Methodologies for synthesizing populations

Bilal Farooq*[1], Kirill Müller†[2], Michel Bierlaire‡[1] and Kay W. Axhausen§[2]

[1]TRANSP-OR, ENAC, EPFL, Lausanne, Switzerland
[2]Institute for Transport Planning and Systems, ETH, Zurich, Switzerland

June 15, 2013

Abstract

This paper reports two different approaches that can be used for the generation of a synthetic population in the context of urban systems microsimulation. First a *fitting and generation* method is proposed to synthesize the agents and their relationships. It works by fitting a large survey to match known zone-level aggregate totals and selecting agents with probability depending on the estimated weight. The fitting is performed by IPF (Iterative Proportional Fitting) or more advanced algorithms which match totals at different aggregation levels (e.g., households and persons). Second, a *Markov chain Monte Carlo* (MCMC) simulation based population synthesis approach is proposed. This approach synthesizes a population by using a Gibbs sampler to simulate drawing from the joint distribution of the agent attributes. We describe different ways for construction of the required conditional distributions.

The added value brought by the fitting and generation method is the simultaneous fitting of different types of agents (households and individuals), while simulation based synthesis takes a different perspective to solve the problem, especially in the case of limited data availability. Both methods proposed here contribute to the research in population synthesis.

1 Introduction

Models for microsimulation of urban systems simulate the interactions of synthetic agents. Due to privacy and other restrictions, detailed data for all agents to be simulated is virtually never available. Synthesis of a population of agents and their attributes is thus one of the first and key steps required for developing an agent based integrated transportation and land use microsimulation. It is anticipated that the agent population (a) matches known aggregate totals for all zones in the study area (e.g., the total number of males and females in each zone), and (b) replicates the correlation structure of the

*Correspondance: bilal.farooq@epfl.ch
†kirill.mueller@ivt.baug.ethz.ch
‡michel.bierlaire@epfl.ch
§axhausen@ivt.baug.ethz.ch

true population (e.g., wealthy people tend to be older than average, own several cars and a detached house, etc.). In order to accurately simulate person interactions (e.g., in the household or in a person's social network), the model also requires information on personal relationships. One of the earliest examples of base year agent synthesis can be found in TRANSIM (LaRon et al. 1996). The agent synthesis was formulated as a fitting problem in which weights of microdata were adjusted to fit the aggregate space level control totals (Beckman et al. 1996). Since then various approaches have been developed that are based on the fitting based optimization.

Here we present two methodologies for synthesis and an application in the Brussels case study of the SustainCity project. As for most applications, the required data was not available in disaggregate form. (For the Swiss case study within this project, the complete disaggregate census data was available, and the agent population has been derived directly from this data set.)

The first approach is based on treating the synthesis as a fitting problem. The fitting of two different types of agents (i.e., household and person) is performed simultaneously. The second approach treats the synthesis as a problem of drawing from a complex joint distribution of agents' attributes. A Markov chain Monte Carlo simulation based procedure is developed. Such an approach is particularly very robust in the case of highly limited data availability.

The rest of the paper is organized as follows. We first present the current state of literature in population synthesis. We then describe two methodologies to synthesize the agent population. The available data sources and input preparation for the two methods is described. An application of the proposed methodologies, and results, are presented. In the end we discuss various aspects of the two approaches.

2 Literature review

Agent-based microsimulation model systems for land use and transportation planning have come into widespread use (UrbanSim 2011, MATSim-T 2011, Bradley et al. 2010, Beckx et al. 2009, Roorda et al. 2008, Bhat et al. 2004, Ben-Akiva et al. 2002, Bowman & Ben-Akiva 2001, de Palma & Marchal 2002, Mahmassani et al. 1995). These models simulate decisions of agents within an urban area, allowing for more detailed simulation and prediction of population distribution, travel demand, and various other indicators than traditional aggregate models. Often, the *agents* represent the individual people living in the study area – grouped in households, as personal decisions are usually influenced by other household members (Jones et al. 1983).

When implementing such a model system, the initial step is the definition of agents and their relationships. This process is called *population synthesis*; also, the term *spatial microsimulation* is sometimes used. Williamson (2013) distinguishes between two practical methods for generating synthetic population: *reweighting* and *imputation*. Reweighting methods construct the population by selecting observations from a microsample. Each record in the sample can be cloned or replicated several times to form agents in the synthetic population. In contrast, imputation methods attempt to generate agents "from scratch", yielding unique characteristics for each agent. The following two subsections review the literature on the two methods. See also (Harland et al. 2012, Hermes & Poulsen 2012, Müller & Axhausen 2011*b*, Ma 2011, Pritchard 2008) for a comprehensive literature review.

After generating the synthetic population, agent relationships such as household

structure (Gargiulo et al. 2010) or social networks (Arentze et al. 2012) can be added to allow simulating interactions between agents. Lenormand & Deffuant (2012) compare the reweighting and imputation methods for generating the household structure.

2.1 Reweighting methods

The main idea here is to combine census microdata (the *reference sample*) with aggregate data at various levels (the *control totals*) in order to generate a set of agents for which the distribution and correlation of the agents' attributes are similar to those in the census microsample, and the number of agents within each category matches the aggregate data. Two approaches can be distinguished: The *fitting and generation* (FG) method (Beckman et al. 1996) generates the synthetic population by drawing from a weighting of the reference sample, whereas the *combinatorial optimization* (CO) method (Voas & Williamson 2000) directly expands the reference sample with an integer vector computed using non-linear integer optimization. See also (Rahman et al. 2010) for a review of reweighting techniques.

Both approaches require a reference sample from which agents are drawn: The generated population contains only realizations that also exist in the reference sample. Furthermore, the weights are generated subject to different optimality criteria, which, depending on the data, sometimes cannot be met. However, reweighting methods usually generate populations with a reasonably good fit to the given control totals.

