Estimating Synthetic Baseline Population Distribution when Only Partial Marginal Information is Available

Natachai WONGCHAVALIDKUL
Graduate Student
School of Civil Engineering and Technology
Sirindhorn International Institute of Technology
Thammasat University, Rangsit Campus
P.O. Box 22, Pathum Thani 12121, Thailand
Fax: +66-2-986-9112
E-mail: natachai_w@siit.tu.ac.th

Mongkut PIANTANAKULCHAI
Assistant Professor
School of Civil Engineering and Technology
Sirindhorn International Institute of Technology
Thammasat University, Rangsit Campus
P.O. Box 22, Pathum Thani 12121, Thailand
Fax: +66-2-986-9112
E-mail: mongkut@siit.tu.ac.th

Abstract: Synthetic baseline population data is one of the most important data required for the activity based travel demand model. The conventional approach to create this baseline population mainly relies on the Iterative Proportional Fitting (IPF) procedure. However, the traditional IPF procedure assumes the known input data from both the observed cell counts and their marginal counts. This paper presents the application of least square procedure for estimating baseline population distribution in the area where only partial marginal distribution data are available. The method concentrates on optimizing the least squares of the errors between the estimated conditional probability and the target conditional probability, given the constraints of underlying population information in the study area (such as total population, total population by gender, and total population by age etc.). Numerical examples and the case study of Phitsanulok city in Thailand are also presented.

Key Words: *synthetic baseline population, contingency table, activity based travel demand model*

1. INTRODUCTION

The activity based travel demand model addresses the view of travel demand analysis as the individual demand rather than the zonal basis. Hence, the detail population data are considered as the important input to the model. However, in most cases, the data are protected under the privacy and only certain kinds of summary or sample data are published. The population synthetic process helps to overcome this limitation by comprising the public data (both summary and sample data sets) and representing them as a single set of synthetic data which realistically represents population characteristics in study areas. In general, the synthetic population data are constructed base on the current population information or baseline population. The baseline synthetic population data are then used to evaluate the existing or current year model. The forecast year of synthetic population can be predicted base on either the changes of land uses or the changes of population in the study area, depending on both the modeling objective and the data availability.

Guo and Bhat (2007) classified the typical population synthesis process into three steps: 1) identifying the socio-demographic attributes desired for the modeling proposes (control variables), 2) Estimate the multi-way joint distribution which satisfies the input marginal distribution data, and 3) Create the synthetic household population data base on the sample data and the distribution gathered in step 2. This paper concentrates on the issue of estimating the multi-way join distribution for synthetic baseline population data in the case which marginal distributions of some control variables are missing. Section two discusses the problem statements of this paper in more details. Section three reviews the literatures which relate to the proposed methodology in the paper. Section four and five present the methodology and the basic numerical example, respectively. Section six presents the case study in the principal town planning area of Phitsanulok province, located in the northern part of Thailand. Section seven discusses the research summary and conclusions, as well as recommendations for further study.

2. PROBLEM STATEMENT

Beckman *et al.* (1996) originally used the Iterative Proportional Fitting (IPF) procedure for synthesizing the baseline population distribution data. The IPF procedure is currently applied for synthesizing population data in several activity-based travel demand modeling (Miller, 1996; Bradley *et al.*, 2001; Bhat *et al.*, 2003; LANL, 2003). However, the traditional IPF procedure assumes the known input data from both the initial cell counts and their marginal counts. This restricted requirement has limited the ability to synthesize the baseline population for study areas in Thailand because the completed marginal distribution data of some control variables are missing or prohibited for the privacy purposes. In other words, only some marginal distribution attributes or partial population information in the area is available.

Further, in Thailand, the data with full joint distribution of control variables are available as the sampling data set or the home interview survey data. Even though this sampling data provides the population attributes that are needed for the activity based travel demand modeling, using only the sampling data and assuming that the data is identically distributed to the real population characteristics without considering other available population information in the study area may lead to unrealistic synthetic population data which is incompatible with other available macro-statistics of the population in the study area. Hence, under this limitation on data availability, the estimation method for synthetic baseline population distribution is needed.

