Australasian Transport Research Forum 2013 Proceedings
2 - 4 October 2013, Brisbane, Australia
Publication website: http://www.patrec.org/atrf.aspx

New Methodology for Synthesizing Population in Metropolitans

Hojjat Rezaee[1], Navid Kalantari[2], Mohsen Babaei[3]

[1]PhD Candidate, School of Civil Engineering, University of Queensland, Brisbane, Australia

[2]PhD, Sharan Transportation Research Center (STRC), Tehran, Iran

[3]PhD, School of Civil Engineering, Iran University of Science and Technology (IUST), Tehran, Iran

Email for correspondence: H.Rezaeestakhruie@uq.edu.au

Abstract

In the Activity Based Modeling (ABM) approach, an activity pattern is specifically assigned to each individual in a household. In this way, one of the fundamental steps in the ABM approach is to project the socio-economic characteristics of all household members while considering some marginal constraints available for the whole of population which is known as population synthesizing. In the current paper, the household characteristics distribution is defined as the probability of having a household with a particular size, number of students and workers. The main goal of current research is to statistically fit the household characteristics distribution of the population to a previously obtained distribution for sample households in a way which satisfies the marginal constraints in traffic zones. It is also followed to solve two main issues in most previous research on population synthesis, one of them is related to the so called “zero cell” and the other is known as the “Integrality” problem. Similarity between characteristics distributions of the sample and all households can be achieved by using Maximum Likelihood Estimation (MLE) with the marginal constraints.

Satisfying all marginal constraints in a single optimization for a real case study involving a huge number of households increases the mathematical complexity of the problem, and likely leads to an infeasible state. In the current paper, a new idea for solving this problem in real cases is proposed. The proposed algorithm using GAMS software is implemented in Mashhad city (population of more than 2.5 million) in Iran.

1. Introduction

Many newly developed travel demand modeling systems such as Travel/Activity Scheduler for Household Agents (TASHA) (Roorda et al., 2007), Comprehensive Econometric Micro-simulator for Daily Activity-travel Patterns (CEMDAP) (Bhat et al., 2004), implemented activity-based modeling in Jakarta (Yagi and Mohammadian, 2006), DAYSIM (Bradley and Bowman, 2006), Coordinated Travel Regional Activity Modeling Platform (CTRAMP) (Davidson et al., 2010 and Vovsha et al., 2011) and Simulator of Activities, Greenhouse Emissions, Networks, and Travel (SimAGENT) (Goulias et al., 2011), highly depend on the individual and household level socioeconomic data. These data are being used as an input to the travel and activity generation process of these models. Household and individual activity decisions heavily depend on all their socioeconomic characteristics (Khan et al., 2012). Meanwhile, micro-simulation procedures require disaggregate demographic data which are not usually available at a zonal level, due to privacy limitations. Therefore population synthesis methods, which generate (synthesize) a complete or nearly complete listing of households with their individual and household level attributes that nearly comply with given marginal data, have been gaining favor in recent years.

The pioneer work of Beckman et al. (1996) is the first study devoted to population synthesis. Many conventional methods of population synthesis are based on the original work of Beckman et al. (1996). This method uses the Iterative Proportional Fitting (IPF) previously used in other fields (Deming and Stephan, 1940), to fit the multi-dimensional distribution of sub-regions to fit the marginal totals. It is assumed in the IPF that the correlation structure of the zones is similar to the correlation structure of the regions to which they belong. IPF has been also implemented in the TRansportation ANalysis and SIMulation System (TRANSIMS) (Weiner and Ducca, 1999) population synthesis module.

As the original IPF method didn't consider the individual characteristics and only used household data to fit the population, Guo and Bhat (2007) proposed an alternative IPF method which also controls the individual level marginal. Similarly, Arentze et al. (2007) and Ye et al. (2009) have proposed methods to fit both individual and household data. In an attempt to fit more attributes, Pritchard and Miller (2008) proposed the IPF with a sparse list-based data structure.

Many population synthesis methods have been implemented into software packages some of the most important ones include; TRANSIMS, PopGen (Pendyala et al., 2011), TRansportation and Environment Strategy Impact Simulator (TRESIS) (Ton and Hensher, 2001) and CEMDAP among others.

