Integrating Bayesian network and generalized raking for population synthesis in Greater Jakarta

Author(s):
Ilahi, Anugrah; Axhausen, Kay W.

Publication Date:
2019-11-20

Permanent Link:
https://doi.org/10.3929/ethz-b-000373535

Originally published in:
Regional Studies, Regional Science 6(1), http://doi.org/10.1080/21681376.2019.1687011

Rights / License:
Creative Commons Attribution 4.0 International
Integrating Bayesian network and generalized raking for population synthesis in Greater Jakarta

Anugrah Ilahi & Kay W. Axhausen

To cite this article: Anugrah Ilahi & Kay W. Axhausen (2019) Integrating Bayesian network and generalized raking for population synthesis in Greater Jakarta, Regional Studies, Regional Science, 6:1, 623-636

To link to this article: https://doi.org/10.1080/21681376.2019.1687011

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 20 Nov 2019.
Integrating Bayesian network and generalized raking for population synthesis in Greater Jakarta

Anugrah Ilahi and Kay W. Axhausen

ABSTRACT
Constructing agent data with detailed information on their sociodemographics is substantially important for agent-based modelling. However, to collect data about the whole population is not efficient, since it requires an expensive and time-consuming survey, especially for a large population. The paper uses a novel approach that integrates Bayesian network (BN) and generalized raking (GR) multilevel iterative proportional fitting (IPF). Furthermore, the approach is applied to construct the population for Greater Jakarta, Indonesia, which consists of 30 million inhabitants. The results show that the BN approach can produce data that represent the probability distribution of sample data and that the IPF can match it against aggregate census data.

ARTICLE HISTORY
Received 3 August 2019; Accepted 7 October 2019

KEYWORDS
Bayesian network; generalized raking; population synthesis; agent-based model

INTRODUCTION
Agent-based transportation models require detailed individual and household information such as sociodemographic and geocoding of activity locations (Balmer, Axhausen, & Nagel, 2006). These data can be collected from multiple data sources. To collect all those data for a population is costly in terms of time and resources. Therefore, several approaches can be used for population synthesis from limited data. One of the initial methods is iterative proportional fitting (IPF), which was first introduced by Deming and Stephan (1940) and first implemented in transport research (Beckman, Baggerly, & McKay, 1996). To tackle the sample problems, Barthelemy and Toint (2013) developed a method for the synthetic population without sample by using the available information of disaggregated-level data and randomly drawing. This method was shown to produce a consistent synthetic population. Basically, IPF consists of iteration steps, where each row and each column are proportionally adjusted to be equal to the marginal row and column totals. The steps are repeated until both row and column convergence or the sum of the rows and columns is relatively similar to their marginal total. However, the control total of IPF is based on marginal individual or marginal household...
totals. Since IPF is widely used for population synthesis, there are several efforts to improve the quality of IPF and to develop new approaches. 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 issue (Guo & Bhat, 2007). Ye, Konduri, Pendyala, Sana, and Waddell (2009) proposed an iterative proportional updating (IPU), Pritchard and Miller (2012) addressed memory consumption issues, and also Casati, Muller, Fourie, Erath, and Axhausen (2015); Zhu and Ferreira (2014) introduced hierarchical and multistage IPF procedures. IPU and hierarchical iterative proportional fitting (HIPF) are advanced methods of IPF that consider household and individual control totals. However, if only using one type of control total, it is considered as IPF.