3. LITERATURE REVIEW

Problems of implementing the traditional IPF procedures have been discussed in several researches. In general, these problems could be classified into three types: type (a) the occurrences of incorrect zero cell values when the sample data is not consistent with the expected population distribution, type (b) the ability of IPF procedure in controlling the distribution of both household level and individual level, and type (c) incomplete marginal distribution data, missing marginal distribution of some control variables (Birkin *et al.,* 2006; Arentze *et al.*, 2007; Guo and Bhat, 2007).

These problems are the barriers in creating the synthetic population database, especially the process of initializing the join distribution data which represents characteristics of population in study areas. From literature review, the methods of estimating population distribution when the prior problems occur can be classified into two main groups: the modified IPF procedure and the application of optimization algorithms. Several modified IPF procedures were proposed in order to solve the problem type (a) and (b). Some proposed methods are also found to be used with the Markov Chain Monte Carlo (MCMC) method in order to generate the final join distribution table (Guo and Bhat, 2007; Arentze *et al.*, 2007; Frick and Axhausen, 2004). Same as the traditional IPF procedure, the modified IPF procedure is mainly based on the proportion of the sampling cell size and the marginal count of each control variables. Hence, applications of these methods are still limited in case of incomplete marginal distribution data, problem type (c). In this case, the optimization algorithms are more applicable than the modified IPF procedure. The applications of optimization procedures are variously used in the literatures. Both the classical optimization and the heuristic optimization procedures are found. For the heuristic optimization, simulate annealing and genetic algorithms were successfully tested and applied. However, the drawback on using these methods is their computational intensity (Williamson *et al.*, 1998; Birkin *et al.*, 2006). Hence, in case that the search area is not too complex, small numbers of required control variables, the classical optimization procedures are the other alternatives to solve the problem. The classical optimization procedures are expected to be simpler in the analysis procedures and less computational intensity than the heuristic procedures. Slavkovic and Fienberg (2004) applied the use of linear programming to estimate cell size in the synthetic contingency table. The results found that the linear programming gave a result table which has a close distribution to the target. Additionally, National Center for Health Statistics (2002) also applied the use of liner programming and MCMC to solve the problem type (a) and (c). Finally, due its simplicity and low computational requirement, this paper explores more alternatives in applying the classical optimization procedures to the above problem.

4. METHODOLOGY

The application of least squares procedure is proposed in this paper for estimating baseline population distribution in the area under the above aforementioned circumstances (section 2). The method concentrates on optimizing the least squares of errors between the estimated conditional probability and the target conditional probability, given the constraints of underlying population information available in the study area. The optimized result is the estimated distribution of area population characteristics.

To illustrate this problem in a simple fashion, the population contingency tables are presented by the table of $I \times J \times K$(3 ways contingency table). However, the method can be extended in the same manner to the general case of k ways contingency table. First, the general notations are summarized as the followings. Let X, Y, Z be discrete random variables with possible values $\{x_1 ,x_2, x_3, \ldots, x_i\}$,$\{y_1, y_2, y_3, \ldots, y_j\}$, and $\{z_1, z_2, z_3, \ldots, z_k\}$ respectively. Let n^*_{ijk} denote the observed cell counts (sampling data) in the table with the size of $I \times J \times K$. Additionally, the joint distribution of the table can be denoted as the $I \times J \times K$ matrix of $P^* = [p^*_{ijk}]$ or the normalized table of the observed cell count (n^*_{ijk}) table. Hence, $p^*_{ijk} = P(X^*=x^*_i,\ Y^*=y^*_j,\ Z^*=z^*_k)$ where $i=\{1, 2, 3,\ldots, I\}$, $j = \{1, 2, 3, \ldots, J\}$, and $k = \{1, 2, 3, \ldots, K\}$. Further, the conditional probability distribution can also represent as the $I \times J \times K$ matrices where $\mathrm{a}^*_{\mathrm{ijk}} = P(X^* =x^*_i|Y^*=y^*_j, Z^*=z^*_k)$, $b^*_{ijk} = P(Y^*=y^*_j|X^*=x^*_I, Z^*=z^*_k)$, $c^*_{ijk} = P(Z^*=z^*_k\ |X^*=x^*_i\ ,Y^*=y^*_j)$.

