Implementing Bayesian Network and Generalized Raking Multilevel IPF for Constructing Population Synthesis in Megacities

Anugrah Ilahi, IVT ETH Zurich
Kay W Axhausen, IVT ETH Zurich

Conference paper STRC 2018

STRC

18th Swiss Transport Research Conference
Monte Verità / Ascona, May 16 -18 2017

Implementing Bayesian Network and Generalized Raking Multilevel IPF for Constructing Population Synthesis in Megacities

Anugrah Ilahi
IVT ETH Zurich
Zurich

Phone: +41 77 974 79 50
Fax: +41 44 633 10 57
email: anugrah.ilahi@ivt.baug.ethz.ch

Kay W Axhausen
IVT ETH Zurich
Zurich

Phone: +41 76 368 02 49
Fax: +41 44 633 10 57
email:axhausen@ivt.baug.ethz.ch

February 2018

Abstract

Constructing agent data with detailed information of their sociodemographics is substantially important for agent-based modelling. However, to collect whole population data is not efficient because it requires an expensive and time-consuming survey, especially for a large populations. Therefore, this research tries to construct the whole population of the Greater Jakarta area using previous surveys. One of them was conducted by JICA in 2009 with a 3% sample of households. This paper uses graphical representation, a Bayesian Network (BN), which allows identifying the best joint probability distribution of the data structure and Iterative Proportional Fitting (IPF) to fit data against aggregate census data. The results show that using BN approach can produce data that represents probability distribution of sample data and IPF to match it against aggregate census data.

Keywords

Bayesian Network, Generalized Raking , Population Synthesis – Agent-Based Model

1. Introduction

Activity-based transportation models require detailed individual and household traveller information such as socio-demographics and information of work and home location (Balmer *et al.*, 2006). Therefore, to collect all data is inefficient in terms of time and resources. However, there are several approaches that can be used for population synthesis from limited data. One of the methods is Iterative Proportional Fitting (IPF), which was first introduced by a mathematician (Deming and Stephan, 1940), and was first implemented in transport research (Beckman *et al.*, 1996).

Basically, IPF method consists iteration steps, where each row and each column are proportionally adjusted to be equal with marginal row and column totals. The step is repeated until both row and column convergence or the sum of the rows and columns is relatively similar to their marginal total. As reviewed by Müller and Axhausen (2011) and Sun and Erath (2015), the IPF has been extended by many researchers, for example, to deal with the zero-cell value (Guo and Bhat, 2007), Ye *et al.* (2009) to propose an iterative proportional Updating (IPU), to address memory consumption issues (Pritchard and Miller, 2012), and to introduce hierarchical and multi-stage IPF procedures (Casati *et al.*, 2015; Zhu and Ferreira, 2014). Furthermore, there are other methods, such as Voas and Williamson (2001) utilize Combinatorial Optimization (CO), Farooq et at. (2013) employ Markov Chain Monte Carlo (MCMC), Hafezi and Habib (2014) introduce Fitness Based Synthesis (FBS), and other methods. Sun and Erath (2015), and also applied by (Zhang *et al.*, 2017) employed Bayesian Networks (BN) for the task.

In this paper, we will use a Bayesian network for the population synthesis in Greater Jakarta Area. The following reasons motivate us to use the Bayesian network in this paper. First, Bayesian Network is the approach that gives better results when the number of data points is limited. For instance, as found by Sun and Erath (2015) that while the sample rate is less than 40%, Bayesian Network outperforms other methods (i.e MCMC, DI, and IPF) in terms of resulting Square Root of The Mean Square Error (SRMSE) and becomes the best model to capture heterogeneity with rates of less than 70%. Therefore, given the limitation of data and the difficulties to get reliable data in Greater Jakarta Area we prefer this approach. Second, this paper will use Bayesian Network for synthetic population in a megacity, of which there are 31 in the world and 26 of them are in the less developed region (United Nations, 2016). However, BN model is only able to reproduce distributions of varibles that are included in BN model as found (Sun and Erath, 2015; Zhang *et al.*, 2017). Therefore, in this research, we add a second step, which is using multilevelIPF (Muller, 2017), to fit the BN generated data to target values from census data.

The remainder of this paper is structured as follows. In Section 2, we review several use cases of population synthesis. In Section 3, we explain the concept of Bayesian network. In Section 4, we apply Bayesian network for conducting a population synthesis in Jakarta and using multilevel IPF to adjust to census data. Conclusions are discussed in Section 5.

2. Population synthesis

There are several use cases that employ population synthesis with several different approaches. Basically, there are two stages of population synthesis which are the fitting and generation stage (Müller and Axhausen, 2011). The fitting process, however, is an important stage that most researchers try to handle the problems and employ different approaches, as mentioned in the introduction.

We summarize the use cases of the various methods for different location and size of sample in Table 1. Furthermore, there are several related software that can be used for the synthetic population generation, such as, PopoSynWin, ILUTE, PopGen, FSUMTS, CEMDAP, ALBATROS , R package sms, R package synthpop, TRANSIMS, Synthia, SMILE as reviewed by (Müller and Axhausen, 2011; Templ *et al.*, 2017).

