Synthetic Population of Greater Jakarta: An Iterative Proportional Updating Approach

Grace O. Kagho, ETH Zurich
Anugrah Ilahi, ETH Zurich
Milos Balac, ETH Zurich
Kay W. Axhausen, ETH Zurich

Conference Paper STRC 2020

STRC | 20th Swiss Transport Research Conference
Monte Verità / Ascona, May 13-15, 2020

Synthetic Population of Greater Jakarta: An Iterative Proportional Updating Approach

Grace O. Kagho
IVT ETH Zürich
CH-8093 Zürich

Anugrah Ilahi
IVT ETH Zürich
CH-8093 Zürich

Milos Balac
IVT ETH Zürich
CH-8093 Zürich

Kay W. Axhausen
IVT ETH Zürich
CH – 8093 Zürich

grace.kagho@ivt.baug.ethz.ch

May 2020

Abstract

Agent based simulation frameworks require synthetic populations to simulate traveller's activity travel patterns. Various methods exist for population synthesis with different advantages and disadvantages. For the Greater Jakarta area, a synthetic population has previously been generated using Bayesian Network (BN) approach and Generalized Raking (GR) multilevel IPF. This approach used very large, city level marginal to fit the population and only captured individual control totals. Here we generate a synthetic population using Iterative Proportional Updating approach to match both individual and household level controls of the population of smallest available census zones, called subdistricts, in Greater Jakarta region, using census-based marginal distributions and existing sample surveys.

Keywords

Population Synthesis, Iterative Proportional Updating, Agent-Based Models

1. Introduction

In agent-based modelling of travel behaviour, where individual travellers, the agents, are simulated as the smallest behavioural unit, one of the core requirement is detailed socio-demographic information about the agents. Therefore, one needs household and person attribute information of the population being modelled. Households, because individuals live together and may share various attributes such as household income, household structure, car ownership, etc. Person attributes, such as age, gender, work status, or location of work, etc. as they affect their activity patterns and mobility choices.

Acquiring such information for a whole population at a disaggregated level for any region can be difficult and usually not possible due to privacy and cost constraints. It is necessary then to conduct surveys for a random sample of the population of the region that would capture detailed disaggregate information of persons and their households' attributes. An example of such survey is the traditional household travel survey (HTS) that surveys the mobility patterns of a random sample of a region's population. These sample data could be used to generate a complete synthetic population by matching the joint distribution of attributes in the sample data to distribution of known aggregates, the marginal distribution obtained from census data. This process is known as Population Synthesis.

Several approaches for generating synthetic population have been developed by many researchers over time, with detailed reviews to be found in Mueller (2010) and Sun and Erath (2015). Some of these include Iterative Proportional Fitting (IPF) (Beckman et al. 1996), Combinatorial Optimization (Voas and Williamson, 2001), Bayesian Networks (Sun and Erath, 2015), Markov Chain Monte Carlo (Farooq et at., 2013), and Iterative Proportional Updating (IPU) (Ye et al., 2009), among others. IPF being the first widely used method was developed by Deming and Stephan (1940) in the field of mathematics and was adapted to transport research by Beckman et al. (1996). It has been improved over time by some of the listed methods above.

In this paper, we apply the IPU method to generate a synthetic population for a large-scale urban area consisting of over 30 million inhabitants, the Greater Jakarta area in Indonesia. Greater Jakarta, also known as Jabodetek, is divided into three provinces comprised of 13 regions, which include Bogor city, Depok city, Bekasi city, Bogor regency, Bekasi Regency, Tangerang city, Tangerang regency, South Tangerang city, and North, South, West, Central and East Jakarta. Previously Ilahi and Axhausen (2019) had generated a synthetic population for the Greater Jakarta Area by combining the two approaches, Bayesian Network (BN) and generalized raking multilevel IPF to create a population for those carrying out activities in the area. In their approach, the BN is used to reproduce the distribution of an HTS data available, expanding the sample survey population of about 300,000 to a population of 22 million. The distribution of the HTS is not necessarily matching the distribution of the population. Hence, Generalized Raking was used to fit the generated synthetic population to the aggregate census

of the region. This method proved to be useful for creating the population based on large areas at the city level, and also given the limitation of data and the difficulties to get reliable data in the Greater Jakarta Area.

