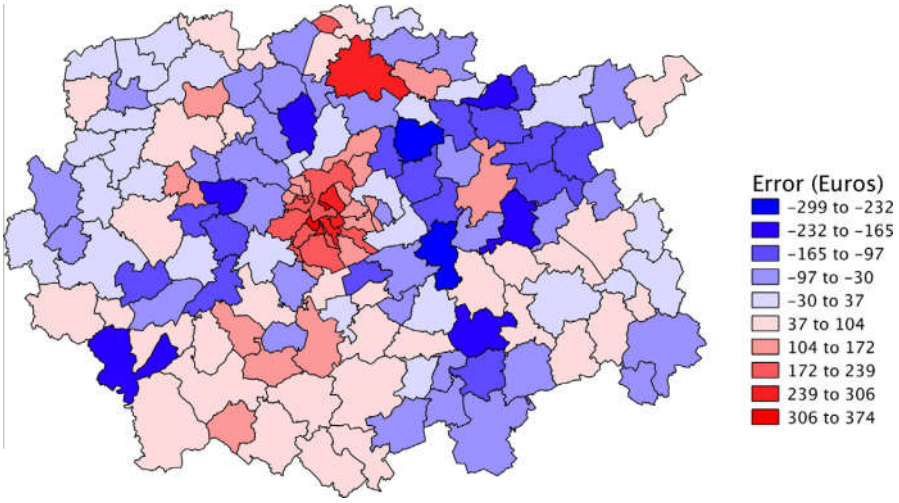


**Fig. 14.** Spatial distribution of error in the simulated average income by commune.

based approaches with at most the same amount of data. More importantly, it convincingly out performed other approaches in retrieving the joint distribution.

In most cases, the available datasets contain incomplete information on the agent attributes. Different datasets may not be consistent in terms of the definitions, spatial aggregation, and time. For privacy reasons they are purposely and systematically tampered using sophisticated statistical techniques. In this context fitting a single optimization based solution does not seem to be an appropriate approach. There can be many possible solutions to the problem. So, it is better to have the ability to draw any number of possible synthetic populations from the original joint distribution of attributes. This approach also enhances the reliability of the outputs coming out of the microsimulation (e.g. MATSim, ILUTE). Instead of running these microsimulation and producing the outputs based on one fitted synthetic population ($O = microsim(p_{syn})$), they can be integrated over all the possible synthetic populations coming out of the joint distribution.

$$O = \int_{p_{syn}} microsim(p_{syn}) f(p_{syn}) dp_{syn}.$$

Moreover, this simulation based approach can become a direct part of these microsimulation as the starting point. This way the population synthesis can be part of the variance analysis of the whole urban systems microsimulation.

The synthetic population coming out of fitting based approaches is generated by cloning of the microsample using computed weights. This means that the resulting population is essentially blowing up of the sample rather than reproducing it from the heterogeneous points in the attribute space. There can be many cases, where a point in the attribute space does not exist in the microsample (e.g. due to sampling zero) and/or in the marginals (e.g. there may not be any marginal available for

certain attribute), but it exists in the real population. Such a point cannot be reproducible in the synthetic population using fitting based methods. On the other hand, the proposed simulation based procedure (with proper specification of the conditionals) can traverse through the entire attribute space. This results in a more heterogenous and representative synthetic population.

Another unique feature of the simulation based synthesis is that the attributes can be both discrete and continuous, or any mix of them. Fitting based approaches are limited to synthesizing discrete attributes only. Due to computational and various other issues, attributes have to be categorized into rather aggregate categories. On the other hand, there are many important attributes (for instance, income, commuting distance, etc.) that behavioral models need as continuous. Simulation based approach can deal with the mix of attributes without any additional cost. In fact, in terms of implementation, it is easier to handle the continuous attribute (e.g. it can be stored as a runtime function rather than an instance of a datastructure).