Table 1. Summarise Use Cases of Population Synthesis

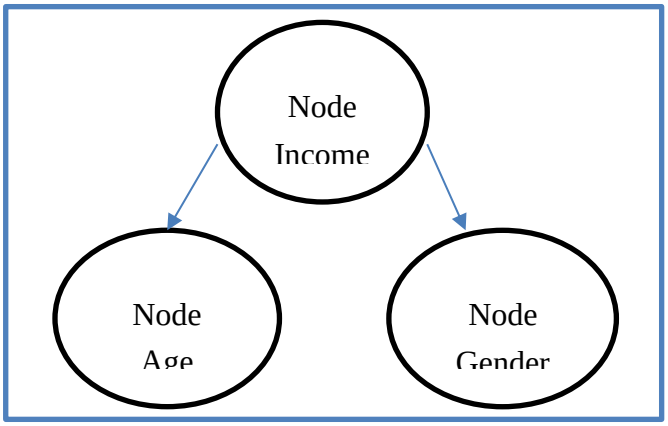
Method	Location	Methods	Sample size (%)	Population (million)
Moeckel *et al.* (2003)	Netanya	IPF with MCMC	6	2.60
Ye *et al.* (2009)	Arizona	IPU	8	3.07
Pritchard and Miller (2012)	Canada	IPF with MCMC	2	3.42
Huynh *et al.* (2013)	Sydney	CO	NA	5
Farooq *et al.* (2013)	Brussels	MCMC	0.1	1.20
Hafezi and Habib (2014)	Canada	FBS	1	1.30
Zhu and Ferreira (2014)	Singapore	IPF	1	4.00
Sun and Erath (2015)	Singapore	Comparing BN, MCMC, IPF, and DI	1*	4.00
Casati *et al.* (2015)	Singapore	IPF	1	4.00
Zhang *et al.* (2017)	San Francisco	BN	6	7

*Test from 1% to 100% of sample size

3. A bayesian network

The Bayesian network uses a graphical method to learn probabilities for a model (Cowel *et al.*, 1999). This method consists two parts: a directed acyclic graph (DAG) and set of the conditional probability distribution (Horný, 2014; Sun and Erath, 2015) where DAG consists of a set of random variables, which are correlated. The variables of graphical structure G = (V, A) are represented by node or vertex (V) and the correlation is represented by the directed edge or arc A. For example, which can be seen in Figure 1, there are variable income, age, and gender. While the directed edges represent probabilities, directed edge from Node Income to Node Age and Node Gender means that Node Age and Node Gender have probabilistic relationship with Node Income. Therefore, the conditional probability distribution of this condition are P[Node Age | Node Income] and P[Node Gender | Node Income].

Figure 1. Directed acyclic graph (DAG) representing potential links between age, gender, and income level



3.1 A learning algorithm for Bayesian network

There are algorithms for learning Bayesian networks as explained by (Scutari, 2010), such as the constraint-based algorithm, score-based algorithm, or hybrid algorithm. Each type of algorithm includes a learning algorithm as explained in Table 2. Here, we used the R package bnlearn (Scutari, 2010), which implements Tabu Search as part of the score-based algorithm. Tabu Search, as generic heuristic procedure, is an iterative searching procedure to obtain the best solution from complex correlation patterns (Glover, 1993) and it also can handle local optima by selecting a very close solution to optimality, which can minimize the score (Scutari, 2010). It supports a whitelist and a blacklist; blacklist means that the arcs will not present be in the network structure, and whitelist means otherwise.

Table 2. Learning algorithm of Bayesian Network

Constraint-based algorithm	Score-based algorithms	Hybrid algorithm
PC Grow-Shrink	Hill-Climbing	Max-min Hill-Climbing
Incremental Association	Tabu Search	Restricted Maximization
Fast Incremental Association		

Source: Scutari, 2010

3.2 Network scores

The step for selecting and measuring, which the candidate of the graphical structure fits the data, is central to ensure our structure is able to produce reliable synthetic population. In this process, several methods were introduced such as maximum likelihood:

$$\lfloor(G^h|D) = \max_G \sup_\Theta l(G, \Theta|D) = \max_G(G, \hat{\Theta}|D), \quad (1)$$

Where $\lfloor(G^h|D) = \max_G \sup_\Theta l(G, \Theta|D)$ is the log-likelihood of a provided pair(G, Θ) given observation D. However, the log-likelihood is not representable as explained by Sun and Erath, (2015) that due to overfitting problem, as this method will always build a fully connected DAG. Thus, most applicable approaches are using Bayesian Information Criterion (BIC) (Rissanen, 1978; Schwarz, 1978) and Akaike Information Criterion (AIC) (Akaike, 1974; Rissanen, 1978):

$$\text{BIC}(G^h|D) = \text{Log}\, P(D|G^h, \hat{\Theta}) - \frac{d}{2}\log m \quad (2)$$

$$\text{AIC}(G^h|D) = \text{Log}\, P(D|G^h, \hat{\Theta}) - d \quad (3)$$

Where Θ, in the first equation, is the maximum likelihood estimate parameter given a hypothetical structure G^h, d are the degree of freedoms in Θ, and m the number of observations. The different of maximum likelihood (1) with BIC (2) and AIC (3) are BIC and AIC give penalty function to the optimal likelihood $\text{Log}\, P(D|G^h, \hat{\Theta})$. Whereas for BIC the penalty is $\frac{d}{2}\log m$ and for AIC is d. Using the scoring function, the best network is used for constructing synthetic population.

4. Constructing the population of greater Jakarta area

The study area is Greater Jakarta Area or Jabodetabek, which consists of Jakarta, parts of West Java, and Banten. There are 31.7 million habitants in this region (see Table 3). The population data used in synthetic population was obtained from the JAPTRAPIS study (Jabodetabek Public Transport Policy Implementation and Strategy) in 2009 (JICA, 2009).

Table 3. The population of Greater Jakarta Area

Province	Region	Male	Female
Jakarta[1]	South Jakarta	1,096,469	1,089,242
	Easts Jakarta	1,436,128	1,407,688
	Central Jakarta	457,025	457,157
	West Jakarta	1,246,288	1,217,272
	North Jakarta	867,727	879,588
Banten[2]	Tangerang City	1,045,113	1,001,992
	Tangerang Regency	1,724,915	1,645,679
	South Tangerang City	777,713	765,496
West Java[3]	Depok	1061900	1044200
	Bogor	532000	515900
	Bogor Regency	2792900	2666800
	Bekasi City	1369600	1654600
	Bekasi Regency	1654600	1345200
Total		31,753,192	

4.1 Data sources

There are two different type of data in JAPTRAPIS study, which are the Household Travel Survey (HTS) and the Activity Diary Survey (ADS). However, we only use HTS data consisting 178,954 households or 334’973 individuals for population synthesis, which are equal to three percent of all households. In HTS data, the respondents are individuals who are going to school or work. Therefore, the aggregate synthetic population from the census is the only from individual who has activities for studying and working (see Table 4).

