Associations Generation in Synthetic Population for Transportation Applications. A Graph-Theoretic Solution

Date of submission: 2013-08-01

Paul Anderson
MSc Student, École Polytechnique Fédérale de Lausanne,
Station 18, Section Génie Civil, 1015 Lausanne
phone: +41 79 881 27 67
fax: +41 21 693 80 60
paul.anderson@epfl.ch

Bilal Farooq
Professor adjoint, École Polytechnique de Montréal,
Département des génies civil, géologique et des mines,
phone: +1 (514) 340-4711 poste 4802
fax: +1 (514) 340-3981
bilal.farooq@polymtl.ca

Dimitrios Efthymiou*
PhD Candidate, National Technical University of Athens,
9 Iroon Polytechneiou, 15780 Athens,
Scientific Assistant, École Polytechnique Fédérale de Lausanne
phone: +30 210 772 16 75
fax: +30 210 772 26 29
defthym@mail.ntua.gr, dimitrios.efthymiou@epfl.ch

Michel Bierlaire
Professor, École Polytechnique Fédérale de Lausanne,
Station 18, INTER, ENAC, TRANSP-OR, 1015 Lausanne
phone: +41 21 693 25 37
fax: +41 21 693 80 60
michel.bierlaire@epfl.ch

4927 Words + 6 Figures + 2 Tables = 6927

Submitted for presentation to the 93rd Annual Meeting of the Transportation Research Board and publication in Transportation Research Record: Journal of the Transportation Research Board.

[*] Corresponding author

ABSTRACT

The generation of synthetic populations through simulation methods is an important research topic and has a key application in transport and land-use agent-based modeling. The next step in this research area is the generation of complete synthetic households, which requires some way to associate synthetic persons with household positions. This work formulates the person to position matching problem as a bipartite graph matching and tests two different models for determining match utility using data from the 2000 Swiss Census. The functions tested are both multinomial logit models, one based on the household size attribute and the other on the household type. Synthetic persons are matched into the head position of real households, and then the remaining population is used to run a second match using a separately calibrated version of the size choice model for the spouse position. This is a long list based approach that keeps the original marginal consistent.

The results show that the size choice model returns the best results for head and spouse positions, although both models provide a good match quality as measured by the distributions of individual attributes in real and matched populations as well as the distributions of unique attribute combinations. Possible extensions include matching to other household positions and evaluating the performance of these synthetic households in modeling applications.

Keywords: Microsimulation, Associations generation, Synthetic population, Integrated land-use and transport models

INTRODUCTION

Agent-based microsimulation is a growing trend in transportation research, fueled by the perpetual increase in computing power as well as greater availability of traffic data. Microsimulation is a broad field and can be traced back to efforts in the 1960s to model individual behavior on a large scale: the effects of taxes, new policy, etc. *(1)*. Later works created a subset called spatial microsimulation, which includes transportation and planning applications *(2)*. Spatial microsimulation grew rapidly after the year 2000 and there is still much potential for future development *(3)*. Microsimulation models are more realistic in their response to network-level changes as compared to zonal aggregation and can be used to evaluate the effects of proposed projects and policy changes *(4)*. One drawback, however, is that a large amount of individual-level data (demographics, potential destinations) is required to accurately reflect behavior. Integrated land-use and transport models, such as UrbanSim *(5)*, have some of the highest data requirements *(6)*.

Individual demographics, as well as home and work locations, are commonly collected in a census, but this data is typically not available for microsimulation due to privacy concerns. This has led to work on population synthesis, which aims to generate virtual people with the same demographics as the real population. Generally two data sets are needed for synthesis: a disaggregate microsample (which typically does not contain any location information) and aggregate measures for the smallest available unit (the name of which varies by country) *(7)*. The idea is that this synthetic population would behave similarly to the real population, allowing for its use in microsimulation and other applications without any of the data confidentiality concerns that arise when real census data is used. Synthetic population generation has applications across many fields of research *(8, 3, 9)*.

Individual attributes are important factors in travel behavior (which includes route choice, mode choice and other decisions), but do not tell the whole story. Households, and particularly an agent's position within the household, play an important role in transportation decisions as well *(10, 11, 12, 13)*. Two-worker households are a particularly significant group and make collective decisions about household location which affect travel behavior *(14)*. Other research papers have considered the influence of social interactions on travel behavior *(*eg. *15, 16, 13)*.

The approach described in this paper is similar to Gargiulo *et al. (17)* and Barthelemy and Toint *(18)* in the sense that the association between synthetic persons and household positions is generated after the generation of individuals and households. That is to say that all three approaches are long-list based rather than fitted table based. (A long-list approach generates a full synthetic population at the beginning. Matching procedure selects the best available person in the population, and people cannot be used more than once. A fitted table approach first creates a contingency table and then in the matching procedure creates a new person by drawing from statistical distributions for each position.) However, it should be noted that only in our work are the agent-level marginals/conditionals before the matching process preserved in the associated population. Both Gargiulo *et al. (17)* and Barthelemy and Toint *(18)* use Monte Carlo simulation to generate the association, which intrinsically means that the marginals at the end of the process will not be an exact match to the individual and household marginals they started with. To avoid unmatched agents, they require a large quantity of oversampling. Also, they are based only on one realization of the simulation. In the case of Gargiulo *et al. (17)*, only age is used to sample the individuals and fit them to households. Their algorithm iteratively draws household size, age of head, household type, spouse age, and children's ages in order from probability distributions conditional to the last attribute chosen. Size and type are the only household attributes considered. Barthelemy and Toint *(18)* also follows a more or less sample procedure where the individual type is used for the match. Our procedure evaluates all attributes simultaneously and is a deterministic solution so the results do not vary from one run to another. We also use more attributes, three for individuals and six for households, compared to three total in Gargiulo *et al. (17)*, which increases the complexity.

A previous work by one of the authors, Farooq *et al. (19)*, looks at inter-agent interactions, modelling the marriage and rental housing markets as bipartite graph matching problems computed by the Hungarian algorithm *(20)*. The association generation problem in this paper is formulated in the same way and also uses the Hungarian algorithm. Persons are matched to household positions in order to have a 1-to-1 matching, required for a minimum weighted bipartite matching. We consider households and individuals to be two independently generated synthetic populations, so there are no prior associations between them. This is another difference compared to Gargiulo *et al. (17)*, who construct households around an identified household head instead of defining all the open positions in the initialization phase.

The rest of the paper is structured as follows. After the introduction, the literature review is presented. The next section describes the developed methodology, the data sources and the model development. The presentation of the results follows. The paper ends with conclusions and recommendations for further research.

LITERATURE REVIEW

Two main methods for population synthesis, are: 'reweighting' and 'imputation' *(21, 22)*. Reweighting is the process of cloning and replicating agents from a small sample to achieve the desired population size. Imputation is the process of generating a completely new population with the same chacateristics as the sample population. Other, more comprehensive literature reviews can be found in Harland *et al. (23),* Hermes and Poulsen *(24),* Müller and Axhausen *(25),* Ma *(26),* Pritchard *(27).*

Reweighting

Reweighting methods can be futher subdivided into *fitting and generation*, and *combinatorial optimization*. These approaches require a reference sample of disaggregate census data as well as aggregate data for the entire population. In the fitting and generation approach, the synthetic population is created by drawing from a weighting of the reference sample *(28)*. Some variants exist, such as a deterministic reweighting using generalized regression *(29, 30)*.