Given the observed cell count $n*_{ijk}$, the conditional probability distribution (A^*, B^*, and C^*) of the sampling size can be calculated. Additionally, the following information of the study area could differently exist (depending on the availability of data in those areas) as the followings:

- total population in the study area: $N = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ijk}$
- total population by some specific category: $n_{i..} = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}$ or $n_{.j.} = \sum_{i=1}^{I}\sum_{k=1}^{K} n_{ijk}$ or

$$n_{..k} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ijk}$$

- expectation of some specific category: $E(x_{i...}) = \sum_{i=1}^{I} n_{i..}p_{i..}$ where $n_{i..} = \sum_{j=1}^{I}\sum_{k=1}^{K} n_{ijk}$ and $p_{i..} = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk}$ or $E(x_{.j.}) = \sum_{j=1}^{J} n_{.j.}p_{.j.}$ where $n_{.j.} = \sum_{i=1}^{I}\sum_{k=1}^{K} n_{ijk}$ and $p_{.j.} = \sum_{i=1}^{I}\sum_{k=1}^{K} p_{ijk}$ or $E(x_{..k}) = \sum_{k=1}^{K} n_{..k}p_{..k}$ where $n_{..k} = \sum_{i=1}^{I}\sum_{k=1}^{K} n_{ijk}$ and $p_{..k} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ijk}$

Hence, the cell count n_{ijk} of the population in the study area can be estimate as the followings.

$$\mathit{Minimize}\ \left\{\left(\sum a^{*}_{ijk} - \sum a_{ijk}\right)^{2} + \left(\sum b^{*}_{ijk} - \sum b_{ijk}\right)^{2} + \left(\sum c^{*}_{ijk} - \sum c_{ijk}\right)^{2}\right\} \qquad (1)$$

Or $$\mathit{Minimize}\ \left\{\left(\sum p^{*}_{ijk} - \sum p_{ijk}\right)^{2}\right\} \qquad (2)$$

where,

$$a_{ijk} = P(X=x_i|Y=y_j, Z=z_k) = \frac{n_{ijk}}{\sum_{k=1}^{K}\sum_{j=1}^{J} n_{ijk}} \qquad (3)$$

$$b_{ijk} = P(Y=y_j|X=x_i, Z=z_k) = \frac{n_{ijk}}{\sum_{k=1}^{K}\sum_{i=1}^{I} n_{ijk}} \qquad (4)$$

$$c_{ijk} = P(Z=z_k\ |X=x_i, Y=y_j) = \frac{n_{ijk}}{\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ijk}} \qquad (5)$$

$$p_{ijk} = P(X=x_i, Y=y_j, Z=z_k) = \frac{n_{ijk}}{\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ijk}} \qquad (6)$$

Subject to:

/* total population in the study area */

$$\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ijk} = N \qquad (7)$$

/* total population by some specific category */

$$\sum_{k=1}^{K}\sum_{j=1}^{J} n_{ijk} = n_{i..}\ ,\ \forall\ i \qquad (8)$$

And/Or $$\sum_{k=1}^{K}\sum_{i=1}^{I} n_{ijk} = n_{.j.}\ ,\ \forall\ j \qquad (9)$$

And/Or $$\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ijk} = n_{..k}\ ,\ \forall\ k \qquad (10)$$