There are two main issues in the most previous research on population synthesis, one of them is related to the so called "zero cell" and the other is known as the "Integrality" problem. The method presented in this paper aims to reduce both these problems. Meanwhile, IPF methods assumed a similar covariance structure between zones which might not necessarily exist. Each of these issues will be discussed in this paper.

The rest of the paper is organized as follows; the methodology presented in this paper is given in section two. In section three the proposed method will be applied to the city of Mashhad and finally section four concludes the paper.

2. Methodology

The "zero cell" problem arises as many household or individual types might not exist in the initial multi-dimensional table. These tables are usually based on a sample of household (Such as Public Use Microdata Sample (PUMS) data or Origin Destination (OD) surveys) which contains the information of about 5% of households in the region. Obviously, this sample does not contain household and individual types with low probability of occurrence. IPF does not have the ability to deal with this problem and the most widely used method implemented to solve the problem is to add a very small number to those cells. This method has been shown to be inefficient by Beckman et al. (1996). In this paper two measures have been considered to deal with the zero cell problem: 1) aggregation 2) probabilistic view. The first method (aggregation) used all the samples in the region to match regional marginal distributions. The use of all the data in the sample reduces the probability of zero cells. In the second approach the problem has been formulated as a maximum likelihood problem in which the probability of each household in the region is estimated for each sub-region. These two mechanisms could be used in order to reduce the zero cell problem. It should be mentioned that the previous methods of dealing with the zero cell issue could be also used in conjunction with the Maximum Likelihood Estimation (MLE) method proposed in this paper.

The problem of interest is to find a multi-dimensional probability matrix, $\mathrm{F}(\mathbf{X})$, which consists of the probability of each attribute combination. For attribute set $\mathbf{X}$, which contains both the household and the individual level attributes, denoted by $\mathbf{X}'_{\mathrm{rs}}$, the subset of attributes are resulted by the omission of attributes r and s. For each *d*-dimensional attribute set $\mathbf{x^d} \subset \mathbf{X}$, denote by $f_{\mathbf{X}'_{\mathbf{x^d}}}(\mathbf{x^d})$, the marginal probability of the *d* attributes, $\mathbf{x^d}$, has been conditioned on

$\mathbf{X}'_{\mathbf{x^d}}$. In the proposed model, each city or region is divided into |k| mutually exclusive and collectively exhaustive sub-regions with unknown marginal distributions. Then the first step is to estimate the marginal distribution of each attribute pair $\mathbf{x}^2 = (\mathrm{r}, \mathrm{s})$ in the sub-region. A Maximum Likelihood Estimation (MLE) procedure has been proposed in this paper to find the distribution of each attribute pair as in Problem 1:

Problem 1:

$$\max_{f_{\mathbf{X}'_{\mathrm{rs}}}(\mathrm{r,s})} \sum_{(\mathrm{r,s})} G^{\mathrm{k}}_{\mathbf{X}'_{\mathrm{rs}}}(\mathrm{r,s}).\ln\left(f^{\mathrm{k}}_{\mathbf{X}'_{\mathrm{rs}}}(\mathrm{r,s})\right)$$

$$E^{k}_{\mathbf{X}'_{\mathrm{s}}}(s) = k_s$$

$$E^{k}_{\mathbf{X}'_{\mathrm{r}}}(r) = k_r$$

$$\sum_{(\mathrm{r,s})} \mathrm{f}^{\mathrm{k}}_{\mathrm{X}'_{\mathrm{rs}}}(\mathrm{r,s}) = 1$$

This problem should be solved for each attribute pair (r,s). In the above problem:

$\mathrm{G}^{\mathrm{k}}_{\mathrm{X}'_{\mathrm{rs}}}(\mathrm{r,s})$= the number of observations (r,s) in region k, which could be obtained from micro data or from OD surveys.

$\mathrm{E}^{\mathrm{k}}_{\mathrm{X}'_{\mathrm{s}}}(\mathrm{s})$ and $\mathrm{E}^{\mathrm{k}}_{\mathrm{X}'_{\mathrm{r}}}(\mathrm{r})$= the expected value of attribute s and r conditioned on X'_{s} and X'_{r} respectively.

k_{r} and k_{s}= the expected marginal number of individual or households with attribute r and s respectively.

