Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 37 (2014) 410 – 415

Procedia Computer Science

The 1st International Workshop on Information Fusion for Smart Mobility Solutions (IFSMS'14)

Synthesizing Population for Microsimulation-Based Integrated Transport Models using Atlantic Canada Micro-Data

Mohammad Hesam Hafezi[a] and Muhammad Ahsanul Habib[b,]*

[a]*Department of Civil and Resource Engineering, Dalhousie University, PO Box: 15000, Halifax, NS, B3H4R2, Canada*
[b]*School of Planning, and Department of Civil and Resource Engineering, Dalhousie University, PO Box: 15000, Halifax, NS, B3H4R2, Canada*

Abstract

Due to the lack of availability of micro-data of population characteristics, the synthesis of individual and household attributes is a necessary step for developing a disaggregate, dynamic travel demand forecasting model. Agent-based micro-simulation models attempt to forecast travel behaviour of individuals and households by simulating the behaviour of the singular actors in the system. The framework for generating synthesis population presented in this paper is a fundamental contribution to the development of an Integrated Transport, Land Use and Environment Modelling System in Nova Scotia, Canada. In this paper, a population is synthesized for individuals and households in Atlantic Canada using the Fitness Based Synthesis (FBS) approach. A synthetic algorithm is designed that allows both individual and household attribute levels to synthesize simultaneously. Unequal probabilities based on the sampling weight are used in the household selection step of the algorithm. In this way, the performance and accuracy of the synthetic population produced has been improved. The synthetic algorithm is tested for two functions: first, using the one level (household) control tables; and second, using two levels (individual and household) control tables. The data used in this study is collected from the 2006 Canadian Census and the 2006 Public Use Micro-data File (PUMF). The algorithm is implemented using a high-level matrix programming language for numerical computation in MATLAB. The results show that the synthetic population with both individual and household level attributes has the best fitness value.

Keywords: Population synthesis; Household level; Individual level; Land Use and Transportation Micro-simulation.

1. Introduction

In the past decade, transportation researchers' interest in developing disaggregate travel demand models (individual- or household-based) has increased in response to the growing importance of complex policy measures,

* Corresponding author. Tel.: 902-4943209; fax: 902-4236672.
E-mail address: ahsan.habib@dal.ca

doi:10.1016/j.procs.2014.08.061

such as travel demand management and road pricing [1]. Advanced agent-based micro-simulation models utilize disaggregate data for a higher level of accuracy and reliability of the models. These agent-based micro-simulation models simulate the behaviour of individuals and households instead of aggregate accounting [2]. As such, the model produced is able to address the impacts of flexible work hours, pricing-based strategies, greenhouse gas (GHG) emissions, interfaces between micro-scale land use changes and travel activity [3]. However, one of the major challenges of disaggregate modelling is the requirement for a large amount of individual-level data [4]. UrbanSim [5], ILUTE [6] and TRANSIMS [2] are some examples of micro-simulation platforms that apply the agent-based model and use individual-data as input for processing. The availability of micro-data is essential for the operation of micro-simulation models. In particular, it is necessary to input details of the individual or household characteristics, as well as home and work locations, for the entire population of the study area to achieve the most fine-grained model results [7]. This information is typically collected in a population census, however, the disaggregate data is not accessible to the public due to privacy concerns [4]. Due to the lack of accessibility and completeness of the micro-data, population synthesis is an essential step for developing a disaggregate travel demand forecasting model. Population synthesis aims to produce virtual individuals by expanding the disaggregate public data to mirror known aggregate sample data with the same demographics as the real population [8]. Ordinarily, two data sets are required for population synthesis: a disaggregate sample micro-data and corresponding aggregate data. Generally, agent-based micro-simulation models perform population synthesis using one of four main methods, which are: Combinatorial Optimization (CO), Iterative Proportional Fitting (IPF), Iterative Proportional Updating (IPU) and Fitness Based Synthesis (FBS). The key difference between these four methods is their ability to simultaneously control for both household and individual attributes of interest in the procedure. Traditional IPF method is the only population synthesis procedure that is incapable of controlling for individual attributes in the process. With the exception of the FBS method, other methods, namely CO and IPU, use the IPF procedure as a part of their process.