/* expectation of some specific category */

$$E(x_{i...}) = \sum_{i=1}^{I} n_{i..}p_{i..}\ ,\ n_{i..} = \sum_{j=1}^{I}\sum_{k=1}^{K} n_{ijk} \text{ and } p_{i..} = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk} \qquad (11)$$

And/Or $E(x_{.j.}) = \sum_{j=1}^{J} n_{.j.} p_{.j.}$, $n_{.j.} = \sum_{i=1}^{I} \sum_{k=1}^{K} n_{ijk}$ and $p_{.j.} = \sum_{i=1}^{I} \sum_{k=1}^{K} p_{ijk}$ (12)

And/Or $E(x_{..k}) = \sum_{k=1}^{K} n_{..k} p_{..k}$, $n_{..k} = \sum_{i=1}^{I} \sum_{k=1}^{K} n_{ijk}$ and $p_{..k} = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ijk}$ (13)

and $n_{ijk} \geq 1, \forall i, j, k$ (14)

From equation (1) to (14), there are two possible objective functions that can be used. However, the empirical results show that the objective function of equation (1) provides a better fit to the results than those of the equation (2). Further illustrations and discussions on this issue are in the next section. Moreover, the objective function of conditional probability, which already exists in the constraints, should be neglected because the relevant terms will become constants in the objective function. Furthermore, the values of cell counts in the contingency table that are initially calculated based on the direct expansion of the sampling data's probability distribution are a good starting point for the optimization process because the values are supposedly close to the optimized distribution.

5. ILLUSTRATIVE EXAMPLES

In this section, data are set in order to illustrate the methodology presented in section 2. Initially, two data set are given: first, the complete population data and, second, the sampling data set. Table 1 and Table 2 present these data, respectively. The illustrative scenario is assumed that, instead of complete data in table 1, only total population by age and total population by gender are given. Table 3 summarizes the illustrative scenarios.

Three different alternatives are illustrated and compared. Firstly, the probability distribution of sampling data set is directly expanded by the total number of complete population data. Secondly, the least squares optimization is performed by setting the objective function to the least squares of the conditional distribution using equation (1). Thirdly, the least squares optimization is performed by setting the objective function to the least squares of the entire sampling probability distribution using equation (2). Results of these alternative formulations are shown in the following sections.

Table 1 Complete population data

	Household Income (k)	Low (k=1)		Medium (k=2)		High (k=3)		Total ($n_{.j.}$)
	Gender of Householder (i)	Male (i=1)	Female (i=2)	Male (i=1)	Female (i=2)	Male (i=1)	Female (i=2)	
Age (j)	15-24 (j=1)	2,004	873	1,165	220	75	16	4,353
	25-34 (j=2)	2,123	2,051	1,759	1,396	954	1,023	9,306
	35-44 (j=3)	4,012	3,143	4,996	4,295	3,150	1,990	21,586
	45-54 (j=4)	1,730	1,628	3,640	4,490	7,620	4,531	23,639
	55-64 (j=5)	1,053	394	6,217	4,734	4,249	2,567	19,214
	65-74 (j=6)	2,707	3,531	4,810	5,522	1,739	2,887	21,196
	>74 (j=7)	1,933	537	4,038	2,877	2,925	1,943	14,253
Total Population by Income and Gender		15,562	12,157	26,625	23,534	20,712	14,957	9,422
Total Population by Income		27,719		50,159		35,669		

Table 2 Sampling data set

	Household Income (k)	Low (k=1)		Medium (k=2)		High (k=3)		Total ($n_{.j.}$)
	Gender of Householder (i)	Male (i=1)	Female (i=2)	Male (i=1)	Female (i=2)	Male (i=1)	Female (i=2)	
Age (j)	15-24 (j=1)	150	60	80	15	5	1	311
	25-34 (j=2)	200	180	156	124	84	90	834
	35-44 (j=3)	315	246	405	345	245	150	1,706
	45-54 (j=4)	113	102	245	312	458	316	1,546
	55-64 (j=5)	106	46	611	463	404	245	1,875
	65-74 (j=6)	245	315	450	513	156	259	1,938
	>74 (j=7)	164	46	345	246	247	164	1,212
Total Population by Income and Gender		1,293	995	2,292	2,018	1,599	1,225	9,422
Total Population by Income		2,288		4,310		2,824		