It should be mentioned that the probabilities of interest in this problem are all discrete probabilities and thus the expectations could be easily calculated by summation. It is worth mentioning that for instance if s is assumed to be the number of students in the household, k_s is the expected number of students in each household in sub-region *k*, which could be obtained from census data.

Then, all attribute combination of interest should by computed using higher order combinations. In general terms problem 1 could be reformulated as in Problem 2.

Problem 2:

$$\max_{f_{\mathbf{X}'_{\mathbf{x^d}}}(\mathbf{x^d})} \sum_{\mathrm{x^d}} G^{k}_{\mathbf{X}'_{\mathbf{x^d}}}(\mathbf{x^d}).\ln\left(f^{k}_{\mathbf{X}'_{\mathbf{x^d}}}(\mathbf{x^d})\right)$$

$$E^{k}_{\mathbf{X}'_{\mathbf{x^h}}}\left(\mathbf{x^h} \epsilon \mathbf{x^d}\right) = k_{|x^h|} \qquad \forall \mathbf{x^h} \subset \mathbf{x^d} \subset \mathbf{X}, \forall 0 < h < d$$

$$\sum_{x^d = \{x_1, \ldots, x_d\} \in \mathbf{x^d}} f^{k}_{\mathbf{X}'_{\mathbf{x^d}}}(x^d) = 1$$

This problem should be solved for each combination $\mathbf{x^d} \subset \mathbf{X} - \mathbf{n}$, where $\mathbf{X} - \mathbf{n}$ is the set of attributes that is obtained by the omission of n attributes of $\mathbf{X}$, until $\mathrm{n} = 0$. $\mathrm{k}_{|\mathbf{x^h}|}$ is the expected number of each attribute $\mathbf{x}^{\mathbf{d}}_{\mathbf{l}}$ in the sub-region k. For instance if the desired attribute is the number of students, $\mathrm{k}_{|\mathbf{x^h}|}$ is the total number of student is the sub-region k. After

estimating the sub-regional marginal probabilities for each attribute pair (r,s), the marginal probabilities should be estimated for each Traffic Analysis Zone (TAZ) $i \epsilon k$ in each sub region k by assuming a similar correlation structure between all TAZs in each given sub-region. A Least Square (LS) model could be used to estimate the marginal probabilities in each TAZ as Problem 3.

Problem 3:

$$\min_{f_i^k(\mathbf{X})} \sum_{i \epsilon k} \sum_{x^n \epsilon \mathrm{X}} \left(f^k(x^n) - f_i^k(x^n) \right)^2$$

$$E^{k,i}_{\mathbf{X}'_{\mathbf{x}^1}}(\mathbf{x}^1) = k_{|\boldsymbol{x}^1|} \qquad \forall \mathbf{x}^1 \epsilon \mathbf{X}, \forall 0 < h < n$$

$$f_i^k(x^n) \leq 1 \qquad \forall x^n \epsilon \mathrm{X}$$

$$\sum_{x^n \epsilon X} f_i^k(x^n) = 1$$

Where $f_i^k(x^n)$ is the probability of x^n in zone i which is located in the sub-region *k*, and $k_{|\boldsymbol{x}^1|}$ is the total marginal number of x^n in zone i which is located in the sub-region *k*. By the use of the above mentioned model the probabilities for each zone will be computed. It is worth mentioning that in some realistic applications, the marginal totals may not exactly fit to the observations (in the first set of constraints). In these cases a small noise could be used to relax the constraints and to allow for a small error in estimation. This noise was considered in the case study applied in this paper. The noise level was 0.001 in the city of Mashhad.

The integrality issue could be circumvented by the used of integer variables. Problem number 3 could be solved in terms of number of individuals or households in the sub-regions and TAZs instead of the probability and solved by an integer programming technique. This method completely resolves the integrality problem of IPF methods. It is also worth mentioning that the mathematical programming framework given in this paper offers a high level of flexibility in the model. Many constraints and attributes could be incorporated in the model. These constraints could be used to balance both household and individual level attributes to any number and type of constraint, upon the modeler's request.