The combinatorial optimization approach involves selecting a population from the reference sample and evaluating the fit to aggregate statistics through the use of tables (one per constraint). The population members are then switched out until an optimal solution is found *(31, 32)*. An advantage of combinatorial optimization is relatively little variation between runs.

A disadvantage of reweighting methods, as Voas and Williamson *(31)* note, is that the resulting synthetic population only produces a good fit to variables which are used as constraints. This implies that this method is oriented towards generating a synthetic population for a specific purpose. Generating a 'general purpose' population is possible, but requires a large number of constraining variables. Combinatorial optimization relies heavily on random number generation, so a sufficiently robust generator is needed since a pseudorandom generator may not be adequate *(33)*. Additionally, the sample population must be sufficiently large and diverse, otherwise reweighting will not be successful *(34)*. More comprehensive reviews of reweighting methods can be found in Huang and Williamson *(34)* and Rahman *et al. (35)*.

Imputation

In general, imputation methods create a population from basic known attributes (such as age and sex) and use probability distributions conditioned to known attributes to sample all other attributes. One such method is the Markov Chain Monte Carlo method, which is an extension of Iterative Proportional Fitting (IPF) proposed by Deming and Stephan *(36)*. This approach first fits a Contingency Table to marginals (which can be calculated from more aggregate demographic data). Farooq *et al. (37)* use a Markov Chain Monte Carlo (MCMC) method to synthesize a population of agents and check the accuracy of this synthetic population by comparing it to full Swiss census data. The procedure in Schafer *(38)* assumes the cell values to be random variables with constrained Dirichlet priors. (A constrained Dirichlet prior is a term introduced to represent uncertainty (prior, or prior probability distribution), which follows a Dirichlet distribution and is subject to the constraints of the contingency table where rows and columns must sum to known values.) The MCMC process is then used to fit cell values from the posterior Dirichlet. The second part of an IPF procedure uses the contingency table and its cell weights to create a synthetic population. Beckman *et al. (28)* use Inverse Transform Sampling in a Monte Carlo simulation to do this.

Validation

Many authors have commented that it is difficult to validate model results without having access to spatially-referenced disaggregate microdata *(39)*. In this study, we do have this kind of data, but find that it is still valuable to review the validation methods which have been used in previous works. One option is to aggregate the synthetic data and compare it with the smallest available level of real data *(30)*. A common measure of fit is the normal Z score, where values over the 5% critical value are judged to be "non-fitting" *(40, 41, 42)*. Total Absolute Error (TAE), Standardized Absolute Error (SAE) and Z score are in common use as measures of fit *(39, 34)*. The best validation methods incorporate both internal (compare input to output) and external (compare to a similar, unused dataset) validation *(43)*.

Other validation methods are regression analysis and the use of R^2 values *(44, 32)* as well as standard error around identity (SEI) *(45, 30, 46)*.

Multinomial Logit

Multinomial logit has seen wide applicability as a mode choice model, and we find that it can also be adapted to a choice between different kinds of households. The method is described in Luce *(47)*, Marschak *(48)* and a literature review can be found in Ben-Akiva and Lerman *(49)*.

METHODOLOGY

An example of a bipartite graph is shown in Figure 1. Set U represents the list of persons while set V represents the list of households positions, since the matching for each position is run separately. These two lists are connected by edges, E, which all have a length, L, associated with them; the length is a transformation of the utility function where smaller values represent better matches. In this application, every element of list U has an edge connecting it to every element of list V. Since we know that every connection is possible, the edge lengths can be stored in a cost matrix. The solution for a bipartite graph is matching such that no two edges share an endpoint. In some algorithms, it is necessary for the two lists to have the same length. We use one where unequal length is permitted; in this case, all elements of the shorter list are connected by an edge while some elements of the longer list are left unconnected.

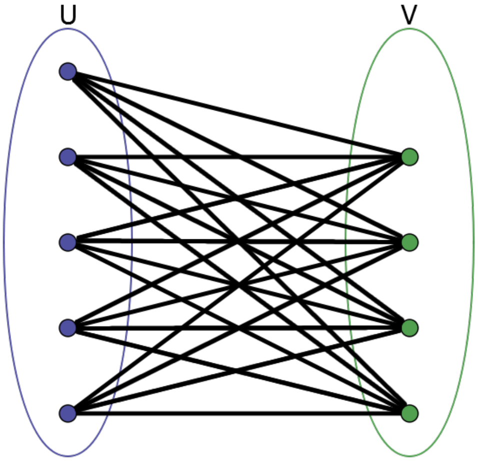


FIGURE 1 Complete Bipartite Graph

The formulation of household-population matching as a bipartite graph problem is similar to the procedure used by Farooq *et al. (19)*. The goal is to determine the one-to-one matching between people and household positions. Mathematically, the problem is to find a graph $G^* = (U, V, E^*, L^*)$ such that the cardinality of E^* is equal to the cardinality of V (but not U, as there is a surplus of people). The sum of lengths L^* is guaranteed to be the minimum possible. This is equivalent to a maximum weighted bipartite matching, which can be solved by linear programming algorithms *(50)*.

The Hungarian algorithm *(20)* is perhaps the most common for bipartite graph matching. However, the Hopcroft-Karp algorithm *(51)* has better performance in terms of time complexity ($O(\sqrt{n})$ vs. O(n^4))[1]. The Hungarian algorithm is currently being used for matching due to the relative ease of implementation. The implementation used is a modified version of the original algorithm *(53)*, which is important because the original requires graphs of equal length, whereas we have a population graph which is larger than the household graph. The modification also reduces the time complexity to $O(n^3)$. A switch to the Hopcroft-Karp algorithm is planned as an extension to this work but has not been completed at this time. The two algorithms use the same problem formulation, so in any case the matching results should be identical. The only advantage is in terms of computation time, which would be significant for a large enough population. For the head matching experiments done in this study (3,100 households), the total computation time is around 5-10 minutes. Computation time was shorter (1-2 minutes) for spouse matching as both graphs were smaller (people matched to head positions were removed from the synthetic population and only 1,200 households had a spouse position).

The population synthesis procedure is the same as Farooq *et al. (37)*, a Markov chain Monte Carlo simulation approach which uses a Gibbs sampler to draw from agent attribute distributions. This study considers the problem of matching synthetic people into household head and spouse positions. Household head is defined in the Swiss Census as “the person

[1] This is big O notation which originates with Bachmann *(52)* and describes time complexity of algorithms

socially and economically responsible for the household." In the case of married couples, one is selected as 'head' and the other as 'spouse' based on economic considerations *(54)*. These are the most important positions, particularly for household-level decision making. Matching people into other household positions is outside the scope of this work, but is planned to be considered in future research. We start from a synthetic population of people and a set of real households from the census data. A sample of 3,000 real households (with the real heads and spouses) is taken from the Lausanne area to mimic the actual data. Not all households have a spouse position so that sample was smaller, around 1,200. This was one of the advantages of having access to full census data, as these records were drawn from the complete population and did not require reweighting. The use of this set for model calibration is equivalent to the use of small-area samples in other studies. A multinomial logit model relating person variables to household variables is fitted to the sample using Biogeme *(55, 56)*. The result is a utility function for possible matches, which is transformed into a shape length attribute for the bipartite match. The bipartite matching algorithm used was written to minimize the distance between potential matches. This, of course, does not work with a utility function since larger values represent a better match. Various transformations are possible to invert the utility function. In this instance, a linear transformation was used to ensures that all functional values are positive and that the best matches are represented by the lowest values. Three different utility functions, described in more detail in the Experiments section, are used. The quality of the matches is evaluated by comparing the total distribution of in the matched and real populations as well as by comparing the count of each unique combination of attributes in the two populations. A flowchart of the method used is shown in Figure 2.