Table 3 Illustrative scenario

	Household Income (k)	Low (k=1)		Medium (k=2)		High (k=3)		Total ($n_{.j.}$)
	Gender of Householder (i)	Male (i=1)	Female (i=2)	Male (i=1)	Female (i=2)	Male (i=1)	Female (i=2)	
Age (j)	15-24 (j=1)							4,353
	25-34 (j=2)							9,306
	35-44 (j=3)	*Given:* p^*_{ijk}, a^*_{ijk}, b^*_{ijk}, c^*_{ijk} *from table 2*						21,586
	45-54 (j=4)							23,639
	55-64 (j=5)							19,214
	65-74 (j=6)							21,196
	>74 (j=7)							14,253
Total Population (N)		113,547						
Total Male Population ($n_{..k=1}$)		62,899						
Total Female Population ($n_{..k=2}$)		50,648						

5.1 Alternative 1: Direct Estimation using Probability Distribution of Sampling Data Set

In this alternative, the probability distribution of the sampling data set is calculated and then used as the expansion factor. Equation (15) to (17) presents the procedure for this alternative.

$$n_{ijk} = p^*_{ijk} \times N \qquad \text{/* estimated cell counts */} \qquad (15)$$

where,

$$p^*_{ijk} = P(X^*=x^*_i,\ Y^*=y^*_j,\ Z^*=z^*_k) \qquad \text{/* probability distribution of sampling data */} \qquad (16)$$

$$N = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ijk} \qquad \text{/* total population in the study area */} \qquad (17)$$

where,

i=1, 2 /* gender */
j = 1, 2, 3, 4, 5, 6, 7 /* age groups of householder */
k = 1, 2, 3 /* family income groups */

5.2 Alternative 2: Least Squares Optimization by Conditional Probability Distribution

The second alternative performs the least squares optimization procedures using the objective function of the least squares of conditional distribution as in equation (1). The following equations can be applied for the alternative 2. The results from the alternative 1 (table 4) are used as the initial solution of the optimization process.

$$Minimize\ \left\{\left(\Sigma\, a^*_{ijk} - \Sigma\, a_{ijk}\right)^2 + \left(\Sigma\, b^*_{ijk} - \Sigma\, b_{ijk}\right)^2 + \left(\Sigma\, c^*_{ijk} - \Sigma\, c_{ijk}\right)^2\right\} \tag{18}$$

where,

$$a_{ijk} = P(X = x_i | Y = y_j, Z = z_k) = \frac{n_{ijk}}{\sum_{k=1}^{K}\sum_{j=1}^{J} n_{ijk}} \tag{19}$$

$$b_{ijk} = P(Y = y_j | X = x_i, Z = z_k) = \frac{n_{ijk}}{\sum_{k=1}^{K}\sum_{i=1}^{I} n_{ijk}} \tag{20}$$

$$c_{ijk} = P(Z = z_k | X = x_i, Y = y_j) = \frac{n_{ijk}}{\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ijk}} \tag{21}$$

Subject to:

$$\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ijk} = N \qquad \text{/* total population in the study area */} \tag{22}$$

$$\sum_{k=1}^{K}\sum_{i=1}^{I} n_{ijk} = n.j.\,,\ \forall\, j \qquad \text{/* total population by age groups */} \tag{23}$$

$$\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ijk} = n..k\ ,\ \forall\, k \qquad \text{/* total population by gender */} \tag{24}$$

where,

i=1, 2 /* gender of householder */
j = 1, 2, 3, 4, 5, 6, 7 /* age groups of householder */
k = 1, 2, 3 /* family income groups */