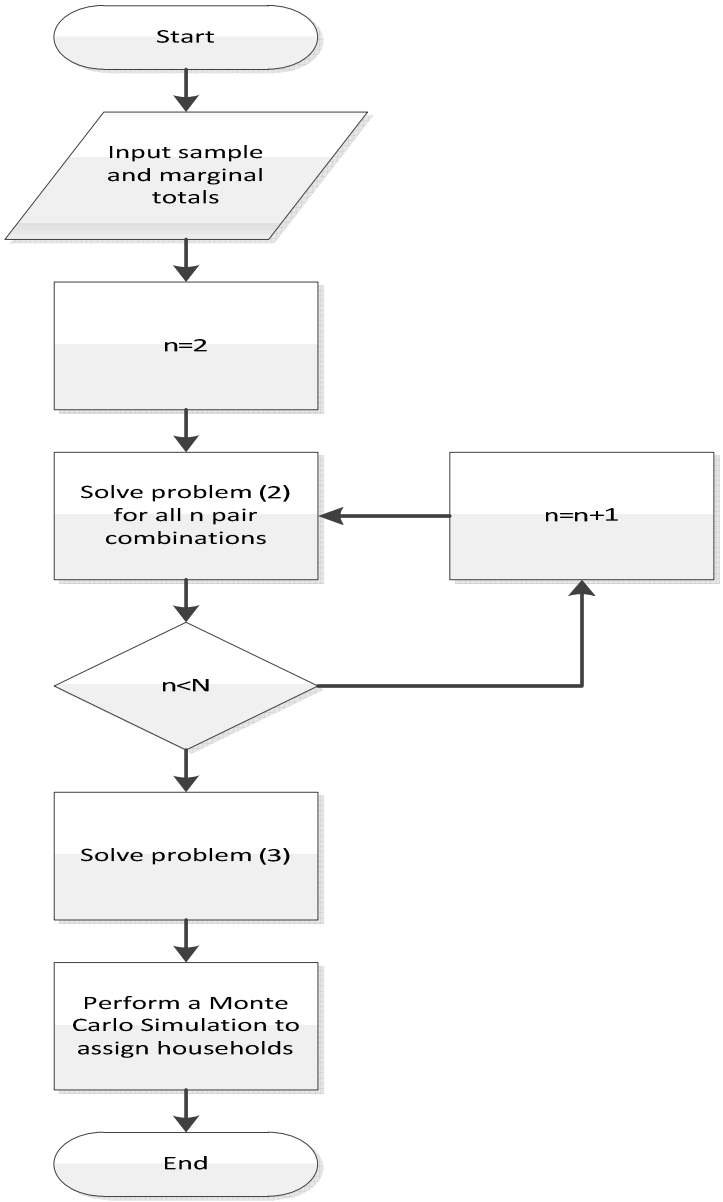
Based on the methodology presented in this section the Population Synthesis Algorithm (PSA) could be summarized as follows in section 2.1.

2.1. Solution Algorithm

- Step 1: Input the sample household data, the marginal total number of attributes in each sub-region and zone.
- Step 2: Compute the marginal probabilities in each sub-region by solving problem 2, respectively for each 2, 3, ..., n combinations of attributes.
- Step 3: Compute the probabilities for each zone using the marginal probabilities computed in step 2 and by solving Problem 3.
- Step 4: Use a Monte Carlo Simulation routine to assign sample households to each zone.

The solution algorithm has been also shown in the flowchart given in Figure 1 for clarity. It should be realized that step 2 is repeated for all combinations of attributes. Although Problem 2 is a linear programming problem, increasing the number of control attributes could increase the burden of the computations. For instance for a problem with 5 control attributes step 2 should be repeated $\sum_{i=2}^{5} \binom{5}{i}$ times.

Figure 1: Flowchart of PSA



3. Case Study

The proposed Algorithm has been applied to the city of Mashhad in Iran. The city has about 2.4 million inhibitions and 639,195 households. It has been divided into 13 sub-regions and 253 TAZs. The model has been implemented as a part of Mashhad's comprehensive traffic studies. As the individual level data were not accurate in Mashhad, the synthesis has been only applied to the synthesis of household level characteristics. Household size, number of students and workers in each household has been used as the control variable in the synthetic population. Thus, the multi-dimensional attribute matrix is a three dimensional matrix in this study. Each dimension is assigned to a different attribute. The second step of PSA was solved four times; three times for each two dimensional probability and once for the three dimensional probability. The two dimensional probabilities include (Size, Student), (Size, Worker) and (Student, Worker) and the three dimensional probability is (Size, Student, Worker). The problem has been solved using the GAMS software package.