The original IPF methods were developed by Beckman et al. (1996) to generate synthetic baseline populations as input data for the TRANSIMS model using sample and census data in the US. The traditional IPF method contains two steps. The first step is fitting unadjusted cell data (called seed data) to a known margin for both the rows and columns of the table (called as control or marginal tables). The second step is generating the synthesized households using Monte Carlo Simulation, where individual household records are drawn from the seed data [2]. However, the traditional IPF method has some limitations, including high-dimensional problems (or memory problem), control for individual attributes, zero cell problem and round-off of the cell values of the joint distribution [9]. Another population synthesizer method that addresses some of the limitations of the IPF method is the combinatorial optimization (CO) method.

Similar to the IPF method, CO also needs the information on population characteristics both at the sample (seed data) and marginal (control table) levels. CO method uses the integer reweighting technique. There are two initial weights of 0 and 1 in the procedure, which are given to the sample. A weight of 0 is assigned to all households at the beginning of the procedure, then, after selecting each household, a weight of 1 is assigned. Subsequently, using the optimization tools such as simulated annealing, genetic algorithm and hill-climbing, these weights will be optimized with the purpose of minimizing the difference between the marginal table and the synthesized population [10, 11].

Another population synthesis method that can simultaneously generate synthetic populations with both household and individual attribute level data is Iterative Population Updating (IPU). Xin et al. (2009) introduce a new algorithm that can synthesize population at a high performance level by matching household and individual level distributions [3]. The IPU method consists of three main steps. First, household and individual level attributes are constrained by selecting 5% of household and individual level attributes from PUMS as seed data to create marginal tables from the census file. Second, the weights of household and individual level joint distributions are estimated. These weights are assigned so that both household and individual level distributions can be matched closely. Finally, households are drawn from the procedure in the previous step, and subsequently synthetic population for the region is generated.

In the all of the three methods of population synthesis, a joint multi-way distribution table is created in the beginning of the procedure. Srinivasan et al. (2008) introduces a method that can generate synthetic populations with the capability of matching several multilevel controls without necessitating a joint multi-way distribution. The population synthesis methodology described in the present paper is comparable to work done by Srinivasan et al. (2008) [12] in the sense that it produces a list of households to match several multilevel controls without the need for a joint multi-way distribution. The Fitness-Based Synthesis (FBS) presented by Srinivasan et al. [12, 13] solves the particular problems of traditional population synthesis such as: zero cell problems, computational resources (memory) and non-integers cell value in the joint-distribution tables [2]. Additionally, the adoption of Monte Carlo simulation in the procedure is waived due to the generation of integer fitness values. In this paper, a population is synthesized for individuals and households in Atlantic Canada using the Fitness Based Synthesis (FBS) approach. The framework for generating synthesis population presented in this paper is a fundamental contribution to the development of an Integrated Transport, Land Use and Environment Modelling System in Nova Scotia, Canada. In this paper, error-percentages and fitting test are used for validation of the synthetic population.

2. Method

The FBS procedure involves selecting a set of households in each iteration from the seed data so that count table closely replicates the control table. The number of count tables is equivalent to the number of control tables in the procedure. In the beginning of the FBS procedure, the count tables for all household attributes are given an initial value of zero. Then in each iteration, one household is added to the count table according to its corresponding fitness value. The population is synthesized in an iterative fashion. This iterative fashion is continued until the count table replicates the control table as closely as possible. The fitness value is given by the following [12]:

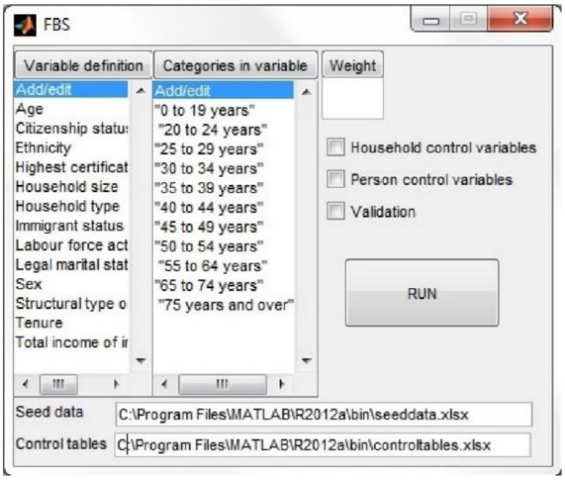
$$F^{xy} = \sum_{q=1}^{Q} \left[\frac{1}{e_q^x} \sum_{w=1}^{W_q} \left[\frac{(\mathrm{G}_{qw}^{y-1})^2}{H_{qw}} - \frac{\left(G_{qw}^{y-1} - MH_{qw}^{x}\right)^2}{H_{qw}} \right] \right] \tag{1}$$