5.3 Alternative 3: Least Squares Optimization by Total Probability Distribution

The third alternative performs the least squares optimization procedures using the objective function of the least squares of probability distribution as in equation (2). Equation (25) to (29) can be applied for the alternative 3. Same as the alternative 2, the results from the alternative 1 (table 4) are used as the initial solution of the optimization process.

$$Minimize\ \left\{\left(\Sigma\, p^*_{ijk} - \Sigma\, p_{ijk}\right)^2\right\} \tag{25}$$

where,

$$p_{ijk} = P(X = x_i\ Y = y_j\ Z = z_k) = \frac{n_{ijk}}{\Sigma_i \Sigma_j \Sigma_k n_{ijk}} \tag{26}$$

Subject to:

$$\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ijk} = N \qquad \text{/* total population in the study area */} \tag{27}$$

$$\sum_{k=1}^{K}\sum_{i=1}^{I} n_{ijk} = n.j.\,,\ \forall\, j \qquad \text{/* total population by age groups */} \tag{28}$$

$$\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ijk} = n..k\ ,\ \forall\, k \qquad \text{/* total population by gender groups */} \tag{29}$$

5.4 Results Comparisons and Conclusions

Goodness of fit tests (χ^2) are individually performed in each alternative (1-3) against the complete population data in table 1. The results show that the alternative 2 (χ^2 = 156.14) provides the best fits to the cell counts in table 1, following by alternative 3 (χ^2 = 216.70) and alternative 1 (X^2 = 2,198.27), respectively. Figure 1 summarizes and compares the results.

In summary, the results show that, by using the objective function on the least squares of conditional distribution, equation (1), the approximated cell counts will provide more promising distribution to the target than those objects the function on the least squares of the entire sampling probability distribution, equation (2). On the other hand, directly using sampling data's probability distribution as the expansion factor will lead to the irrelevant results between approximated population characteristics and the real population characteristics in the area.

Alternative 1	Alternative 2	Alternative 3
χ^2 = 2,198.27	χ^2 = 156.14	χ^2 = 216.70

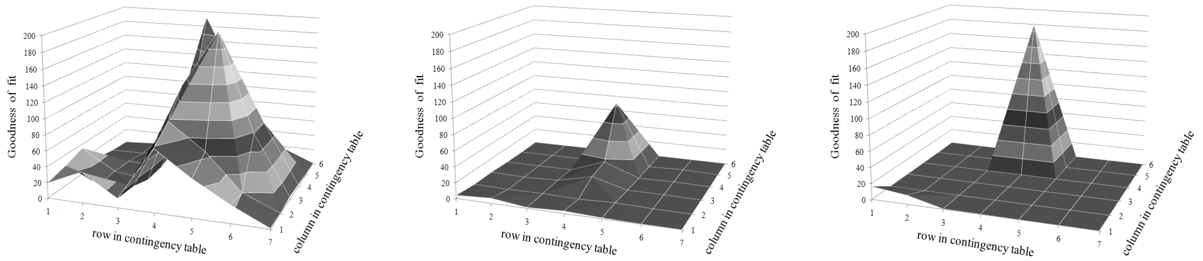

Figure 1 Error distributions from the contingency table of alternative 1, 2, and 3

6. CASE STUDY

6.1 General

Phitsanulok city, located in the northern part of Thailand, was selected as the pilot site for the first activity based travel demand modeling in Thailand. The city covers approximately 709.61 square kilometer. In year 2007, the number of population under the registration record in this area is 271,110 in totals, which 130, 830 and 140,280 are male and female respectively. This number of population was approximately 2 percent decreasing from the previous year. Figure 2 presents the Phitsanulok city boundary, percent of population in each jurisdiction, principal city planning boundary, and the study area. Further, the summary records of the population data show that 1,758 missing data (unrecorded) on age of the population. Therefore, the population data in the study area used for estimating the new baseline population distribution are 267,148 in total, with 128,508 and 138,640 are male and female respectively. Table 4 summarizes the available population statistics from the National Statistical Office (NSO) which could be gathered and used in this case study.

