Population Synthesis Literature Review

Jin Zhou

August 10, 2017

1 Goal

Since the step of urban typology has been done, we need to fill in the blank between city clustering results and a complete Simmobility input of virtual city. The city model that Simmobility takes as input has two components. One is the demand and the other is the road network. To model the demand, a methodology has to be developed to construct the population with desired demographics and allocate the population in space. Next to it, we also need to figure out how to use the synthesized households and individuals in disaggregate level to get the travel demand in the city. The literature review at this time is to gain a better idea of previous work people have done to address the problem of population synthesis.

1.1 Questions:

1. What should the input of travel demand generation look like? This guides the adaptation of population synthesis method from literature.
2. What attributes should be included to characterize the population? Age, sex, educational level, and there must be some attributes reflect the traveling behavior of people.
3. Do we synthesize households or personnels? Or how to match individuals to households?
4. We need to come up with a solution to cities don't have PUMS data and virtual cities that don't even have census data for demographics.

2 Prior Work

Household assignment and spatial allocation are two big issues we need to solve when facing the problem of population synthesis. Specifically, the first one addresses the problem of how to match individuals to the household where they belong, while the spatial allocation distributes households in geographical space. It is necessary because the ultimate goal of our work is to do transportation simulation, which can only be achieved by knowing the location of each agent.

There has been lots of literature on population synthesis. The methods they use varies from Iterative Proportional Fitting (IPF), MCMC-based sampling [2], to Bayesian network approach [6]. The original work in this field is presented in [1]. It used a method called Iterative Proportional Fitting (IPF) to synthesize population based on data from two sources. What they do is iteratively fitting the cells of a multi-way contingency table expanded by attributes that characterize a household/individual so that it satisfies the constraints of marginal distribution

of demographics and also preserves the correlation structure carried by sample data of the whole population. Here the marginals of demographics comes from census data in the level of census tract or block group while the correlation structure need to be derived from survey data of target population. In case of the absence of disaggregated sample data, IPF can still give a solution by iteratively adjust the cells in accordance only with the marginal constraints. However, lots of different joint distributions can share the same marginals and IPF may converge to anyone of them. This is certainly a downside of traditional IPF, and is also the reason why we need the correlation structure of sample data to make the solution captures the true population distribution. After the fitting of the table, a desired number of synthetical population is sampled according to individual or household weights in the table. And this is where the sensitivity of IPF to the quality of data and the sample size stems from. In this method, spatial allocation is solved concurrently with the population synthesis, since it generates the population for each zone and then adds up to the whole population. On contrary to this, one can also synthesize the total population first and then allocate them to different zones. However, this includes implementing a detailed matching method between households and zones, like what's developed in [4], which is also very time-consuming.

Ever since Beckham et al. used IPF to do population synthesis, limitation of this method and the targeted modification has been published by other researchers. By IPF alone, one can only separately generate a set of weights either for individual or household. According to [7], the synthetic population generated based on the application of household weights can give a joint distribution of person attributes far from the given person marginal distribution. This is due to the fact that all persons are just simply selected from chosen households according to household weights. [7] has improved this problem of IPF by a method called Iterative Proportional Updating (IPU), and they claim that the algorithm they proposed can match the person-level distribution and household-level attributes as closely as possible.

MCMC-based sampling is brought up by [2]. The problem is defined as how to use partial views, including samples and conditional distributions, to estimate the true joint distribution of population attributes, where the attributes are defined as $X = (X^1, X^2, ...X^n)$. Here in this paper, they used Gibbs Sampling to generate certain amount of population based on prior knowledge of the joint distribution and proved that the sampled population is very much similar to the true simulation in statistical senses. It has been pointed out that, since the task of getting full conditionals won't be trivial in reality, they replaced some of conditional distributions by marginal distributions. So, if one wants to implement MCMC, the sample data of the entire population is fundamental. On top of that, the derivation of conditional distribution is not trivial especially when the number of attributes gets large.

