Simulation based generation of synthetic population for Brussels case study

Bilal Farooq*[1], Ricardo Hurtubia†[2] and Michel Bierlaire‡[3]

[1]Départment des Génies Civil, Géologique et des Mines, École Polytechnique de Montréal, Canada
[2]Facultad de Arquitectura y Urbanismo, Universidad de Chile, Santiago, Chile
[3]Transport and Mobility Lab, École Polytechnique Féderale de Lausanne, Switzerland

Wednesday 21st August, 2013

Abstract

In the Brussels case study, we had access to very limited amount of data (e.g. a microsample of approximately 0.1% of the total households). Initial experiments suggested that conventional approaches, for instance the Iterated Proportional Fitting (IPF), will fail to generate a population with the required attributes. To overcome this difficulty, a Markov Chain Monte Carlo (MCMC) simulation based synthesis method was developed. We used Gibbs sampler to simulate the draws from the joint distribution of household attributes. As an input, this required generating conditional distributions for the attributes over others.

Household agents with eight attributes (income, household size, workers, adults with university, children, cars, dwelling type, and sector) were synthesized for the Brussels case study. A mix of marginals, partial conditionals, discrete choice models, and assumptions were used to construct the input to the synthesis process. Here we present the input data preparation process, brief overview of the methodology, operationalization of a software (*SimPSynz*), and analysis of the quality of the generated population. The Census commune level comparison of the marginals from real and simulated population for average income showed a maximum error of €375.

*Correspondance: bilal.farooq@polymtl.ca
†rhurtubia@uchilefau.cl
‡michel.bierlaire@epfl.ch

1 Introduction

The microsimulation of urban systems evolution using UrbanSim and MATSim requires an agent population for the base year. As a first step in Brussels case study a set of household agents with rich attributes was synthesized. The main challenge here was to overcome constraint of very limited data availability. At sector and commune level, we had access to up to three-way cross-classification tables from 2001 Census. The Census statistics contained aggregate information for the 1.2 million households of the area of study. While on micro-level, we only had access to an approximately 0.1% sample of the household (1367 records) in the study area. This sample came from the 1999 household survey conducted by Hubert & Toint (2002). The area of study consists of 151 communes that are further divided in 4945 sectors.

In the initial attempt we tried to use the Iterative Proportional Fitting (IPF) approach (Beckman et al. 1996) to synthesize the household population, but given the small size of the sample, it was a challenge. In our further investigation on another case study where we had access to the real population, we realized that to have a reasonable fit with the real population a sample size of at least 5% was required. We thus focused our attention to developing a new methodology for synthesis that can overcome the above mentioned challenges. This methodology involved using Markov Chain Monte Carlo (MCMC) simulation to draw agents directly from the joint distribution of their attributes rather than searching of an optimal weighting scheme using some fitting process. In particular, we used Gibbs sampler to simulate the draws. In each iteration, the potential new agent is generated by i) copying the agent generated in last iteration ii) selecting one attribute (X_i) of the agent using a uniform draw iii) using antithetic draw on the conditional-marginal distribution of X_i over the values of all the other attributes of the agent, the new value x_i is generated. This process is repeatedly run to generate the agent population. The details of the methodology and algorithm can be found in Chapter Farooq, Müler, Bierlaire & Axhausen (2013) and Farooq, Bierlaire, Hurtubia & Flötteröd (2013).

The major advantage of using a simulation based approach here was that data from various sources at various spatial and agent aggregation, domain knowledge, and assumptions can be consistently and seamlessly used to construct the conditional-marginal distributions required as input for the Gibbs sampler. MCMC based procedures allow for generation of a mix of discrete and continuous attributes. The proposed methodology does not suffer from the scalability issues with respect of the number of attributes. The only information that is required to generated the next agent is the previous agents synthesized and the input conditionals.

In next section we first describe the attributes that were needed to be synthesized for the Brussels case study. The available data sources and input preparation process is described. The implementation details of the proposed methodology are discussed. We analyse the generated population and compare it to the available dataset. In the end, a concluding discussion is presented.

2 Data

The models estimated in UrbanSIM and MATSim (e.g. dwelling or job location models) needed attributes of household and its members to evaluate their decisions during the simulation period. For this, as a starting point, the base year (2001) population with these attributes had to be synthesized. The agent type synthesized was *Household* that contained attributes representing resources (dwelling type/sector, and number of cars) and information on the person agents associated to it. The complete list of these attributes is described in Table 1, while details on models that used these attributes as explanatory variables, the parameter values, and other statistics can be found in Efthymiou et al. (2013).

Table 1: Household attributes synthesized for Greater Brussels Areas in 2001

Attribute	levels
Income level of the household (inc_h)	1 (0-1859 Euros)
	2 (745-1859 Euros)
	3 (1860-3099 Euros)
	4 (3100-4958 Euros)
	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
Zone[a] (i)	4945

a: We used Belgian Census Sectors as zones

The study area consists of 151 communes that are further divided in 4945 sectors. Data from three different sources were available for the study area. From 2001 Census sector and commune level statistics were available in form of cross-tabulations. Belgium Land Registry data provided a cadastre of real estate goods. These datasets were obtained at sector or commune level from the Belgian Statistical Authority (SPF - Economie, http://www.economie.fgov.be). While on micro level, we had access to an approximately 0.1% sample of the household (1367) in the study area. This sample came from the 1999 household survey of Belgium, called MOBEL Hubert & Toint (2002).

To generate the new state in a Markov chain, Gibbs sampler needs access to conditional distribution of one attribute on all the remaining attributes. This means that in total we needed 8 conditionals of as input. Table 2 lists out the exact details of the all these conditional-marginal distributions.