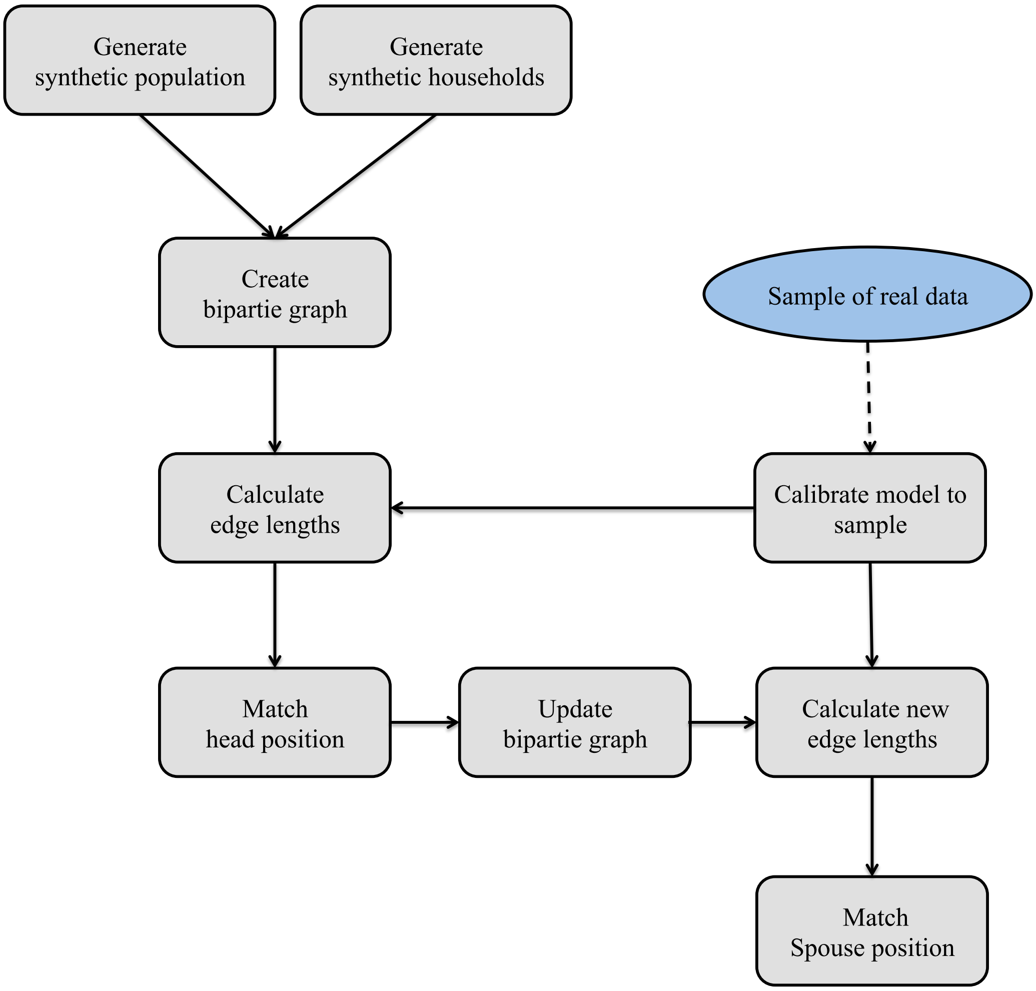


FIGURE 2 Flowchart of Associations Generation Method

DATA

The main data source for this project is the 2000 Swiss Census, which has been made available for research purposes. The person population used in the matching experiments was generated from marginals calculated from the census data using the procedures described in *(37)*. The household list used is all households in postal code 1004, which is located in Lausanne. This zone has 3,100 households, and 1,205 of these households have a spouse position. The census data is at the individual level, which means that household level variables had to be derived based on the unique household ids. Since the synthetic population was very large (over 1,000,000 records), a sample of 6,458 records (equivalent to the population of zone 1004) was taken to shorten the runtime. After the head matching is complete, individuals which have been matched into head positions are removed from the synthetic population for the spouse matching.

Households have seven variables: size, number of workers, dwelling type, number of cars, type, nationality, and language spoken. All of these were calculated from the census data by performing a dissolve on the household ID field and taking a sum/minimum of the other fields. Size has categories for 1, 2, 3, 4, 5, and 6+ persons. Workers is an integer field with a range of 0 to 6. Dwelling contains a unique building ID. This field is unused at the moment but will be incorporated in future work when a list of building IDs and types can be obtained. Cars is an integer field with a range of 0 to 4. This was actually calculated by looking at the number of household members who drive a car to work, since there was no field for car ownership. We believe this to be a reasonable substitute because it is safe to assume that every person who drives a car to work is a car owner. However, there may be cars unaccounted for with this method because they are not used for work trips. Household type has three categories: one person, family, and non-family. All married couples are contained in the family category, regardless of children. Nationality is binary variable with choices Swiss or other. Language variables exist for all of the official languages of Switzerland. Since the study area is in the French-speaking area, we use the French language variable which is also binary (French spoken at home, or any other language spoken at home).

Persons have four variables: age, sex, household size, and education. Age has 8 categories: <15, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, and 75+. Sex is either male or female. Size has the same categories as on the household side: 1, 2, 3, 4, 5, 6+. Education has four categories: none, primary, secondary, and tertiary.

One limitation of the Swiss Census is that there is no field which directly measures income, however number of workers, number of cars, and building type (presently unused) but especially education have some correlation with income.

MODEL DEVELOPMENT

Head Matching

For the head matching, two different models are used: one which constructs the operation as a choice between different household sizes, and a second which chooses between different household types.

Size Choice

Each value in the household size category (ranging from 1 to 6+ persons) is considered as an independent choice. One person households are the base case and have a utility (after transformation for the bipartite match) of:

$$C_{ij} = -ASC_j + L \tag{1}$$

where i represents people and j households. L is a constant parameter to ensure that edge length remains positive. For all other household sizes:

$$\begin{aligned} C_{ij} = -(ASC_j + AgeConst[Age_i] - 0.328 * Sex_i + EducConst[Educ_i] - 0.419 * Nationality_j \\ - 0.393 * Language_j + WorkersConst[Workers_j] + CarsConst[Cars_j]) + L \end{aligned} \tag{2}$$

where AgeConst, EducConst, WorkersConst, and CarsConst are arrays which contain constants for different values of the respective attribute. The values in the arrays can be found in Table 2. The attribute value is used as the index. After trying various values, the parameter L was set to 20.

Type Choice

Each value in the household type category (one person, family, and non-family) is used as an independent choice. One person households are the base case and have a utility (after transformation for the bipartite match) of:

$$C_{ij} = -ASC_j + L \tag{3}$$

For the two other types:

$$\begin{aligned} C_{ij} = -(&ASC_j + AgeConst[Age_i] - 0.390 * Sex_i + EducConst[Educ_i] - 0.327 * Nationality_j \\ &- 0.376 * Language_j + WorkersConst[Workers_j] + CarsConst[Cars_j]) + L \end{aligned} \tag{4}$$

This is very similar in construction to the Size Choice model but the use of a different attribute changes some of the constants and leads to a different fit as can be seen in Table 2.. The parameter L is also set to 20 for this model.