$$G_{qw}^{y-1} = H_{qw} - AH_{qw}^{y-1} \tag{2}$$

In formulas 1 and 2, F^{xy} is the fitness value; x is the selected household; y is the iteration number; q is the index representing both the control and count tables; Q is the total number of both control/count tables; w is the index representing the different cells in the count table; W is the index representing the different cells in the control table; H_{qw} represents the value of cell w in the control table q ; AH_{qw}^{y-1} represents the value of cell w in the count table q ; G_{qw}^{y-1} is the difference between control and count tables for cell w in control table q; MH_{qw}^{x} is the contribution of the x^{th} household in the seed data of the w^{th} cell in control table q ; e_q^x is the coefficient of the equation that takes the value of 1 for a household control table and takes the size of household x for an individual control table. FBS procedure continues by calculating the fitness value for each household in the seed data for each iteration. Households with maximum fitness values are selected in each iteration, then using random selection, one of the households is selected and added to the list of updated synthetic population list. After this process, all the values in the count tables are updated. The next iteration repeats the previous steps by calculating fitness value, selecting households and updating tables in the algorithm. The termination criteria occurs when there are no positive fitness values left in the procedure. The FBS method for this paper was coded in Matlab R2012a and uses xlsx files for input and output. The tests were performed on a personal computer with Inter(R) Core(TM) i7-4770 CPU @ 3.40GHz, and 16.00 GB RAM in the environment of Microsoft Windows 7 professional 64-bit. The sparse matrix technique was used in Matlab to improve the computational time due to the varying number of zero cells within the matrices. The prototype version of GUI is created with GUIDE-Matlab and is shown as Figure 1.

3. Data

Two main data sources are used in this study. First, the 2006 Canadian Census Hierarchical Public Use Microdata File (PUMF); and, second, the 2006 Canadian Census (Statistics Canada). Initially, this study planned to generate synthetic population for Halifax region only. However, hierarchical source micro-data was not available for this purpose [14]. Therefore this paper uses the PUMF for the Atlantic Canada. Since this sample is quite large (1,325,000 records), a sample of 1% (9,142 household records and 4,108 individual records) is considered to reduce runtime for evaluating the algorithm.