The input conditionals for the Gibbs sampler are prepared using marginals, partial conditionals, and estimated models from various sources. Where there is no data available, bringing in assumptions based on domain knowledge. Due to this the issue of inconsistency may arise. That is to say, there may not be any unique stationary distribution that Gibbs sampler can arrive to, which corresponds to all the constructed conditional. Buuren (2007) and Chen et al. (2011) investigated the behavior of Gibbs sampler in the case of inconsistent conditionals. Buuren (2007) concluded that essentially the estimates were still unbiased and had a good practical performance. They however, suggested that theoretically it is still a research question and deserve further study. Chen et al. (2011) proposed a technique called Gibbs Ensembles to retrieve the joint distribution that deviate the least from all the inconsistent conditionals. Our objective however is to propose a new methodology and not the various improvements on it, so we restricted to standard Gibbs sampling only. In any future studies, various refinements of Gibbs sampling and other sampling procedures can be investigated.

An important aspect that is not covered in this paper, but is part of ongoing work, is the generation of associations between different types of agents (household, family, and person). For any realization, the associations can be treated as edges (with weights), between two different types of vertices in a bi-partite graph. In the case of associations between the households and persons agents, for example, there are set of positions available in the household agents (mother, father, first child, etc.) to which an association may exist for person agents. So, the household positions and person agents form the two types of vertices in the bi-partite graph and we are interested in reconstructing the joint distribution of the association weights. Again, a simulation based technique can be used for reconstructing this distribution. The actual association is then a realization out of this distribution. The next step of this research is to operationalize the methodology for association generation.

In the case of IPF, if the future year synthetic population is needed, the base CT is refitted to the future years marginals. The assumption here is that the correlation structure of attributes within the population for future year remains the same as the base year. The equivalent can be done in simulation based synthesis by resampling based of future years marginals and conditionals. Another feature of the proposed methodology is that the data preparation is completely separate from the actual agent generation. The simulation requires conditionals that can be generated by any possible mean. It can be direct counts from the data, a behavioral model, assumptions, parametric distribution, etc. Here we have only used Gibbs sampling to generate the population, but various other sampling procedures and their combinations can be used.

In terms of the implementation, IPF severely suffers from the curse of dimensionality. In the fitting processes IPF has to maintain and iterate through the entire contingency table. That limits the number of attributes that can be synthesized. While, Markov chain Monte Carlo Simulation based procedures on the other hand only needs to keep the last synthesized agent to synthesize the next agent.

In terms of application, we were able to generate a decent synthetic population for Greater Brussels Area with a very limited information (microsample of only 0.1% of the population and very partial marginals) as outlined in Section 5. It would not have been possible to generate a synthetic population using IPF in this case, as we have already shown in sub Section 4.3.2 that even with 1% sample the synthesized population is not representative of the true population. Section 5 also illustrates the high flexibility in which the conditionals can be constructed for the attributes in the data preparation stage.

## Acknowledgements

Research in this article has been funded by the European Commission's Seventh Framework Programme, Swiss National Science Foundation, and Danish Council of Strategic Research. Authors are grateful for this generous support. We are also very thankful to the reviewers for providing us insightful feedback.

## Appendix A. Correlation analysis

In this section we report the correlation analysis of the real and simulated population from the Swiss Census. Pair-wise uncontrolled and higher order partial correlations are computed in Tables A.6, A.7, and A.8. As the joint distribution is not known to be multivariate Gaussian and the pair-wise relationship may not be linear, we used the procedure outlined in Kendall (1938) to compute the correlations. The reported partial correlations are significant with more than 99% statistical confidence. The comparison between the values for real and the simulation populations shows a close match. The simulation based approach was able to avoid generating any unnecessary higher-order correlation, while having a close fit for the highly correlated attributes.