Spouse Matching

Since all households with a spouse position are assigned the family type by the census, the type choice model would be meaningless for the spouse choice problem. Therefore, a variant of the size choice model is used for all spouse matching. It is run twice because the available population depends on the head matching results.

The model is similar to the head size choice model. For two person households:

$$C_{ij} = -ASC_j + L \tag{5}$$

where i represents people and j households. For all other household sizes:

$$\begin{aligned} C_{ij} = -(&ASC_j + AgeConst[Age_i] + 0.534 * Sex_i + EducConst[Educ_i] - 0.410 * Nationality_j \\ &- 0.146 * Language_j + WorkersConst[Workers_j] + CarsConst[Cars_j]) + L \end{aligned} \tag{6}$$

Array values can be found in Table 2. The parameter L is also set to 20 for this model.

Implementation

The implementation used is a modified version *(53)* of the original Hungarian Algorithm written in Java. As mentioned earlier, there is assumed to be an edge connecting every element of U, the person set, to every element of V, the household set. This allows edge lengths, E, to be stored in a cost matrix. The code iterates through persons and households, filling in the cost matrix according to the formulas given above. This matrix is passed to the Hungarian algorithm, and a list of edges is returned as the solution.

RESULTS

The specification of variables is contained in Table 1 and the estimated values along with the t-test values in the three models are shown in Table 2. Nearly all variables are significant at the 95% level. The four exceptions were kept in the models, as suggested by Ziliak *et al. (57)*, because of the desire to have a constant for each attribute value and to have the same attributes involved in all models. The ρ values suggest that the Type Choice model is the best fit, followed by Size Choice and Spouse (although the lower value for Spouse may be a result of the smaller number of observations).

We note that, while the age variables are highly significant, the variable $Education_1$ in Type Choice, and the variables $Workers_2$ and Car_2 in both Size and Type Choice have similar or higher t-test values. This suggests that the approach of Gargiulo *et al. (17)*, which uses only age, household size, and household type, omits several highly significant variables that could be incorporated to improve the match quality. Their choice of attributes may be explained by the model's application, demographic evolution, since most other attributes (education, marital status, employment, nationality, language spoken at home) are variable on a long enough time scale. But even for this application, the existence of more person and household attributes in the present state can only help to predict transitions into future states (and if not, the extra variable can be ignored).

Many different approaches are possible for measuring match quality. In this paper, we have chosen to look at two different measures: the population-level distributions of attributes, and the fit in terms of unique combinations of attributes. We look at attribute distributions over the whole population because the area from which the households are taken is actually somewhat small. In an aggregate model, this might be one zone. And in a broader sense, the goal is to produce a synthetic population which will behave similarly. If the attribute distributions match, then the real and synthetic populations have similar characteristics and diversity and should have similar behavior as well. The other measure, which looks at attribute combinations, can help to show if there is systematic bias in the model. A good model would have a trendline slope close to 1 and a high R^2 value. If this is skewed in either direction, there is an overrepresentation or underrepresentation of some types of people. In that case, the behavior of the synthetic population may not match the real even if the attribute distributions are a close match.

Figure 3 shows a comparison of the attribute distributions for real household heads to the matched populations from Size Choice and Type Choice models. In Figure 3(a), we observe that both models produce distributions which are close to that of the real population for all age groups. Size Choice produces a slightly older population, overestimating ages 45-74 while underestimating ages 15-44. Type Choice significantly overestimates ages 65-74 but has no overall trend and is relatively close on all other groups. Looking at Figure 3(c) we see that both models have the correct trend (more men than women in head positions). Size Choice slightly overestimates men while Type Choice is almost exactly equal to the real population. For education, Figure 3(b), Size Choice overestimates secondary education and underestimates primary while Type Choice does the inverse.

Figure 4 looks at the same attribute distributions for the spouse results. Although the model used is the same for both Size and Type Choice, results differ because the remaining population is dependent on the results of the head matching. In the age results, Figure 4(a), we observe a shift towards the younger age groups with both matched populations underestimating ages 55-74 and overestimating 25-34. The sex distribution, Figure 4(c), shows that both matched populations are close to the true distribution with Type Choice performing slightly better. The spouse education distribution, Figure 4(b) is not a particularly good fit, with both populations under- or overestimating all categories to some degree.

Figures 5 and 6 compare the count of each unique combination of attributes (such as: age 25-44, female, tertiary education) between the model and the real population.

Figure 5(a) shows the fitting results for the Size Choice model. The head results, show that most points are relatively well spread out on either side of the y=x line. The absence of points on the x-axis shows that all unique combinations of attributes which exist in the real population are represented to some degree in the fitted population. The shape of the trendline shows a slight trend toward underrepresentation among unique attribute combinations. The spouse results, showed in Figure 6(a) are also good. The trendline slope and the R^2 are better than the head position. However, there are some points on the x-axis which indicates combinations unrepresented in the matched population.

Figure 5(b) shows the fitting results for the Type Choice model. Comparing the head results presented in Figure 5(b) to the Size Choice presented in Figure 5(a), we observe that the trendline is more skewed away from the y=x line. This indicates that there are more undersampled groups in the Type Choice model. Results for the spouse model, Figure 6(b) are similar to those for Size Choice, which is expected because the model is the same.

Overall, the four fitting plots show that the Size Choice model performs better than the Type Choice model in this application. Although the difference is not so great as to preclude the use of the Type Choice model; in fact, it might perform better than Size Choice in other cases.

