Population synthesis of Jakarta

Population Synthesis of Jakarta

Anugrah Ilahi[1], Kay W. Axhausen[1]
[1]ETH Zurich, Institute for Transport Planning and Systems, Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland

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 [1]. 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 [2] and was first implemented in transport research [3].

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 [4]. 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.

3 A Bayesian network

The Bayesian network uses a graphical method to learn probabilities for a model [5]. This method consists two parts: a directed acyclic graph (DAG) and set of the conditional probability distribution [6, 7] 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 Fig. 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 conditional probability with Node Income. Therefore, the conditional probability distribution of this condition are P[Node Age | Node Income] and P[Node Gender | Node Income].

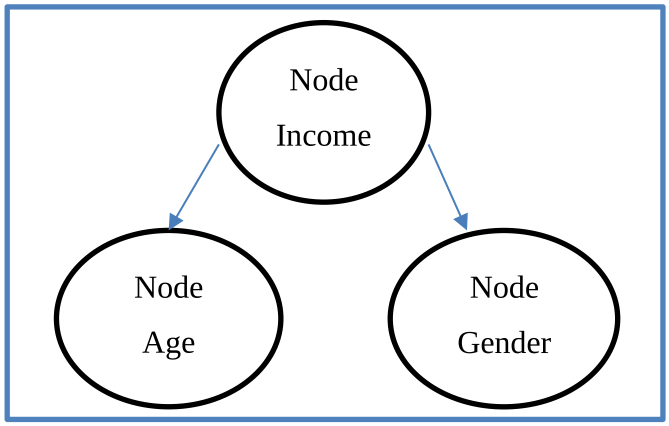


Fig. 1. Directed acyclic graph (DAG) representing potential links between age, gender, and income level

3.1 A learning algorithm for Bayesian network

There are algorithms for learning Bayesian networks as explained by [8], such as the constraint-based algorithm, score-based algorithm, or hybrid algorithm. Each type of algorithm includes a learning algorithm as explained, see Table 2. Here, we used the R package bnlearn [8], 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 [9] and it also can handle local optima by selecting a very close solution to optimality, which can minimize the score [8]. 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: [8]

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:

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

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) [10, 11] and Akaike Information Criterion (AIC) [10, 12].

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

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

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 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 Value of Travel Time Savings

The study area is Greater Jakarta Area or Jabodetabek, which consists of Jakarta province, parts of West Java province, and Banten province. There are 31.7 million habitants in this region [13-15], 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 [16, 17].

Table 3
The population of greater Jakarta area

Province	Region	Male	Female
Jakarta	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	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	Depok	1,061,900	1,044,200
	Bogor	532,000	515,900
	Bogor Regency	2,792,900	2,666,800
	Bekasi City	1,369,600	1,654,600
	Bekasi Regency	1,654,600	1,345,200
	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 with total more than 20 million, see Table 4.

Table 4
The population of Jakarta greater area with activities from census

Province	Region	Male	Female
Jakarta[1]	South Jakarta	877,679	626,495
	Easts Jakarta	1,141,943	815,130
	Central Jakarta	367,092	262,034
	West Jakarta	989,250	706,136
	North Jakarta	701,639	500,837
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	20,086,633	

4.2 Model estimation

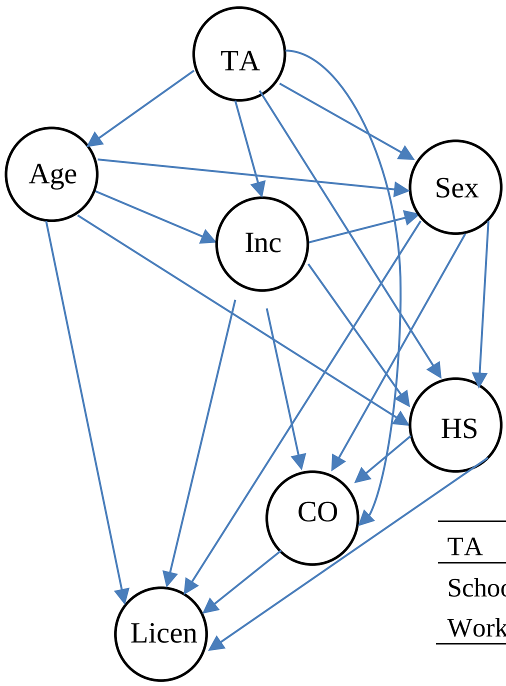
4.2.1 Bayesian network step

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 the *R* package *bnlearn* [8, 18].