Table 4. The Population of Jakarta Greater Area With Activities from Census

Province	Region	Male	Female
Jakarta[1]	South Jakarta	1,122,151	867,078
	Easts Jakarta	1,460,024	1,128,150
	Central Jakarta	469,344	362,659
	West Jakarta	1,264,799	977,301
	North Jakarta	897,077	693,165
Banten[2]	Tangerang City	999,131	590,688
	Tangerang Regency	1,645,087	972,578
	South Tangerang City	753,194	445,290
West Java[3]	Depok	711,791	395,011
	Bogor	354,155	196,539
	Bogor Regency	1,845,195	1,023,997
	Bekasi City	1,022,078	567,205
	Bekasi Regency	1,013,831	562,629
Total		22,340,145	

4.2 Model estimation

4.2.1 Bayesian network

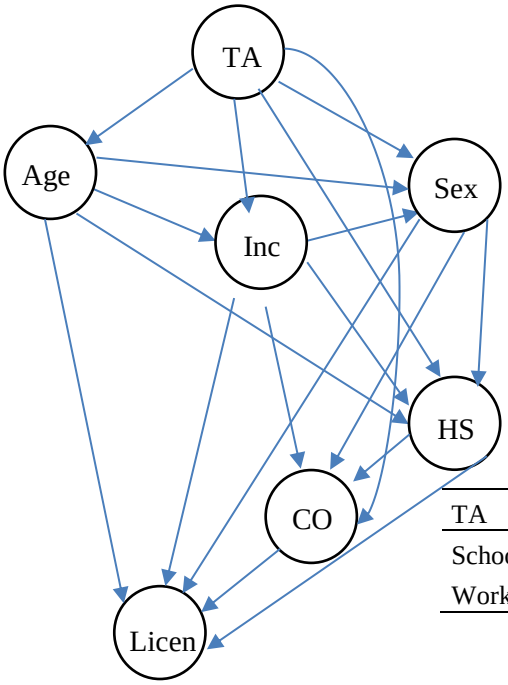
We consider seven variables for each individual from HTS data for the synthesis as presented in Table 5, such as type of activities, age, sex, income, housing, car ownership, and driving license. However, for the variable of income, housing status, and car ownership for each individual the household information is used. The chosen network structure is based on the result from score function AIC employing the Tabu search algorithm to learn the structure of the BN, as implemented in bnlearn package.

Table 4. The Population of Jakarta Greater Area With Activities from Census

Table 5 Attributes of individual

Variable	Definition [number of categories]	Values
TA	Type of activities of individual [2]	School; Work
Age	Age of individual [8]	< 6; 6-12, 12-18;18-24;24-32;32-42;42-60,>60
Sex	Gender of individual [2]	Male; Female
Inc	Income of household [7]	NA; < 1, 1-3; 3-5; 5-8; 8-15; >15
HS	Housing status of household [2]	Owned; Rented
CO	Car ownership of household [2]	Yes; No
Licen	License of individual [2]	Yes; No

Figure 2. Final model structure G



			Sex	
Inc	TA	Age	Male	Female
5-8	School	<6	0.43	0.57
5-8	Working	12-18	0.78	0.22
5-8	School	18-24	0.50	0.50
5-8	Working	24-32	0.52	0.48

	Age							
TA	<6	6-12	12-18	18-24	24-32	32-42	42-60	>60
School	0.0696	0.4695	0.3768	0.0699	0.0073	0.0037	0.0030	0.0003
Working	0.0002	0.0013	0.0181	0.1250	0.2484	0.2978	0.2810	0.0280

There are two steps tabu search in this scenario; without using whitelist and blacklist G structures for initial search and with using whitelist and blacklist for final search, where there are 256 number initialize graph in the initial search and 64 for the final search. The final structure is obtained by an iterative process after measuring the error of each arch. The arch, which gives smallest error, is included in the network using the whitelist command and the arch, which gives highest error, is never included in the network using the blacklist command. In final search, we found the value of AIC is -1679334. The model structure can be seen in Figure 2. After we find the best structure, we generate data for the total population, where, in our case, we extend to differ in size from 4 million, 8 million, 16million, and 22 million observations. Figure 3 shows the marginal distribution of each variable and Figure 4 shows the joints distribution comparison, which in each size of data BN can give similar distribution. It shows from initial (network.1) and final search (network.fin) that there is no meaningful difference distribution with HTS survey, which means that the BN structure is able to generate data with similar distribution.