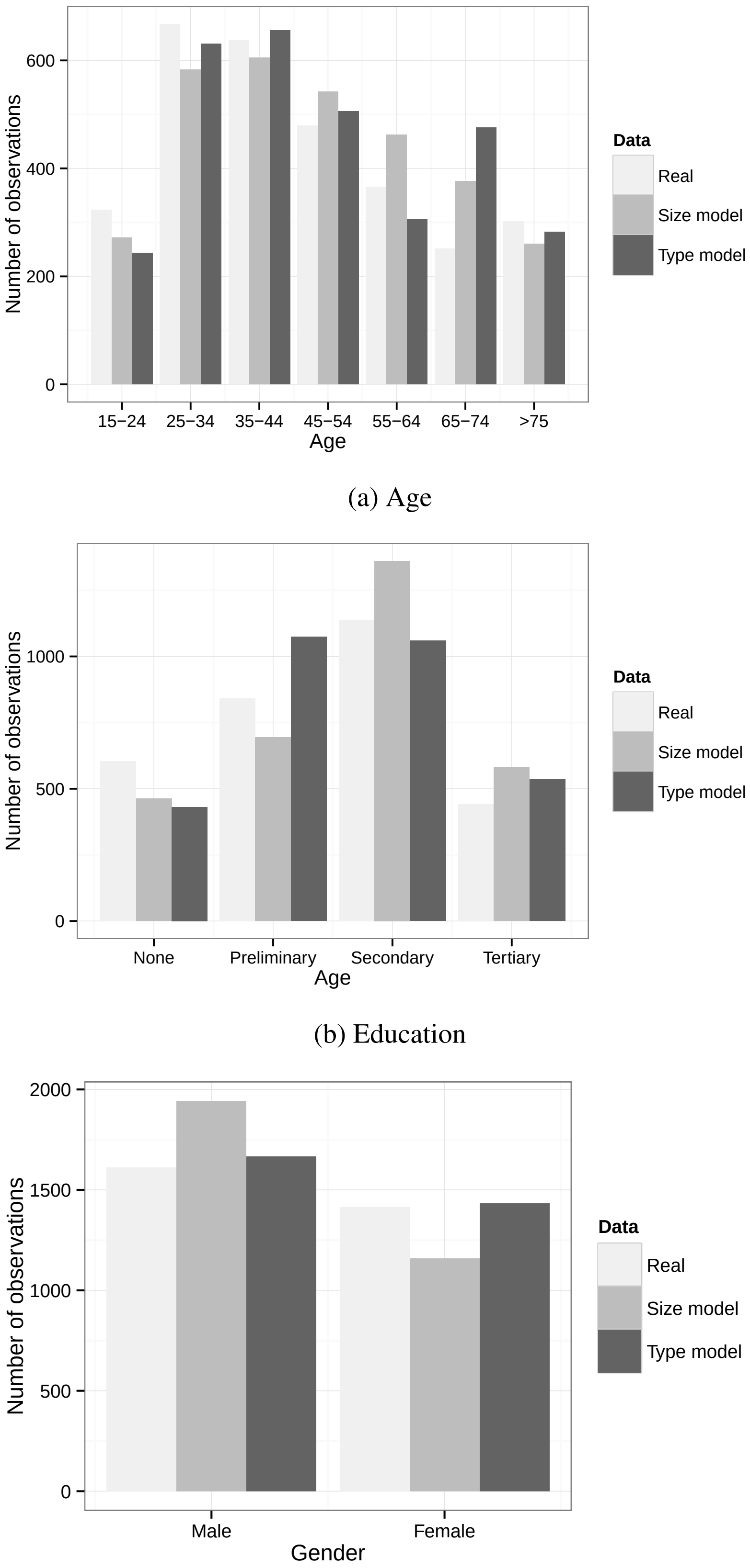
TABLE 1 Specification of variables

Variable	Specification	Name
Alternative Specific Constant		ASC_1
Alternative Specific Constant		ASC_2
Alternative Specific Constant		ASC_3
Alternative Specific Constant		ASC_4
Alternative Specific Constant		ASC_5
Alternative Specific Constant		ASC_6
Alternative Specific Constant		ASC_{one}
Alternative Specific Constant		ASC_{notfam}
Age 15-24	dummy, 1 if TRUE	Age_1
Age 25-44	dummy, 1 if TRUE	Age_{23}
Age 45-54	dummy, 1 if TRUE	Age_4
Age 55-64	dummy, 1 if TRUE	Age_5
Age 65-74	dummy, 1 if TRUE	Age_6
Age 55-74	dummy, 1 if TRUE	Age_{56}
Age >75	dummy, 1 if TRUE	Age_7
Age >65	dummy, 1 if TRUE	Age_{67}
Education 0	dummy, 1 if TRUE	Education_0
Education 1	dummy, 1 if TRUE	Education_1
Education 1 and 2	dummy, 1 if TRUE	Education_{12}
Education 3	dummy, 1 if TRUE	Education_3
Female	dummy, 1 if TRUE	Female
Workers 1	dummy, 1 if TRUE	Workers_1
Workers 2	dummy, 1 if TRUE	Workers_2
French speaking	dummy, 1 if TRUE	Fr
Swiss Nationality	dummy, 1 if TRUE	Nat
Car 2	dummy, 1 if TRUE	Car_2

TABLE 2 Model estimations

	Head				**Spouse**	
	Size n=3026		**Type** n=3026		**Size and Type** n=1202	
Variable	**Value**	**t-test**	**Value**	**t-test**	**Value**	**t-test**
ASC$_1$	1.91	9.90	-	-		
ASC$_2$	-	-	-	-	1.12	3.16
ASC$_3$	-0.737	-11.77	-	-	-	-
ASC$_4$	-0.993	-14.49	-	-	0.027	0.31*
ASC$_5$	-1.86	-19.16	-	-	-0.821	-7.30
ASC$_6$	-2.63	-19.15	-	-	-1.63	-10.61
ASC$_{one}$	-	-	1.33	7.18	-	-
ASC$_{notfam}$	-	-	-2.62	-26.66	-	-
Age$_1$	-	-	-	-	-1.35	-4.56
Age$_{23}$	1.59	8.89	1.49	8.24	0.681	3.32
Age$_4$	1.57	7.78	1.44	7.08	-	-
Age$_5$	-	-	1.17	5.70	-1.73	-6.03
Age$_6$	-	-	1.49	6.89	-	-
Age$_{56}$	1.47	7.94	-	-	-	-
Age$_7$	1.16	5.39	0.987	4.44	-	-
Age$_{67}$	-	-	-	-	-2.57	-6.46
Education$_0$	-	-	0.142	1.05*	-0.502	2.16
Education$_1$	-	-	0.590	5.00	0.380	2.02
Education$_{12}$	0.150	1.28*	-	-	-	-
Education$_3$	-0.657	-3.91	-0.577	-3.89	-0.648	-2.45
Female	-0.328	-3.52	-0.390	-4.13	0.534	2.77
Workers$_1$	0.393	3.63	0.473	4.17	0.731	2.83
Workers$_2$	5.40	10.53	5.49	10.67	0.563	2.15
Fr	-0.393	-3.69	-0.376	-3.48	-0.146	-0.82*
Nat	-0.419	-4.19	-0.327	-3.18	-0.410	-2.32
Car$_2$	6.96	13.26	7.77	14.78	-	-
	ρ =0.328		ρ =0.431		ρ =0.268	