**Table A.6**
Correlation between attributes.

| Attribute pair | Census | Full_Cond | Partial_1 | Partial_2 | Partial_3 |
|---|---|---|---|---|---|
| Age-Sex | 0.083 | 0.087 | 0.059 | 0.060 | 0.050 |
| Age-Hhld_Size | −0.314 | −0.317 | −0.312 | −0.313 | −0.312 |
| Age-Edu_Lvl | −0.432 | −0.434 | −0.435 | −0.434 | −0.430 |
| Sex-Hhld_Size | −0.046 | −0.049 | −0.052 | −0.027 | −0.035 |
| Sex-Edu_Lvl | −0.042 | −0.033 | −0.045 | −0.036 | −0.029 |
| Hhld_Size-Edu_Lvl | 0.214 | 0.216 | 0.220 | 0.212 | 0.221 |

**Table A.7**
1st order partial correlation between attributes.

| Attribute pair | Control | Census | Full_Cond | Partial_1 | Partial_2 | Partial_3 |
|---|---|---|---|---|---|---|
| Age-Sex | Hhld_Size | 0.072 | 0.075 | 0.045 | 0.054 | 0.041 |
| Age-Sex | Edu_Lvl | 0.072 | 0.081 | 0.044 | 0.049 | 0.042 |
| Age-Hhld_Size | Sex | −0.312 | −0.314 | −0.310 | −0.312 | −0.311 |
| Age-Hhld_Size | Edu_Lvl | −0.251 | −0.254 | −0.246 | −0.251 | −0.246 |
| Age-Edu_Lvl | Sex | −0.430 | −0.433 | −0.434 | −0.433 | −0.429 |
| Age-Edu_Lvl | Hhld_Size | −0.393 | −0.395 | −0.395 | −0.396 | −0.390 |
| Sex-Hhld_Size | Age | −0.021 | −0.023 | −0.035 | 0 | −0.02 |
| Sex-Hhld_Size | Edu_Lvl | −0.038 | −0.043 | −0.043 | −0.020 | −0.029 |
| Sex-Edu_Lvl | Age | 0 | 0 | −0.022 | 0 | 0 |
| Sex-Edu_Lvl | Hhld_Size | −0.033 | −0.023 | −0.034 | −0.031 | −0.022 |
| Hhld_Size-Edu_Lvl | Age | 0.092 | 0.092 | 0.99 | 0.89 | 0.11 |
| Hhld_Size-Edu_Lvl | Sex | 0.212 | 0.215 | 0.218 | 0.211 | 0.22 |

**Table A.8**
2nd Order partial correlation between attributes.

| Attribute pair | Control | Census | Full_Cond | Partial_1 | Partial_2 | Partial_3 |
|---|---|---|---|---|---|---|
| Age-Sex | Hhld_Size-Edu_Lvl | 0.064 | 0.072 | 0.034 | 0.045 | 0.035 |
| Age-Hhld_Size | Sex-Edu_Lvl | −0.250 | −0.251 | −0.245 | −0.250 | −0.246 |
| Age-Edu_Lvl | Sex-Hhld_Size | −0.392 | −0.394 | −0.395 | −0.395 | −0.389 |
| Sex-Hhld_Size | Age-Edu_Lvl | 0 | 0 | −0.033 | 0 | 0 |
| Sex-Edu_Lvl | Age-Hhld_Size | 0 | 0 | 0 | 0 | 0 |
| Hhld_Size-Edu_Lvl | Age-Sex | 0.092 | 0.092 | 0.098 | 0.089 | 0.101 |

## Appendix B. Higher moments analysis

Table B.9, B.10, B.11, B.12, and B.13 reports the first seven raw standardized moments of the real and simulation based synthesized population. The Kruskal–Wallis test (95% confidence level) on these set of moments revealed that they represent the same population (Kruskal and Wallis, 1952).