Furthermore, there are other methods, such as combinatorial optimization (CO) (Voas & Williamson, 2001), which estimates the integer weight of a sample and replicates the desired marginal target total. It claims to have less variance than IPF. Farooq, Bierlaire, Hurtubia, and Flutterod (2013) and Saadi, Mustafa, Teller, and Cools (2016) proposed Markov chain Monte Carlo (MCMC), which deals with zero-cell issue. There is an extension of MCMC, called hierarchical MCMC, that uses individual and household attributes simultaneously. Hafezi and Habib (2014) further employed fitness-based synthesis (FBS), which allows the consideration of multilevel controls by selecting variables that have maximum fitness. However, this approach gave less accuracy when there is more detailed classification of attributes. Sun, Erath, and Cai (2018) developed a hierarchical mixture model, which combines three different models: probabilistic tensor factorization, multilevel latent class modelling and rejection sampling. This approach can produce a pool of agents that consider the interdependencies of household and individual structure. This method is complicated and not widely applicable, though.

Several use cases employ population synthesis with different approaches. We now summarize the use cases of the various methods for different locations and population sizes. As can be seen in Table 1, most of the locations are in developed countries and have populations < 10 million. Furthermore, related software can be used for synthetic population generation, such as, PopoSyn-Win, ILUTE, PopGen, FSUMTS, CEMDAP, ALBATROS, R package sms, R package synthpop, TRANSIMS, Synthia and SMILE (Müller & Axhausen, 2011; Templ, Meindl, Kowarik, & Dupriez, 2017).

As an alternative, Sun and Erath (2015) and Zhang et al. (2017) employed Bayesian networks (BN) for the task. They claim it can capture the complex interaction and hierarchical household structure of the sample and show that it is better than other methods (i.e., MCMC, directly inflating (DI) and IPF). Moreover, Saadi, Mustafa, Teller, Farooq, and Cools (2016) used the hidden Markov model (HMM), which is a stochastic model that learns a complex joint distribution sample structure or known as the simplest BN. However, this model has less consideration for hierarchical household structure. Since HMM cannot fit the marginal totals, Saadi, Farooq, Mustafa, Teller, and Cools (2018) improved the method that integrated an HMM and IPF. After the combination, the model can give a quasi-perfect result.

In this paper we use a BN. However, similar to an HMM, a BN can only generate synthetic data and gives a similar distribution to the sample (Sun & Erath, 2015; Zhang et al., 2017). Therefore, we add a similar type of integration as in Saadi et al. (2018). However, in this case, we integrated a BN and a generalized raking (GR) multilevel IPF to fit the aggregate census data. The following reasons motivate this choice. First, when the sample is < 40%, BN is known to give better results to model the interdependencies of individual and household attributes (Sun & Erath, 2015). Second, BN outperforms other methods in terms of the resulting square root of the mean square error (SRMSE) and becomes an excellent model with which to capture heterogeneity when the sample is < 70%.
This study makes the following contributions:

- We combined two methodological approaches: BN integrated with GR multilevel IPF.
- We applied the model to developing countries and a megacity, one of a total of 31 locations worldwide, of which 26 are in the less developed regions (United Nations, 2016) that have lower data availability.
- This paper adds to the growing literature on BN for population synthesis and its application in developing countries and megacities.

The paper is structured as follows. The next section explains the concept of BN. The third section discusses the study area and framework model used. The fourth section applies the approach to a population synthesis using BN and GR multilevel IPF to fit aggregate census data. Discussions and conclusions follow in the fifth and sixth sections.

**BAYESIAN NETWORK**

The BN uses a graphical method to learn probabilities for a model (Cowell, Dawid, Lauritzen, & Spiegelhalter, 2006). It consists of two parts: a directed acyclic graph (DAG) and a set of conditional probability distributions (Sun & Erath, 2015; Horný, 2014), where DAG consists of a set of correlated random variables. The variables of a graphical structure \( G = (V, A) \) are represented by node or