Table 2: Conditionals needed for synthesizing population of Great Brussels Area in 2001

No.	Conditional
1	$\pi(\text{inc}_h \mid \text{size}_h, \text{children}_h, \text{workers}_h, \text{cars}_h, \text{univ}_h, \text{v}_h, \text{i})$
2	$\pi(\text{size}_h \mid \text{inc}_h, \text{children}_h, \text{workers}_h, \text{cars}_h, \text{univ}_h, \text{v}_h, \text{i})$
3	$\pi(\text{children}_h \mid \text{inc}_h, \text{size}_h, \text{workers}_h, \text{cars}_h, \text{univ}_h, \text{v}_h, \text{i})$
4	$\pi(\text{workers}_h \mid \text{inc}_h, \text{size}_h, \text{children}_h, \text{cars}_h, \text{univ}_h, \text{v}_h, \text{i})$
5	$\pi(\text{cars}_h \mid \text{inc}_h, \text{size}_h, \text{children}_h, \text{workers}_h, \text{univ}_h, \text{v}_h, \text{i})$
6	$\pi(\text{univ}_h \mid \text{inc}_h, \text{size}_h, \text{children}_h, \text{workers}_h, \text{cars}_h, \text{v}_h, \text{i})$
7	$\pi(\text{v}_h \mid \text{inc}_h, \text{size}_h, \text{children}_h, \text{workers}_h, \text{cars}_h, \text{univ}_h, \text{i})$
8	$\pi(\text{i} \mid \text{inc}_h, \text{size}_h, \text{children}_h, \text{workers}_h, \text{cars}_h, \text{univ}_h, \text{v}_h)$

We chose to synthesize the population zone-by-zone. This meant that the first seven conditionals in Table 2 were required for each zone and last one was not needed. The full-conditionals were constructed using assumptions and discrete conditional models. The process is explained in the following sections.

2.1 Estimated models

Four models accounting for conditional relationships between characteristics and attributes were estimated, as described by Table 3 . Of all these models, only the dwelling type model describes a choice, where the alternative (the type of unit) has attributes. The rest of the models (income, car ownership and education) describe the membership of the household to a particular level of the characteristic/variable that can not be described itself in terms of additional attributes. The estimated models do not pretend to provide a behavioral or causal explanation of the choices and characteristics of the households. Instead, they try to find structural (and significant) correlations between characteristics and attributes in order to build the conditional relationships needed for the simulation process.

Table 3: Modelled conditionals for Great Brussels Area

attribute	household variables	spatial variables
inc_h	$\text{cars}_h, \text{workers}_h, \text{univ}_h, \text{v}_h$	income_i
cars_h	$\text{inc}_h, \text{workers}_h, \text{univ}_h, \text{v}_h, \text{children}_h$	income_i, car_ownership_i
univ_h	$\text{inc}_h, \text{workers}_h$	univ_i
v_h	$\text{cars}_h, \text{size}_h$	surface_{vi}

All models consider alternative-specific or level-specific utility functions where the membership to a specific level is described as a function of other socioeconomic characteristic or zonal attributes:

$$V_{in} = \sum_k \beta_k^i X_{nz_{(n)}k} \tag{1}$$

where i is the alternative or the "level" or the attribute (e.g. dwelling type or income level) and $X_{nz_{(n)}k}$ is the k-th attribute of household n, or zone $z_{(n)}$ (the residential location of household n). In some cases the attributes are related only to the household (X_n), only to the zone ($X_{z_{(n)}}$) or are an interaction of both dimensions ($X_{nz_{(n)}} = X_n \cdot X_{z_{(n)}}$). β_k^i is a parameter to be estimated which is specific to alternative or attribute-level i. The introduction of spatial information (z) produces a richer set of conditional distributions, allowing to have a different conditional relationship for each zone.

2.1.1 Dwelling type model

Household can live in four types of dwellings: Isolated houses, semi-attached houses, attached houses and apartments. The choice of type of dwelling depends on characteristics of the household like size and car ownership, and zonal attributes like the average size of each type of dwelling in each zone. Estimation results are shown in Table 4

Table 4: Dwelling type model ^{*a}

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
$\beta_{\text{surf}\times\text{h2}}$	dummy for hh size=2 x zonal avg surface of dwelling^b	0.0146	0.00533	2.74
$\beta_{\text{surf}\times\text{h3}}$	dummy for hh size=3 x zonal avg surface of dwelling^b	0.0194	0.00597	3.25
$\beta_{\text{surf}\times\text{h4+}}$	dummy for hh size>3 x zonal avg surface of dwelling^b	0.0249	0.00299	8.31
β^2_{cars}	number of cars in the household	-0.279	0.182	-1.53
β^3_{cars}	number of cars in the household	-0.593	0.207	-2.86
β^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$^{b\,c}$	1	-	-
A^2	ln of number of dwellings of type 2 in zone$^{b\,c}$	1	-	-
A^3	ln of number of dwellings of type 3 in zone$^{b\,c}$	1	-	-
A^4	ln of number of dwellings of type 4 in zone$^{b\,c}$	1	-	-

*: dwelling types are: isolated house (1), semi-attached house (2), attached house(3) and apartment(4)
a: The alternative of isolated house is used as a reference ($\beta^1 = 0$ in all cases with the exception of the surface parameters)
b: spatial variable
c: expansion factor used to account for the (un)availability of different types of dwelling in different zones. Not an estimated parameter

All parameters have the expected sign and most of them are significant at the 95% level with the exception of β^2_{cars}, which is significant at the 88% level. To avoid

under identification, all parameters are set to zero for the isolated house type, with the exception of the surface-related parameters ($\beta_{\text{surf}\times\text{h2}}$,$\beta_{\text{surf}\times\text{h3}}$ and $\beta_{\text{surf}\times\text{h4+}}$) which are generic and can be fully identified for all alternatives since the average surface of the dwelling varies between different types of units in the same zone. We observe that larger households tend to prefer units with a bigger surface, meaning that we should observe a concentration of large households in zones where the average size of units is big and also in unit types with bigger average surface. The number of cars in the household has a negative effect for (semi)attached houses and apartments, which is expected since this type of housing tends to be located in denser areas of the city, where parking spaces are scarcer. The utility function is corrected with the expansion factor A^i in order to take into account the fact that, for each zone, households are distributed across types of dwelling according to their availability. This also makes sure that the probability of a household locating in a type of dwelling that does not exist in a zone will be zero.

