A Preliminary Analysis on Generating Synthetic Population in South Korea

Abstract

As the initial step to an activity-based travel demand model, population synthesis is essential to the accuracy of the results. The aim of population synthesis is generating virtual agents that can represent the behavior and decision-making characteristics of the actual population. A fair number of software packages, mostly based on the Iterative Proportional Fitting (IPF) algorithm, have been proposed to synthesize agents. Among these packages, PopGen is the only one using an approach other than IPF, which unlike its counterparts allows for simultaneously matching both household and person level attributes, resulting in a more representative synthetic population. This paper aims at using this software package to generate a synthetic population for a district of Seoul, South Korea. The data in this project was taken from Seoul Metropolitan Area Household Travel Survey (SMAHTS) and Korea Population and Household Census (KPHC). The results show a good level of fit between the synthesized and actual population.

1. Introduction

With the advent of activity-based models, a new era has been introduced in the travel demand modeling field. The conventional four-step travel demand models, although less data demanding and computationally lighter, have distinct shortcomings and weak points that many of the agencies, specifically in the United States, have moved to supplant such aggregate models with their agent-based microsimulation counterparts. Basically, these models work on the basis of simulating agents’ behavior and decisions so as to forecast the state of transportation system in the future [1].

The initial step to an agent-based microsimulation model is generating the “agents”, or the population, aiming to use the transportation network. Obviously, generating agents that can match the real population as closely as possible is only made possible by the availability of accurate data at a disaggregate level. However, such micro level data is often not accessible, either due to confidentiality reasons, or simply the high cost of obtaining them. “Population synthesis”, as first pioneered by Beckman et al. [2], aims at using available disaggregate data along with available
aggregate data to produce a synthetic population that can represent the characteristics and number of the actual corresponding population.

There have been generally two approaches to population synthesis. The first, known as “synthetic reconstruction”, was first used by Beckman et al. [2] and later continued to be employed by other researchers. The second approach, or the “combinatorial optimization” technique, was employed and its results were demonstrated by Voas and Williamson [3]. Ryan et al. [4] have provided a comparison of these two approaches.

Over the years, a fair number of population synthesis software packages have been introduced and some have found their way into practice. PopGen, presented by Ye et al. [5], is a standalone software package which uses a novel technique, called Iterative Proportional Updating (IPU), to simultaneously fit for both household and person level characteristics. This package has been initially tested for Maricopa County, Arizona, but is also gaining traction in other areas of United States such as Atlanta, Georgia. CEMDAP model [6], incorporates a population synthesizer based on the work of Guo and Bhat [7], and has been used in Dallas-Fort Worth, Texas. The ILUTE model [8] takes advantage of a population synthesizer by Pritchard and Miller [9], and is used for Toronto, Canada. ALBATROSS model [10] uses a synthesizing process in which household distribution is computed based on the person-level distribution.

Although the use of activity-based models have been gaining traction in America, Canada, and Europe, Asian countries are mostly left behind in this state-of-the-art demand modeling. The aim of this paper is to generate a synthetic population, for the first time, in a region of South Korea. To accomplish this, we have taken advantage of PopGen software package and data from Seoul Metropolitan Area Household Travel Survey (SMAHTS) and Korea Population and Household Census (KPHC). To further pave the way for presenting the results, the remainder of this paper is organized as follows. Section 2 discusses the fitting procedures of population synthesis algorithms and specifically the IPU procedure used in PopGen which makes it distinct from the other models. Next, section 3 briefly discusses the procedure of drawing the households from the fitted sample. Sections 4 and 5, respectively, introduce the region for which population is synthesized and discuss the data used. Section 6 elaborates on the methodology and the results are presented in section 7. Finally, section 8 provides the summary and concluding remarks.
2. Fitting procedure in population synthesis

The term fitting, in population synthesis literature, refers to fitting a random disaggregate sample to an aggregate constraint. In other words, we aim to match the joint distribution of the critical characteristics of the synthesized population with their known aggregate distributions of their counterparts [5]. This random sample, also called reference sample or seed, can come from microcensus, public-use micro sample (PUMS) data, or a travel survey. The aggregate constraint, also referred to as marginal sums, are often obtained from the available aggregate data for the base year of the population synthesis [1].

The problem that researchers often face in this regard is the unavailability of the joint distribution of the critical attributes of interest. Thus, such joint distribution usually has to be generated by researchers and practitioners themselves.