*Not significant at 0.05 level

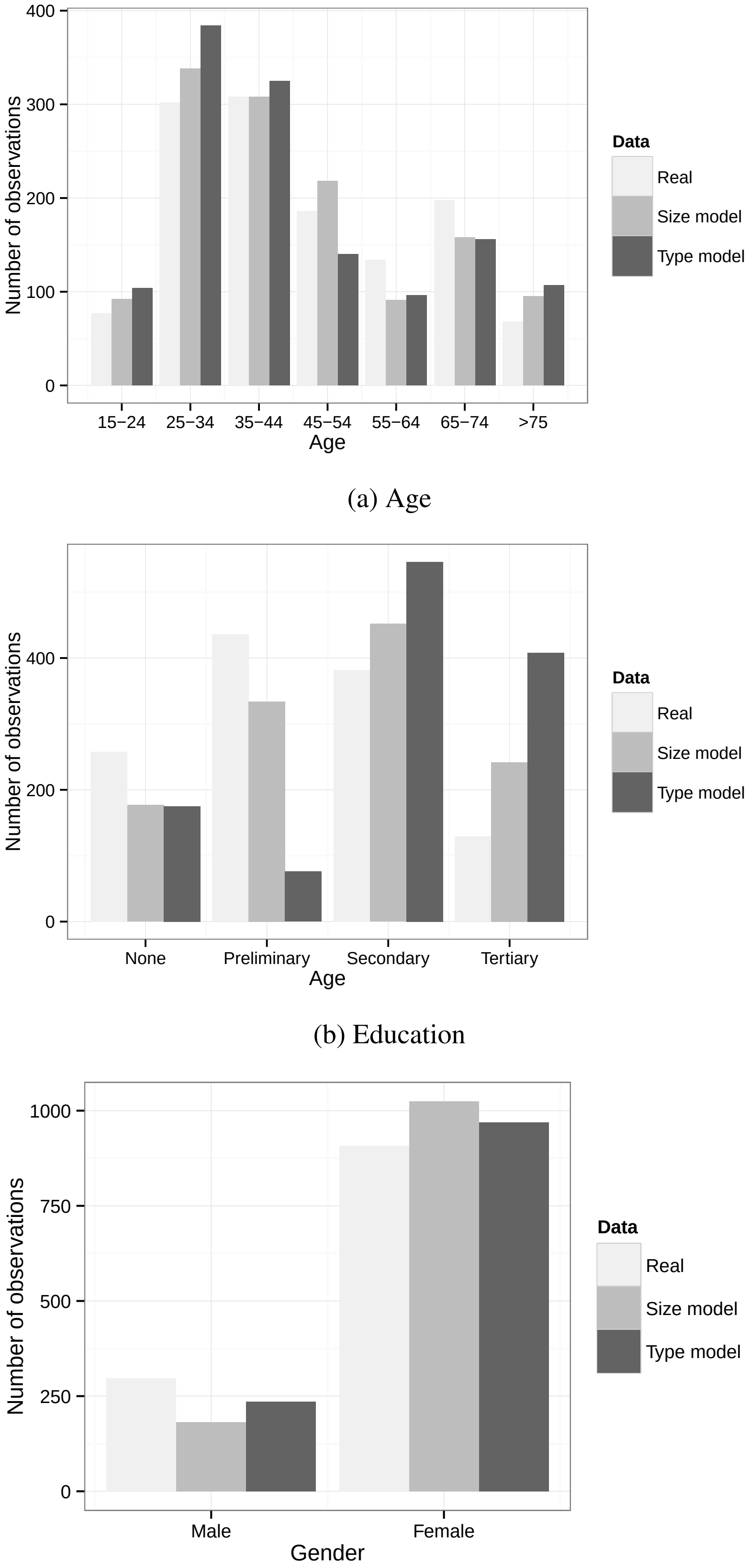


(a) Age

(b) Education

(c) Education

FIGURE 3 Household Head Distribution Comparison

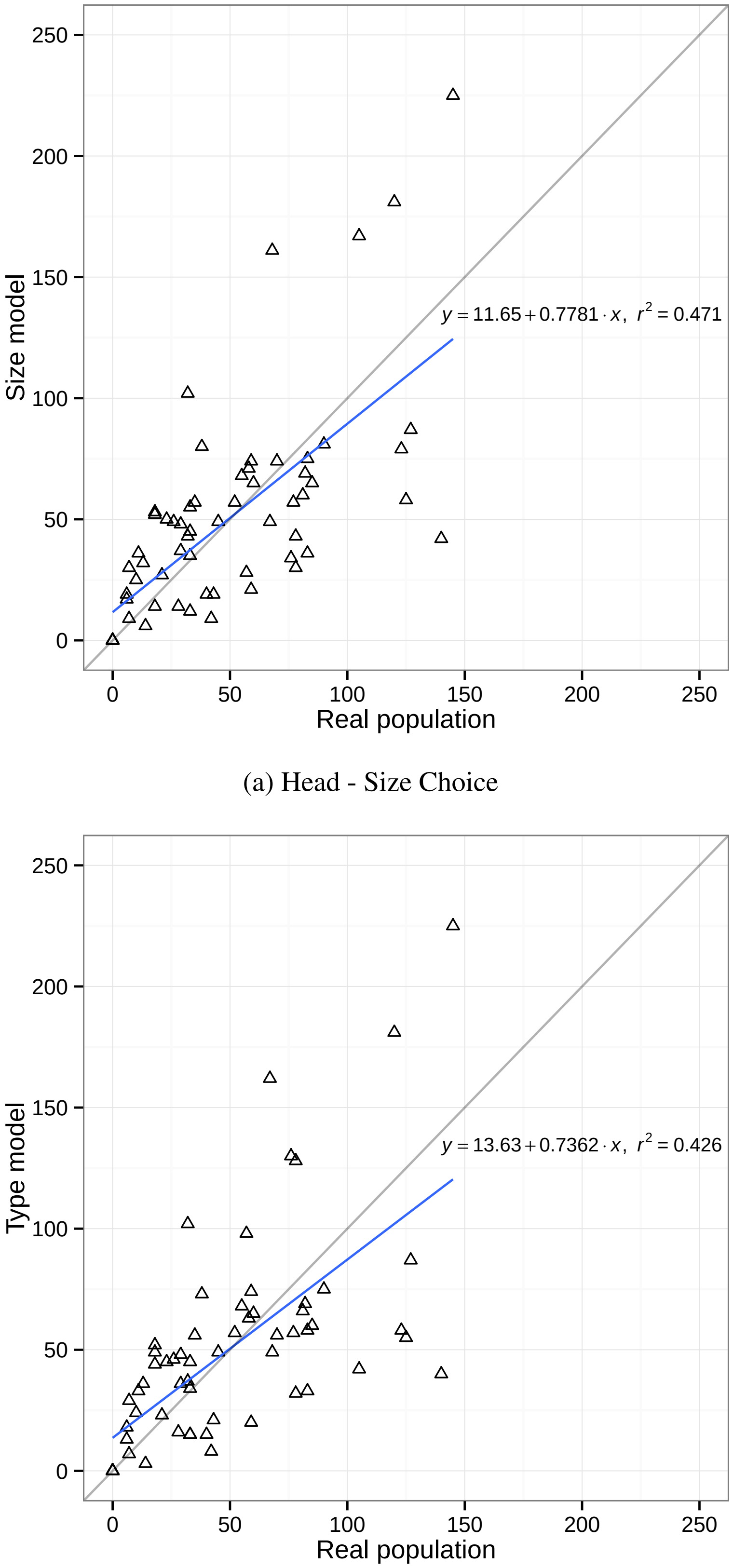


(a) Age

(b) Education

(c) Sex

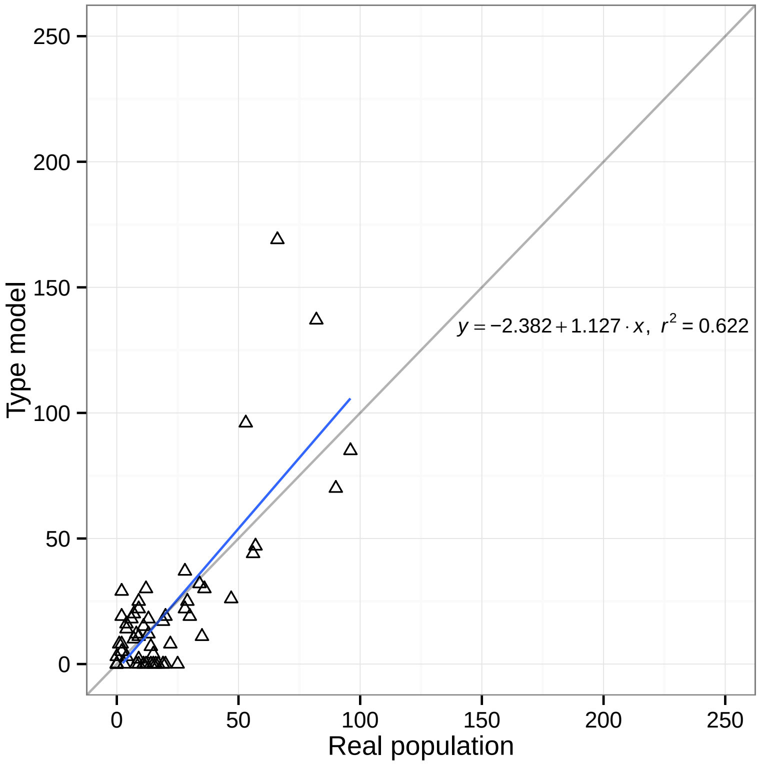
FIGURE 4 Household Spouse Distribution Comparison