2.1.2 Income level model

Membership to an income level is modeled as a function of household characteristics and the average zonal income of the residential location. There are five possible income groups: from low (level 1) to high (level 5). Results for this model are shown in table 5.

All parameters show the expected sign and are significant at the 90% level. Because income levels do not have attributes to identify them, and to avoid under-identification, all parameters are set to zero for the lowest income category (level 1). Households that have members with university degree and that are located in high income zones are more likely to belong to higher income levels, as expected due to observed socioeconomic agglomeration. The number of cars in the household also has a positive effect on the income that is stronger for higher income levels. Households living in a house have a higher probability of belonging to the higher income levels than those living in apartments. Finally, the number of workers in the household also explain a higher income level, although it has a lower effect for the highest level than for mid and mid-high income level, this is probably due to the fact that the highest income levels depend more on the type of jobs than on the number of active persons in the households.

2.1.3 Car ownership model

Car ownership is modeled as a function of household socioeconomics, such as income, education level, presence of children and type of dwelling, and zonal attributes like the average income and the number of households by car ownership-level by zone. Four possible levels of car ownership are modeled: no cars (level 0), one car (level 1), two cars (level 2) and three or more cars per household (level 3). Results are shown in Table 6.

Table 5: 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
β^3_{educ}	dummy for presence of people with higher educ in the hh	0.831	0.177	4.69
β^4_{educ}	dummy for presence of people with higher educ in the hh	1.72	0.314	5.49
β^5_{educ}	dummy for presence of people with higher educ in the hh	1.92	0.656	2.93
$\beta^2_{\text{zonal\_inc}}$	average zonal income[b]	0.0008	0.0004	1.84
$\beta^3_{\text{zonal\_inc}}$	average zonal income[b]	0.0012	0.0005	2.55
$\beta^4_{\text{zonal\_inc}}$	average zonal income[b]	0.0016	0.0005	3.09
$\beta^5_{\text{zonal\_inc}}$	average zonal income[b]	0.0016	0.0006	2.47
β^2_{cars}	number of cars in the household	1.16	0.265	4.39
β^3_{cars}	number of cars in the household	1.92	0.299	6.41
β^4_{cars}	number of cars in the household	2.33	0.341	6.83
β^5_{cars}	number of cars in the household	3.2	0.466	6.87
β^3_{house}	dummy for dwelling being a house	0.45	0.193	2.34
β^4_{house}	dummy for dwelling being a house	0.485	0.294	1.65
β^5_{house}	dummy for dwelling being a house	0.485	0.294	1.65
β^2_{workers}	number of workers in the household	1.14	0.277	4.11
β^3_{workers}	number of workers in the household	2.22	0.295	7.53
β^4_{workers}	number of workers in the household	2.46	0.345	7.13
β^5_{workers}	number of workers in the household	1.74	0.428	4.07

a: the super-index 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

All parameters show the expected sign and are significant at the 94% level. Household with members having a university diploma and with higher income levels are more likely to have a larger number of vehicles. Households located in zones where the average income is high and where households tend to have more cars are more likely to own a larger number of vehicles. Households with children living in houses and with working members have a higher chance of owning more cars.

2.1.4 Education level model

The number of individuals with university degree in a household is modeled as a function of socioeconomics like income, car ownership and number of workers. The presence of other highly educated households in the residential location is also considered as a spatial explanatory variable. Three household education levels are considered: no one in the household has a diploma (level 0), only one person in the household has a diploma (level 1) and 2 or more people in the household have an

Table 6: 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
β^1_{educ}	dummy for presence of people with higher educ in the hh	0.504	0.196	2.57
β^2_{educ}	dummy for presence of people with higher educ in the hh	0.933	0.267	3.49
β^3_{educ}	dummy for presence of people with higher educ in the hh	1.07	0.552	1.94
$\beta^1_{\text{high\_inc}}$	dummy for households with high income (>3)	0.977	0.499	1.96
$\beta^2_{\text{high\_inc}}$	dummy for households with high income (>3)	2.43	0.569	4.28
$\beta^3_{\text{high\_inc}}$	dummy for households with high income (>3)	3.24	0.801	4.05
$\beta^1_{\text{mid\_inc}}$	dummy for households with mid income (=3)	0.858	0.267	3.22
$\beta^2_{\text{mid\_inc}}$	dummy for households with mid income (=3)	1.87	0.342	5.46
$\beta^3_{\text{mid\_inc}}$	dummy for households with mid income (=3)	1.42	0.676	2.11
$\beta^1_{\text{zonal\_inc}}$	average zonal income[b]	0.001	0.0003	3.51
$\beta^2_{\text{zonal\_inc}}$	average zonal income[b]	0.0013	0.0004	3.29
$\beta^3_{\text{zonal\_inc}}$	average zonal income[b]	0.0013	0.0004	3.29
$\beta^1_{\text{car1\_zone}}$	percentage of hh's with 1 car in the zone[b]	0.498	0.172	2.89
$\beta^2_{\text{car2\_zone}}$	percentage of hh's with 2 cars in the zone[b]	2.13	0.842	2.53
$\beta^3_{\text{car3\_zone}}$	percentage of hh's with 3+ cars in the zone[b]	14.1	7.57	1.86
$\beta^1_{\text{children}}$	dummy for presence of children in the household	0.457	0.24	1.9
$\beta^2_{\text{children}}$	dummy for presence of children in the household	0.8	0.276	2.9
β^1_{house}	dummy for dwelling being a house	0.841	0.191	4.4
β^2_{house}	dummy for dwelling being a house	1.86	0.289	6.42
β^3_{house}	dummy for dwelling being a house	2.66	0.776	3.43
β^1_{workers}	number of workers in the household	0.437	0.139	3.15
β^2_{workers}	number of workers in the household	1.24	0.193	6.42
β^3_{workers}	number of workers in the household	1.6	0.358	4.46

a: the super-index 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

university diploma (level 2). Results are shown in Table 7.