2.1.1 Fitting and generation

Given a reference sample of households that contains detailed data for all persons, and constraints at both household and person levels, two options are available for the FG method:

1. The weights obey only household-level constraints, person-level constraints are considered when selecting households (*single-level fitting*, cf. (Auld & Mohammadian 2010, Pritchard & Miller 2009, Srinivasan & Ma 2009, Guo & Bhat 2007)), or

2. The weights obey constraints at both person and household levels, household selection is unconstrained (*multi-level fitting*, cf. (Ye et al. 2009, Bar-Gera et al. 2009, Lee & Fu 2011, Müller & Axhausen 2011*a*)).

The multi-level strategy greatly simplifies the construction of the final synthetic population by using more complicated fitting algorithms that have become available only recently, and hence can be considered superior to the single-level strategy. Pendyala et al. (2012) describe the practical application of a multi-level fitting algorithm.

The FG method requires careful preparation of the data. Especially the *zero-cell* or *missing observation* problem has been noted early (Beckman et al. 1996); flexible aggregation schemes are suggested in (Auld et al. 2008, Otani et al. 2012). The problem of conflicting control totals has been addressed by Rich & Mulalic (2012).

2.1.2 Combinatorial optimization

In this method the sample is replicated zone by zone so as to optimize the weight (where $w = \{0, 1\}$) of each observation under zonal marginal constraints (Voas &

Williamson (2000); Ma (2011)). Openshaw & Rao (1995), Williamson et al. (1998), and Voas & Williamson (2000) used Simulated Annealing (SA) as an optimization tool to produce the synthetic population. Ryan et al. (2009) compared CO and FG based methods 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 slow, especially in the case of SA based optimization.

Barthelemy & Toint (2012) developed a 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 hierarchical fitting steps (for each aggregation level where the data is available), entropy maximization, tabu search, and various ad hoc matching rules. 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.

Abraham et al. (2012) have implemented an analogy to multi-level fitting in a CO scenario: Constraints at both household and person levels are part of the optimization process.

2.2 Imputation methods

While imputation methods can make use of a microsample, it is not required. The observed population is assumed to be a realization of a (multivariate) distribution, and this distribution is simulated to generate the synthetic population. In contrast to reweighting approaches, no cloning or replication occurs – when many attributes are generated, the characteristics of each generated agent are unique. Usually, imputation refers to substituting missing values for an existing attribute based on the data. In extreme cases, the attribute is missing for all observations, or the dataset is constructed from scratch. This justifies the usage of the term "imputation" in a very general sense in the context of population synthesis.

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 where 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.

An example of a multistage approach where attributes are generated sequentially and attributes depend on other attributes generated earlier is described in (Williamson et al. 1998); also some earlier studies are referenced therein. In a recent effort, Frazier & Alfons (2012) have adapted this methodology to generate a synthetic population of Ghana. Barthelemy & Toint (2012) imputed the associations by sampling the empty household structure from the distribution and fitting the generated persons into it; various ad hoc rules were used to avoid incompatible associations.

To the authors' knowledge, this is for the first time when the agents are created directly from the underlying joint distribution.

3 Methodology

Here we describe two different approaches to synthesize the population. For both approaches, the agents are defined by their attributes $X = \{X^1, X^2, \ldots, X^k\}$, where k is the number of attributes to be synthesized. There are n agents in the synthesized population. In the first approach, the synthesis problem is treated as an optimization problem. The second approach treats it as a problem of drawing from a joint distribution which is only partially known and is very hard to directly draw from.

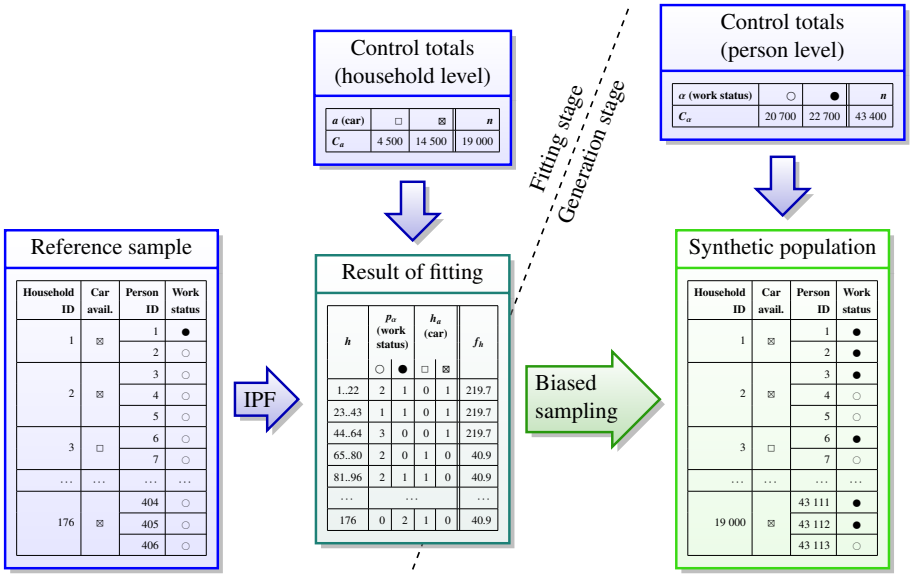
3.1 Fitting and generation based method

The fitting and generation (FG) method belongs to the class of reweighting methods and consists of two principal stages. In the *fitting stage*, a disaggregate sample of agents (the *reference sample*) is reweighted, yielding a fractional *expansion factor* for each household. The reweighted reference sample corresponds to the full population of the study area and is required to satisfy aggregate constraints (referred to as *control totals* or *controls*). After that, in the *generation stage*, this expansion factor is used to draw agents from the reference sample.