Their approach, while successfully matching the synthetic population with the census distribution of the regions, however, has its shortcomings. Firstly, the fitting was done only to match the distribution for person attributes, ignoring the distribution of the households of the travellers. Secondly, the fitting procedure did not take cognizance of the smallest geographic area, which means, while the distributions may have matched for the larger regions, this might not have been the case for smaller zonal level. Thus, this study extends their work, using the IPU method to create synthetic population matching both the person and household attributes for smaller geographical zones of the Greater Jakarta Area.

The rest of the paper is organised as follows: a brief description is given of the IPU method, followed by an outline of the data used, results and the final section concludes with a discussion.

2. Iterative Proportional Updating

The IPU proposed by Ye et al. (2009) is selected for this study because of its advantage over the IPF method. IPU controls for multiple hierarchy levels i.e. person-level and household-level attributes at the same time during fitting, in a computationally efficient manner, using an algorithm that iteratively adjusts and reallocates weights of households, until both the household and person level attributes are matched.

This is different from the IPF method where only one attribute level is matched, either the household or the person attributes, resulting in different weights. For example, Beckman et al. (1996) first applied the IPF to generate synthetic households by matching joint distribution of household attributes to their marginal distribution in the census data. And then generated a synthetic population made up of persons from the randomly drawn households based on the estimated joint distributions. However, the IPF method creates an issue where the distribution of person attributes of interest in the synthetic population is not necessarily matched with the marginal distribution of person level attributes in the census data, as the person weights are forced to be equal to the corresponding household weights, even though they are different. The IPU not only addresses this issue, but also provides a practical approach to generate populations for small geographical zones, with better computational efficiency.

IPU extends the IPF method by adjusting the household weights based on the person weights obtained from an IPF procedure. This is done iteratively using a frequency matrix where a row in the matrix represents a single household record containing the composition of the household, the household attributes e.g. household size and the person attributes e.g. age, gender. There is an additional column of the household weight initially set to the value of one. A weighted sum is calculated by summing each column weighted by the weight column. The weights are

adjusted iteratively with the difference between the weighted sum and the marginal distribution, used for setting a convergence criterion for fitting the marginal distribution for both households and person attributes. The closer to zero, the better the fit, and the number of iterations to determine an optimal point is left to an analyst who is to observe and monitor the performance of simulation. A clearer example of how this process works, along with the mathematical details, are illustrated in Ye et al. (2009).

The IPU method also solves the zero-cell problem and the zero-marginal problem that occurs when using IPF for small geographical zones. These two problems can arise for small geographical zones where a sample survey or aggregate census data for a zone could result in a zero value for certain attributes and their dimensions. For example, there might be no person of age 65 years and above in a particular small geographical zone in the census data hence, a zero-marginal problem occurs. Or due to the small number of people in that zone, the sample survey did not capture anyone; hence, a zero-cell problem occurs. When weights are then computed to match the zero constraints, the denominator will take on a zero value when adjusting the weights making the algorithm fail. To overcome the zero-marginal problem, a small positive value such as 0.001 can be assigned to all zero-marginal attributes. For the zero-cell problem, the constraint for a larger region can be applied for attributes with zero cells.

In this study, to avoid the zero cell problem, we use the joint distributions of the sample data for the entire Greater Jakarta rather than for each small geographical zone where the zero-cell problem could occur. Also, attributes such as age that could have given a zero marginal problem were grouped into wider age intervals.

3. Data description and preparation

The data required are from two sources, an HTS sample data, and census aggregates. The HTS data used is from the JAPTRAPIS study conducted in 2012 by the Japan International corporation Agency (JICA). It consists of 657,165 individuals and 178,954 households, approximately 3% of the household population of Greater Jakarta. The survey collected socio-economic information of households, and workers and students in the various households were further interviewed. The variables of the household attribute information collected include income, housing status, vehicle ownership and location. The individual attributes include age, gender, education and employment status.