---

Table 1. Summary of use cases of population synthesis.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Location</th>
<th>Methods</th>
<th>Sample size (%)</th>
<th>Population (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barthelemy and Toint (2013)</td>
<td>Belgium</td>
<td>Randomly drawing</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Casati et al. (2015)</td>
<td>Singapore</td>
<td>IPF</td>
<td>1</td>
<td>4.0</td>
</tr>
<tr>
<td>Farooq et al. (2013)</td>
<td>Brussels, Belgium</td>
<td>MCMC</td>
<td>0.1</td>
<td>1.20</td>
</tr>
<tr>
<td>Hafezi and Habib (2014)</td>
<td>Canada</td>
<td>FBS</td>
<td>1</td>
<td>1.30</td>
</tr>
<tr>
<td>Huynh, Namazi-Rad, Perez, Berryman, and Chen (2013)</td>
<td>Sydney, Australia</td>
<td>CO</td>
<td>n.a.</td>
<td>5.00</td>
</tr>
<tr>
<td>Moekel, Spiekermann, and Wegener (2003)</td>
<td>Netanya, Israel</td>
<td>IPF with MCMC</td>
<td>6</td>
<td>2.60</td>
</tr>
<tr>
<td>Müllner and Axhausen (2011)</td>
<td>Switzerland</td>
<td>IPF</td>
<td>5</td>
<td>8.00</td>
</tr>
<tr>
<td>Pritchard and Miller (2012)</td>
<td>Canada</td>
<td>IPF with MCMC</td>
<td>2</td>
<td>3.42</td>
</tr>
<tr>
<td>Saadi, Mustafa, Teller, and Cools (2016)</td>
<td>Brussels</td>
<td>MCMC, IPF</td>
<td>0.1</td>
<td>1.20</td>
</tr>
<tr>
<td>Sun and Erath (2015)</td>
<td>Singapore</td>
<td>Comparing BN, MCMC, IPF</td>
<td>1(^b)</td>
<td>4.00</td>
</tr>
<tr>
<td>Ye et al. (2009)</td>
<td>Arizona, USA</td>
<td>IPU</td>
<td>8</td>
<td>3.07</td>
</tr>
<tr>
<td>Zhang, Cao, Feygin, Dounan, and Pozdnozhkov (2017)</td>
<td>San Francisco, USA</td>
<td>BN</td>
<td>6</td>
<td>7.00</td>
</tr>
<tr>
<td>Zhu and Ferreira (2014)</td>
<td>Singapore</td>
<td>IPF</td>
<td>1</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Note: \(^a\)For abbreviations, see the text.
\(^b\)Tests from 1% to 100% of sample size.
vertex \((V)\), and the correlation is represented by the directed edge or arc \(A\). In the example shown in Figure 1, there are variables including income, age and gender. The directed edge from \(NodeIncome\) to \(NodeAge\) and \(NodeIncome\) to \(NodeGender\) indicates that \(NodeAge\) and \(NodeGender\) are linked by a conditional probability with \(NodeIncome\). Therefore, the conditional probability distribution of this condition are \(P[NodeAge|NodeIncome]\) and \(P[NodeGender|NodeIncome]\).

**A learning algorithm for BNs**

There are different algorithms for learning BNs, as explained by Scutari (2010), such as the constraint-based algorithm, the score-based algorithm or the hybrid algorithm. Each type includes a learning algorithm. Here we used the R package \texttt{bnlearn} (Scutari, 2010), which implements Tabu Search as part of the score-based algorithm. Tabu Search, as a generic heuristic procedure, is an iterative searching procedure used to obtain the best solution from complex correlation patterns (Glover & Taillard, 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; a blacklist means that those arcs will not present be in the network structure; a whitelist is the reverse.