Figure 3. Marginal Distribution of Variable Structure

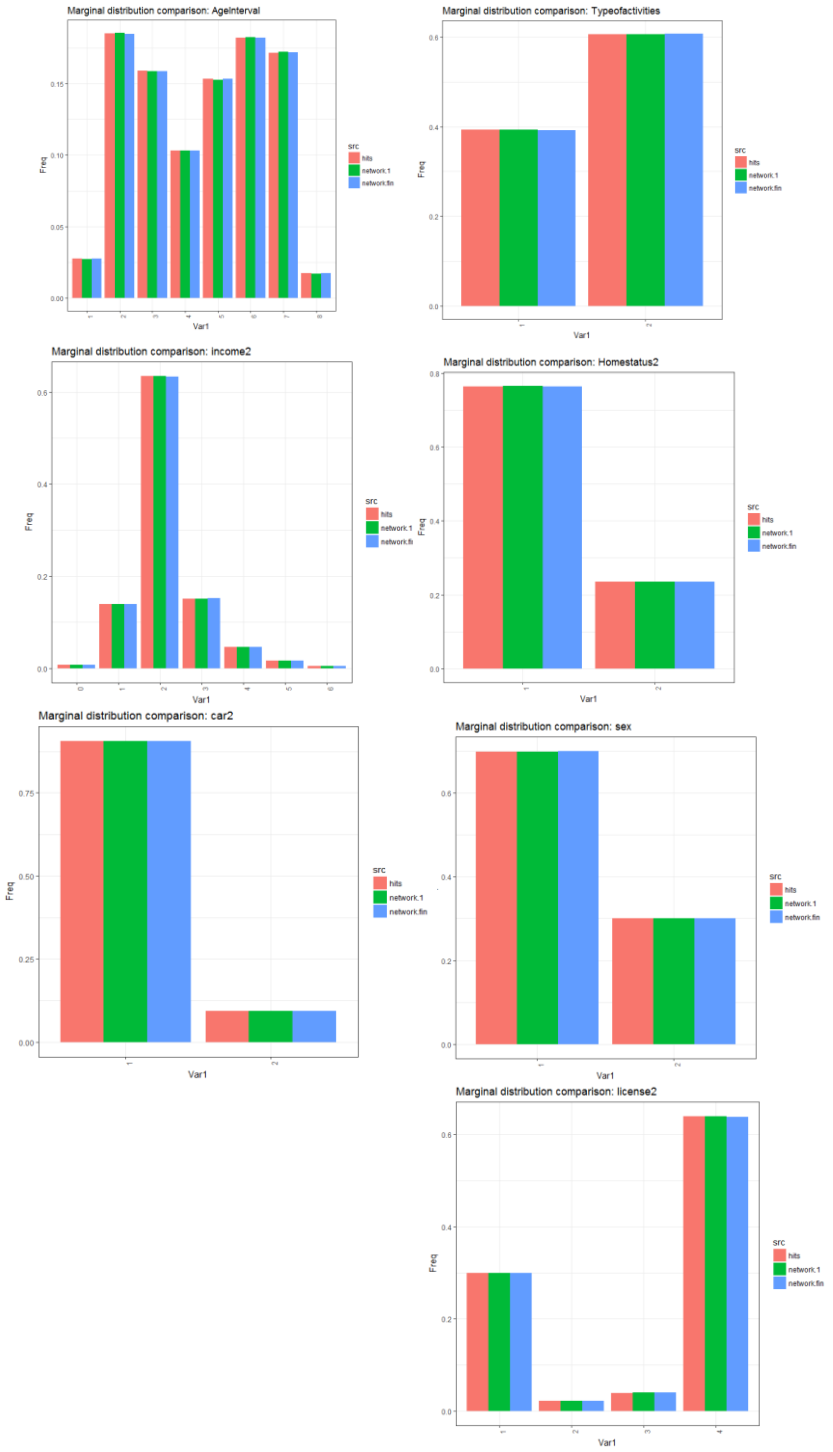
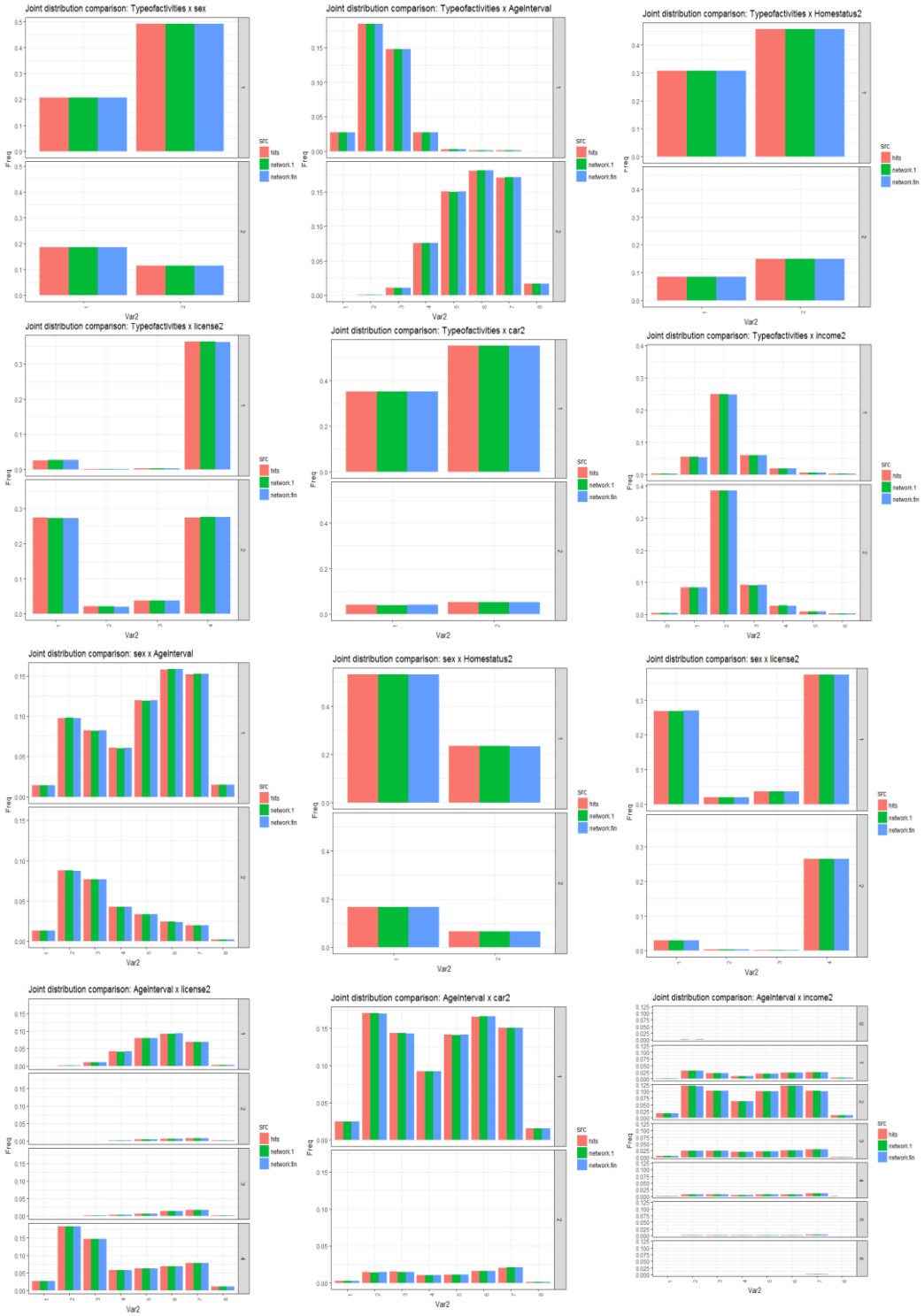