The census dataset is collected from census publications of different districts of the Greater Jakarta region spanning across three years, 2016 to 2018, to make up for missing data. The census data represents information about population gender, age group and household numbers in different subdistricts, the smallest geographical zones. There were 1,336 subdistricts with an average of 10,000 people per zone. For the population synthesis, age and gender were the person level attributes, and whether a person belonged to a household, as the household level attribute accounting for the household number variable in the census data. Other attributes available in

the survey data, such as income and car ownership, were not available in the census data, and were later matched based on the fitting from the available attributes used in generating the population synthesis. For the age attribute, the census data was presented in 16 age groups, however due to missing data and possibility of zero marginal occurring, the age group was collated into seven bins (see Table 2) and the HTS data was grouped accordingly.

Treatment of missing data

The table below provides a summary of data issues encountered in the data preparation process. There were cases of missing values, wrongly imputed values, and mismatched totals. For subdistricts with missing information of age and gender attributes, the distribution of the larger region that the subdistrict belonged to has been used. With the subdistrict at the smallest zone, the next larger region is the district level and when not available, the city level distribution is used and so on. For cases of wrongly imputed values, such as where the total number of households in a subdistrict, is larger than the population or is so low that the average household size is above 10, neighbouring zones within same district, as well as previous years, were checked to verify the household sizes. When the information is not found in previous years, the district average household size is used to compute the household number for the zone. Furthermore, in the cases where there were mismatches in census data between total of gender population and total of age population, the gender sum was used to fit the age total based on the existing age distribution.

Table 1: Summary of Missing Data

Number of subdistricts	Age	Gender	Household
407	√	√	√
900	X	√	√
29	X	√	X

4. Population Synthesis

For applying the IPU, the PopGen 2.0, Synthetic Population Generator developed by Mobility Analytics Research Group (MARG, 2016) was used. PopGen was developed as an open-source heuristic algorithm for iteratively generating a synthetic population that is representative of the actual population. It provides both IPU and entropy-based method for generating a synthetic population. Using PopGen, one can control for various attributes at different geographical resolution simultaneously. The PopGen requires the use of Python27 and the use of the

Command Line Interface for interaction. The data is prepared in a format that maps the different spatial resolutions. For example in this study, the spatial resolution of the sample data obtained from the JAPTRAPIS study is that of the whole Greater Jakarta region, an upper spatial resolution, while the census data is at the smallest geographical unit, the subdistricts, a lower spatial resolution referred to as the geo level in PopGen. The mapping is one-to-many, whereby households are drawn from the sample data, - an upper resolution, and then assigned to the geographical resolution observed in the census data for the synthetic population. The total runtime for generating a synthetic population of approximately 30 million people took 27 hours for a 1000 iteration run of the simulation.

5. Results

Fig 1 compares the total population of the census population and the synthetic population for each subdistrict. It produces a near perfect fit between census population and synthesized population with a near perfect slope and R2 of 0.999. The IPU is only slightly under predicting the distribution, with the synthetic population having a slightly lower total population. This can been seen in Fig 3 and Fig 4 showing a comparison of the total population by person attributes between the census population and the synthetic population. For the households (Fig 2), it is practically perfect with slope of 1 and R2 value of 0.999.

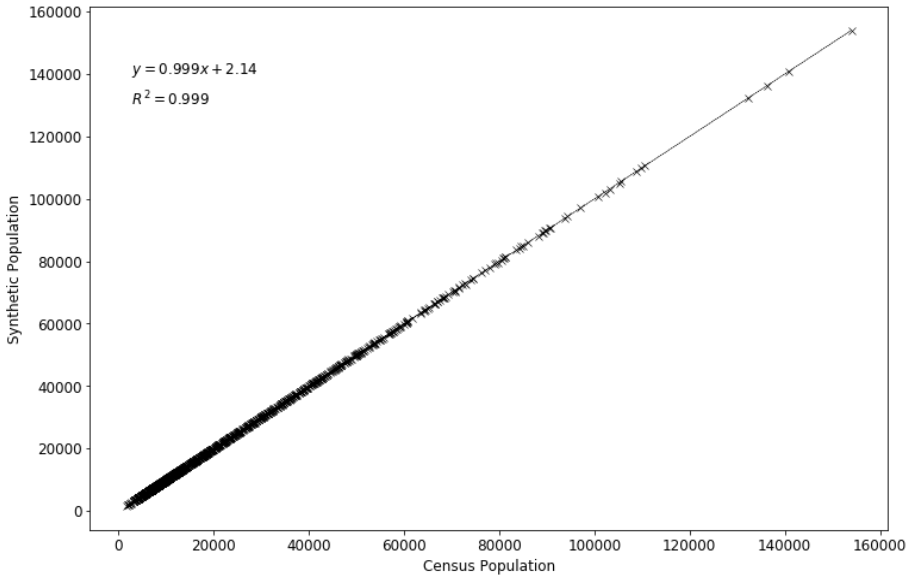

Fig 1: Fit between census population and synthetic population (Total Population)