☐ Household control variables
☐ Person control variables
☐ Validation

Seed data: C:\Program Files\MATLAB\R2012a\bin\seeddata.xlsx

Control tables: C:\Program Files\MATLAB\R2012a\bin\controltables.xlsx

Fig. 1. FBS GUI interface

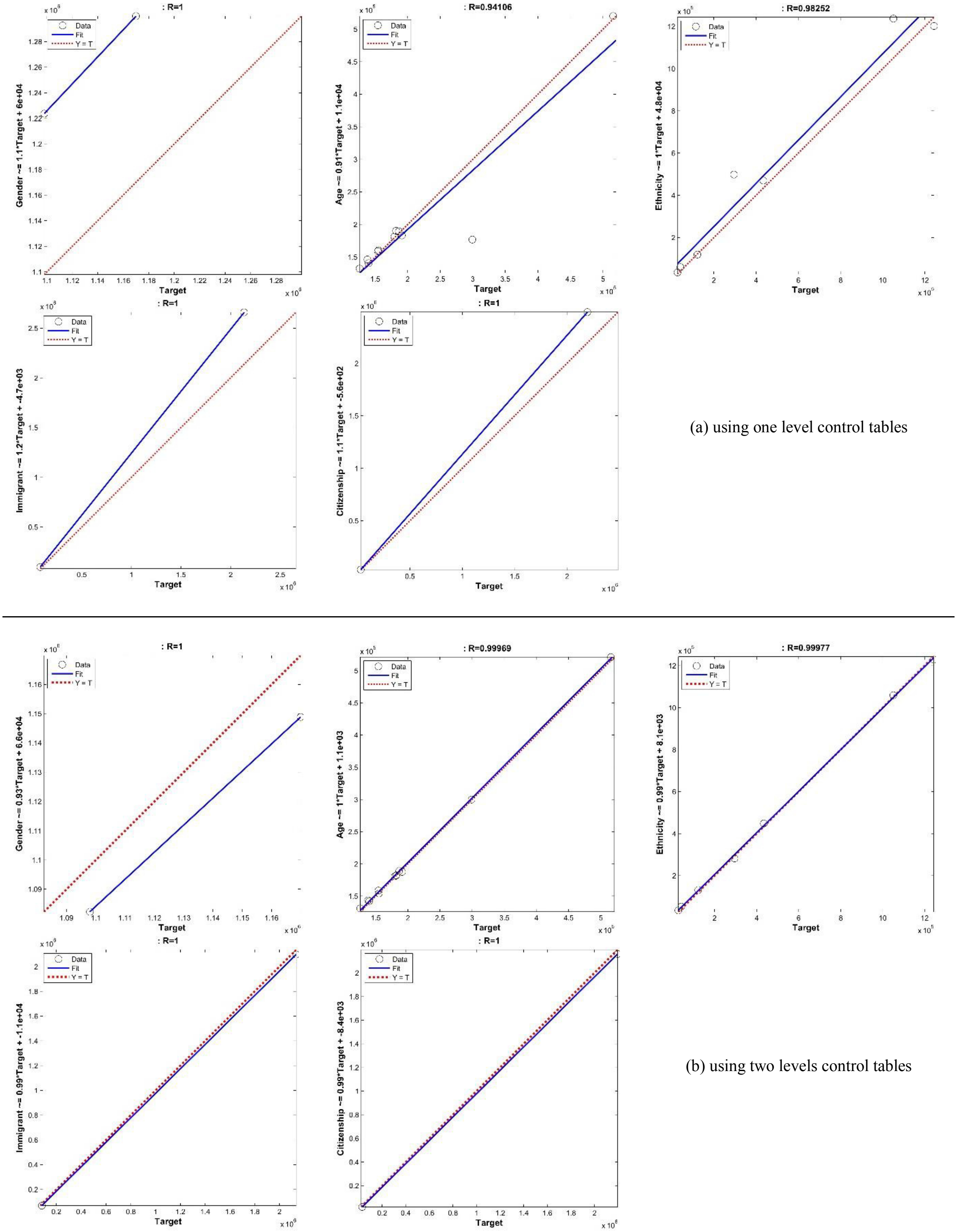
Individual and household attributes are two groups of explanatory variables that are used in this study. Household variables included: household size, tenure, dwelling type and household income. Household size has categories for 1, 2, 3, 4, 5 and 6+ persons. Tenure has two categories, including owned or rented properties. Dwelling type has eight categories: single-detached house, semi-detached or double house, row house, apartment/flat in a duplex, apartment in a building that has five or more storeys, apartment in a building that has fewer than five storeys, other single-attached house and mobile home, and, other movable dwelling. Household income has categories for under $ 2000; between $ 2000 and $ 6,999; between $ 7,000 and $ 14,999; between $ 15,000 and $ 29,999; between $ 30,000 and $ 79,999; and, $ 80,000 and over. Individual variables include: gender, age, ethnicity, immigrant status and citizenship. Gender has two categories: male and female. Age has eleven categories: ≤19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-64, 65-74 and 75+. Ethnicity has seven categories: British Isles origins, French origins, aboriginal origins, Canadian, European origins, Asian origins and other origins. Immigrant status has two categories: non-immigrant and immigrant. Citizenship status has two categories: non-citizen and citizen.

4. Discussion of Results

There are different techniques for measuring the accuracy level of synthetic population. In this paper, the error-percentage and goodness-of-fit for every unique control and count table are computed. For each of the control tables, error percentages using one level (group 1) and two levels (group 2) control tables are summarized in Table 1. There are two main differences in the error percentages of the two groups in Table 1. First, error percentages for individual level control tables in group 2 are significantly smaller than in group 1. This difference demonstrates that using two levels control tables results in improvements to the accuracy of synthetic population because of the increased precision of the fitness value calculation by involving more control tables in each iteration. Second, the values of the error percentages for household level control tables in group 2 are slightly larger than in group 1. This is because an increase in the total number of control tables in the procedure decreases the ability to replicate each table during the household selection in the procedure. Figures 2 compares the fit of each individual control and count table.