**Table B.9**
Raw standardized moments in real population.

| Moment | Age | Sex | Hhld_Size | Edu_Lvl |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
| 2 | 3.11 | 0.53 | 1.61 | 0.36 |
| 3 | 14.38 | 0.53 | 4.66 | 0.86 |
| 4 | 77.11 | 0.53 | 16.18 | 2.30 |
| 5 | 449.42 | 0.53 | 63.00 | 6.48 |
| 6 | 2754.99 | 0.53 | 263.73 | 18.80 |
| 7 | 17450.03 | 0.53 | 1159.58 | 55.25 |

**Table B.10**
Raw standardized moments in Full_Cond.

| Moment | Age | Sex | Hhld_Size | Edu_Lvl |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
| 2 | 3.13 | 0.53 | 1.61 | 0.36 |
| 3 | 14.44 | 0.53 | 4.66 | 0.86 |
| 4 | 77.31 | 0.53 | 16.18 | 2.29 |
| 5 | 449.82 | 0.53 | 62.96 | 6.45 |
| 6 | 2753.26 | 0.53 | 263.57 | 18.68 |
| 7 | 17417.60 | 0.53 | 1159.42 | 54.87 |

**Table B.11**
Raw standardized moments in Partial_1.

| Moment | Age | Sex | Hhld_Size | Edu_Lvl |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
| 2 | 3.10 | 0.53 | 1.62 | 0.37 |
| 3 | 14.27 | 0.53 | 4.67 | 0.86 |
| 4 | 76.27 | 0.53 | 16.22 | 2.29 |
| 5 | 443.18 | 0.53 | 63.20 | 6.45 |
| 6 | 2709.57 | 0.53 | 264.86 | 18.68 |
| 7 | 17122.80 | 0.53 | 1165.98 | 54.89 |

**Table B.12**
Raw standardized moments in Partial_2.

| Moment | Age | Sex | Hhld_Size | Edu_Lvl |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
| 2 | 3.11 | 0.53 | 1.63 | 0.36 |
| 3 | 14.38 | 0.53 | 4.71 | 0.86 |
| 4 | 76.99 | 0.53 | 16.38 | 2.28 |
| 5 | 447.99 | 0.53 | 63.74 | 6.44 |
| 6 | 2741.66 | 0.53 | 266.50 | 18.69 |
| 7 | 17338.72 | 0.53 | 1169.96 | 54.96 |

**Table B.13**
Raw standardized moments in Partial_3.

| Moment | Age | Sex | Hhld_Size | Edu_Lvl |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
| 2 | 3.13 | 0.53 | 1.61 | 0.35 |
| 3 | 14.45 | 0.53 | 4.63 | 0.81 |
| 4 | 77.27 | 0.53 | 16.08 | 2.16 |
| 5 | 449.20 | 0.53 | 62.52 | 6.07 |
| 6 | 2746.81 | 0.53 | 261.56 | 17.56 |
| 7 | 17358.86 | 0.53 | 1149.58 | 51.56 |

## Appendix C. Estimated models

In this section we describe the parameters for the models that were estimated for the Brussels case study. (See Table C.14, C.15, C.16, C.17).

**Table C.14**
Dwelling type model.[a,b]

| Parameter | Variable | Value | Std err | *t*-test |
|---|---|---|---|---|
| $ASC^2$ | Constant for dwelling type 2 | 0.423 | 0.297 | 1.42 |
| $ASC^3$ | Constant for dwelling type 3 | 0.87 | 0.305 | 2.86 |
| $ASC^4$ | Constant for dwelling type 4 | 1.2 | 0.327 | 3.68 |
| $\beta_{surf\times h2}$ | Dummy for hh size = 2 × zonal avg surface of dwelling[c] | 0.0146 | 0.00533 | 2.74 |
| $\beta_{surf\times h3}$ | Dummy for hh size = 3 × zonal avg surface of dwelling[c] | 0.0194 | 0.00597 | 3.25 |
| $\beta_{surf\times h4+}$ | Dummy for hh size > 3 × zonal avg surface of dwelling[c] | 0.0249 | 0.00299 | 8.31 |
| $\beta^2_{cars}$ | Number of cars in the household | −0.279 | 0.182 | −1.53 |
| $\beta^3_{cars}$ | Number of cars in the household | −0.593 | 0.207 | −2.86 |
| $\beta^4_{cars}$ | Number of cars in the household | −0.948 | 0.233 | −4.07 |
| $A^1$ | ln of number of dwellings of type 1 in zone[c, d] | 1 | – | – |
| $A^2$ | ln of number of dwellings of type 2 in zone[c, d] | 1 | – | – |
| $A^3$ | ln of number of dwellings of type 3 in zone[c, d] | 1 | – | – |
| $A^4$ | ln of number of dwellings of type 4 in zone[c,d] | 1 | – | – |