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 [4]	Motor cycle; Car, Motorcycle and Car, No License

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.



			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

Fig. 2. Final model structure G

The model structure and probability distribution of variables can be seen in Fig. 2. After we find the best structure, we generate data as close as possible to aggregate of census data, where, in our case, we extend to differ in size from 4 million, 8 million, 16million, and 22 million observations. Fig. 3 shows the marginal distribution of each variable and Fig. 4 represents the joints distribution of variables, which give similar distribution to HTS data. Furthermore, it shows that there is no meaningful difference distribution from initial and final search with HTS data, which means that the BN structure is able to generate data with similar distribution. Discussion and 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 joining process of data also gives similar distributions. Our results, as also be found in [7, 19], confirm that BN approach can be used to increase number of sample data with similar distribution, which is useful for researcher who has limited sample.

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.

References

[1] M. Balmer, K.W. Axhausen, K. Nagel, Agent-based demand-modeling framework for large-scale microsimulations, Transport Res Rec, DOI (2006) 125-134.
[2] W.E. Deming, F.F. Stephan, On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, Ann Math Stat, 11 (1940) 427-444.
[3] R.J. Beckman, K.A. Baggerly, M.D. McKay, Creating synthetic baseline populations, Transport Res a-Pol, 30 (1996) 415-429.
[4] K. Müller, K.W. Axhausen, Population synthesis for microsimulation: state of the art., In: Transportation Research Board 90th Annual Meeting, Washington, D.C, 2011.
[5] R.G. Cowel, A.P. Dawid, S.L. Lauritzen, D.J. Spiegelhalter, Probabilistic Networks and Expert Systems, Springer-Verlag New York, 1999.
[6] M. Horný, Bayesian Networks, Boston University School of Public Health, 2014.
[7] L.J. Sun, A. Erath, A Bayesian network approach for population synthesis, Transport Res C-Emer, 61 (2015) 49-62.
[8] M. Scutari, Learning Bayesian Networks with the bnlearn R Package, J Stat Softw, 35 (2010) 1-22.
[9] F. Glover, A user's guide to tabu search, Annals of Operations Research, 41 (1993) 3-28.
[10] J. Rissanen, Modeling by Shortest Data Description, Automatica, 14 (1978) 465-471.
[11] G. Schwarz, Estimating Dimension of a Model, Ann Stat, 6 (1978) 461-464.
[12] H. Akaike, New Look at Statistical-Model Identification, Ieee T Automat Contr, Ac19 (1974) 716-723.
[13] BPS-Statistics, West Java in figures, BPS-Statistics of West Java Province2016.
[14] BPS-Statistics, Banten in Figures, BPS-Statistics of Banten Province2016.
[15] BPS-Statistics, Jakarta in Figures, BPS-Statistics of DKI Jakarta Province, 2016.
[16] JICA, Traffic Data Collected Under "The Jabodetabek Urban Transport Policy integration", in: J.I.C.A. (JICA) (Ed.), Indonesia, 2009.
[17] A. Ilahi, K.W. Axhausen, Measuring accessibility using an activity based model approach in Jabodetabek, 17th Swiss Transport Research Conference, Ascona, 2017.
[18] R.C. Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2010.
[19] D. Zhang, J. Cao, S. Feygin, T. Dounan, Pozdnoukhov, Connected Population Synthesis for Urban Simulation, 2017.

1

9

APPENDIX 2. Gender by region

Region	Gender	Population
South Jakarta	Male	877,679
	Female	626,495
Easts Jakarta	Male	1,141,943
	Female	815,130
Central Jakarta	Male	367,092
	Female	262,034
West Jakarta	Male	989,210
	Female	706,136
North Jakarta	Male	701,639
	Female	500,837
Tangerang City	Male	999,131
	Female	590,678
Tangerang regency	Male	1,645,087
	Female	972,578
South Tangerang City	Male	753,194
	Female	445,290
Depok	Male	711,791
	Female	395,011
Bogor	Male	354,155
	Female	196,559
Bogor Regency	Male	1,845,495
	Female	1,023,997
Bekasi City	Male	1,022,078
	Female	567,205
Bekasi Regency	Male	1,013,821
	Female	562,639