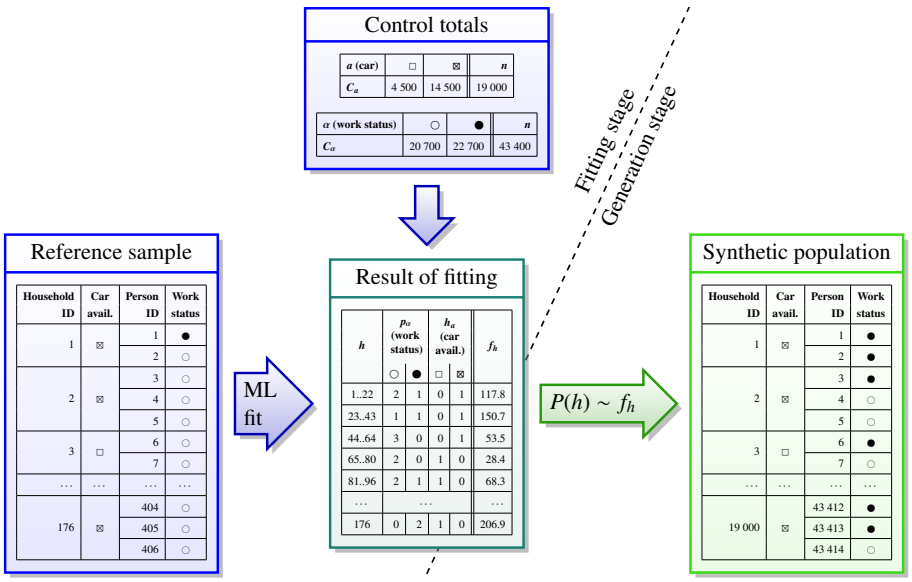
Expansion factors that satisfy the household-level control totals can be estimated using the well-known IPF algorithm (Deming & Stephan 1940). In order to also satisfy the person-level controls, one can employ a biased selection procedure that prefers households with persons in still underrepresented categories (Auld & Mohammadian 2010, Srinivasan & Ma 2009, Guo & Bhat 2007). This approach is sketched in Figure 1a. However, apart from the additional bias, the biased selection complicates the generation stage and sometimes requires time-consuming computations not suitable for frequent repetition. In addition, it seems to be difficult to specify a mathematical model for the population generated by this procedure.

In contrast, *multi-level algorithms* estimate household-level expansion factors that satisfy the controls at both household and person levels. These procedures include *Iterative Proportional Updating* (IPU) by Ye et al. (2009), *entropy maximization* (Ent) by Bar-Gera et al. (2009) and independently by Lee & Fu (2011), and *Hierarchical Iterative Proportional Fitting* (HIPF) by Müller & Axhausen (2011*a*). For all three approaches, the weights define a probability distribution of the households, and construction of the final population is possible using, e.g., weighted random sampling with replacement. Figure 1b illustrates this strategy. This simple model allows generating thousands of instances of similar yet different populations, e.g., in the context of multiple imputation (Rubin 1987). In the remainder of the paper, only the multi-level fitting approach is considered.

Formally, the reference sample is a $n \times k$ matrix $R = \left(R^1, R^2, \ldots, R^k\right)$, and the control totals are a vector x_j, $j = 1 \ldots c \leq k$. For this, a weight vector w_i, $i = 1 \ldots n$ is estimated so that $\sum_{i=1}^{n} (w^\intercal \cdot R)_{ij} = x_j$ for all $1 \leq j \leq c$. Then, $w^\intercal \cdot R \sim X$ is treated as the multivariate distribution that defines the population of interest, and a realization $\hat{X}$ is sampled from this distribution. The estimation of the weight vector and the format of R and x differs between the algorithms. For single-level algorithms, R contains either households or persons, and w is usually estimated using IPF. Multi-level algorithms always use households in R and additionally store person-level attributes as count variables; the estimation of w is carried out by repeatedly fitting all constraints in a round-robin fashion, or by entropy maximization.



(a) Single-level fitting



(b) Multi-level fitting

Figure 1: Illustration of single- and multi-level fitting algorithms

3.2 Markov chain Monte Carlo simulation based method

The agents' attributes X can be treated as random variables having a unique joint distribution $\pi_X(x)$ in the real population. In order to synthesize agent population, we are interested in drawing from this underlying joint distribution of agent attributes. This distribution is not completely known and is very hard to directly draw from. We only have partial views of the $\pi_X(x)$ through samples, marginals, and conditional-marginals. Using these partial views, a technique is developed here that lets us draw as if we were drawing from the joint distribution $\pi_X(x)$.

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/distributions (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; implement assumptions in a coherent manner, where the data is not available; 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 to use MCMC techniques to simulate the drawing from the real population distribution $\pi_X(x)$ and thus creating the synthetic population. Given the fact that we only have partial information about the joint distribution, there can be any number of synthetic populations $\hat{X}$ representing the real population X. The proposed technique lets us draw any number of the possible synthetic populations. The rest of the section describes how a specific MCMC technique can be used to generate synthetic populations.

3.3 Gibbs sampling basics

In particular here in this paper we have used Gibbs sampling (Geman & Geman 1984) procedure to develop the drawing of agents from the joint distribution. We are interested in generating synthetic populations $\hat{X}$ by sampling from the joint distribution of the attributes in the real population X. The problem is thus reduced to first retrieving the best representation $\hat{\pi}_X(x)$ of the joint distribution $\pi_X(x)$ that can then be sampled to generate any number of possible synthetic populations $\hat{x}$ representing X.