The result of applying the fourth step of PSA to the sub-regions is given in Figures 2, 3, 4. Figure 2 shows the percent distribution of the sample, synthetic population and real data in

all 13 sub-regions. As can be seen, although the sample distribution differs from reality the synthesized data perfectly match the real data in the sub-regions. The same analysis has been performed on the number of households in the sub-regions in Figure 3.

Figure 2: Percent of population in each sub-region

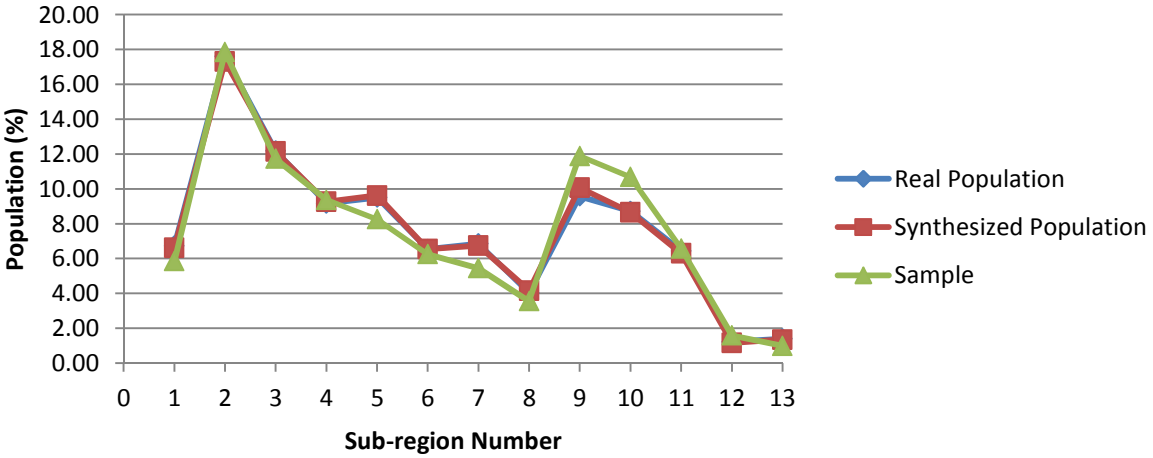


Figure 3: Percent of households in each sub-region

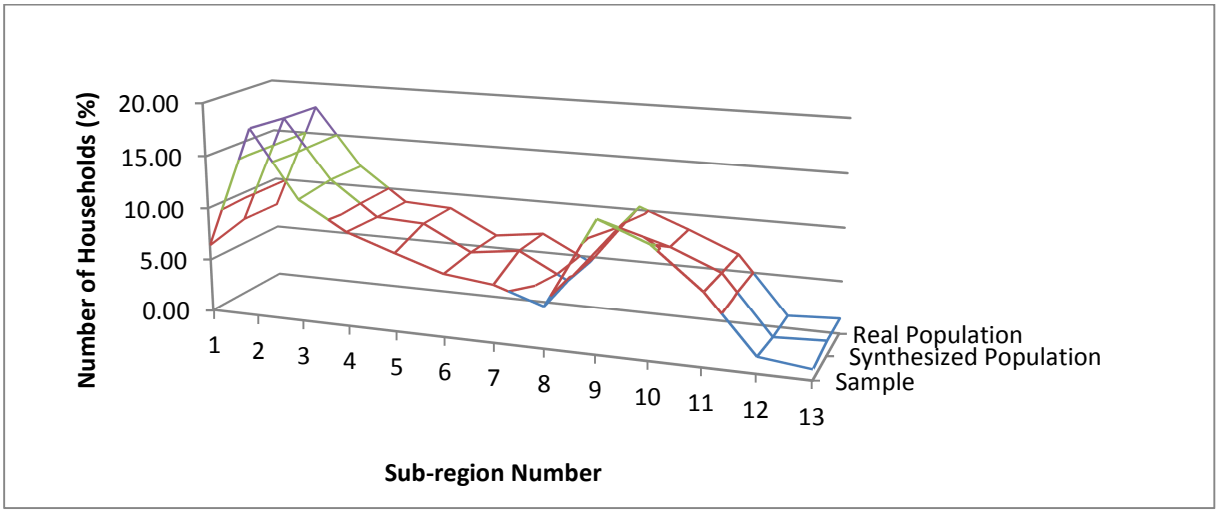
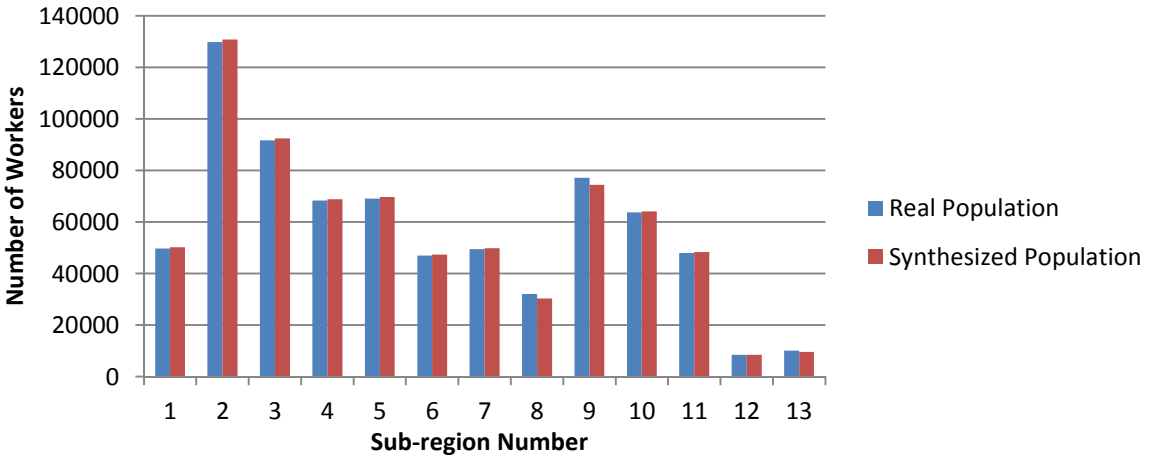