[a] Dwelling types are: isolated house (1), semi-attached house (2), attached house (3), and apartment (4).
[b] The superindex in the parameter indicates to which household education level (0, 1, 2) it is specific. The alternative of isolated house is used as a reference ($\beta^1_* = 0$).
[c] Spatial variable.
[d] Expansion factor used to account for the (un)availability of different types of dwelling in different zones. Not an estimated parameter.

**Table C.15**
Income level model.[a]

| Parameter | Variable | Value | Std err | *t*-test |
|---|---|---|---|---|
| $ASC^2$ | Constant for income level 2 | −0.86 | 0.789 | −1.09 |
| $ASC^3$ | Constant for income level 3 | −4.64 | 0.901 | −5.14 |
| $ASC^4$ | Constant for income level 4 | −8.31 | 1.12 | −7.39 |
| $ASC^5$ | Constant for income level 5 | −10.6 | 1.55 | −6.82 |
| $\beta^3_{educ}$ | Dummy for presence of people with higher educ in the hh | 0.831 | 0.177 | 4.69 |
| $\beta^4_{educ}$ | Dummy for presence of people with higher educ in the hh | 1.72 | 0.314 | 5.49 |
| $\beta^5_{educ}$ | Dummy for presence of people with higher educ in the hh | 1.92 | 0.656 | 2.93 |
| $\beta^2_{zonal\_inc}$ | Average zonal income[b] | 0.0008 | 0.0004 | 1.84 |
| $\beta^3_{zonal\_inc}$ | Average zonal income[b] | 0.0012 | 0.0005 | 2.55 |
| $\beta^4_{zonal\_inc}$ | Average zonal income[b] | 0.0016 | 0.0005 | 3.09 |
| $\beta^5_{zonal\_inc}$ | Average zonal income[b] | 0.0016 | 0.0006 | 2.47 |
| $\beta^2_{cars}$ | Number of cars in the household | 1.16 | 0.265 | 4.39 |
| $\beta^3_{cars}$ | Number of cars in the household | 1.92 | 0.299 | 6.41 |
| $\beta^4_{cars}$ | Number of cars in the household | 2.33 | 0.341 | 6.83 |
| $\beta^5_{cars}$ | Number of cars in the household | 3.2 | 0.466 | 6.87 |
| $\beta^3_{house}$ | Dummy for dwelling being a house | 0.45 | 0.193 | 2.34 |
| $\beta^4_{house}$ | Dummy for dwelling being a house | 0.485 | 0.294 | 1.65 |
| $\beta^5_{house}$ | Dummy for dwelling being a house | 0.485 | 0.294 | 1.65 |
| $\beta^2_{workers}$ | Number of workers in the household | 1.14 | 0.277 | 4.11 |
| $\beta^3_{workers}$ | Number of workers in the household | 2.22 | 0.295 | 7.53 |
| $\beta^4_{workers}$ | Number of workers in the household | 2.46 | 0.345 | 7.13 |
| $\beta^5_{workers}$ | Number of workers in the household | 1.74 | 0.428 | 4.07 |

[a] The superindex in the parameter indicates to which income level (1, 2, 3, 4, 5) it is specific. Income level 1 is used as a reference ($\beta^1_* = 0$).
[b] Spatial variable.