Let $X = (X^1, X^2, ..., X^k)$ be a set of random variables with a joint distribution $\pi(x)$ that is hard to retrieve. Instead we have the conditionals $\pi(X^i|X^j = x^j$, for $j = 1...k \ \& \ i \neq j), i = 1, ..., k$ available. The RVs themselves can be discrete or continuous. Then draws from the distribution $\pi(x)$ can be retrieved by using the procedure described in Algorithm 1 and Figure 2.

To reach a stationary state, the Gibbs sampler first has to run for an extended amount of iterations. At that point any draw from Algorithm 1 will be as if the draw was from $\pi(x)$ (Train 2003). To avoid the correlation between consecutive draws, few draws between two recorded draws are skipped.

3.4 Generation of synthetic agents

To generate a population, we simply have to warm up the Gibbs sampler and then draw the number of agents that are needed. For computational efficiency the state can be saved at the end of the draw. This way the sampler does not need to be warmed up every time a new population is needed. Another option is to pre-fetch a very large

Data:
$\pi(X^i|X^j = x^j$, for $j = 1...k \ \& \ i \neq j), i = 1, ..., k$
iterations (integer): Size of the population pool
interval (integer): Acceptance interval
Result: Draws from $\pi(x)$
initialize X_{prev};
initialize X_pool;
initialize counter;
for $size\_pool \times interval$ **do**
  Generate a random number from $r = U(1, k)$;
  Generate x^r_{curr} using **Inverse Transform** on
  $\pi(X^r_{curr}|X^j = x^j_{prev}$, for $j = 1...n \ \& \ r \neq j)$;
  X_{curr} = X_{prev} with x^r_{prev} replaced by x^r_{curr};
  if *counter equals interval* **then**
    X_pool.Add(X_{curr});
  end
  X_{prev} = X_{curr};
end

Algorithm 1: Gibbs Sampling

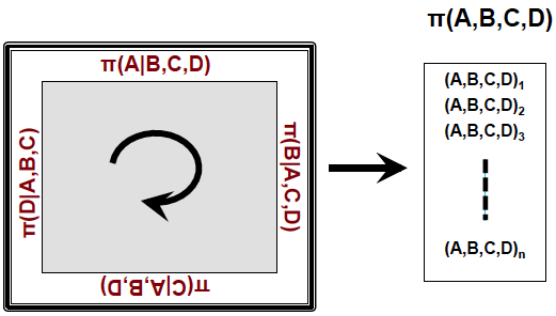


Figure 2: Illustration of the working of Gibbs sampler

pool of agents using Gibbs sampler. This pool can then be used to draw any number of synthetic populations. The procedure is described in Algorithm 2.

4 Data

4.1 Available data sources

Census and travel surveys have been the primary data sources to construct a synthetic population. Other sources, for instance a household spending survey, a labour force survey etc., can also be used. However, 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.

Data:
- *n (integer)*: number of synthetic populations
- *size (integer)*: size of a population
- *cond (integer)*: conditions (usually marginals) based on which a population should be drawn. By default the population is drawn from the pool using a uniform distribution

Result: *n* synthetic populations
if *population pool does not exist* **then**
  Call Gibbs Sampler;
end
while $k < n$ **do**
  while $l <$ *size* **do**
    Draw from pool using *cond* ;
  end
end

Algorithm 2: Synthetic population generator

4.1.1 Zoning systems

The data is available at certain aggregations of space that are defined by a zoning system. The aggregation may be based on certain maximum density levels, physical obstacles (river, street etc.), and political boundaries. There is a hierarchy of aggregations within each zoning system. For instance, in the case the of Canadian Census, the lowest level of zone is called dissemination area where 400 to 700 persons are living/working. One level above is the census tract where the limit is 2,500 to 8,000 persons. Further aggregations are census sub-division, division, and municipality, respectively. The zoning system also changes with time, i.e., a discrimination area in the year 2001 census may have been divided into two in the 2010 census, so as to satisfy the number of person constraints.

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

In agent based microsimulations, usually one zoning system is set as base; a mapping between different systems is defined; and all the data is made consistent to that system. An example of such a procedure can be found in Farooq et al. (2008).

4.1.2 Sample of individual agents

The Statistics bureau of a country, among other surveys, also conducts periodic censuses of the entire population. The periodicity of this census ranges from 5 (in the case of Canada) to 10 years (USA, UK, etc.). While the whole dataset is hardly ever available for the research (with some exemptions, like 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 UK and a few other countries, Sample of Anonymised Records (SARs). In this paper we will use the term PUMS. This sample is available at a large spatial resolution (for instance, City of Toronto,

London etc.), so as to make sure that the privacy of individuals is protected. The size of sample may range from 1% to 5% of the total population. The sample may contain a 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, the census bureau may hide information on certain individuals, if they deem it exposing the identity of these individuals.

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 socioeconomics and demographics of agents. There is more fluctuation here, in terms of the detail, size, and periodicity of travel survey among the regions/countries. Moreover, the periodicity of travel surveys of neighbouring regions may not coincide and it may also not coincide with that of the census.

4.1.3 Cross-classification tables

At various zoning levels (for instance, dissemination area, and census tracts in case of Canada, and sectors, and communes in case of Belgium and France), the Statistics bureau also releases the cross-classification tables for socioeconomics and demographics (for instance, income by age at sector level). These tables are usually restricted to 1 to 3 dimensions due to privacy concerns—the cell values of tables with more dimensions might be small enough to reproduce the population. In the available tables, very small cell values are usually 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, among others, in Sweeney (2002), Dalenius & Reiss (1982), Brand (2002). The point that should be noted here is that the available data has already been treated with various such anonymization techniques. Thus, a population synthesized based only on counts data may be not completely representative of the real population.

4.2 Data preparation

The two synthesis procedures described in Section 3 require the above mentioned data to be preprocessed as input in different forms. Here we first describe the data preparation process for the two processes and then illustrate the results.