Figure 4: Number of Workers in each sub-region

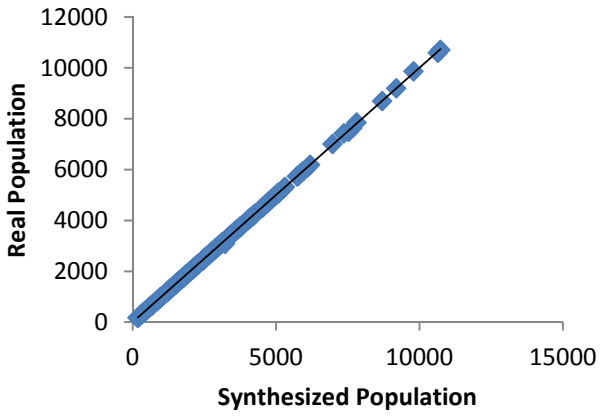


After fitting the sub-regional data and computation of their probability distributions, the zonal attributes should be computed. Figure 5 shows the accuracy of the fitted data in each zone to the attributes under consideration. Obviously the control data have shown a good fit to the marginal totals in each zone.

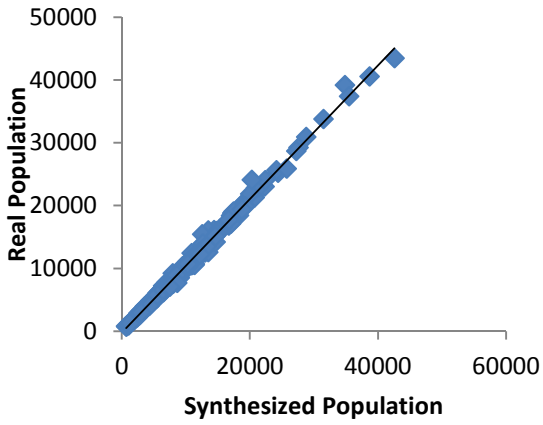
In order to validate the proposed model the household size distribution, which has been treated as an uncontrolled variable in this study has been used. Figure 6 shows the distribution of household in the city both in the synthetic and real population.