Table 1 Error percentages

	Person level					Households level			
Group	Gender	Age	Ethnicity	Immigrant	Citizenship	Household size	Tenure	Dwelling Type	Household Income
1	11.28%	8.25%	15.43%	24.32%	13.31%	0.05%	0.03%	0.11%	0.27%
2	1.88%	1.39%	4.91%	3.18%	4.50%	1.34%	1.83%	2.14%	2.49%



(a) using one level control tables

(b) using two levels control tables

Fig. 2. Synthesized population compared to control table

Examining the fit for every unique control and count table is another method used in this study to evaluate the accuracy of the synthesized population. A trend-line slope close to 1 and a high R-square value demonstrate that the synthesized population has a good fit with the control table. Comparing the two sets of graphs in Figure 2 demonstrates that the synthetic population using two levels control tables has a better fit than when using one level control tables. Goodness-of-fit results also indicate that the attributes with less classifications have better fitting results in the algorithm that, when used, cause an increase the accuracy of synthesized population.

5. Conclusions

Agent-based micro-simulation models attempt to forecast behaviour of individuals and households by simulating the behaviour of a disaggregate sample of individual and households instead of aggregate accounting. One of the main challenges in the agent-based models is the generation of an appropriate disaggregate data set for modeling. Most significantly, population synthesis is the primary step in micro-simulation modeling. The goal is to produce virtual people by expanding the disaggregate sample data to mirror known aggregate sample data with the same demographics as the real population. In this paper, a framework for generating synthetic population for Atlantic Canada using the fitness based synthesis (FBS) approach is achieved. This is a first step towards developing a microsimulation-based Integrated Transport, Land Use and Environment Modelling System for the Halifax Regional Municipality (HRM), Nova Scotia, Canada. The algorithm was implemented in Matlab using the sparse matrix technique, which performed well with reasonable runtime for 9 attributes (at both household and individual levels) and represented reality for the study area. As validated by the error percentages and goodness-of-fit evaluation, the FBS can efficiently obtain a satisfactory result using two levels control tables. Future works include running the model for 100% population of Atlantic Canada as well as examination of alternative approaches to derive synthesized population for the smaller municipal level, such as the Halifax Regional Municipality (HRM).

Acknowledgements

This research is funded by Natural Science and Engineering Research Council (NSERC)-Discovery Grant and Nova Scotia Department of Energy. The authors would like to thank Siobhan Witherbee for proof reading the article.

References

1. Jovicic, G., Activity based travel demand modelling - a literature study. *Danmarks TransportForskning*. 2001.
2. Beckman, R.J., K.A. Baggerly, and M.D. McKay, Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 1996. **30**(6): p. 415-429.
3. Ye, X., et al. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. *in 88th Annual Meeting of the Transportation Research Board*, Washington, DC. 2009.
4. Anderson, P., et al. Associations Generation in Synthetic Population for Transportation Applications. A Graph-Theoretic Solution. *in Transportation Research Board 93rd Annual Meeting*. 2014.
5. Waddell, P., Integrated land use and transportation planning and modelling: addressing challenges in research and practice. *Transport Reviews*, 2011. **31**(2): p. 209-229.
6. Salvini, P. and E.J. Miller, ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems. *Networks and Spatial Economics*, 2005. **5**(2): p. 217-234.
7. Arentze, T.A. and H.J. Timmermans, A learning-based transportation oriented simulation system. *Transportation Research Part B: Methodological*, 2004. **38**(7): p. 613-633.
8. Guo, J.Y. and C.R. Bhat, Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board*, 2007. **2014**(1): p. 92-101.
9. Pritchard, D.R. and E.J. Miller, Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, 2012. **39**(3): p. 685-704.
10. Voas, D. and P. Williamson, An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography*, 2000. **6**(5): p. 349-366.
11. Abraham, J.E., K.J. Stefan, and J. Hunt. Population synthesis using combinatorial optimization at multiple levels. *in 91th Annual Meeting of Transportation Research Board*, Washington DC. 2012.
12. Srinivasan, S., L. Ma, and K. Yathindra, Procedure for forecasting household characteristics for input to travel-demand models. 2008.
13. Ma, L. and S. Srinivasan, Synthetic Population Generation with Multilevel Controls: A Fitness-Based Synthesis Approach and Validations. *Computer-Aided Civil and Infrastructure Engineering*, 2014.
14. 2006 census public use microdata file (PUMF), hierarchical file: Documentation and user guide. Statistics Canada. URL http://equinox.uwo.ca/docfiles/2006_Census/pumf/hier/pumf%20user%20guide.pdf, 2011.