The most widely used procedure to generate the joint distribution of the desired attributes is the Iterative Proportional Fitting (IPF). This iterative method, first described by Deming and Stephan [11] in the context of adjusting sample frequency tables to the known marginal distributions, was later adopted by Beckman et al. [2] to estimate the joint distributions of household attributes. Essentially, IPF estimates the distribution of the control variables (characteristics of interest) in a way that the structure of the seed is maintained while the total number of agents in a category also matches the marginal sum. With the exception of PopGen, all other software packages mentioned in section 1 use the IPF procedure for the fitting purpose. The one caveat that comes with using this method is that it only controls for one level of attributes (either household level or person level). It is often the case that the packages employing this method only control for household level variables, leaving the person level variables unmatched.

To remedy this situation, Ye et al. [5] have proposed a new iterative algorithm called Iterative proportional Updating (IPU) that controls both for the household level and person level attributes. This algorithm is shown to perform quite well both from a standpoint of computation time and matching household and person attributes. PopGen is the only population synthesizer taking advantage of IPU procedure.

3. Allocation procedure in population synthesis
The major purpose of this step is drawing households from the fitted reference sample to form the synthetic population. In PopGen, unlike other procedures using the IPF algorithm, households that fall in the same category may have different weights due to matching attributes at both household and person levels. This makes the drawing procedure of PopGen a little different from other software packages. Ye et al. [5] describe the probability of a household being chosen as equal to its weight divided by the sum of weights of all households belonging to a particular household type. Based on this probability, households are randomly drawn from fitted sample.

4. Study Area

This project’s target area is the city of Seoul, which is the capital of South Korea. It has around 10 million population and is spread over an area of 233.7 square miles. The city has the largest and most diverse population composition across the country, so the region is the best area to, for the first time, generate a synthetic population in South Korea. More specifically, the study chooses Gangnam-gu, one of the local government districts in Seoul, as a study area, as shown in Figure 1. Gangnam-gu is one of the three central business districts (CBDs) in Seoul, and possesses a highly mixed land-use pattern, including evenly distributed residential, business, commercial, and cultural areas. For this reason, it has been preferable for Korean transportation researchers to choose the region as their preliminary studies.

![Figure 1. Narrowing down the study area](image)

5. Data
This study takes advantage of two data sources, KPHC and SMAHTS, for setting up the marginal and sample data, respectively. The KPHC collects household and population data across the entire Korea region using specified small-sized census tracts (873 and 16,617 census tracts in Gangnam-gu and Seoul, respectively) every five years. This study uses the latest version of KPHC dataset released in the year 2010. The KPHC collects a spectrum of household and population data as shown in Table 1 below. According to the KPHC, Gangnam-gu region has 200,965 households and 515,165 people, which translates into 33,737 persons per square mile.

The SMAHTS is conducted by Seoul Metropolitan Transportation Authority every five years in order to collect people’s travel histories with several characteristics of the household and individuals as shown in Table 1 below. The SMAHTS has data on around 1 to 2 percent of households over Seoul metropolitan area. Especially, it collected 6,145 households (16,450 people) in the year of 2012, which is equivalent to 3.1 percent of total households (3.2 percent of total population). The SMAHTS is based on 424 special Travel Analysis Zones (TAZs) across the city of Seoul, 22 of which define the Gangnam-gu district.

**Table 1. Survey data collected by KPHC and SMAHTS**

<table>
<thead>
<tr>
<th>Classification</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>KPHC</strong></td>
<td>Household: The number of households in each of the following categories: Number of rooms, Heating type, Renting type, Family composition</td>
</tr>
<tr>
<td></td>
<td>person: The number of population in each of the following categories: Age, Gender, Level of Education, Marriage status, Religion</td>
</tr>
<tr>
<td><strong>SMAHTS</strong></td>
<td>Household: Household characteristics in each of the following categories: Household type, Income, Number of household members, Number of ages under six, Household ownership</td>
</tr>
<tr>
<td></td>
<td>Person: Person characteristics in each of the following categories: Age, Date of birth, Gender, Occupation type and status, Driver’s license</td>
</tr>
</tbody>
</table>

**6. Methodology**
This study, as mentioned, uses Population synthesizer software package PopGen v. 1.0 to. One of the advantages of PopGen is its straightforward process when it comes to the data format to be used as the input, and also its run time and friendly Graphical User Interface (GUI) in a way that researchers even in other countries than United States could apply their own dataset in the software. Therefore, the study will show the practical ways for the future researchers how to build the input data for the software. As an example, the study chose to use the Korean dataset which is fairly comprehensive, but somewhat different than the U.S. dataset which the PopGen is initially design for. The flow of the study is briefly described in Figure 2 below.

![Figure 2. The flow of this study](image)

Considering the common data between KPHC and SMHTS datasets, this study suggests two cases of population synthesis. The first case only uses KPHC dataset to come up with the marginals for items included in both KPHC and SMAHTS. It is recommended, in general, to extract the marginal sets from census because it is the right way to represent the actual totals. However, this method has a limitation when census and other data sources such as the travel survey don’t share the same variables. Therefore, this study suggests that additional marginal sets be extracted from SMAHTS by researchers to selectively utilize those for their research purposes. As an example, we estimated the additional marginal sets by adjusting them with adjustment factors to match the marginal sum to the total population in KPHC dataset.