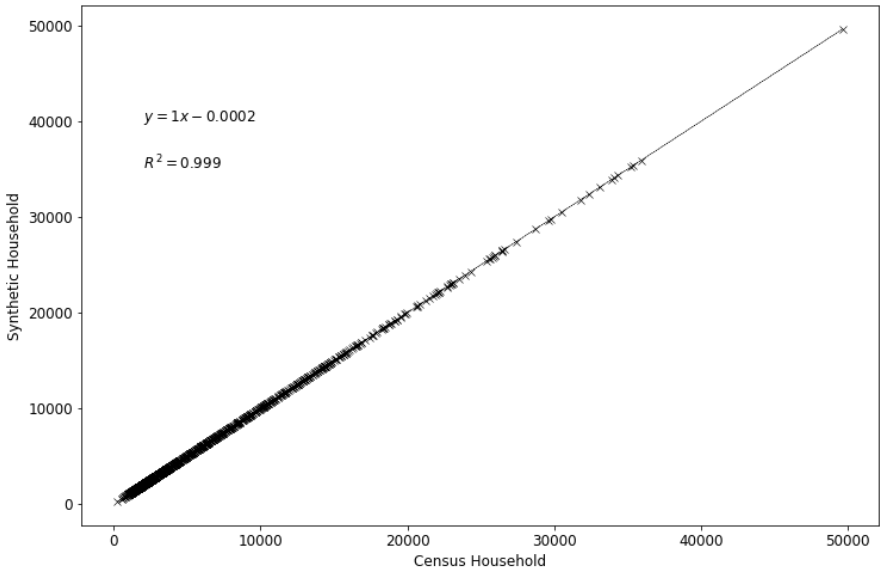

Fig 2: Fit between census population and synthetic population (Households)