**Network scores**

The step of measuring the candidate of the graphical structure that fits to the data is important to ensure the structure can produce a fit synthetic population. In this process, several methods were introduced, such as maximum likelihood:

\[
\mathcal{L}(G^h|D) = \sup_{\Theta} \mathcal{L}(G^h, \Theta|D) = \max_{G^h} \mathcal{L}(G^h, \hat{\Theta}|D) \tag{1}
\]

where \(\mathcal{L}(G^h|D) = \sup_{\Theta} \mathcal{L}(G^h, \Theta|D)\) is the log-likelihood of a provided pair \((G, \Theta)\) given observation \(D\). However, the log-likelihood is not re-presentable, as explained by Sun and Erath (2015), due to overfitting, since this method will always build a fully connected DAG. Therefore, most applicable approaches are using the Bayesian information criterion (BIC) (Rissanen, 1978; Schwarz, 1978) and Akaike information criterion (AIC) (Rissanen, 1978; Akaike, 1974):

\[
\text{BIC}(G^h|D) = \log P(D|G^h, \hat{\Theta}) - \frac{d}{2} \log m \tag{2}
\]

\[
\text{AIC}(G^h|D) = \log P(G^h, \Theta|D) - d \tag{3}
\]
where $\Theta$, in the first equation, is the maximum likelihood estimate parameter given a hypothetical structure $G^0$; $d$ is the degree of freedoms in $\Theta$; and $m$ is the number of observations. In contrast to maximum likelihood (1), BIC (2) and AIC (3) have a penalty function for the optimal likelihood $\log P(G^0, \Theta | D)$. For BIC the penalty is $\frac{d}{2} \log m$; and for AIC is $d$. Using the scoring function, the best network is selected for constructing the synthetic population.

**Constructing the population of the Greater Jakarta Area**

The study area is the Greater Jakarta Area, or Jabodetabek, which consists of Jakarta province, parts of West Java province and Banten province. It has 31.7 million inhabitants (BPS-Statistics, 2016a, 2016b, 2016c) (Table 2). The population data used for the synthetic population generation were partly obtained from the JAPTRAPIS study (Jabodetabek Public Transport Policy Implementation and Strategy) in 2009 (JICA, 2009; Ilahi & Axhausen, 2017).

There are two different types of data in the JAPTRAPIS study: the Household Travel Survey (HTS) and the Activity Diary Survey (ADS). However, we only use HTS data consisting of 178,954 households, or 334,973 individuals, for the population synthesis, which are equal to 3% of all households. In the HTS, the respondents are individuals who are going to school or the office. Therefore, the aggregate synthetic population is limited to the census population of individuals who have activities for studying or working with a total of > 20 million (Table 3).

The data used for this approach are from multiple sources within the HTS data. The variables of the individuals include age, gender, education and employment status. The variables of the households are income, housing status, vehicle ownership and address. In the activities-based model, we also need geocoding of activity locations. Therefore, we geocode first based on Google Application Programming Interface (API). Therefore, there are three sources of data in the model. We use the R package dplyr for data joint. Figure 2 shows the framework model of data combination and integration between BN and GR. Further detail of the BN is presented in Figure 4.

**MODEL ESTIMATION**

**BN step**

We consider seven variables by combining individual and household data for the population synthesis using BN, as presented in Table 4: type of activities, age, sex, income, housing, car