All parameters show the expected sign and are significant at the 93% level. Households with more cars, higher income and more workers are more likely to have more individual with a high education level. Households located in zones with a higher percentage of households with highly educated members are more likely to have a higher education level.

2.2 Assumptions-based conditionals

For the rest of the attributes (i.e. household size, number of children, and number of workers), we did not have enough information to estimate the discrete-conditional

Table 7: Household education level model[a]

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
β^1_{cars}	number of cars in the household	0.238	0.133	1.79
β^2_{cars}	number of cars in the household	0.701	0.156	4.51
$\beta^1_{\text{educ\_zone}}$	percentage of hh's with educ level 1 in zone[b]	3.34	0.566	5.91
$\beta^2_{\text{educ\_zone}}$	percentage of hh's with educ level in zone[b]	4.34	0.708	6.13
β^1_{income}	income level of the household	0.24	0.129	1.87
β^2_{income}	income level of the household	1.09	0.152	7.18
β^1_{workers}	number of workers in the household	0.393	0.113	3.47
β^2_{workers}	number of workers in the household	0.851	0.154	5.52

a: the super-index 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 ($\beta^0_* = 0$)

b: spatial variable

models. Here we had to rely on the assumptions and domain knowledge to construct the full-conditionals. The only dataset that has information on these three attributes was MOBEL. As mentioned earlier, this dataset is very small and various zones (Census sectors) are either not represented or have very few observations. We thus decided to pool the dataset. The actual conditionals were generated by counting each category of an attribute for rest of the attributes combinations. For example, counting the number of households with size equal 1, 2, 3, 4, and 5+ for all the possible category combinations of income level, children, workers, cars, education, and dwelling type. We assumed that the household structure for the household size, number of children, and number of workers follow the same conditional distributions in all the sectors of the study area. We understand that in general there are differences in the structure among households living in city centres, uptown, suburbs, etc. But in the absence of any other information, we are not able to make the conditionals sensitive to the spatial heterogeneity for these three attributes.

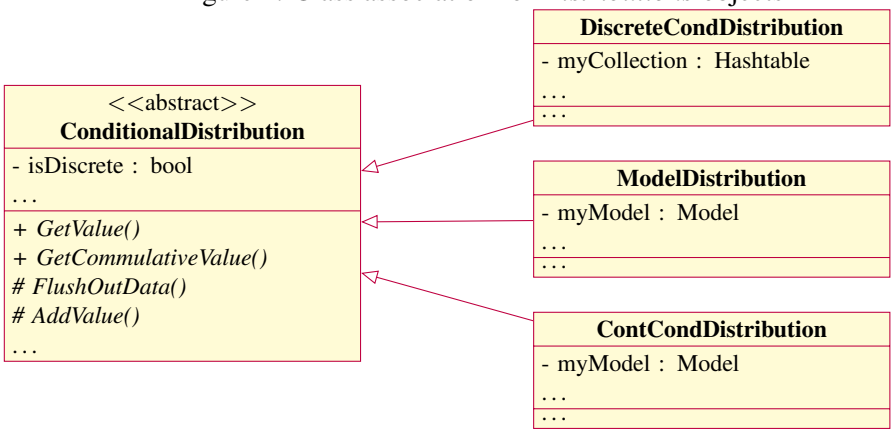
3 Population generation

To operationalize the simulation based methodology for the generation of synthetic population, we developed a generic object oriented software called **Sim**ulation based **P**opulation **Syn**thesi**z**er (**SimPSynz**). This software can i) pre-process the data to generate full-conditionals of the attributes ii) generate a large pool of synthetic agents using different types of samplers iii) produce any number of realization of synthetic populations from that large pool. These realizations can be based on user defined marginals or just uniform sampling. The implementation is available upon request in $C^{\#}$ and $Java$ programming languages.

3.1 Software architecture

An object oriented class structure was designed[1]. The software architecture was divided into three interacting group of class libraries: *Distributions*, *Samplers*, and *SimulationObjects*. A top level handler class (PopSyn) was implemented in order to integrate the functionality from these libraries into a coherent software. Here we will only describe the main classes of these three libraries. The first library maintains the conditional distributions in form of counts (DiscreteCondDistribution), discrete-conditional models (ModelDistribution) and continuous functions (ContCondDistribution). The structure and association of this library is described in Figure 1. Note that at the moment, functionality of the ContCondDistribution class is not completely implemented.

Figure 1: Class association for *Distributions* objects[a]



[a] : "..." denotes that there are additional attributes and functionalities of a class that we did not listed here for clarity reasons.