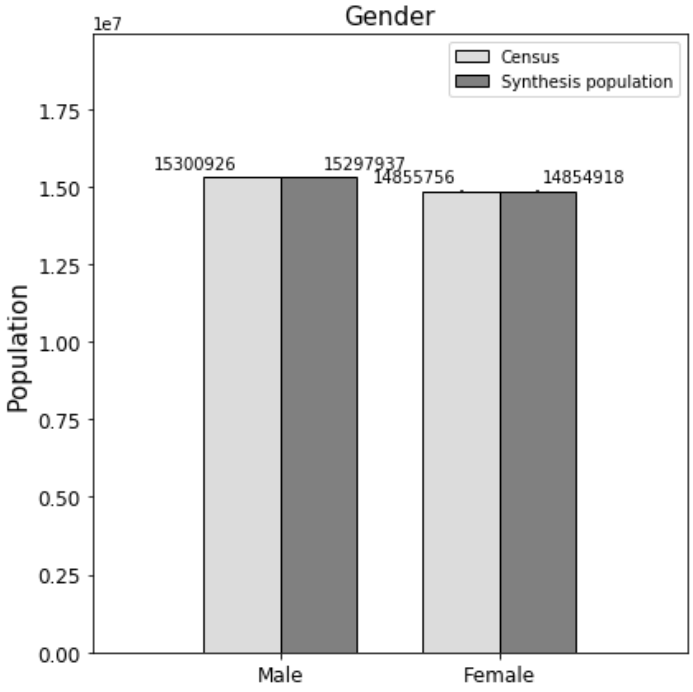

Fig 3: comparing census population and synthetic population for gender

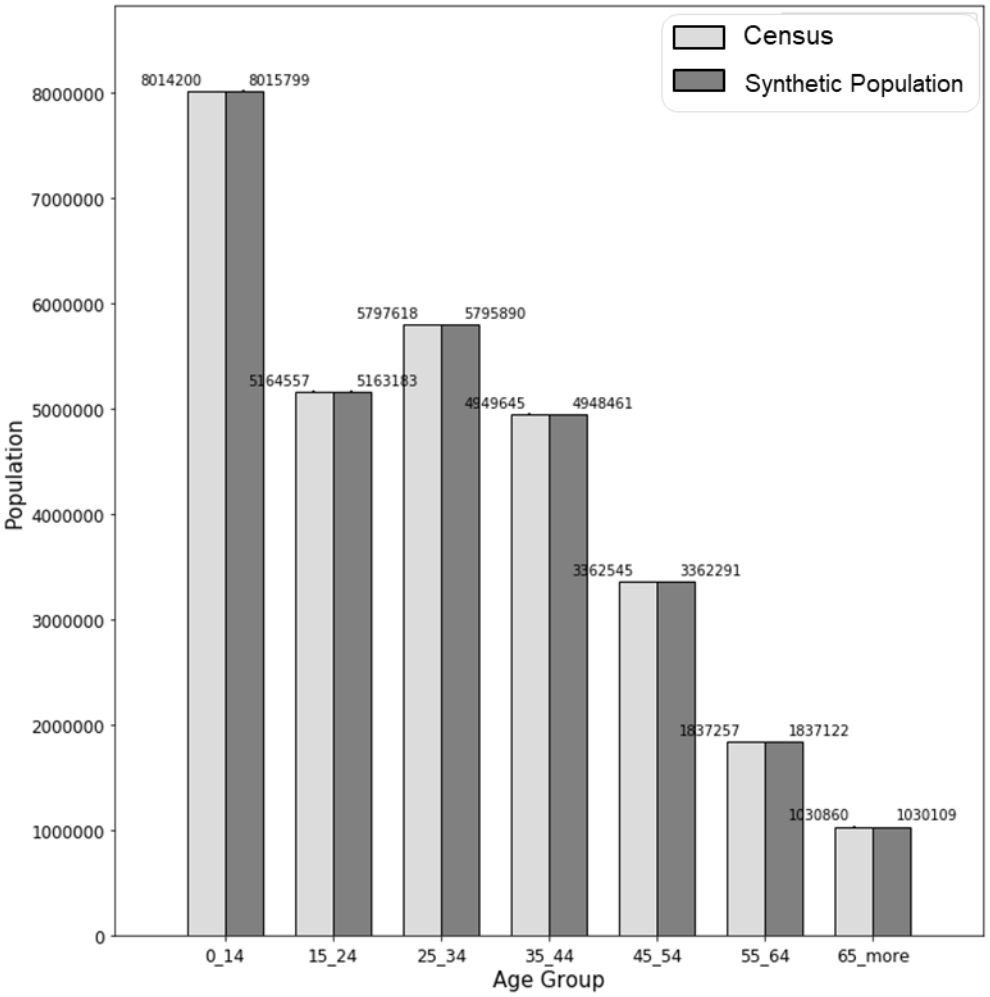

Fig 4: comparing census population and synthetic population for different age groups

The synthetic population has approximately 3827 (0.013%) people less than the census. If we compare the joint distribution at each geographical unit using histogram of the errors, to show the relative difference between the census population and synthetic population, we would see the variations better. Fig 5 - 8 show the distribution of the relative differences and a zoom in on the number of subdistricts with more than 1% in difference.

For the total population, the relative differences are less than 2% for 62 subdistricts (4.7% of total subdistricts) having an error greater than 1%.

The error difference between census gender totals and the synthetic population gender totals, are generally less than 3% with only one subdistrict having a difference greater than 3% for female totals. 147 subdistricts (11% of total subdistricts) have difference greater than 1% for male, and 174 subdistricts (13% of total subdistricts) for females.