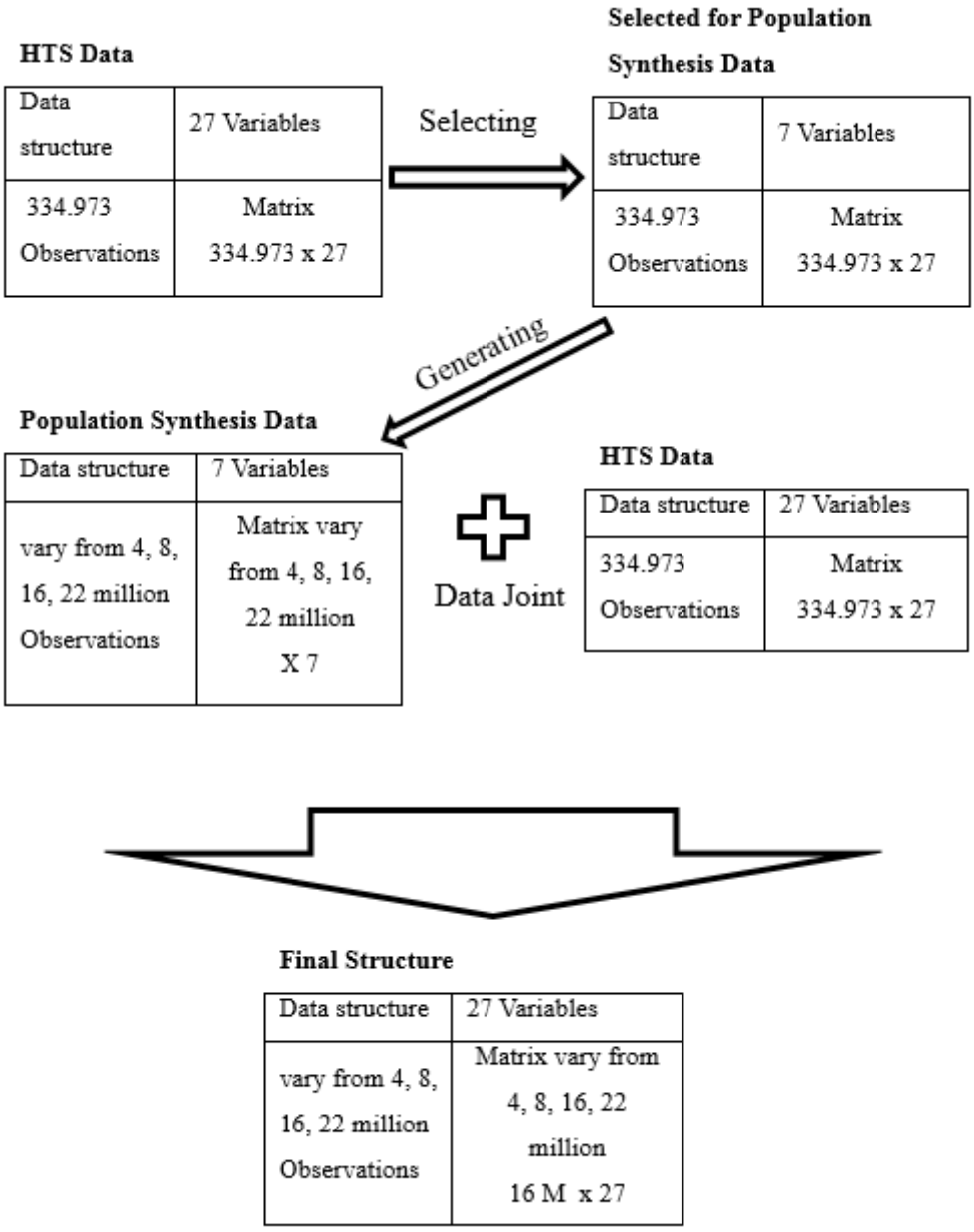
Figure 4. Join Distribution of Variable Structure



Based on the generated population to extended target, we join them to the HTS data, which has 20 variables and 334.000 observation. Thus, we get joint data consisting 27 variables with vary from extended target observation million observations as seen in Figure 5. For extended target 16 million observations, this joint operation takes 3 days 13 hours. Moreover,

for 22 million, this joint operation takes more than 5 days. Joint operation uses the servers of ETH Zurich computer cluster (https://scicomp.ethz.ch/wiki/Euler). This operation takes considerable amount of time, which because of the large size of the data and complicated data structure.