Suppose that we are interested in synthesizing the person agents with attributes age A, sex S, household size H, eduction level E, and location in terms of spatial zone Z. In addition, the persons should be grouped into households, and the car availability C is of interest.

4.2.1 Fitting and generation based method

As mentioned earlier, multi-level FG algorithms operate on the following input data:

- a representative reference sample that contains the characteristics of sampled households and all constituent persons, and
- control totals for selected attributes on both household and person levels.

The objective is to estimate a positive weight (or *expansion factor*) for each household so that all control totals are satisfied.

For the given example, the reference sample should contain the household attributes H, C and, for each person in a household, the attributes A, S and E, obtained from a Public Use Microsample or from a large transportation survey (cf. Section 4.1.2). If no dataset with all attributes can be found, statistical matching (D'Orazio et al. 2006) or multiple imputation (Rubin 1987) can be used to create one from multiple datasets.

Cross-classification tables (cf. Section 4.1.3) are used as control totals. Here, at least one cross-classification that contains the Z attribute is required to distribute the population over all zones. The attributes in the control totals must be compatible with those in the reference sample. In addition, for each non-zero cell in each cross-classification table there must be at least one corresponding item in the reference sample to avoid the *missing observation problem* (or *zero-cell problem*). One possible remedy to both issues is collapsing categories. This approach has been implemented in a general fashion by Auld et al. (2008) and formalized by Otani et al. (2012). Note that collapsing two categories Y_1 and Y_2 is safe only if they are "similar", i.e. if $P(X|Y_1) \approx P(X|Y_2)$ for all other attributes $X = \{X_1, X_2, \ldots\}$.

Moreover, the cross-classification tables must be consistent with each other (in particular, the grand total and marginal totals for all attributes must match). Rich & Mulalic (2012) present a technique to cope with inconsistent cross-classification tables.

Usually it is assumed that the reference sample is equally representative of all zones in the study area. Hence, even if the reference sample contains precise location information, it is either not used or at best substituted by some higher-order categorization (e.g., a coarser zoning system, a zone classification into urban, periurban or rural, etc.) to avoid the missing observation problem.

4.2.2 MCMC based method

As proposed in Section 3.2, we are interested in using a Gibbs sampler to draw from the underlying distribution of the agents attributes. The Gibbs sampler needs the conditional distribution of one attribute over all the other attributes (Figure 2). These conditionals can be generated from the available data sources, i.e., cross-tabulations and microsample. In the simplest case, when the complete crossing is available in the cross-tables, they can be converted into full-conditionals. However, as pointed out in Section 4.1.3, it is rarely the case that the complete crossing is available. So, the conditionals have to be constructed using all the available data, domain knowledge, and bringing in assumptions where there is no information available.

Suppose we are interested in synthesizing person agents (A, S, H, E, Z). This means drawing from $\pi(A, S, H, E, Z)$. For the synthesis process to work, we need to compute the conditionals $\pi(A|S, H, E, Z)$, $\pi(S|A, H, E, Z)$, $\pi(H|A, S, E, Z)$, $\pi(E|A, S, H, Z)$, and $\pi(Z|A, S, H, E)$. In order to generate the population of person agents, suppose that we only have access to a 3% microsample of the region, but it is missing the attributes age and education level. Thus from sample we only have information about $\pi(A, H, Z)$. At zonal level, the marginals for sex $\pi(S)$ and household size $\pi(H)$ are available. In addition to that the conditional marginals of age on education level $\pi(A|E)$ are also available. Note that these marginals are available as counts of persons for each category.

Using the available data as described in Table 1, we can generate the population for either a) the entire area simultaneously or b) zone by zone. The complete conditionals

Table 1: Description of inputs

Attribute	Required	Available[a]
Z	$\pi(Z\|A,S,H,E)$	$\pi(Z\|A,H)$ $\pi(Z\|A)$, $\pi(Z\|S)$, $\pi(Z\|H)$, $\pi(Z\|E)$
A	$\pi(A\|S,H,E,Z)$	$\pi(A\|H,Z)$ $\pi(A\|Z)$, $\pi(A\|E,Z)$
S	$\pi(S\|A,H,E,Z)$	$-$ $\pi(S\|Z)$
H	$\pi(H\|A,S,E,Z)$	$\pi(H\|A,Z)$ $\pi(H\|Z)$
E	$\pi(E\|A,S,H,Z)$	$-$ $\pi(E\|Z)$, $\pi(E\|A,Z)$

a: The source for first row in this column is the sample and zonal cross-tabs for the second.

can thus be prepared as follows:

1. $\pi(Z|A,S,H,E)$
   Here we have two choices. First, we can invert the available marginal counts to develop $\pi(Z|A,E)$, $\pi(Z|H)$, $\pi(Z|E)$. Second option is to construct the person population zone-by-zone. In first case, we can take one of the conditionals and assume that the value of Z is uniform across other variables, given the selected one. If we do that then we are not using all the information available to us in form of the other marginals. In the case of second option we can construct the other marginals for each zone, but this way we will be making best use of all the information available to us. This also means that the population will have to be synthesized zone by zone.

2. $\pi(A|S,H,E,Z)$
   We have the counts available for $(A|H,Z)$ from the microsample and $(A|Z)$ and $(A|E,Z)$ from the zonal data. In case of zone-by-zone synthesis we can estimate an individual level RUM based conditional model for A that uses H at individual level and E at zonal level as explanatory variables. In addition to that it can use other zonal information, e.g., average age or share of age categories in the zone, or information coming from other zonal statistics, e.g., number of schools, land use in the zone, etc. The quality of this model will depend on the availability of the explanatory variables, but here we want to emphasize that this type of conditional model is a coherent way of using information from different sources and constructing the conditional marginals for the attributes.