The population synthesis method that [5] used adapted ideas from probabilistic graphical models. In their graphical model, there are 7 socio-demographical attributes, which are Age, Gender, Employment, Income, Car ownership, Purpose of trip, and Who pay for the trip, to represent a person. Once the dependencies between variables are specified, we need to calculate or estimate conditional probability of one node given its parents. The flexibility of this method is that those conditional probabilities can be derived from different sources like census data or surveys on individuals. Finally, the synthetic population is just random draws from the probabilistic graphical model. Though we can take marginal distributions of all variables as supplement for the graphical model, the population generated by Bayesian network model can only precisely match the marginals of mother nodes. This is because the distribution of a node is uniquely defined by its parent node and the conditional distribution based on PGM. And this point should be taken into consideration when choosing the population synthesis method.

3 Data

Most paper in population synthesis used the Public Use Microdata Sample (PUMS), which is available both in individual leve and household level. It includes multiple socio-economic parameters of a household or an individual in the region. The sample rate of this dataset is about 1% to 10% of the total population in the region. We found PUMS data is complete in US cities but haven't found microdata in other cities in the world. Currently, we have found aggregated information of population, age, sex, household, education in districts in Beijing.

The choice of algorithm should be based both on the data we have and the appearance of the final result we want. What's also important for us is the simplicity and novelty of methodology.

4 Iterative Proportional Fitting

Using IPF to synthesize population includes two major steps. In the first step, the disaggregate samples of population are used to initialize the contingency table expanded by attributes, and the fitting is carried out to meet marginal constraints from aggregated data. In the next step, the fitted contingency table is used to sample the population. To describe IPF, we need to define the contingency table it fits. Suppose a person or household is described by m attributes (or demographics), then we need to develop a m-way table. For each demographic i, there are n_i categories. So basically, a cell $(i_1, i_2, ..., i_m)$ represents a kind of person/household, where $i_j = 1, 2, ...n_j$ is the observed value of the jth demographic with n_j categories it can choose from. From sample data like PUMS, n is the total number of observations while $n_{i_1,i_2,...,i_m}$ is the counts for people/household of this type. Hence, the proportion of this type in the sample is denoted by

$$p_{i_1,i_2,...,i_m} = \frac{n_{i_1,i_2,...,i_m}}{n}$$

The constraints is defined as T_k^j, which is the marginal totals for the kth category of attribute i from census. We can easily get the equation $\sum_i \sum_{k=1}^{n_i} T_k^j = n$. Let $p^{(t)}_{i_1,i_2,...,i_m}$ denote estimated proportion of cell $(i_1, i_2, ..., i_m)$ in iteration and

$$p_{...,i_j=k,...} = \sum_{i_1=1}^{n_1} ... \sum_{i_m=1}^{n_m} p^{(t)}_{i_1,i_2,...,i_j=k,...,i_m}$$

The IPF starts with initiating all weights in the contingency table by

$$p^{(0)}_{i_1,i_2,...,i_m} = p_{i_1,i_2,...,i_m} \tag{1}$$

and in each iteration, the proportions are updated by the formula

$$p^{(new)}_{i_1,i_2,...,i_j=k,...,i_m} = p^{(old)}_{i_1,i_2,...,i_j=k,...,i_m} * \frac{T_k^j * p^{(new)}_{...,i_j=k,...}}{n} \tag{2}$$

for all categories of each attribute, until it converges. Within an iteration, $p^{(old)}$ for the first marginal corresponds to $p^{(t-1)}$, which is the result from last iteration. And for marginals come next, $p^{(old)}$ should be set to proportions calculated from previous marginals. According to [1], the algorithm converges in 10 to 20 iterations.

In the next step, the synthetic population shall be constructed by selecting households from PUMS in proportion to the estimated weights in the multiway table. Specifically, the number of households of each demographic type in a census tract can be obtained by multiplying the

total number of households by the probability in the cell representing this household type, or by drawing the numbers at random according to these probabilities. However, the household type represent by a cell in multiway table is not exactly that of PUMS data. The fitted multi-way table can only be span by control variables, which is defined as variables which have marginal distribution in census data. However, households and individuals in disaggregate data are usually described by additional desired variables. So, the probability of a household type in PUMS data being chosen shall be assigned according to the distance between such a household p and a household type c in multi-way table. In [1], this probability is defined as

$$D(p,c) = w_p \prod_{i \in J} \left(1 - |(d_i^p - d_i^c)/r_i|^k\right) * \prod_{i \notin J} \left(1 - \Delta(d_i^p, d_i^c)\right) \tag{3}$$