### Table 2. Population of the Greater Jakarta Area.

<table>
<thead>
<tr>
<th>Province</th>
<th>Region</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jakarta</td>
<td>South Jakarta</td>
<td>1,096,469</td>
<td>1,089,242</td>
</tr>
<tr>
<td></td>
<td>East Jakarta</td>
<td>1,436,128</td>
<td>1,407,688</td>
</tr>
<tr>
<td></td>
<td>Central Jakarta</td>
<td>457,025</td>
<td>457,157</td>
</tr>
<tr>
<td></td>
<td>West Jakarta</td>
<td>1,246,288</td>
<td>1,217,272</td>
</tr>
<tr>
<td></td>
<td>North Jakarta</td>
<td>867,727</td>
<td>879,588</td>
</tr>
<tr>
<td>Banten</td>
<td>Tangerang City</td>
<td>1,045,113</td>
<td>1,001,992</td>
</tr>
<tr>
<td></td>
<td>Tangerang Regency</td>
<td>1,724,915</td>
<td>1,645,679</td>
</tr>
<tr>
<td></td>
<td>South Tangerang City</td>
<td>777,713</td>
<td>765,496</td>
</tr>
<tr>
<td>West Java</td>
<td>Depok</td>
<td>1,061,900</td>
<td>1,044,200</td>
</tr>
<tr>
<td></td>
<td>Bogor</td>
<td>532,000</td>
<td>515,900</td>
</tr>
<tr>
<td></td>
<td>Bogor Regency</td>
<td>2,792,900</td>
<td>2,666,800</td>
</tr>
<tr>
<td></td>
<td>Bekasi City</td>
<td>1,369,600</td>
<td>1,345,200</td>
</tr>
<tr>
<td></td>
<td>Bekasi Regency</td>
<td>1,654,600</td>
<td>1,591,400</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>31,689,992</td>
<td></td>
</tr>
</tbody>
</table>
ownership and driving licence. These variables are considered as the main sociodemographic variables to reduce the complexity of the network structure. However, for the variables income, housing status and car ownership, the household data are used. The chosen network structure is based on the AIC score employing the Tabu Search algorithm to learn the network structure of the BN, as implemented in the R package bnlearn (Scutari, 2010; R Core Team and others, 2013).

There are two steps in the Tabu Search in this scenario; without using the whitelist and blacklist G structures for the initial search and with using the whitelist and blacklist for the final search, where there are 256 search in the initial search and 64 search for the final search.

Table 3. Population of the Greater Jakarta Area with the relevant activities in the census.

<table>
<thead>
<tr>
<th>Province</th>
<th>Region</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jakarta</td>
<td>South Jakarta</td>
<td>877,679</td>
<td>626,495</td>
</tr>
<tr>
<td></td>
<td>East Jakarta</td>
<td>1,141,943</td>
<td>815,130</td>
</tr>
<tr>
<td></td>
<td>Central Jakarta</td>
<td>367,092</td>
<td>262,034</td>
</tr>
<tr>
<td></td>
<td>West Jakarta</td>
<td>989,250</td>
<td>706,136</td>
</tr>
<tr>
<td></td>
<td>North Jakarta</td>
<td>701,639</td>
<td>500,837</td>
</tr>
<tr>
<td>Banten</td>
<td>Tangerang City</td>
<td>999,131</td>
<td>590,688</td>
</tr>
<tr>
<td></td>
<td>Tangerang Regency</td>
<td>1,645,087</td>
<td>972,578</td>
</tr>
<tr>
<td></td>
<td>South Tangerang City</td>
<td>753,194</td>
<td>445,290</td>
</tr>
<tr>
<td>West Java</td>
<td>Depok</td>
<td>711,791</td>
<td>394,497</td>
</tr>
<tr>
<td></td>
<td>Bogor</td>
<td>354,155</td>
<td>196,284</td>
</tr>
<tr>
<td></td>
<td>Bogor Regency</td>
<td>1,845,195</td>
<td>1,022,666</td>
</tr>
<tr>
<td></td>
<td>Bekasi City</td>
<td>917,511</td>
<td>508,514</td>
</tr>
<tr>
<td></td>
<td>Bekasi Regency</td>
<td>1,097,039</td>
<td>608,014</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>20,049,867</td>
<td>12,394,887</td>
</tr>
</tbody>
</table>

Figure 2. Framework model.
structure is obtained by an iterative process after measuring the error for 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 an AIC of $-1,679,334$ for the selected BN.

The model structure and the conditional probabilities of the variables can be seen in Figure 3. The tables accompanying Figure 3 presents the conditional probabilities. In this case, we show variables sex and income. There is a directed arc of NodeTA and NodeAge that goes to NodeSex. Therefore, the conditional probability of an income between five and eight and type of activity (TA) school for females is 0.43 and for males is 0.57. This way of interpretation is the same for other variables, and the total of conditional probability is always equal to 1. After we identified the best structure, we generated the data as close as possible to the aggregate census data. In this research, the data are generated for 4, 8, 16 and 22 million agents.