3. $\pi(S|A,H,E,Z)$
   Here we only have $\pi(S|Z)$. In case of no other information we can assume that given Z, S is uniform across A, E, and H. Under this assumption $\pi(S|A,E,H,Z) = \pi(S|Z)$.

4. $\pi(H|A,S,E,Z)$
   As shown for $\pi(A|S,H,E,Z)$, here too the conditional models can be estimated by mixing the individual and zonal data.

5. $\pi(E|A, S, H, Z)$
   From the available sources, we only have counts for $(E|A, Z)$. As there is no further information available, we can assume that given Z and A, E is uniform across S, and H. Under this assumption, $\pi(E|A, S, H, Z) = (E|A, Z)$. We can also bring in any possible domain knowledge to justify or improve the conditionals. Here, for instance, we know that the education level is mostly dependent on age rather than sex or household size.

5 Application and results

5.1 Fitting and generation based method

In this section we summarize the findings of two previous contributions to the FG method (Müller & Axhausen 2011*a*, 2012). For the Zurich case study described in this book, the full year 2000 census was available (Swiss Federal Statistical Office (BFS) 2000*a*), and no synthetic population has been generated.

In (Müller & Axhausen 2011*a*), the HIPF algorithm has been first presented and compared to two other multi-level fitting algorithms (Entropy (Bar-Gera et al. 2009) and IPU (Ye et al. 2009)). This was not a real case study but a simulation study. A 5% sample has been drawn from the complete census and used as input for the three algorithms. Each canton's population has been synthesized separately. All resulting three-dimensional joint distributions of the generated attributes have been compared to the true distribution (from the complete census) using two distance metrics. The results of the comparison (see Figure 3) indicate an advantage of the new HIPF algorithm over the other two, especially for the less populated cantons.

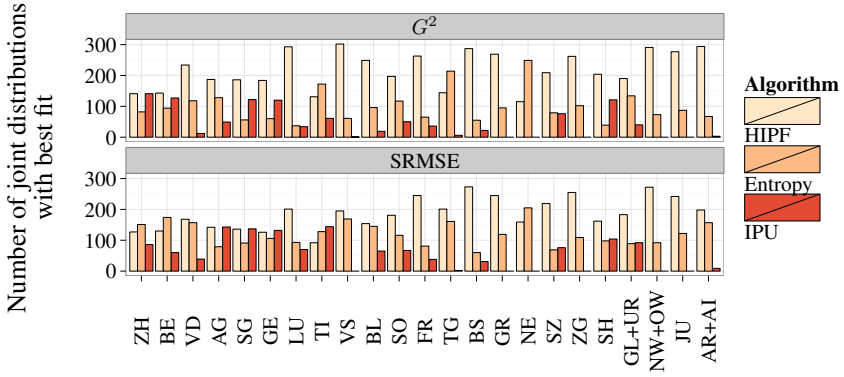


Figure 3: Results of the comparison between HIPF, Entropy and IPU

Müller & Axhausen (2012) describe the preparation process of a household sample with detailed person information from the Swiss PUMS (Swiss Federal Statistical Office (BFS) 2000*b*), a freely available 5 % sample of the Swiss population census that has been anonymized (cf. Section 4.1.2). Due to this anonymization, the data preparation was considerably more difficult than initially anticipated. However, in this

data preparation process, almost all attributes in the PUS were treated, without a specific application in mind. This suggests that the effort required for the data preparation should not be underestimated, and should focus only on attributes really required in the target population. The data preparation has been carried out entirely in the R programming language (R Development Core Team 2013) following the paradigm of reproducible research.

Figure 4 shows an excerpt of preliminary results from synthesizing a population using the PUS and freely available aggregated totals at both person and household levels. The mean number of full-time workers in a household is compared between a population synthesized from the PUS and the full census. This indicator is similar for all household sizes and independent of the presence of children, although work status has not been controlled for in the fitting stage.

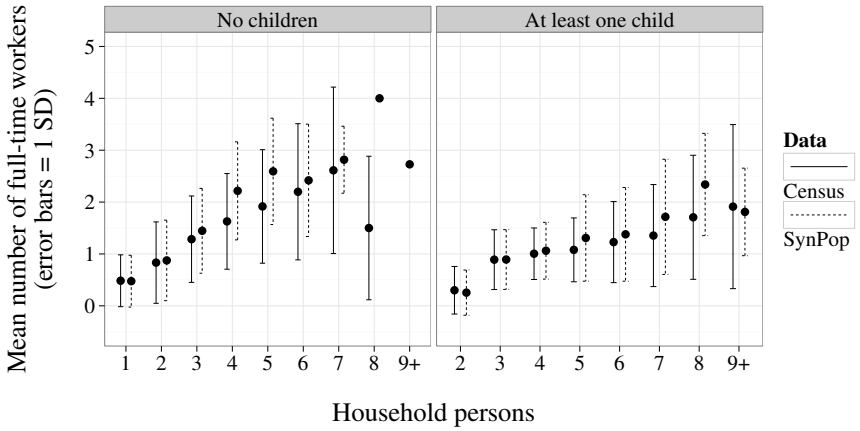


Figure 4: Excerpt of preliminary results for the generation of a synthetic population from the Swiss PUS

5.2 MCMC based method

In the Brussels case study the data availability was highly limited. The size of the available micro-sample was 0.1% only, and the zonal level conditionals were also only partially available. The simulation based procedure was used to synthesize a population out of this information. Details on the application of simulation based methodology to Brussels case study can be found in the case study chapter of this book.

6 Discussion

In this paper, two different strategies for generating synthetic populations have been presented: The fitting and generation (FG) and the Markov chain Monte Carlo (MCMC) methods. The FG method requires a sample of the target population which is reweighted and sampled from. In contrast, the MCMC method constructs the target population from scratch by repeatedly sampling from conditional probabilities at the attribute level.