5 Proposed method

Step 1: Network: modify OSM network file, make it legal input to Simobility

1. links has to be unidirectional
2. shorten the links and add turning points
3. maintain the attributes of each object in the network

Step 2: Population Synthesis

1. Data collection: survey data, aggregate census data for zones in each city
2. Learn the parameter and the structure of Bayesian Network which captures the structure of population
3. sampling from Bayesian Network to get whole population in the city
4. Clustering the synthetical population in a city by K-means
5. Divide the city into zones according to clustering result, and form the aggregate census data for zones

Although applying IPF to population synthesis is prevailing and the idea is straightforward, the algorithm itself is super time consuming and expansive once the combinations of demographics gets large. Suppose we have 2 sex, 7 age groups, 5 income groups, 5 levels of education, 9 types of occupation, 2 kinds of car ownership, and 5 household types, it's already 31500 types of individuals. Not to mention describing individuals with more demographics. Another issue with published methods is that the time we spend to generate population for one zone has to be multiplied by the number of zones in a city in order to get the whole population for the city, which means the time for a complete process is hundreds of that for a single zone if the zone is in the sense of census tract. Since what we want is something less time consuming, a new methodology has to be brought up.

Considering the accessibility of the two sources of data that the synthesizer needs, one can always get sampled survey data in a city level but not necessarily census data for zones in a city. In the worldwide, which is the scope of our problem, there are cities whose zones are even not defined properly. So we will face the problem of defining the zones of a city by our own in case of the data is in shortage. In developing the zones in a city, we will borrow the idea of zones from [3], where a city is divided into one CBD and subcenters along with their peripheral. It has to be admitted that in this way, we won't have aggregate census data for zones. This is

a problem we will address later.

In regards of assigning population to zones, we will cluster the population in city level with K-means first. The reason is that similar and highly correlated people will most likely live in the same area within a city. Then we get grouped people with similar attributes. We chose K-means because it's more intuitive if we want a spatial allocation and the feature of hidden centroid for each cluster will hopefully capture the characteristic of the zone.

6 References

[1] Richard J Beckman, Keith A Baggerly, and Michael D McKay. "Creating synthetic baseline populations". In: *Transportation Research Part A: Policy and Practice* 30.6 (1996), pp. 415–429.

[2] Bilal Farooq et al. "Simulation based population synthesis". In: *Transportation Research Part B: Methodological* 58 (2013), pp. 243 –263. ISSN: 0191-2615. DOI: http://dx.doi.org/10.1016/j.trb.2013.09.012. URL: http://www.sciencedirect.com/science/article/pii/S0191261513001720.

[3] Andrés Fielbaum, Sergio Jara-Diaz, and Antonio Gschwender. "A Parametric Description of Cities for the Normative Analysis of Transport Systems". In: *Networks and Spatial Economics* 17.2 (2017), pp. 343–365. ISSN: 1572-9427. DOI: 10.1007/s11067-016-9329-7. URL: https://doi.org/10.1007/s11067-016-9329-7.

[4] Yuanzheng Ge et al. "Virtual city: An individual-based digital environment for human mobility and interactive behavior". In: *Simulation* 90.8 (2014), pp. 917–935.

[5] Olga Petrik, Filipe Moura, and João de Abreu e Silva. "Measuring uncertainty in discrete choice travel demand forecasting models". In: *Transportation Planning and Technology* 39.2 (2016), pp. 218–237. DOI: 10.1080/03081060.2015.1127542. eprint: http://dx.doi.org/10.1080/03081060.2015.1127542. URL: http://dx.doi.org/10.1080/03081060.2015.1127542.

[6] Lijun Sun and Alexander Erath. "A Bayesian network approach for population synthesis". In: *Transportation Research Part C: Emerging Technologies* 61 (2015), pp. 49 –62. ISSN: 0968-090X. DOI: http://dx.doi.org/10.1016/j.trc.2015.10.010. URL: http://www.sciencedirect.com/science/article/pii/S0968090X15003599.

[7] Xin Ye et al. "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*. 2009.