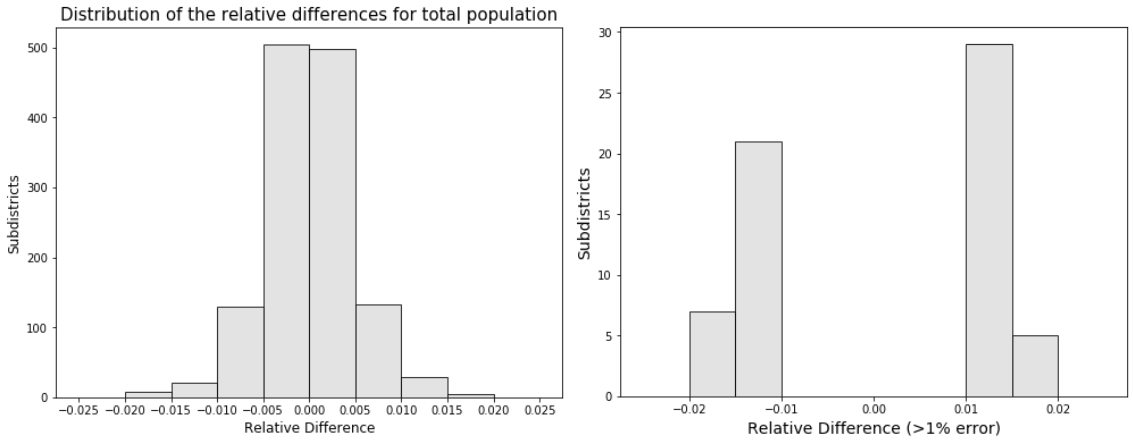

Fig 5: Distribution of the relative differences for (a) total population and (b) with relative difference greater than 1%

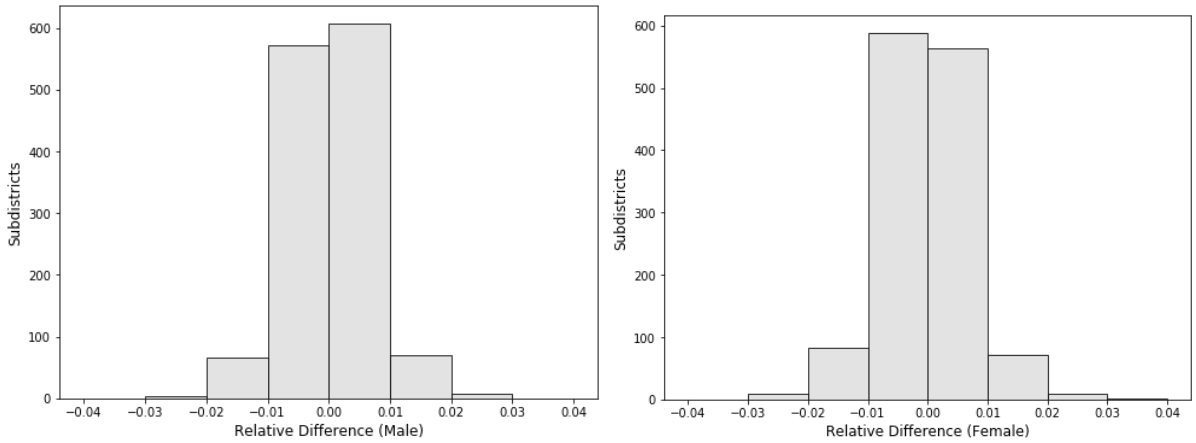

Fig 6: Distribution of the relative differences for (a) male population and (b) female population