While the FG method has been widely adopted and enhanced for almost two decades, the MCMC method is a new and entirely different approach to the problem.

Usually, the limiting factor is data availability: Large samples of individual data are difficult or impossible to obtain, and the anonymization process further complicates their usage especially for FG methods. However, once the data has been obtained and prepared, recent FG methods allow generating synthetic populations controlled at different aggregation levels (households and persons). On the other hand, the MCMC method overcomes these restrictions by employing a simple and robust method that allows using data and even models of different kinds as input for the synthetic population.

As an illustration, simulation based methodology described how a Gibbs sampler can be used for sampling from the joint distribution. Using an example, it was shown how a mix of models/counts, assumptions, and domain knowledge can be used to prepare the input. It was also shown how data from various sources can be coherently merged using discrete conditional models. Such an approach is particularly useful when the data from source (e.g., microsample) is very limited. The MCMC approach can synthesize a mix of discrete (e.g., marital status) and continuous (e.g., income) attributes. Other samplers (e.g., Metropolis Hastings) or any combination of them can also be used.

In the context of large scale models for European cities in general and SustainCity project in particular, this research has played a critical role in developing methodologies for generating a reliable base line population. A large part of the available literature focuses on data from the US. As seen in the Brussels case study (cf. Chapter ???), this is sometimes difficult to adopt to other study areas, and more generic approaches are called for. This research provides solutions that are more generic and are easily applicable in the European context.

In future we plan on focusing in two directions: first extending the simulation based approach to accommodate the synthesis of associations among agents (e.g. households and persons). Secondly, we are interested in developing a hybrid approach in which concepts from FG are used to post-process the population generated from MCMC. This way a better fit with available marginals and conditionals can be maintained.

References

Abraham, J. E., Stefan, K. J. & Hunt, J. D. (2012), Population synthesis using combinatorial optimization at multiple levels, *in* '91st Annual Meeting of the Transportation Research Board'.

Arentze, T., van den Berg, P. & Timmermans, H. (2012), 'Modeling social networks in geographic space: approach and empirical application', *Environment and Planning A* **44**(5), 1101–1120.

Auld, J. & Mohammadian, A. K. (2010), 'Efficient methodology for generating synthetic populations with multiple control levels', *Transportation Research Record* **2183**, 19–28.

Auld, J., Mohammadian, A. K. & Wies, K. (2008), Population synthesis with region-level control variable aggregation, *in* '87th Annual Meeting of the Transportation Research Board', Transportation Research Board, Washington, D.C.

Bar-Gera, H., Konduri, K., Sana, B., Ye, X. & Pendyala, R. M. (2009), Estimating survey weights with multiple constraints using entropy optimization methods, *in* '88th Annual Meeting of the Transportation Research Board', Transportation Research Board, Washington, D.C.

Barthelemy, J. & Toint, P. L. (2012), 'Synthetic population generation without a sample', *Transportation Science* .

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

Beckx, C., Arentze, T. A., Int Panis, L., Janssens, D., Vankerkorn, J. & Wets, G. (2009), 'An integrated activity-based modelling framework to assess vehicle emissions: Approach and application', *Environment and Planning B* **36**(6), 1086–1102.

Ben-Akiva, M. E., Bierlaire, M., Koutsopoulos, H. & Mishalani, R. (2002), Real time simulation of traffic demand-supply interactions with DynaMIT, *in* M. Gendreau & P. Marcotte, eds, 'Transportation and Network Analysis: Current Trends', Kluwer, Dordrecht, pp. 19–36.

Bhat, C. R., Guo, J. Y., Srinivasan, S. & Sivakumar, A. (2004), 'A comprehensive econometric microsimulator for daily activity-travel patterns (cemdap)', *Transportation Research Record* **1894**, 57–66.

Bowman, J. L. & Ben-Akiva, M. E. (2001), 'Activity-based disaggregate travel demand model system with activity schedules', *Transportation Research Part A: Policy and Practice* **35**(1), 1–28.

Bradley, M. A., Bowman, J. L. & Griesenbeck, B. (2010), 'SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution', *Journal of Choice Modelling* **3**(1), 5–31.

Brand, R. (2002), 'Microdata protection through noise addition, inference control in statistical databases: From theory to practice', *Lecture Notes in Computer Science* **2316**.

Dalenius, T. & Reiss, S. P. (1982), 'Dataswapping: A technique for disclosure limitation', *Journal of Statistical Planning and Inference* **6**, 73 – 85.

de Palma, A. & Marchal, F. (2002), 'Real cases applications of the fully dynamic METROPOLIS tool-box: An advocacy for large-scale mesoscopic transportation systems', *Networks and Spatial Economics* **2**(4), 347–369.

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

D'Orazio, M., Zio, M. D. & Scanu, M. (2006), *Statistical Matching: Theory and Practice*.
URL: *http://onlinelibrary.wiley.com/book/10.1002/0470023554*

Farooq, B., Salvini, P. J. & Miller, E. J. (2008), Development of an operational integrated urban model system: Software documentation, Technical report, Urban Transportation Research and Advancement Centre, University of Toronto.

Frazier, T. & Alfons, A. (2012), Generating a close-to-reality synthetic population of ghana, Open access publications from katholieke universiteit leuven, Katholieke Universiteit Leuven.

Gargiulo, F., Ternes, S., Huet, S. & Deffuant, G. (2010), 'An iterative approach for generating statistically realistic populations of households', *PloS one* **5**(1), e8828. PMID: 20107505.

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

Guo, J. Y. & Bhat, C. R. (2007), 'Population synthesis for microsimulating travel behavior', *Transportation Research Record* **2014**(12), 92–101.