In order to test for validity four different tests has been used. The first two tests are parametric test which assume normality in the distributions. Using the correlation test, the correlation between the two set is 0.995 which is highly significant. By testing for the differences using the t-test, the computed t is equal to 0 which again suggests the equality of the two data with high level of significance. The other two tests are nonparametric test. The Wilcoxon Signed Ranks Test was first used. The asymptotic two tail significance level of the test is 0.721 which highly recommends a correlation (>>0.05) and the other test is the Sign Test which showed a significance 0.344 (>>0.05) that also shows a high level of significance.

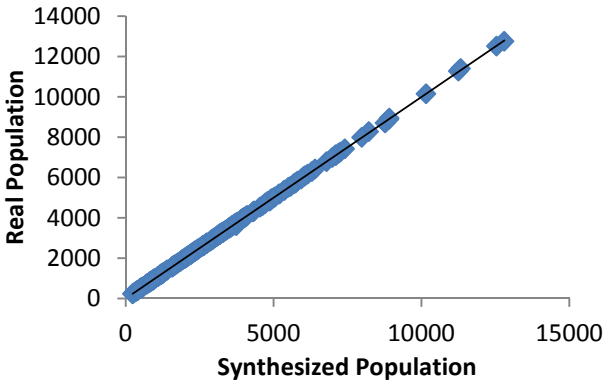
Figure 5: Investigation of control totals in each TAZ