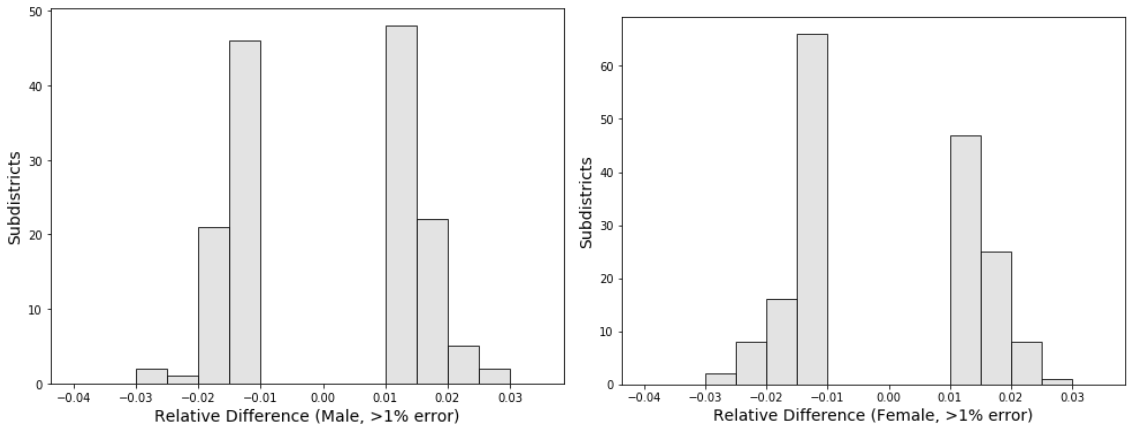

Fig 7: Distribution of the relative differences greater than 1% for (a) male population and (b) female population

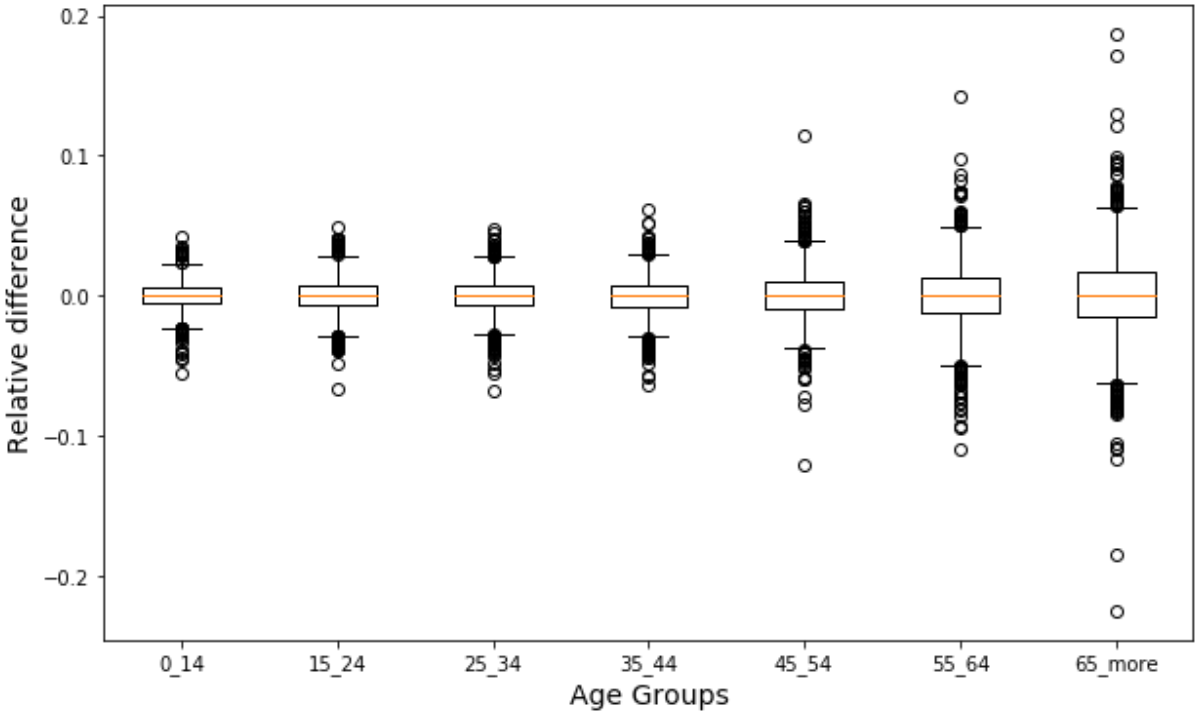

Fig 8: Distribution of the relative differences for population based on age groups

A boxplot is used to depict the relative difference between census population and the synthetic population per age group for different subdistricts. The age groups have more subdistricts with more than 3% relative difference between census and synthetic population. This difference varies per age group, for example, there are more subdistricts with people aged 55 and above, not well matched in the synthetic population. This may be due to the samples in the age classes

not being representative enough for certain locations and hence a poor fit. Table 2 below gives a summary of the number of subdistricts having an error above 3% and 10%.