**Table C.16**
Car ownership model.[a]

| Parameter | Variable | Value | Std err | *t*-test |
|---|---|---|---|---|
| $ASC^1$ | Constant for 1 car | −2.75 | 0.611 | −4.5 |
| $ASC^2$ | Constant for 2 cars | −7.02 | 0.812 | −8.65 |
| $ASC^3$ | Constant for 3+ cars | −10.1 | 1.23 | −8.2 |
| $\beta^1_{educ}$ | Dummy for presence of people with higher educ in the hh | 0.504 | 0.196 | 2.57 |
| $\beta^2_{educ}$ | Dummy for presence of people with higher educ in the hh | 0.933 | 0.267 | 3.49 |
| $\beta^3_{educ}$ | Dummy for presence of people with higher educ in the hh | 1.07 | 0.552 | 1.94 |
| $\beta^1_{high\_inc}$ | Dummy for households with high income (>3) | 0.977 | 0.499 | 1.96 |
| $\beta^2_{high\_inc}$ | Dummy for households with high income (>3) | 2.43 | 0.569 | 4.28 |
| $\beta^3_{high\_inc}$ | Dummy for households with high income (>3) | 3.24 | 0.801 | 4.05 |
| $\beta^1_{mid\_inc}$ | Dummy for households with mid income (=3) | 0.858 | 0.267 | 3.22 |
| $\beta^2_{mid\_inc}$ | Dummy for households with mid income (=3) | 1.87 | 0.342 | 5.46 |
| $\beta^3_{mid\_inc}$ | Dummy for households with mid income (=3) | 1.42 | 0.676 | 2.11 |
| $\beta^1_{zonal\_inc}$ | Average zonal income[b] | 0.001 | 0.0003 | 3.51 |
| $\beta^2_{zonal\_inc}$ | Average zonal income[b] | 0.0013 | 0.0004 | 3.29 |
| $\beta^3_{zonal\_inc}$ | Average zonal income[b] | 0.0013 | 0.0004 | 3.29 |
| $\beta^1_{car1\_zone}$ | Percentage of hh's with 1 car in the zone[b] | 0.498 | 0.172 | 2.89 |
| $\beta^2_{car2\_zone}$ | Percentage of hh's with 2 cars in the zone[b] | 2.13 | 0.842 | 2.53 |
| $\beta^3_{car3\_zone}$ | Percentage of hh's with 3+ cars in the zone[b] | 14.1 | 7.57 | 1.86 |
| $\beta^1_{children}$ | Dummy for presence of children in the household | 0.457 | 0.24 | 1.9 |
| $\beta^2_{children}$ | Dummy for presence of children in the household | 0.8 | 0.276 | 2.9 |
| $\beta^1_{house}$ | Dummy for dwelling being a house | 0.841 | 0.191 | 4.4 |
| $\beta^2_{house}$ | Dummy for dwelling being a house | 1.86 | 0.289 | 6.42 |
| $\beta^3_{house}$ | Dummy for dwelling being a house | 2.66 | 0.776 | 3.43 |
| $\beta^1_{workers}$ | Number of workers in the household | 0.437 | 0.139 | 3.15 |
| $\beta^2_{workers}$ | Number of workers in the household | 1.24 | 0.193 | 6.42 |
| $\beta^3_{workers}$ | Number of workers in the household | 1.6 | 0.358 | 4.46 |

[a] The superindex in the parameter indicates to which car ownership level (0, 1, 2, 3+) it is specific. The alternative of 0 cars is used as a reference ($\beta^0_* = 0$).
[b] Spatial variable.

**Table C.17**
Houshold education level model.[a,b]

| Parameter | Variable | Value | Std err | *t*-test |
|---|---|---|---|---|
| $ASC^1$ | Constant for 1 person with high educ in hh | −2.96 | 0.34 | −8.72 |
| $ASC^2$ | Constant for 2+ persons with high educ in hh | −7.19 | 0.547 | −13.14 |

(*continued on next page*)