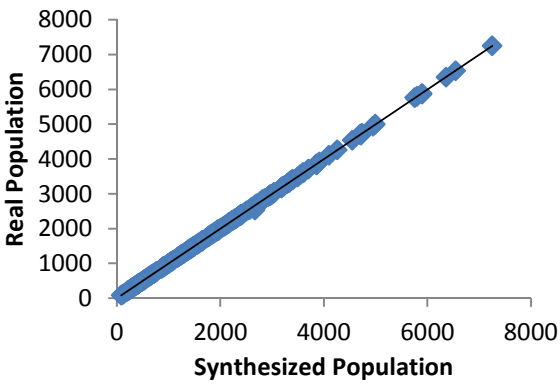
(a) Number of households in each zone



(b) Population in each zone

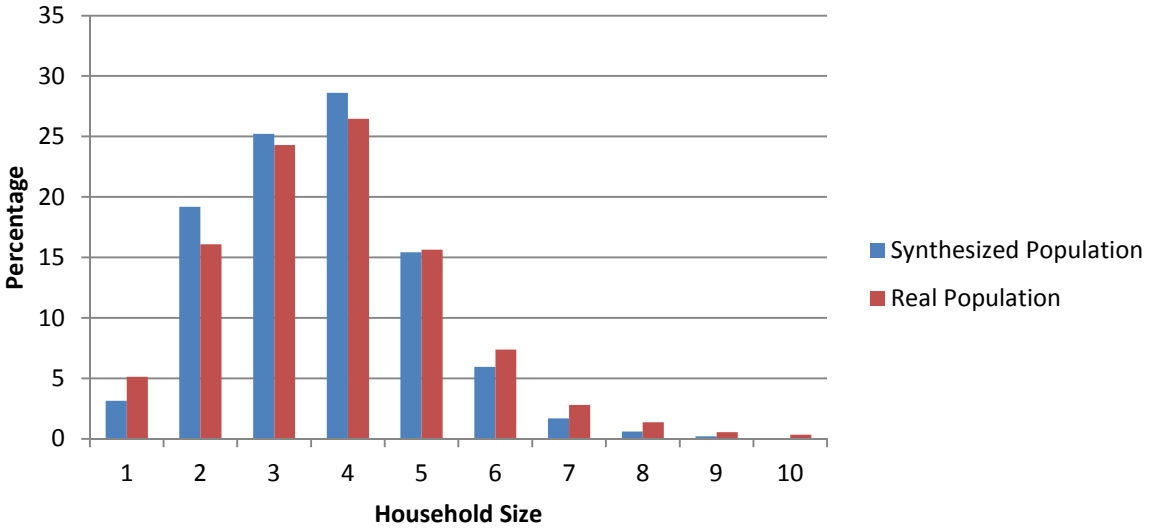


(c) Number of workers in each zone



(d) Number of student in each zone

Figure 6: Distribution of household size in the synthesized and real population



4. Conclusion

In this paper an operations research based method has been proposed for population synthesis. The method has been applied to the city of Mashhad and has exhibited promising results. The accuracy of the proposed model has been evaluated by the use of an uncontrolled attribute, for which statistical testing has proven its accuracy. The statistical test has shown a highly significant correlation between the synthetic and real population. The method needs some modification which could reduce its computational efficiency which has been left for future research. This paper did not give any recommendation on the last step, which is the Monte Carlo simulation method used previously in many studies and has been applied in this study as well. The use of Spatial Statistics is also an ongoing research which is being undertaken by the authors. This is because Monte Carlo Simulation gives an identical weight to all households and does not consider the socio-economic similarity between the residents living close to each other.

References

Arentze, T A, Timmermans, H J P and Hofman, F (2007) Population synthesis for microsimulating travel behavior. *Journal of Transportation Research Record 2014 (11),* 85–91

Beckman, R J, Baggerly, K A and McKay, M D (1996) Creating synthetic baseline populations. *Journal of Transportation Research Part A 30 (6),* 415–435

Bhat, C R, Guo, J Y, Srinivasan, S and Sivakumar, A (2004) Comprehensive econometric microsimulator for daily activity-travel patterns. *Journal of Transportation Research Record 1894,* 57-66

Davidson, W, Vovsha, P, Freedman, J and Donnelly, R (2010) CT-RAMP family of activity-based models. *The 33rd Australasian Transport Research Forum* Canberra: ATRF

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

Gaulias, K G, Bhat, C R, Pendyala, R M, Chen, Y, Paleti, R, Konduri, K C, Huang, G and Hu, H H (2011) Simulator of activities, greenhouse emissions, networks, and travel (SimAGENT) in Southern California: Design, implementation, preliminary findings, and integration plans. *IEEE Forum on Integrated and Sustainable Transportation Systems* Vienna

Guo, J Y and Bhat, C R (2007) Population synthesis for microsimulating travel behavior. *Journal of Transportation Research Record 2014 (12)*, 92–101

Khan, M, Paleti, R, Bhat, C and Pendyala, R (2012) Joint household-level analysis of individuals' work arrangement choices. Journal of Transportation Research Record 2323, 56–66

Pendyala, R M, Christian, K P and Konduri, K C (2011) *PopGen 1.1 User's Guide.* Raleigh, N.C.: Lulu Publishers

Pritchard, D R and Miller, E J (2009) Advances in agent population synthesis and application in an integrated land use and transportation model. *The 88th Annual Meeting of the Transportation Research Board* Washington, D.C.: TRB

Roorda, M J, Miller, E J and Habib, K M N (2007) Validation of TASHA: a 24-hour activity scheduling microsimulation model. *The 86th annual meeting of the Transportation Research Board* Washington, D.C.: TRB

SACSIM/05 Activity-based travel forecasting model for SACOG: Featuring Daysim – The Person Day Activity and Travel Simulator (2006) *Report* (M Bradley and J Bowman) prepared for Sacramento Area Council of Governments (SACOG).

Ton, T and Hensher, D A (2001) Synthesising Population Data: The Specification and Generation of Synthetic Households in TRESIS. *The 9th World Conference on Transport Research* Seoul

Vovsha, P, Freedman, J, Livshits, V and Sun, W (2011) Design features of activity-based models in practice: Coordinated travel-regional activity modeling platform. *Journal of the Transportation Research Record 2254*, 19-27

Weiner, E and Ducca. F (1999) Upgrading Travel Demand Forecasting Capabilities. *Journal of Institute of Transportation Engineers (ITE) 69(7)*, 28–33

Yagi, S and Mohammadian, A (2006) An activity-based microsimulation model of travel demand in the Jakarta Metropolitan Area. *The 11th International conference for Travel Behavior Research* Kyoto: IATBR

Ye, X, Konduri, K C, Pendyala, R M, Sana, B and Waddell, P (2009) Methodology to Match Distributions of Both Household and Person Attributes in Generation of Synthetic Populations. *The 88th Annual Meeting of the Transportation Research Board* Washington, D.C.: TRB