Table 2: Number of subdistricts with high error margin for different age groups

Age Groups	Subdistricts with Error Above 10%	Subdistricts with Error Above 3%	Max Error (%)
0-14	0	17	4.3%
15-24	0	47	4.9%
25-34	0	46	4.8%
35-44	0	54	6.2%
45-54	2	104	11.5%
55-64	2	213	14.3%
65+	10	338	18.8%

6. Discussion and Conclusion

Here we presented the use of IPU approach for synthetic population generation for the Greater Jakarta Area. This approach allowed the matching of both household level and person attributes of the synthetic population to the census population. The results highlighted above has shown that the IPU simulation was able to reproduce the marginal distribution with minimal error. However, this error varies across the person attributes for different geographical zones, with up to 18% error for the age group of 65 and above.

Presence of multi-dimensional attributes such as the age attribute has been known to prevent a perfect fit of both the household and person attributes as it becomes more difficult for the IPU algorithm to converge (Ye et al. 2009). Additional reduction of the number of age classes could reduce the accuracy of the population and may even lead to worse fits as the amount of information required to select the households would have been further reduced. Since less than 1% of the subdistricts have relative difference above 10%, the quality of the generated population is acceptable for further use in agent based microsimulation of travel behaviour using Multi-Agent Transport Simulation (MATSim) (Horni et al., 2016).

The next step of this study was to include more household attributes needed to be able to generate a richer input for the agent-based model. Due to the limited availability of census data with household attributes, only one household attribute, the household size has been used to fit the synthetic population to the census population for each subdistrict. The other household attributes, which include income, car ownership, and motorcycle ownership, have been matched based on the characteristics of the drawn synthetic population from the sample data. Similarly, social status (whether a person is working, schooling or retired) has been added as person attribute to the synthetic population.

Before concluding, there is a need to consider that the cleaning and transformation of the input data, used to generate the synthetic population, can affect the simulation and possibly lead to uncertainty in the model output (Kagho et al. 2020). As previously mentioned, the census data used in this study was not readily available. The information was extracted from tables of published reports in PDF format. Obtaining this information was a cumbersome process. Besides having to extract the tables manually, the formatting of the figures was different across different districts and regions, i.e. a thousand separator varied between a coma, space and point. This introduced the possibility of making mistakes and required going over the data many times to ensure it was in order. Furthermore, the reports had missing values, wrong data and data from different years for different attributes. These input data issues have been treated using ad hoc means as highlighted above and will be taken into consideration when analysing the model output.

Acknowledgement

The authors wish to acknowledge JICA for allowing the use of the JAPATRIS survey data in this study. The authors also would like to acknowledge Airbus Urban Mobility GmbH for partially funding this research.

7. References

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice,* 30(6), 415–429. doi: 10.1016/0965-8564(96)00004-3

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics,* 11(4), 427-444.

Farooq, B., Bierlaire, M., Hurtubia, R., & Flotterod, G. (2013). Simulation based population synthesis. Transportation Research Part B: Methodological, 58, 243–263. doi: 10.1016/j.trb.2013.09.012

MARG (2016) PopGen: Synthetic Population Generator [online]. Mobility Analytics Research Group. Available at: http://www.mobilityanalytics.org/popgen.html

Ilahi, A., & Axhausen, K. W. (2019). Integrating Bayesian network and generalized raking for population synthesis in Greater Jakarta. Regional Studies, Regional Science, 6(1), 623-636.

JICA. (2009). Traffic data collected under the Jabodetabek urban transport policy integration. Tokyo: JICA.

Kagho, G. O., Balac, M., & Axhausen, K. W. (2020). Agent-based models in transport planning: Current state, issues, and expectations. Procedia Computer Science, 170, 726-732.

Müller, K., & Axhausen, K. W. (2010). Population synthesis for microsimulation: State of the art. Arbeitsberichte Verkehrs-und Raumplanung, 638.

Sun, L., & Erath, A. (2015). A Bayesian network approach for population synthesis. Transportation Research Part C: Emerging Technologies, 61, 49-62.

Voas, D., & Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. Geographical and Environmental Modelling, 5(2), 177–200. doi: 10.1080/13615930120086078

Ye, X., Konduri, K., Pendyala, R. M., Sana, B., & Waddell, P. (2009, January). 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.