Figure 5. Generating Population Synthesis Step



As a result, to visualize the goodness of fit after joining the data. We compare HTS data and final structure from BN. However, in this case, we only visualize two different type joint distributions. While Figure 6 shows joint distribution of income and age, Figure 7 shows joint distribution of income and region. The left figure shows the joint distribution data from HTS,

the middle figure shows the joint distribution of the data from BN, and the right shows the fit of joint distribution of both HTS and BN. We consider region, which was not used in synthesis data in Figure 7, to ensure joining process gives similar distribution.

Figure 6. Joint Distribution of Income and Age

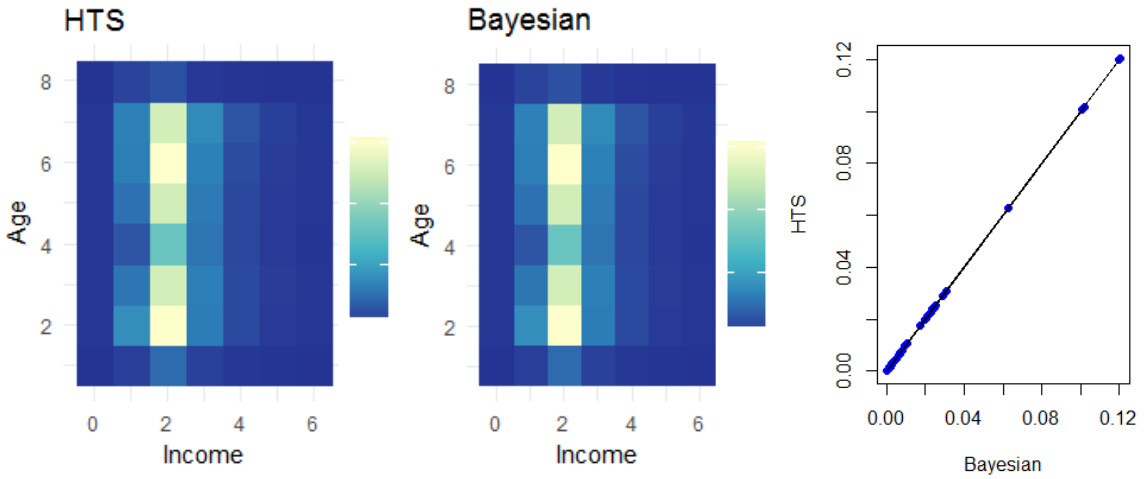
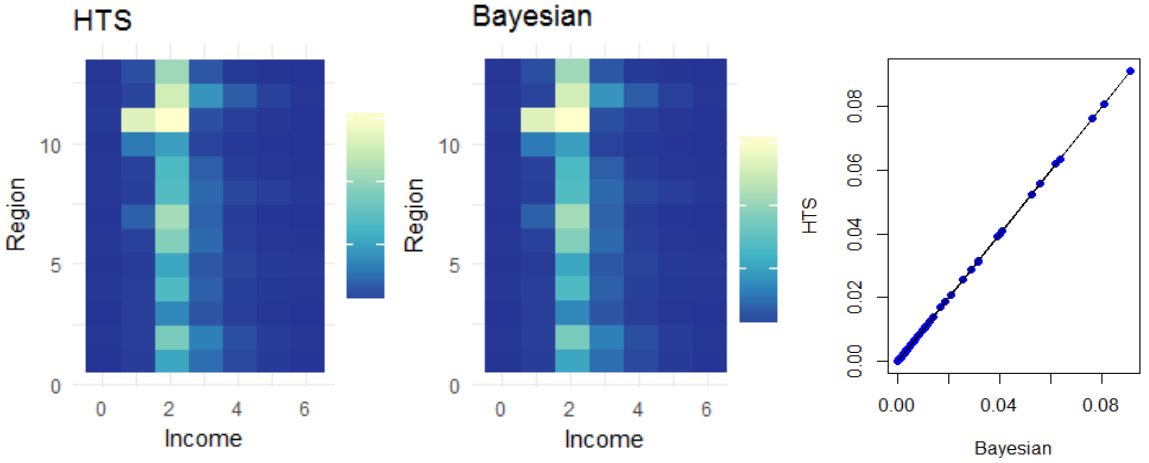


Figure 7. Joint Distribution of Income and Region



Furthermore, as found in (Sun and Erath, 2015; Zhang *et al.*, 2017) that the BN model is only able to give similar distributions for the selected variables in BN model. That is because we can only assign to the target value based on the best BN structure. Therefore, in this research, we do Multilevel IPF using MultilevelIPF packed in R (Müller, 2017) to adjust to the total population.

4.2.2 Multilevel IPF for fitting against census cata

The fitting against census data is done by using MultilevelIPF package in R (Müller, 2017). The are several algorithms, that have been implemented such as Hierarchical Iterative

Proportional Fitting (HIPF), Iterative Proportional Updating (IPU), Entropy Optimization (Ent) and Generalized Raking (GR). Here, we use a GR-raking algorithm that has been shown to outperform the other algorithms (Müller, 2017). There are two group control that we used: region and gender, and region ang age in this scenario.

Furthermore, the weakness of IPF is that it produces non-integer weights (Lovalance and Ballas, 2012; Müller, 2017), while agent-based models require integer weights for the simulation. Integerisation is the process to convert the value of decimal weights (related with how many times each agent is replicated) to integer values. Therefore, integerisation is important to produce the best fitting result with less overspecified to marginal census data. This can be done using weighted random sampling without replacement using the wrswoR package (Müller, 2018).