To gather more details of population data, the study area is divided into 78 zones and the home interview survey was taken, in year 2005. Figure 3 presents boundaries and the total number of households in each zone. Further, 1,894 households out of 60,677 households (approximately 3.1 percent) were randomly (and spatially) interviewed in the area. Even though the size of this sampling data set is quite promising for the entire study area, the sampling data is not appropriate for some zones. This is the result from in adequate sampling in some zones. Hence, some area has very small percentage of sampling data. Note that, the percentage of sampling data ranges from 0.40 percent to 7.47 percent. Figure 4 presents the percentage of sampling households to the total household in each zone. Furthermore, all statistics for the study area are mainly available on the area basis, rather than the zonal basis.

To this limitation, the population distribution from a sampling data set are first required to fit to the marginal distribution of available population information in the study area then the individual population characteristics will be rather generated and spatially distributed in the area.

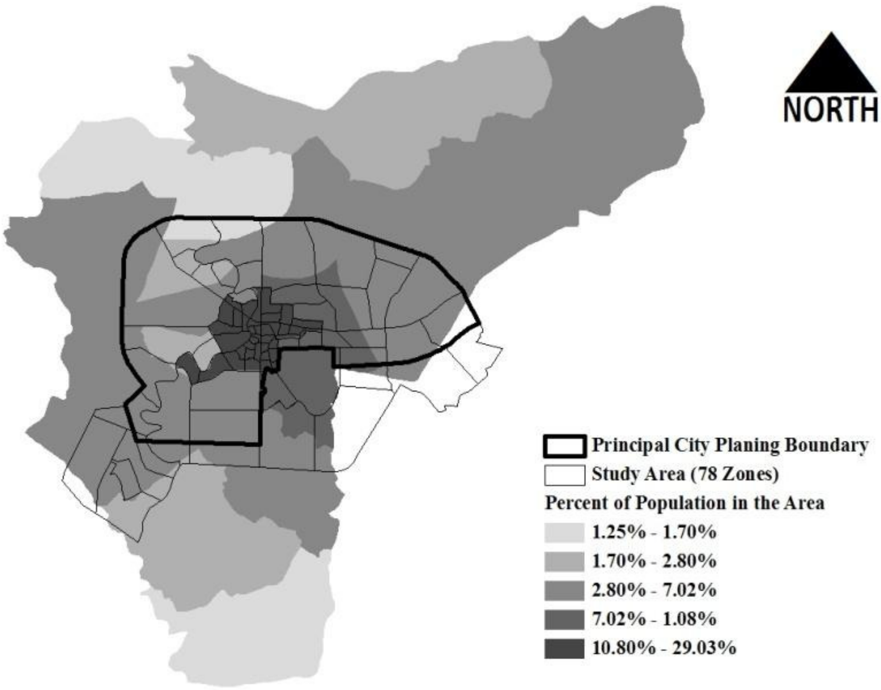

Figure 2 Percent of population in each jurisdiction comparing to the study area

Table 4 Available population statistics in the study area

Age	*Male*	*Female*	*Total*
0-4	7,115	6,735	13,850
5-9	7,803	7,586	15,389
10-14	10,083	9,668	19,751
15-19	10,898	11,898	22,796
20-24	12,302	10,727	23,029
25-29	10,312	10,479	20,791
30-34	10,005	10,885	20,890
35-39	10,267	11,465	21,732
40-44	11,132	12,889	24,021
45-49	11,088	11,918	23,006
50-54	8,390	9,730	18,120
55-59	6,109	7,460	13,569
60-64	3,930	4,802	8,732
65-69	3,552	4,312	7,864
70-74	2,651	3,535	6,186
75-79	1,643	2,398	4,041
80-84	751	1,329	2,080
sed>= 85	477	824	1,301
Total	128,508	138,640	267,148