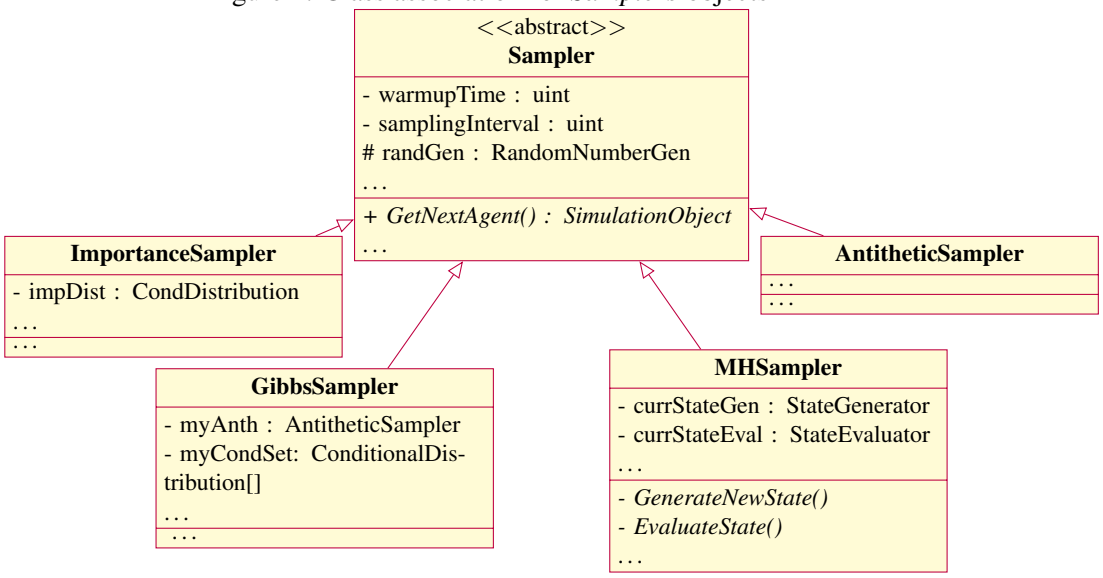
The second set of classes represents the four different types of samplers. The main class used to synthesize the population for the Brussels study area was the Gibbs sampler (*GibbsSampler*). However, for further options, we also implemented Metropolis Hasting (*MHSampler*), importance (*ImportanceSampler*), and antithetic (*AntitheticSampler*) samplers. All these classes inherit basic functionality from the main abstract class called *Sampler*. The structure and association of this library is described in Figure 2. *GibbsSampler* operates on various conditionals and with the help of a contained *AntitheticSampler* object, it generates the agents iteratively. The iteration parameters (e.g. warm up time) are set in the parent class. The *MHSampler* needs an object of type *StateGenerator* and an evaluator (*StateEvaluator*) to generate an agent. It also work in iterative manner and inherits paramters from the parent class, *Sampler*.

[1] Here, it is assumed that the reader is familiar with the basics of Unified Modelling Language (UML)

The classes that are involved in the actual simulation are defined in *SimulationObjects* class library. The basic functionality is defined in the *SimulationObject* abstract class, while all the other classes inherit this class. Figure 3 represents the class diagram of the library. *World* maintains and manages the states of various simulation objects. Furthermore, it also controls the execution of different samplers and the state of available information. In essence it represents the functionality of the actual World in virtuality. Space is represented by *SpatialZone* class. It contains the information and operations relevant to specific space. The class is designed such that it can be extended in hierarchy to accommodate different types of zoning systems (sectors, communes, etc.). Currently there are two different types of agent classes, *Person* and *Household*. In the Brussels case study, only the *Household* class was utilized. This library also maintains supporting enumerations and structures that represents the agents attributes, types, and associations.

In addition to the three main libraries, *Utilities* class library was defined that contained support-classes. They covered the functionality related to random number generations (based on various distributions), file reading/writing, generation of the conditionals from the raw data, and various conversion utilities.

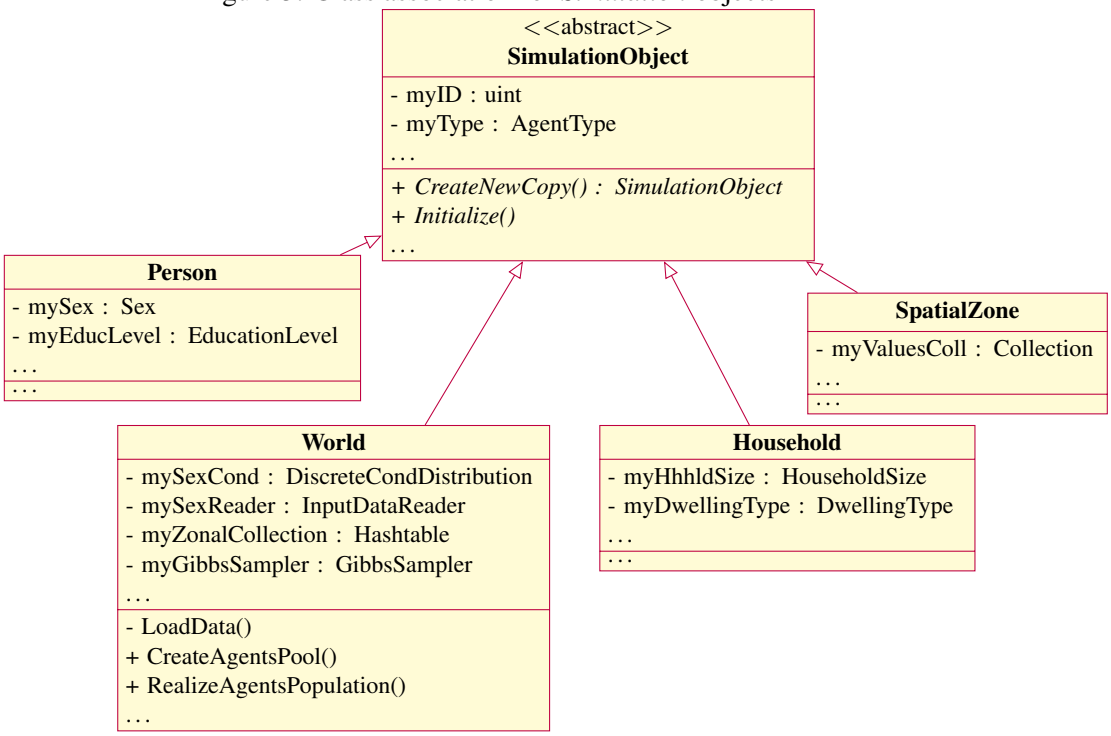
Figure 2: Class association for *Samplers* objects