**Table C.17** (*continued*)

| Parameter | Variable | Value | Std err | *t*-test |
|---|---|---|---|---|
| $\beta^1_{cars}$ | Number of cars in the household | 0.238 | 0.133 | 1.79 |
| $\beta^2_{cars}$ | Number of cars in the household | 0.701 | 0.156 | 4.51 |
| $\beta^1_{educ\_zone}$ | Percentage of hh's with educ level 1 in zone[c] | 3.34 | 0.566 | 5.91 |
| $\beta^2_{educ\_zone}$ | Percentage of hh's with educ level in zone[c] | 4.34 | 0.708 | 6.13 |
| $\beta^1_{income}$ | Income level of the household | 0.24 | 0.129 | 1.87 |
| $\beta^2_{income}$ | Income level of the household | 1.09 | 0.152 | 7.18 |
| $\beta^1_{workers}$ | Number of workers in the household | 0.393 | 0.113 | 3.47 |
| $\beta^2_{workers}$ | Number of workers in the household | 0.851 | 0.154 | 5.52 |

[a] Household education level refers to the number of people with a university diploma in the household (0, 1, 2+).
[b] The superindex in the parameter indicates to which household education level (0, 1, 2+) it is specific. The alternative of 0 persons is used as a reference ($\beta^0_* = 0$).
[c] Spatial variable.

## References


Arentze, T.A., Timmermans, H.J.P., 2004. A learning-based transportation oriented simulation system. Transportation Research Part B 38 (7), 613–633.

Arentze, T.A., Timmermans, H.J.P., Hofman, F., 2007. Creating synthetic household populations: problem and approach. Transportation Research Record 2014, 85–91.

Auld, J., Mohammadian, A., Wies, K., 2009. Population synthesis with subregion-level control variable aggregation. Journal of Transportation Engineering 135 (9), 632–639.

Balmer, M., Nagel, K., Raney, B., 2006. Agent-based demand modeling framework for large scale micro-simulations. Transportation Research Record 1985, 125–134.

Barthelemy, J., Toint, P.L., 2013. Synthetic population generation without a sample. Transportation Science 47 (2), 266–279, <http://transci.journal.informs.org/content/early/2012/04/05/trsc.1120.0408.abstract>.

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

Bishop, Y.M.M., Fienberg, S.E., Holl, P.W., Light, R.J., Mosteller, F., Imrey, P.B., Bishop, Y.M.M., Fienberg, S.E., Holl, P.W., 1975. In: Discrete Multivariate Analysis: Theory and Practice, second ed. MIT Press, Cambridge, MA.

Brand, R., 2002. Microdata protection through noise addition, inference control in statistical databases: from theory to practice. Lecture Notes in Computer Science 2316, 97–116.

Brown, M.B., Fuchs, C., 1983. On maximum likelihood estimation in sparse contingency tables. Computational Statistics & Data Analysis 1, 3–15, <http://deepblue.lib.umich. edu/bitstream/handle/2027.42/25269/0000712.pdf?sequence=1>.

Buuren, S.V., 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research 16 (3), 219–242.

Chen, S.H., Ip, E.H., Wang, Y.J., April, 2011. Gibbs ensembles for nearly compatible and incompatible conditional models. Computational Statistics & Data Analysis 55 (4), 1760–1769, <http://dx.doi.org/10.1016/j.csda.2010.11.006>.

Dalenius, T., Reiss, S.P., 1982. Dataswapping: a technique for disclosure limitation. Journal of Statistical Planning and Inference 6, 73–85.

Deming, W.E., Stephan, F.F., 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics 11 (4), 427–444.

Duguay, G., Jung, W., McFadden, D., 1976. SYNSAM: A Methodology for Synthesizing Household Transportation Survey Data. Working paper. Urban Travel Demand Forecasting Project, Institute of Transportation Studies. <http://books.google.ch/books?id=g4liHQAACAAJ>.

Farooq, B., Miller, E.J., Chingcuanco, F., 2009. A dynamic microsimulation model for demographic update. In: Paper Presented at 56th Annual North American Meetings of the Regional Science Association International.

Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6 (6), 721–741.