We randomly joined the generated BN population to the remaining 20 variables left in the HTS data, based on the available seven variables in both data sets. Complete sets after both data were joined are used for the agent-based model. We obtain a joint data set consisting of 27 variables (Figure 4). For the extended target of 16 million agents, this joint operation takes three days and 13 h. Moreover, for 22 million agents, this joint operation takes more than five days on the servers of ETH Zürich computer cluster (ETH Zürich, 2016).

**Figure 3.** Final model structure G.

**Table 4.** Population of the Greater Jakarta Area.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition (number of categories)</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA</td>
<td>Type of activities of individual (2)</td>
<td>School; work</td>
</tr>
<tr>
<td>Age</td>
<td>Age of the individual (8)</td>
<td>0–5; 6–12; 13–18; 19–24; 25–32; 33–42; 43–60; 61/+</td>
</tr>
<tr>
<td>Sex</td>
<td>Gender of the individual (2)</td>
<td>Male; female</td>
</tr>
<tr>
<td>Inc</td>
<td>Income of the household (IDR, millions) (7)</td>
<td>n.a.; 0–0.9; 1–3; 4–5; 6–8; 9–15; 16/+</td>
</tr>
<tr>
<td>HS</td>
<td>Housing status of the household (2)</td>
<td>Owned; rented</td>
</tr>
<tr>
<td>CO</td>
<td>Car ownership of the household (2)</td>
<td>No; yes</td>
</tr>
<tr>
<td>Licen</td>
<td>Licence of the individual (4)</td>
<td>Motorcycle; car; motorcycle and car; no licence</td>
</tr>
</tbody>
</table>
To visualize the goodness of fit after the BN approach and the joint data, we compared HTS data and the final structure from BN. However, in this case, we only visualize two different type of joint distributions. Figure 5 shows the joint distribution of income and age. Figure 6 shows the joint distribution of income and region. The left figures show the joint distribution data from HTS, the middle figures show the joint distribution of the data from BN, and the right figures show the fit of joint distribution of both HTS and BN. We considered the variable region, which was not used as a variable in the BN approach in Figure 6 to ensure that a joint data procedure gives a good fit with unselected variables. We found a similar distribution after both HTS and BN data were joined.

Furthermore, as also found in Sun and Erath (2015) and Zhang et al. (2017) that the BN model can reproduce the distribution of the HTS data. This is because we can only assign to this target value, whether we estimate 4 million or 22 million, and it is not based on the control total of aggregate census data in Table 3. Therefore, to solve this issue, we employ multilevel IPF with the R-package MultilevelIPF (Müller, 2017) to adjust the artificial sample to the aggregate census data (Table 3).

**GR multilevel IPF for fitting against census data**

Several multilevel IPF algorithms have been implemented such as HIPF, IPU, entropy optimization (Ent) and GR. Nevertheless, we use a GR algorithm that has been shown to outperform the other algorithms, which generates weights with minimal squared error and is faster. It matches the
data with the aggregate control groups by estimating a weight for each observation in the sample (Müller, 2017). There are two group controls used: region and gender, and region and age. Furthermore, the weakness of IPF is its non-integer weights (Müller, 2017; Lovelace & Ballas, 2013), while it requires integer weights for generating the target population. Integerization is the process to convert the decimal weights (related with how many times each agent is replicated) to integer values. Therefore, integerization is an important step to produce the best fit to the marginal census data. This can be done using weighted random sampling without replacement using the wrswoR package (Müller, 2016), and also using the truncate, replicate, sample (TRS) method (Lovelace & Ballas, 2013). However, we employ random sampling without replacement, which consists of three steps. The first is removing the decimal values. The second is calculating the decimal remainders that will be used as a vector of probability weights. The third is implementing weighted sampling without replacement. A crank algorithm is used in this operation for faster results (Müller, 2016).