There are three steps in the weighted random sampling without replacement. The first step is removing the decimal values (floorweight). Then, calculating the decimal remainders (prob) that will be used as a vector probability weights, and implementing to weighted sampling without replacement (indexes) using wrswoR package. Algorithm crank is used in this operation because it gives a faster result (Müller, 2018). The following command is used for this operation:

```
floorweight <- floor(weights)

Prob <- weight - floorweight

indexes <- sample_int_crank(length(weights), sum(weights) - sum(floorweight), prob)
```

The results can be seen in Figure 8 and Figure 9 that the synthetic population data has less overspecified fitting problem after conducting multilevel IPF to target census data. The data that we used from BN for fitting with IPF differ in size from 4 million, 8 million, 16million, and 22 million observations. However, the higher data used gives more overspecified result. Thus, in this research, we use 4 million data observations for fitting operations that gives less overspecified less than 1%.

Figure 8. Comparison between population synthesis and census data of region and gender

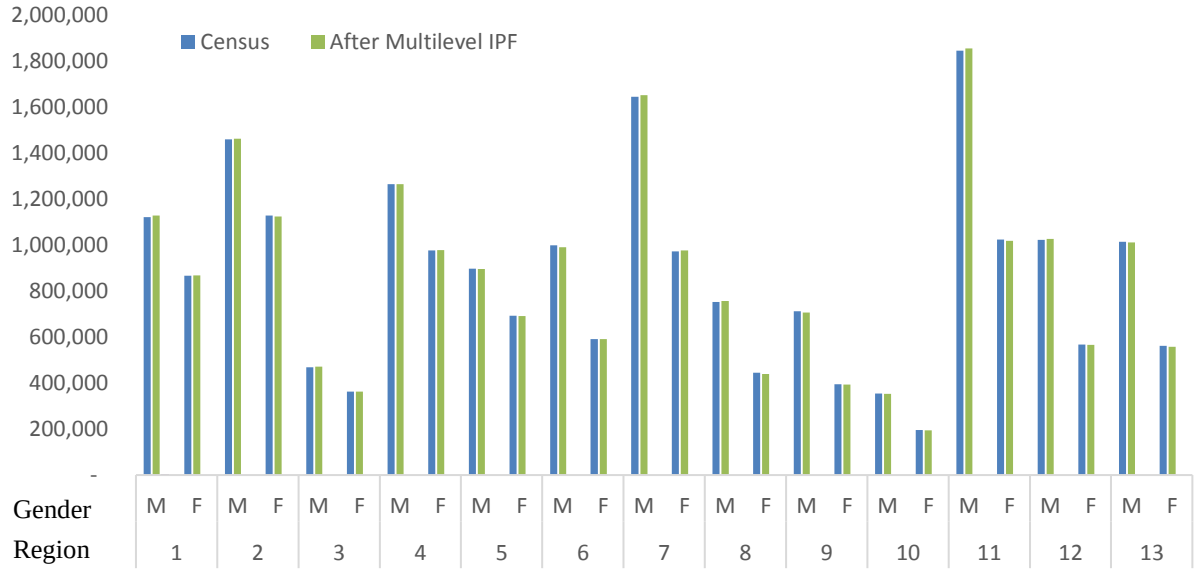
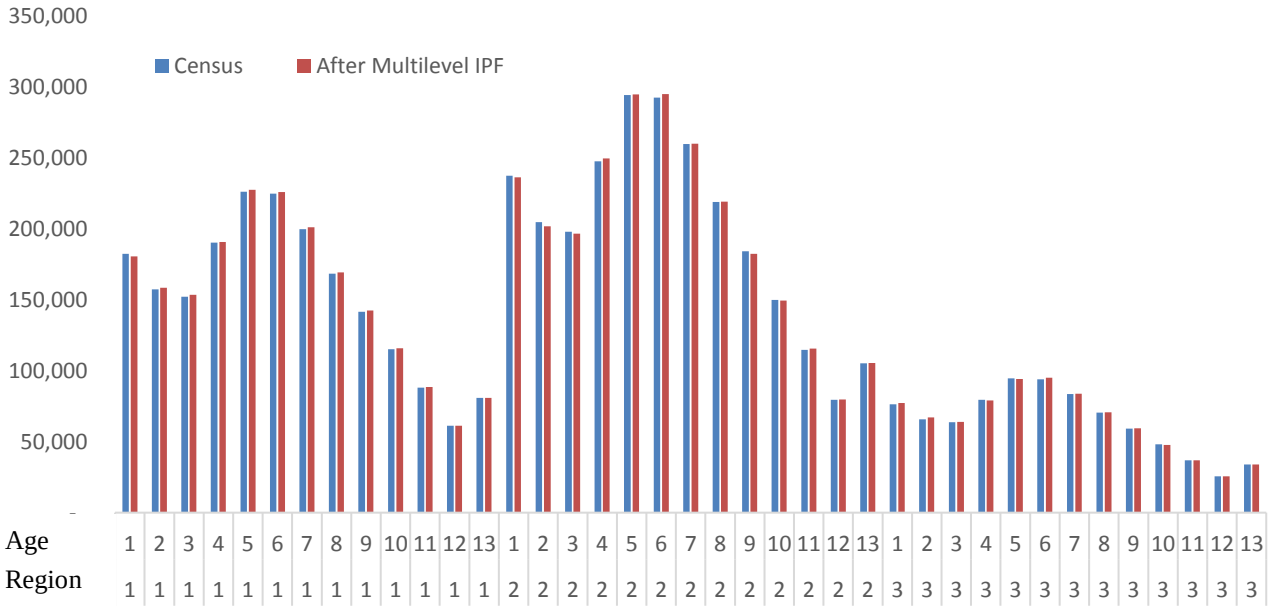


Figure 9 Comparison between population synthesis and census data of region and age



5. Conclusion

BN approach can be used to preform population synthesis in many cases. In our case, we found that this approach is able to construct a synthetic population and give a similar distribution with HTS data. On one hand, there is no difference if we look at marginal distribution and joint distribution of variables against HTS data. Moreover, the joint data also gives similar distributions. On the other hand, we have some differences vis-à-vis census data. Therefore, we need fit against the marginal distributions of census data. However, the differences with census data are addressed with GR.