3.2 Simulation runs

The actual synthesis of the population for Brussels study area involved first preprocessing the available datasets using the utilities developed in SimPSynz. Once the conditionals were available via preprocessing process, the Gibbs sampler in SimP-

Figure 3: Class association for *Simulation* objects



Synz was used to generate a household agent pool. It contained around 100 million agents with 10,000 or more agents in one sector. Synthesis was done zone-by-zone. UrbanSim requires the number of dwellings to match exactly with the number of households in the zones. The sector level dwelling type marginals were thus used to sample the synthetic population from the pool of agents. The run-time to generate the agents pool on an Intel Core 2 Duo processor with 4 gigabytes of memory was approximately 8 hours.

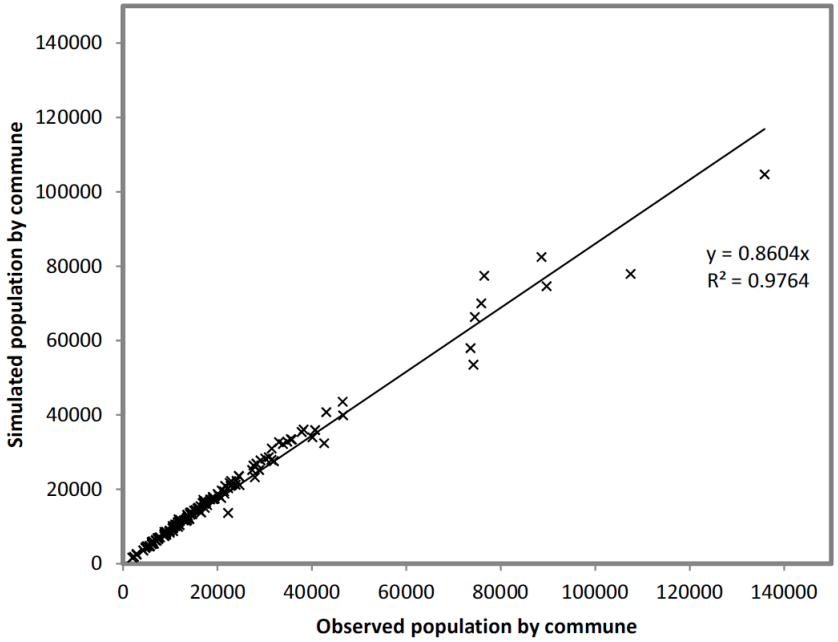
4 Results

After the simulation-based synthesis process is finalized, results are validated by comparing them with observed commune level statistics obtained from the Population Census and the Belgian tax authority.

To analyze the goodness of the synthetic population in terms of distribution of household size, Figure 4 shows the simulated population (number of people) aggregated at the commune level in the Y-axis and the census statistic in the X-axis. The plot also shows the tendency line (solution to the equation $Y = \alpha X$) and the fit (R^2) for the data. The fit between the simulated and observed data is good

(R^2=0.98). However, the synthetic population overall tends to underestimate the number of people by commune ($\alpha = 0.86$), specially for large communes with more than 70000 people. This difference is expected since the census data accounts for a slightly larger total number of households in the whole region, since it considers households located in dwelling of type "other" that was not included in our model.

Figure 4: Population by commune



Figures 5, 6 and 7 compare the synthetic population with the observed number of households with no car, one car and two or more cars respectively. The fit for all three car-ownership levels is considerably good, with a R^2 value of 0.98, 0.99 and 0.94 respectively. While households with no cars tend to be underestimated in the synthetic population, household with one, two and more cars are only slightly overestimated.

Figure 8 shows the number of households with one or more people having a university degree by commune. While the fit is good, the synthetic population tends to overestimate the number of university degrees. The trend is, however, correctly followed.

Finally, Figure 9 shows the observed average household-income by commune, obtained from the Belgian tax authority compared with the simulated income. The average income from the synthetic population is computed as the number of households by income level multiplied by the median value of each income level. Although this is a much harder variable to fit, results are reasonably good, with a R^2 value of 0.76. The plot shows that the synthetic population tends to overestimate the average income for the poorer communes, however the overall trend is correctly