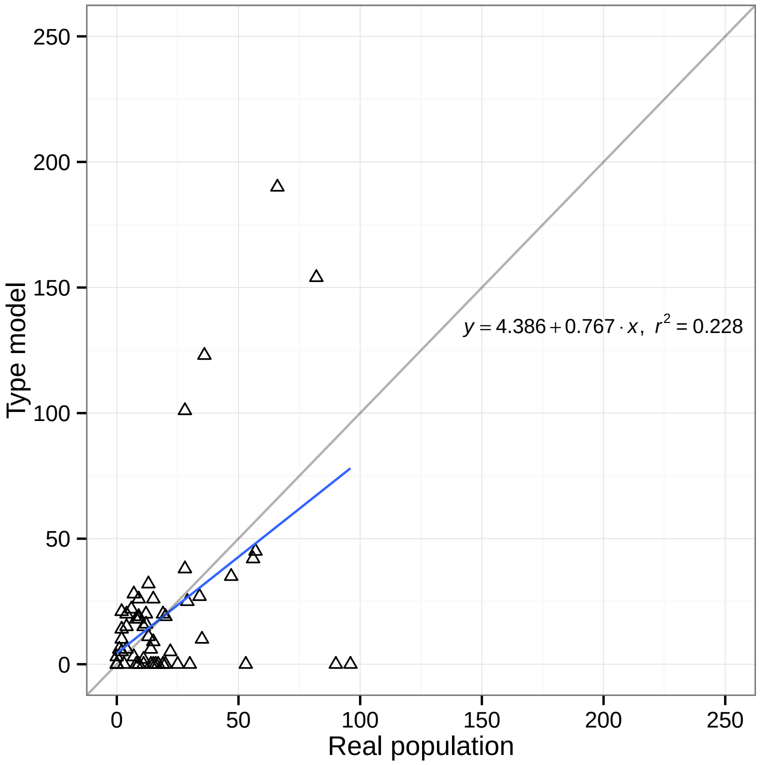
(a) Head - Size Choice

(b) Head - Type Choice

FIGURE 5 Household Distribution Comparison - Head Position



(a) Spouse - Size Choice



(b) Spouse - Type Choice

FIGURE 6 Household Distribution Comparison - Spouse Position

CONCLUSIONS

In this research, bipartite graph based method for associations generation in synthetic population is proposed. It is then applied to matching of head and spouse positions for the case study of Switzerland. The presented methodology could be proven valuable for the development of Integrated Land-Use and Transport Models. Two approaches of associations generation were tested: one which formulated the problem as matching people to a particular household size, and a second which looked at matching people to household type. Both of these approaches were used to match people into household head positions, whereas only the size model could be used for spouse positions because the type was not meaningful for this problem. All three models (as size choice was calibrated separately for head and spouse positions) are multinomial logit models calibrated to a sample of the real population using Biogeme. The resulting utility functions were then transformed to a length attribute and the head matching was set up as a bipartite graph. A modified Hungarian Algorithm was used to solve the graph, and the remaining unmatched population was used to run a second bipartite match for the spouse position.

Overall, the attribute distributions are a relatively good fit for the head position. Some attributes are overrepresented in the fitted population and others are underrepresented, but all attributes are present and the general shape of the age, sex, and education distributions is correct. For the spouse position, the distributions are similarly good with the exception of education. The four education categories are over- or undersampled to some degree. This could be caused by propagation of error. The population available for the spouse matching depends on which individuals are used in the head matching. Consequently, if a particular attribute is oversampled in the head matching, less individuals will be available for the spouse matching. We can see that primary education is oversampled in the head results for Type Choice. In the spouse results, this is reversed, which suggests that there may be a correlation.

The fitting results show that all combinations of attributes in the real population are also present in the matched populations for the head position. There is room for improvement, but with results like this we would expect the real and matched populations to behave similarly in simulation. The spouse models result in good fit, but there are some attribute combinations which are absent in the matched populations. This is an area for improvement in future models.

These results are encouraging, and show that the idea of first generating a population of synthetic individuals and then matching them into household positions is feasible. Although disaggregate data was used in this study, the methodology is applicable without it. Model calibration could be done with a small area sample, and results could be evaluated by comparison to aggregate distributions, as we have done in Figures 3 and 4. We have demonstrated the matching procedure for the head and spouse positions. There are many possible extensions of this topic. The next logical step would be to extend this procedure to all household positions. Once that is done, we would have a complete synthetic population of people and households and could begin to use it in microsimulation models and evaluate its performance by comparing to results obtained using other data sources. This method is capable of generating the large amount of individual and household-level data required for detailed transportation models while balancing the privacy concerns inherent with personal data.

REFERENCES

1. Orcutt, G. H., M. Greenberger, J. J. Korbel, A. M. Rivlin *et al.* (1961) *Microanalysis of socioeconomic systems: A simulation study*, vol. 6, Harper New York.

2. Clarke, M. and E. Holm (1987) Microsimulation methods in spatial analysis and planning, *Geografiska Annaler. Series B. Human Geography*, 145–164.

3. Zaidi, A., A. Harding and P. Williamson (2009) *New frontiers in microsimulation modelling*, Ashgate Vienna.

4. Patterson, Z., M. Kryvobokov, F. Marchal and M. Bierlaire (2010) Disaggregate models with aggregate data: Two urbansim applications, *The Journal of Transport and Land Use*, **3** (2) 5–37.

5. Waddell, P. (2002) UrbanSim: Modelling urban development for land use, transportation and environmental planning, *Journal of the American Planning Association*, **68** (3) 297–314.

6. Patterson, Z. and M. Bierlaire (2010) Development of prototype urbansim models, *Environment and planning. B, Planning & design*, **37** (2) 344.

7. Williamson, P., M. Birkin, P. H. Rees *et al.* (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.

8. Fotheringham, A. S. and P. A. Rogerson (2009) *The SAGE handbook of spatial analysis*, SAGE Publications Limited.

9. Birkin, M. and M. Clarke (2011) Spatial microsimulation models: A review and a glimpse into the future, 193–208, Springer.

10. Jones, P. M., M. C. Dix, M. I. Clarke and I. G. Heggie (1983) *Understanding travel behaviour*, Monograph.

11. Ballas, D. and G. Clarke (1999) Regional versus local multipliers of economic change? a microsimulation approach, paper presented at the *39th European Regional Science Association Congress, University College Dublin, Dublin, Ireland*, 23–27.

12. Fisher, K., M. Egerton, J. I. Gershuny and J. P. Robinson (2007) Gender convergence in the american heritage time use study (ahtus), *Social Indicators Research*, **82** (1) 1–33.