Furthermore, the result of this synthetic population will be used to develop an agent-based model using Multi-Agent Transport Simulation (MATSim) (Horni *et al.*, 2016). This result will be first case scenario of an agent-based model for Greater Jakarta.

Acknowledgements

The authors wish to acknowledge the scholarship from LPDP (Indonesia Endowment Fund for Education) Ministry of Economic. Special thanks also to JICA (Japan International Cooperation Agency) for allowing us to use survey data in this study.

6. References

Akaike, H. (1974) A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control,* AC-**19**(6) 716-723.

Balmer, M., K. W. Axhausen and K. Nagel (2006) Agent-Based Demand-Modeling Framework for Large-Scale Microsimulations, *Transportation Research Record: Journal of the Transportation Research Board*, **1985**, 125-134.

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.

BPS-Statistics of Banten Province. (2016) *Banten in Figures.*

BPS-Statistics of DKI Jakarta Province. (2016) *Jakarta in Figures.*

BPS-Statistics of West Java Province. (2016) *West Java in figures.*

Casati, D., K. Müller, P. Fourie, A. Erath and K. W. Axhausen (2015) Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking, paper presented at the *94th Annual Meeting of Transportation Research Board*, Washington, D.C., January 2015.

Cowel, R. G., A. P. Dawid, S. L. Lauritzen and D. J. Spiegelhalter (1999) *Probabilistic Networks and Expert Systems*, Springer-Verlag New York.

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, *Ann. Math. Statist.*, **11**(4) 427-444.

ETH Zurich. (2018, February 14). *the Scientific IT Services (SIS)*, Retrieved from https://scicomp.ethz.ch/wiki/Euler

Farooq, B., M. Bierlaire, R. Hurtubia and G. Flötteröd (2013) Simulation based population synthesis, *Transportation Research Part B: Methodological*, **58**, 243-263.

Glover, F. (1993) A user's guide to tabu search. *Annals of Operations Research,* **41**, 3-28.

Guo, J. and C. Bhat (2007) Population Synthesis for Microsimulating Travel Behavior, *Transportation Research Record: Journal of the Transportation Research Board*, **2014**, 92-101.

Hafezi, M. H. and M. A. Habib (2014) Synthesizing Population for Microsimulation-Based Integrated Transport Models using Atlantic Canada Micro-Data, *The 1st International Workshop on Information Fusion for Smart Mobility Solutions (IFSMS'14)*, **37**, pp. 410-415. Procedia Computer Science.

Horni, A., K. Nagel, and K. W. Axhausen (2016) *The Multi-Agent Transport Simulation MATSim*, London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/baw. License: CC-BY 4.0.

Horný, M. (2014) *Bayesian Networks.* Technical Report, Boston University School of Public Health, Department of Health Policy and Management.

Huynh, N., M. Namazi-Rad, P. Perez, M. J. Berryman, and Q. Chen (2013) Generating a synthetic population in support of agent-based modeling of transportation in Sydney, *20th International Congress on Modelling and Simulation (MODSIM 2013)*, 1357-1363.

JICA. (2009) *Traffic Data Collected Under "The Jabodetabek Urban Transport Policy integration"*, Japan International Cooperation Agency (JICA), Indonesia.

Lovelace, R., and B. Ballas (2013) 'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation, *Computersm Environment and Urban Systems*, **41,** 1-11.

Moeckel, R., M. Wegner and K. Spiekermann (2003) Creating a Synthetic Population, paper presented at *the 8th International Conference on Computers in Urban Planning and Urban Management (CUPUM)*, Sendai.

Müller, K. (2017) *A generalized approach to population synthesis*, Doctoral Thesis, ETH Zürich.

Müller, K. (2018) *Weighted Random Sampling without Replacement*, https://cran.r-project.org/web/packages/wrswoR/wrswoR.pdf.

Müller, K. and K. W. Axhausen (2011) Population synthesis for microsimulation: state of the art. *In: Transportation Research Board 90th Annual Meeting*, Washington, D.C.

Pritchard, D. R. and E. J. Miller (2012) Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins, *Transportation*, **39**(13) 685-714.

Rissanen, J. (1978) Modeling By Shortest Data Description, *Automatica*, **14,** 465-471.

Schwarz, G. (1978) Estimation the Dimension of a Model, *The Annals of Statistic*, **6**(2) 461-464.

Scutari, M. (2010) Learning Bayesian Networks with the bnlearn R Package, *Journal of Statistical Software*, **35**(3).

Sun, L., and A. Erath (2015) A Bayesian network approach for population synthesis, *Transportation Research Part C*, **61**, 49-62.

Templ, M., B. Meindl, A. Kowarik and O. Dupriez (2017) Simulation of Synthetic Complex Data: The R Package simPop, *Journal of Statistical Software*, **79**(10).

United Nations. (2016) *The World's Cities in 2016*, Data Booklet.

Voas, D., and P. Williamson (2001) Evaluating Goodness-of-Fit Measures for Synthetic Microdata, *Geographical and Environmental Modelling*, **5**(2) 177-200.

Ye, X., K. Konduri, R. M. Pendyala, B. Sana, and P. Waddell (2009) Methodology to match distributions of both household and person attributes in the generation, paper presented at the *88th Annual Meeting of Transportation Research Board*, Washington, D.C., January 2009.

Zhang, D., J. Cao, S. Feygin, T. Dounan, T and A. Pozdnoukhov (2017) *Connected Population Synthesis for Urban Simulation*, The report is available in the link of UC Berkeley.

Zhu, Y. and J. Ferreira (2014) Synthetic Population Generation at Disaggregated Spatial Scales for Land Use and Transportation Microsimulation, *Transportation Research Record: Journal of the Transportation Research Board*, **2429**, 168-177.