The data used from BN for fitting using GR differ in size from 4, 8, 16 and 22 million agents. Less fit result was acquired when we use data from 8, 16 and 22 million. It can be because of the huge size and complex data structure that creates errors in implementation. Moreover, integerization has also created an error in replication. Therefore, we used the 4 million agent data set that fitted with < 0.2% data difference compared with the target census data. The synthetic population data fits after conducting GR multilevel IPF to the target of 20 million marginal census data and the control totals in Table 3.
which are based on age and gender for each region. The results are shown in Table 5 and Figure 7.

**DISCUSSION**

Population synthesis is an important step when we develop an agent-based model. Data become an important factor to ensure the population synthesis gives a precise description of
the agents. There are several methods available. Some tried to improve the previous methods; others tried to use new methods, as mentioned above. Each region has different difficulties, since each country has different regulation regarding data issues and data availability. Less developed countries may have less accessibility and quality of the data. One of the studies generated a synthetic population without a sample to address the availability of sample issue (Barthelemy & Toint, 2013). However, it becomes questionable at which level of disaggregation this can give quasi-perfect distribution. On the other hand, it would be a solution for the initial development of an agent-based model.

The size of the data set also influences the processing. As mentioned, it takes more than a day to join the different data sets. It will be an interesting topic to accelerate the joining process for huge data sets. Moreover, a BN alone cannot fit the marginal census data. However, using GR only also fails to make sure that the dependency of each variable is perfectly reproduced. On the other hand, the combination of BN and GR is a solution to maintain the stability of the dependencies and to ensure a quasi-perfect distribution. Saadi et al. (2018) also combined HMM and IPF in the same spirit.

CONCLUSIONS

In the present case, we found that the BN can construct a synthetic population and reproduce the HTS data. The result after data joining between population synthesis and HTS data gives well-fitting distributions. However, BN has some differences vis-à-vis marginal data from the census. Therefore, we need to fit against the marginal distribution of the census data.

The differences are addressed with GR. We made an effort to remain as close as possible to the target census data when we fit with GR. We started from 4, 8, 16 and 22 million agents produced by the BN approach. However, it gives different distribution against census data when we tried with 8, 16 and 22 million. Therefore, we used 4 million agents from the BN to fit with the 20 million target agents of the marginal census data/control total using GR, which is equal to 20% of final data. Based on the results described above, we can conclude the following:

- The results confirm that the BN approach can be used to produce large samples with well-fitting distributions, which is useful for any researcher who has a limited sample to start with.
- The result from GR can fit to the control totals or marginal census data.
- Combing BN and GR can help researchers produce data that fit to control totals or marginal census data.

This synthetic population will be used for our further research to develop an agent-based model using multi-agent transport simulation (MATSim) (Horni, Nagel, & Axhausen, 2016). This is the first scenario of an agent-based model for Greater Jakarta (Ilahi, Balac, Li, & Axhausen, 2019). Several variables from the synthetic population will be used in the agent-based modelling, such as age, gender, income, coordinates of both home and office, activities, licence, and car ownership. Furthermore, we will integrate secondary activities in our population synthesis, since no data have been available so far. However, the steps will be similar with what we have explained in this paper. In addition, another method to address joint time use should also be investigated in future work.
ACKNOWLEDGEMENTS

Special thanks also to the Japan International Cooperation Agency (JICA) for allowing the authors to use the survey data in the study. The authors are grateful to the anonymous reviewers whose comments helped to improve the paper. The remaining errors are the authors’ own.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

FUNDING

The authors acknowledge the funding from the Lembaga Pembiayaan Dana Pendidikan (LPDP), Indonesia [grant number 2016042201598].

ORCID

Anugrah Ilahi @ http://orcid.org/0000-0001-7387-4516
Kay W. Axhausen @ http://orcid.org/0000-0003-3331-1318

REFERENCES