13. Ben-Akiva, M., A. de Palma, D. McFadden, M. Abou-Zeid, P.-A. Chiappori, M. de Lapparent, S. N. Durlauf, M. Fosgerau, D. Fukuda, S. Hess *et al.* (2012) Process and context in choice models, *Marketing Letters*, **23** (2) 439–456.

14. Surprenant-Legault, J., Z. Patterson and A. M. El-Geneidy (2013) Commuting trade-offs and distance reduction in two-worker households, *Transportation Research Part A: Policy and Practice*, **51**, 12–28.

15. Dugundji, E. R. and J. L. Walker (2005) Discrete choice with social and spatial network interdependencies: An empirical example using mixed generalized extreme value models with field and panel effects, *Transportation Research Record: Journal of the Transportation Research Board*, **1921** (1) 70–78.

16. Arentze, T., P. van den Berg and H. Timmermans (2012) Modeling social networks in geographic space: approach and empirical application, *Environment and Planning-Part A*, **44** (5) 1101.

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

18. Barthelemy, J. and P. L. Toint (2013) Synthetic population generation without a sample, *Transportation Science*, **47** (2) 266–279.

19. Farooq, B., E. J. Miller, F. Chingcuanco and M. Cook (2013) Microsimulation framework for urban price-taker markets, *Journal of Transport and Land Use*.

20. Kuhn, H. W. (1955) The hungarian method for the assignment problem, *Naval research logistics quarterly*, **2** (1-2) 83–97.

21. Williamson, P. (2013) An evaluation of two synthetic small-area microdata simulation methodologies: Synthetic reconstruction and combinatorial optimisation, in *Spatial Microsimulation: A Reference Guide for Users*, 19–47, Springer.

22. Lenormand, M. and G. Deffuant (2012) Generating a synthetic population of individuals in households: Sample-free vs sample-based methods, *arXiv preprint arXiv:1208.6403*.

23. Harland, K., A. Heppenstall, D. Smith and M. Birkin (2012) Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques, *Journal of Artifical Societies and Social Simulation*, **15** (1) 1–15.

24. Hermes, K. and M. Poulsen (2012) A review of current methods to generate synthetic spatial microdata using reweighting and future directions, *Computers, Environment and Urban Systems*.

25. Müller, K. and K. W. Axhausen (2010) *Population synthesis for microsimulation: State of the art*, ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT).

26. Ma, L. (2011) Generating disaggregate population characteristics for input to travel-demand models, Ph.D. Thesis, University of Florida.

27. Pritchard, D. R. (2008) Synthesizing agents and relationships for land use/transportation modelling, Ph.D. Thesis, University of Toronto.

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

29. Harding, A., R. Lloyd, A. Bill and A. King (2004) *Assessing poverty and inequality at a detailed regional level: New advances in spatial microsimulation*, no. 2004/26, Research Paper, UNU-WIDER, United Nations University (UNU).

30. Tanton, R., Y. Vidyattama, B. Nepal and J. McNamara (2011) Small area estimation using a reweighting algorithm, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174** (4) 931–951.

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

32. Edwards, K. L. and G. P. Clarke (2009) The design and validation of a spatial microsimulation model of obesogenic environments for children in leeds, uk: Simobesity, *Social Science & Medicine*, **69** (7) 1127–1134.

33. Voas, D. and P. Williamson (1998) Testing the acceptability of random number generators, *Technical Report*, Working Paper 1998/2). Liverpool: Population Microdata Unit, Department of Geography, University of Liverpool.(Available from: http://pcwww. liv. ac. uk/microdata).

34. Huang, Z. and P. Williamson (2001) A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata, *Department of Geography, University of Liverpool*.

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

36. Deming, W. E. and F. F. Stephan (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.

37. Farooq, B., M. Bierlaire, R. Hurtubia and G. Flötteröd (2012) Simulation based approach for agents synthesis in large-scale urban systems modelling, *Proceedings of the 13th Conference of International Association of Travel Behavior Research (IATBR), July 15-20, 2012*.

38. Schafer, J. (2010) *Analysis of incomplete multivariate data*, London: Chapman & Hall.

39. Ballas, D., G. Clarke and I. Turton (1999) Exploring microsimulation methodologies for the estimation of household attributes, paper presented at the *4th International Conference on GeoComputation, Mary Washington College, Virginia, USA*.

40. Voas, D. and P. Williamson (2001) Evaluating goodness-of-fit measures for synthetic microdata, *Geographical and Environmental Modelling*, **5** (2) 177–200.

41. Hynes, S., K. Morrissey, C. O'Donoghue and G. Clarke (2009) Building a static farm level spatial microsimulation model for rural development and agricultural policy analysis in ireland, *International journal of agricultural resources, governance and ecology*, **8** (2) 282–299.

42. Rahman, A., A. Harding, R. Tanton and S. Liu (2013) Simulating the characteristics of populations at the small area level: New validation techniques for a spatial microsimulation model in australia, *Computational Statistics & Data Analysis*, **57** (1) 149–165.

43. Oketch, T. and M. Carrick (2005) Calibration and validation of a micro-simulation model in network analysis, paper presented at the *Proceedings of the 84th TRB Annual Meeting, Washington, DC*.

44. Ballas, D. (2005) *Geography matters: simulating the local impacts of national social policies*, Joseph Rowntree Foundation.

45. Ballas, D., G. Clarke, D. Dorling and D. Rossiter (2007) Using simbritain to model the geographical impact of national government policies, *Geographical Analysis*, **39** (1) 44–77.

46. Tanton, R. and K. Edwards (2013) *Spatial microsimulation: a reference guide for users*, vol. 6, Springer.

47. Luce, R. D. (1959) *Individual Choice Behavior a Theoretical Analysis*, John Wiley and sons.

48. Marschak, J. (1960) Binary-choice constraints and random utility indicators, paper presented at the *Proceedings of a Symposium on Mathematical Methods in the Social Sciences*.

49. Ben-Akiva, M. and S. R. Lerman (1985) *Discrete choice analysis: theory and application to predict travel demand*, vol. 9, The MIT press.

50. Burkard, R. E., M. Dell'Amico and S. Martello (2009) *Assignment problems*, Siam.

51. Hopcroft, J. E. and R. M. Karp (1973) An n^5/2 algorithm for maximum matchings in bipartite graphs, *SIAM Journal on computing*, **2** (4) 225–231.

52. Bachmann, P. (1894) *Die Analytische Zahlentheorie*, Teubner, Leipzig, Germany.

53. Munkres, J. (1957) Algorithms for the assignment and transportation problems, *Journal of the Society for Industrial & Applied Mathematics*, **5** (1) 32–38.

54. Renaud, A. (2004) Coverage estimation for the swiss population census 2000: Estimation methodology and results, *Technical Report*, Swiss Federal Statistical Office.

55. Bierlaire, M. (2003) BIOGEME: A free package for the estimation of discrete choice models, paper presented at the *Proceedings of the 3rd Swiss Transportation Research Conference*, Ascona, Switzerland.

56. Bierlaire, M. and M. Fetiarison (2009) Estimation of discrete choice models: extending biogeme, paper presented at the *Proceedings of the 9th Swiss Transport Research Conference. Ascona, Switzerland*.

57. Ziliak, S. T., D. N. McCloskey and N. Deirdre (2008) *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*, University of Michigan Press.