In the next step, we built the sample dataset of household and population based on SMAHTS that can be fed as the seed into PopGen. Each sample has its own household characteristics such as household type, income, renting type and children presence, and person characteristics such as
age, gender, license presence, and employment status. To run the PopGen, those sample characteristics were categorized in order to match the classifications of KPHC.

The marginal distributions and samples are obtained by aggregating the number of households and population in each category to the TAZ level whose size, as shown schematically in Figure 2, is much larger than that of a census tract. The reason for doing this aggregation was that samples from SMAHTS are basically based on TAZs, and also it would highly increase the run time of PopGen if we opted for using the highly sophisticated small-sized census tracts. Considering the zero-cell problem, we then checked for TAZs with no households or persons. Such TAZs, however, were not encountered due to the high density of the selected region. However, it is recommended for future researchers to check the solution to solve the zero-cell problem as described in Ye et al [5] if they come across such a problem.

Figure 2. Aggregation of household and population from census tracts to TAZs

7. Results

Using the marginals and samples discussed in the previous section, we ran PogGen to fit the samples of SMAHTS proportional to the marginal distributions based on the IPU algorithm.
Overall, the results of the study show a high goodness-of-fit for most variables in both cases. As demonstrated in Figures 3 and 4, case 2 proves to have performed better than case 1 in terms of p-values and average absolute relative differences (AARD); however, both cases seem to have acceptable results. It should be noted that, even though case 2 shows a much higher goodness-of-fit, future researchers should be careful when using the marginals made by data from a source other than census. It should be kept in mind that because those marginals are based on the dataset that is also being used for the samples, the results, consequently, possess a higher goodness-of-fit level.

As shown in Figures 5 and 6, the results of the variables in both cases also show a good replication power. Especially, the renting type variable shows a perfect match to the actual marginal values as shown on the left bar graph in the Figure 5. The reason for this is that only one variable, renting type, was controlled for in case 1. Furthermore, the study also checked the distributions of every variable in the TAZ level, and found that almost every one of them are close to the actual ones.
Figure 5. Distributions of household and personal variables of case 1

Figure 6. Distributions of household and personal variables of case 2

On the other hand, the total numbers of synthesized households and population are somewhat underestimated in both cases. The study found that this discrepancy rises from the fact that the marginal total of renting type variable, as obtained from the census data, is less than the household total that the census data accounts for. In order to solve the problem, the study suggests two types of solutions. The first one is to fix the marginal by multiplying it by the ratio of it to the household total and the second solution is to replicate the randomly-selected synthetic population proportional to the ratio to the household total.
Table 2. Comparison the number of household and people between synthesized and actual

<table>
<thead>
<tr>
<th>Classification</th>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Synthetic</td>
<td>Actual</td>
</tr>
<tr>
<td>Household</td>
<td>197,956</td>
<td>200,965</td>
</tr>
<tr>
<td>People</td>
<td>497,327</td>
<td>515,165</td>
</tr>
</tbody>
</table>

8. Conclusion

In this project, we aimed to generate, for the first time, a synthetic population for a district in Seoul, South Korea. PopGen software v. 1.0 was used due to its high capability in synthesizing agents whose attributes match the actual population on both household and person levels.

The data were taken from Seoul Metropolitan Area Household Travel Survey (SMAHTS) and Korea Population and Household Census (KPHC). The latter is a countrywide survey collecting data on household and person levels. The former is travel survey carried out by Seoul Metropolitan Transportation Authority from 1 to 2 percent of the population in Seoul metro area.

PopGen is run for two cases. In the first case, the marginals are solely based on the KPHC dataset. Case 2 also uses marginals from the SMAHTS. This was because that not all the attributes collected in the SMATHS exist in it larger counterpart.

The results show a good level of fit for both cases. Case 2, however, shows a better level of fit compared to case 1. This is obviously due to case 1 using marginals from SMAHTS.

The future research can investigate using a population synthesizer based on the IPF procedure and compare the result to that of PopGen. In addition, the synthesized population may be used as the input to an activity-based model in South Korea. This would of course require running n PopGen for all the districts in the metro area.
9. References


6- Pinjari, A. R., N. Eluru, R. B. Copperman, I. N. Sener, J. Y. Guo, S. Srinivasan and C. R. Bhat, 2006, “Activity-based travel-demand analysis for metropolitan areas in Texas: CEMDAP models, framework, software architecture and application results”, Research Report, 4080-8, Texas Department of Transportation, Department of Civil, Architectural and Environmental Engineering, University of Texas Austin, Austin,