Figure 5: Households with 0 cars by commune

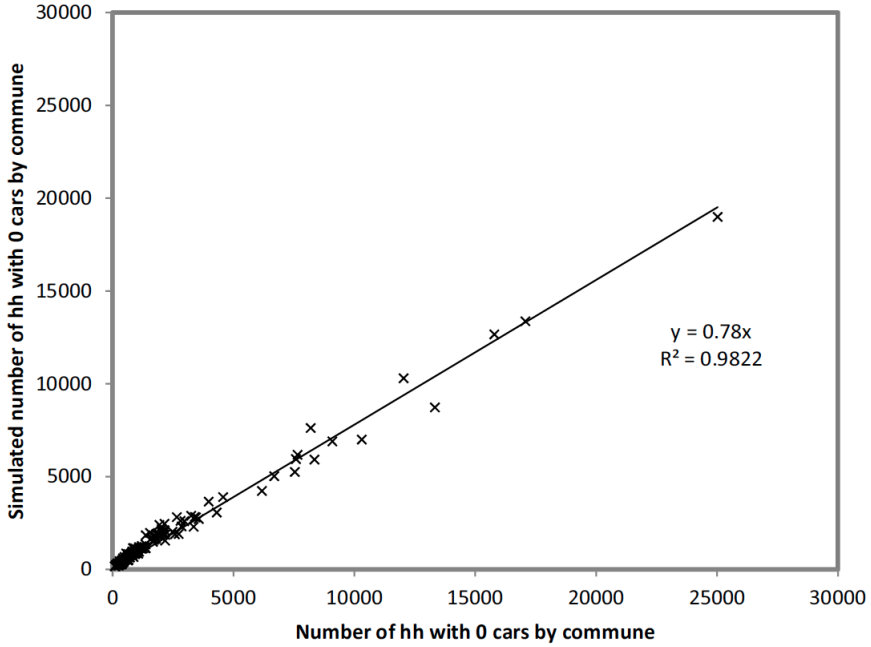
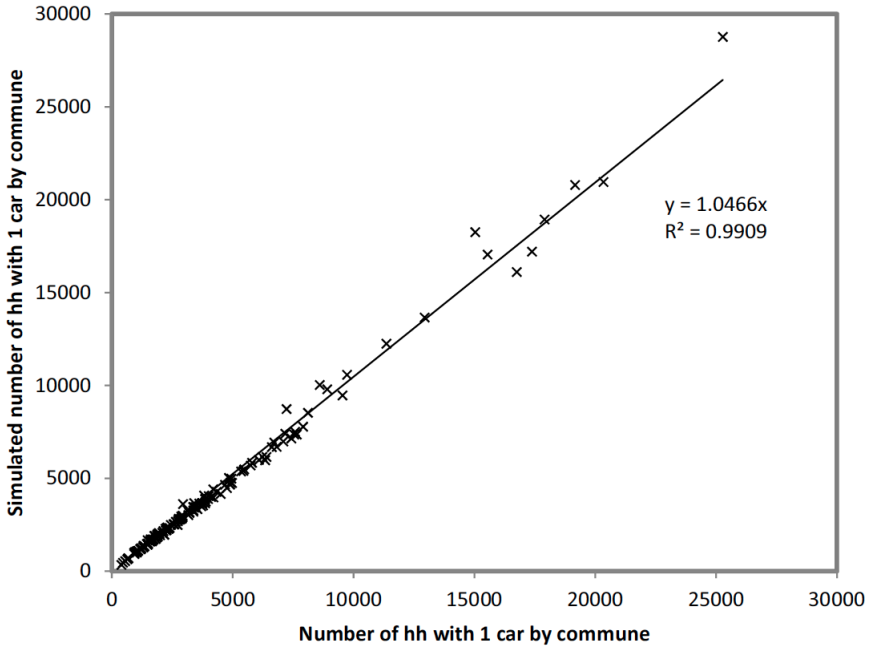


Figure 6: Households with 1 car by commune



followed.