Guo, J.Y., Bhat, C.R., 2007. Population synthesis for microsimulation: state of the art. Transportation Research Record 2014, 92–101.

Hastings, W.K., 1970. Monte carlo sampling methods using markov chains and their applications. Biometrika 57 (1), 97–109.

Holford, T.R., 1980. The analysis of rates and of survivorship using log-linear models. Biometrics 36 (2), 299–305, <http://www.jstor.org/stable/2529982>.

Hubert, J.P., Toint, P.L., 2002. La mobilite quotidienne des Belges. Presses Universitaires de Namur, Namur, vol. 1.

Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30 (1/2), pp. 81–93. <http://www.jstor.org/stable/2332226>.

Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47 (260), pp. 583–621. <http://www.jstor.org/stable/2280779>.

LaRon, S., Beckman, R., Baggerly, K., Anson, D., Williams, M., 1996. Transims transportation analysis and simulation system project summary and status. NASA Open Source Agreement Version 1.3.

Lu, M., 2011. Generating disaggregate population characteristics for input to travel-demand models. Dissertation, University of Florida. <http://gradworks.umi.com/35/14/3514962.html>.

Miller, E.J., Farooq, B., Chingcuanco, F., Wang, D., 2011. Historical validation of integrated transport-land use model system. Transportation Research Record 2255, 91–99.

Miller, E.J., Noehammer, P.J., Ross, D.R., 1987. A micro-simulation model of residential mobility. In: Paper Presented at International Symposium on Transport, Communication and Urban Form: 2, Analytical Techniques and Case Studies.

Miller, E.J., Roorda, M., 2003. Prototype model of household activity and travel scheduling. Transportation Research Record 1831, 114–121.

Müller, K., Axhausen, K.W., 2011. Population synthesis for microsimulation: state of the art. In: Proceeding of Transportation Research Board 90th Annual Meeting.

Openshaw, S., Rao, L., 1995. Algorithms for reengineering 1991 census geography. Environment and Planning A 27 (3), 425–446.

Pitfield, D.E., 1978. Sub-optimality in freight distribution. Transportation Research 12 (6), 403–409.

Pritchard, D., Miller, E.J., 2012. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. Transportation 39, 685–704, <http://dx.doi.org/10.1007/s11116-011-9367-4>.

Pritchard, D.R., 2008. Synthesizing Agents and Relationships for Land Use/transportation Modelling. Canadian theses, University of Toronto. <http://davidpritchard.org/archives/111>.

Ryan, J., Maoh, H., Kanaroglou, P., 2009. Population synthesis: comparing the major techniques using a small, complete population of firms. Geographical Analysis 41 (2), 181–203.

Schafer, J., 1997. Analysis of Incomplete Multivariate Data. Monographs on Statistics and Applied Probability, vol. 72. Chapman & Hall, London, <http://www.amazon.com/dp/0412040611>.

Sweeney, L., 2002. Achieving k-anonymity privacy protection using generalization and suppression. Journal on Uncertainty, Fuzziness, and Knowledge-based Systems 10 (5), 571–588.

Train, K.E., 2003. Discrete Choice Methods with Simulation. Discrete Choice Methods with Simulation. Cambridge University Press, <http://books.google.ch/books?id=F-_gYALlfR4C>.

Voas, D., Williamson, P., 2000. An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. International Journal of Population Geography 6 (5), 349–366.

Waddell, P., 2002. Urbansim: modeling urban development for land use, transportation and environmental planning. Journal of American Planning Association 68 (3), 297–314.

Williamson, P., Birkin, M., Rees, P.H., 1998. The estimation of population microdata by using data from small area statistics and samples of anonymised records. Environment and Planning A 30 (5), 785–816.

Ye, X., Konduri, K.C., Pendyala, R.M., Sana, B., Waddell, P., 2009. Methodology to match distributions of both household and person attributes in generation of synthetic populations. In: Proceeding of Transportation Research Board 88th Annual Meeting. <http://trid.trb.org/view.aspx?id=881554>.