Harland, K., Heppenstall, A., Smith, D. & Birkin, M. (2012), 'Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques', *Journal of Artificial Societies and Social Simulation* .

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**(1), 97–109.

Hermes, K. & Poulsen, M. (2012), 'A review of current methods to generate synthetic spatial microdata using reweighting and future directions', *Computers, Environment and Urban Systems* **36**(4), 281–290.

Jones, P. M., Dix, M. C., Clarke, M. I. & Heggie, I. G. (1983), *Understanding Travel Behaviour*, Gower, Aldershot.

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.
URL: *http://ntl.bts.gov/DOCS/466.html*

Lee, D.-H. & Fu, Y. (2011), 'Cross-entropy optimization model for population synthesis in activity-based microsimulation models', *Transportation Research Record* **2255**, 20–27.

Lenormand, M. & Deffuant, G. (2012), 'Generating a synthetic population of individuals in households: Sample-free vs sample-based methods', *arXiv:1208.6403* .
URL: *http://arxiv.org/abs/1208.6403*

Ma, L. (2011), *Generating disaggregate population characteristics for input to travel-demand models*, Dissertation, University of Florida.
URL: *http://gradworks.umi.com/35/14/3514962.html*

Mahmassani, H. S., Hu, T. & Jayakrishnan, R. (1995), Dynamic traffic assignment and simulation for advanced network informatics (DYNASMART), *in* N. H. Gartner & G. Improta, eds, 'Urban traffic networks: dynamic flow modeling and control', Springer, Berlin.

Müller, K. & Axhausen, K. W. (2011*a*), Hierarchical IPF: Generating a synthetic population for Switzerland, *in* ERSA, ed., '51st Congress of the European Regional Science Association', University of Barcelona, Barcelona.

Müller, K. & Axhausen, K. W. (2011*b*), Population synthesis for microsimulation: State of the art, *in* TRB, ed., '90th Annual Meeting of the Transportation Research Board', Transportation Research Board, Washington, D.C.

Müller, K. & Axhausen, K. W. (2012), Preparing the Swiss Public-Use Sample for generating a synthetic population of Switzerland, *in* STRC, ed., '12th Swiss Transport Research Conference', Ascona.

MATSim-T (2011), 'Multi Agent Transportation Simulation Toolkit', webpage.
URL: *http://www.matsim.org*

Miller, E. J., Noehammer, P. & Ross, D. R. (1987), "a micro-simulation model of residential mobility", *in* '"International Symposium on Transport, Communication and Urban Form: 2, Analytical Techniques and Case Studies"'.

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

Otani, N., Sugiki, N., Vichiensan, V. & Miyamoto, K. (2012), Modifiable attribute cell problem and a method of solution for population synthesis in land-use microsimulation.

Pendyala, R. M., Bhat, C. R., Goulias, K. G., Paleti, R., Konduri, K. C., Sidharthan, R., Hu, H.-h., Huang, G. & Christian, K. P. (2012), The application of a socioeconomic model system for activity-based modeling: Experience from southern california.

Pritchard, D. R. (2008), Synthesizing agents and relationships for land use/transportation modelling, Master's thesis, University of Toronto.
URL: *http://hdl.handle.net/1807/17214*

Pritchard, D. R. & Miller, E. J. (2009), Advances in agent population synthesis and application in an integrated land use and transportation model, *in* TRB, ed., '88th Annual Meeting of the Transportation Research Board', Transportation Research Board, Washington, D.C.

R Development Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna.
URL: *http://www.r-project.org*

Rahman, A., Harding, A. M., Tanton, R. & Liu, S. (2010), 'Methodological issues in spatial microsimulation modelling for small area estimation', *International Journal of Microsimulation* **3**(2), 3–22.

Rich, J. & Mulalic, I. (2012), 'Generating synthetic baseline populations from register data', *Transportation Research Part A: Policy and Practice* **46**(3), 467–479.

Roorda, M. J., Miller, E. J. & Habib, K. M. N. (2008), 'Validation of TASHA: A 24-h activity scheduling microsimulation model', *Transportation Research Part A: Policy and Practice* **42**(2), 360–375.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

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.

Srinivasan, S. & Ma, L. (2009), Synthetic population generation: A heuristic data-fitting approach and validations, *in* IATBR, ed., '12th International Conference on Travel Behaviour Research (IATBR)', Jaipur.

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

Swiss Federal Statistical Office (BFS) (2000*a*), 'Eidgenössische Volkszählung 2000'. **URL:** *http://www.bfs.admin.ch/bfs/portal/de/index/infothek/erhebungen__quellen/blank/blank/vz/uebersicht.html*

Swiss Federal Statistical Office (BFS) (2000*b*), 'Public use samples (PUS): Excerpts for general use from the Swiss federal population censuses 1970-2000'. **URL:** *http://www.portal-stat.admin.ch/pus/files/index_e.html*

Train, K. (2003), *Discrete Choice Methods with Simulation*, Discrete Choice Methods with Simulation, Cambridge University Press.

UrbanSim (2011), 'Open Platform for Urban Simulation', webpage. **URL:** *http://www.urbansim.org*

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.

Williamson, P. (2013), An evaluation of two synthetic small-area microdata simulation methodologies: Synthetic reconstruction and combinatorial optimisation, *in* R. Tanton & K. Edwards, eds, 'Spatial Microsimulation: A Reference Guide for Users', number 6 *in* 'Understanding Population Trends and Processes', Springer Netherlands, pp. 19–47.

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

Ye, X., Konduri, K., Pendyala, R. M., Sana, B. & Waddell, P. A. (2009), A methodology to match distributions of both household and person attributes in the generation of synthetic populations, *in* TRB, ed., '88th Annual Meeting of the Transportation Research Board', Transportation Research Board, Washington, D.C.