Figure 7: Households with 2+ cars by commune

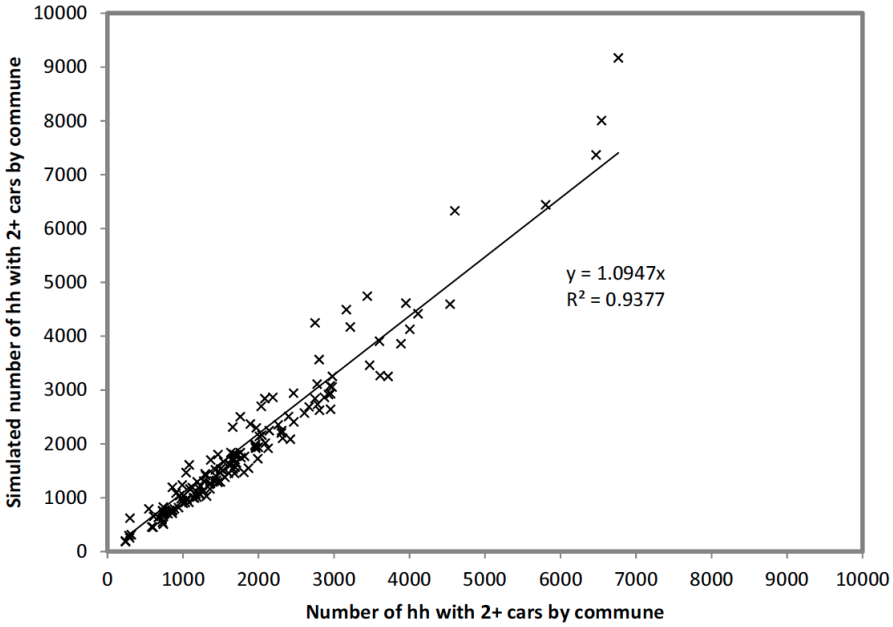


Figure 8: Households with 1+ university degree by commune

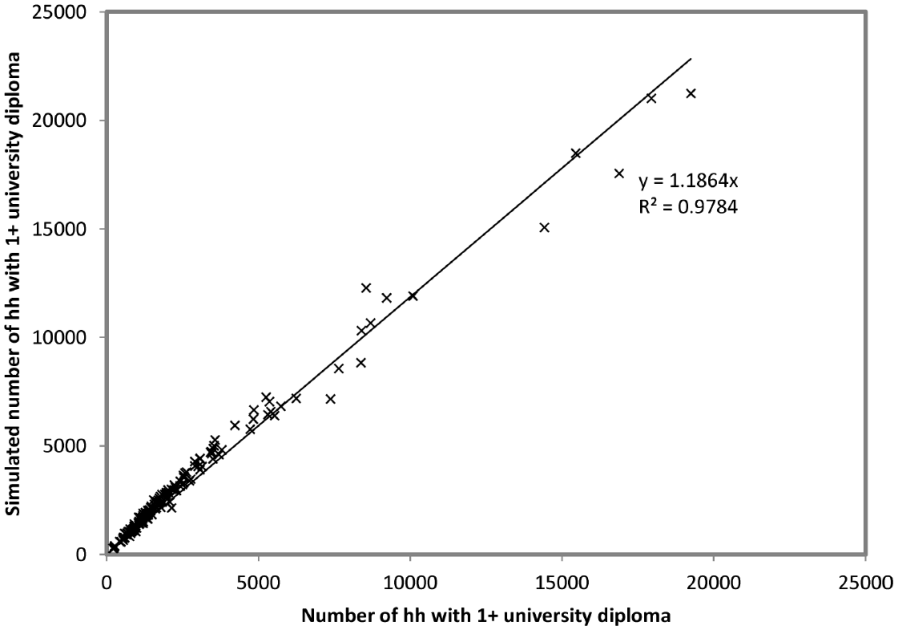
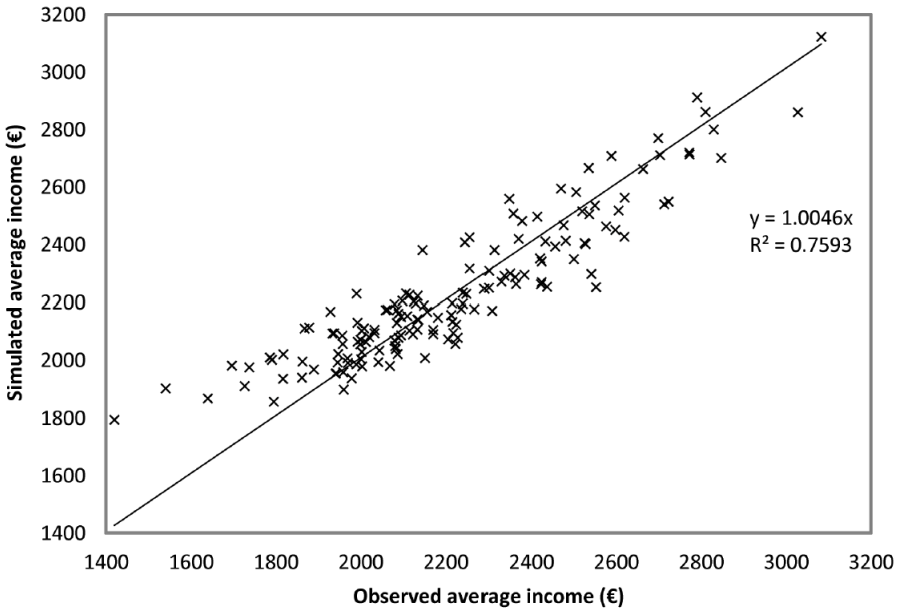


Figure 9: Average income by commune



5 Concluding remarks

Most of the effort in terms of methodologies for population synthesis is concentrated on fitting based approaches. Farooq, Bierlaire, Hurtubia & Flötteröd (2013) has pointed out various shortcomings of such approaches, including: i) fitting of one contingency table, while there may be other solutions matching the available data ii) due to cloning rather than true synthesis of the population, losing the heterogeneity that may not have been captured in the microdata iii) over reliance on the accuracy of the data to determine the cloning weights iv) poor scalability with respect to the increase in number of attributes of the synthesized agents. In this regard, the simulation based approach was proposed. Here we showed an application of the methodology by synthesizing a population for the Brussels case study. The available dataset was highly limited. We had access to only a 0.1% microsample for the area. Conventional fitting based approaches were not very feasible. We thus exploited the flexible input requirements of the simulation based approach to make the best out of the available datasets.

A software tool called *SimPSynz* was developed for the operationalization of the methodology. Household agents with eight attributes, including: location, dwelling type, car ownership, income level, household size, number of children, number of worker, and education level. A mix of conditional-choice models, assumptions, and domain knowledge were used to prepare the input. The population was synthesized

zone-by-zone. The analysis of the generated population indicated a good fit with the available data.

The literature on methodology for population synthesis is predominantly based on North American datasets (Pritchard (2008); Müller & Axhausen (2011)). Extending these methodologies in order to use census formats from other geographic regions (including Europe) is not trivial. In the context of SustainCity project the simulation based methodology and its application to Brussels case study provides a step forward towards approaches that are more flexible to adapt and are directly applicable in the European context.

In future we intend to extend the simulation based methodology to other cities in Europe and other parts of the world. We intend to develop SimPSynz as a generic tool for the synthesis of agent population and their associations. Moreover, due to its simulation based nature SimPSynz is planned to be embedded directly in the microsimulation frameworks for modelling the urban systems dynamics.

References

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.

Efthymiou, D., Hurtubia, R., Bierlaire, M. & Antoniou, C. (2013), The integrated land use and transport model of brussels, *in* 'Proceedings of the Swiss Transport Research Conference', Ascona, Switzerland.

Farooq, B., Bierlaire, M., Hurtubia, R. & Flötteröd, G. (2013), Simulation based population synthesis, *in* 'Proceedings of the 13th Conference of Iternational Association of Travel Behaviour Research', Toronto, Canada.

Farooq, B., Müler, K., Bierlaire, M. & Axhausen, K. W. (2013), *Methodologies for synthesizing populations*, SustainCity Hand book, EPFL Press, Lausanne.

Hubert, J. P. & Toint, P. L. (2002), 'La mobilite quotidienne des belges', *Mobilite et Transports* **1**.

Müller, K. & Axhausen, K. W. (2011), 'Population synthesis for microsimulation: State of the art', *Proceeding of Transportation Research Board 90th Annual Meeting* .

Pritchard, D. R. (2008), *Synthesizing Agents and Relationships for Land Use/transportation Modelling*, Canadian theses, University of Toronto.