Source: Modified from National Statistical Office (NSO) population data

6.2 Data Manipulation and Model Formulation

6.2.1 Treatment of Missing Sampling Data

Table 5 presents population variables that are gathered from the survey data. After reviewing the data completeness, there are total 1,824 records missing on both the household income and the individual income or either of these attributes. There are numbers of techniques to treat the problem of missing data. In this paper, the listwise deletion which is considered as the simplest method is used. Further, the missing data of these income variables are found to be randomness in nature. Hence, even though, listwise deletion method causes the decreasing in sample size, it ensures that the sample size used to estimate baseline population distribution in this study will not lead to unbiased estimates. As a result, there are 4,948 complete sample records left for synthesizing the baseline distribution population in the study area.

Table 5 Population variables from survey data

Variables	Class	Category	Description
total_mem	Household	1-10	Number of member in the household
total_hhinc		1-4	Total household incomes 1:low (<5000) 2:medium (5000-14999) 3:high(15000-29999) 4:overly high (>30000)
total_veh		0-12	Number of vehicles available in the household
triph		0-24	Number of total trip in the household
pc		0-12	Number of passenger car available in the household
lt		0-12	Number of pickup available in the household
van		0-12	Number of van available in the household
mc		0-12	Number of motorcycle available in the household
ht		0-12	Number of truck available in the household
sb		0-12	Number of small bus available in the household
bus		0-12	Number of bus available in the household
bc		0-12	Number of bicycle/tricycle available in the household
tc		0-12	Number of tuk tuk available in the household
ot		0-12	Number of other vehicles available in the household
relation	Individual	1-6	Relation to the householder 1: householder, 2:spouse, 3:son or daughter / son or daughter - inlaw, 4:father/mother, 5: relatives, 6: servant
person		1-10	Order of memberships in the household
sex		1-2	Gender 1:male 2:female
age		1-18	Age 1: 0-4, 2:5-9, 3:10-14, 4:15-19, 5:20-24, 6:25-29, 7:30-34, 8:35-39, 9:40-44, 10:45-49, 11:50-54, 12:55-59, 13:60-64, 14:65-69, 15:70-74, 16:75-79, 17:80-84, 18: 85 or Greater
education		1-10	Education 1:non, 2:kindergarten or lower, 3:elementrary, 4:junior high school, 5: senior high school/ vocational cert, 6: diaploma or bachelor, 7: graudated diaploma, 8: graudated bachelor, 9: graduated higher than bachelor, 10: others
career		1-11	Career 1:unemployment, 2:government, 3:private sector, 4: worker, 5:technician, 6:commerce, 7:student, 8:business owner, 9:broker, 10:agriculture, 11:retirement
inc		1-4	Individual income 1:low (<5000) 2:midium (5000-14999) 3:fairly high (15000-29999) 4:high (>30000)
drivec		1-2	Ability to drive a car 1: can, 2: can not
drivemc		1-2	Ability to drive a motorcycle 1: can, 2: can not
trip		0-10	Total trips

6.2.2 Data Manipulation

From table 8, there are 28 variables with high categories in each variable. Additionally, the survey data of 1,894 sampling households contains total 6,772 individual sampling populations. By ignoring the 0 cell combination, the data produces 6,686 population and household group of characteristics. On the other hand, the data produces 28 ways contingency table with 6,686 cells. As for the limited space and for illustrative purpose of this paper, the relevant variables in table 8 and its appropriate categories are manipulated. Table 9 presents the selected variables. Additionally, by ignoring the 0 cell count, the variables in table 9 could be able to group the population characteristics into 193 groups from total 6,772 sampling population (5 ways contingency table with 193 cell counts). As a results there are two sources of data in this case study, first, the total population data from NSO presented in table 7 and, second, the manipulated sampling data set which the variables presented in table 6.

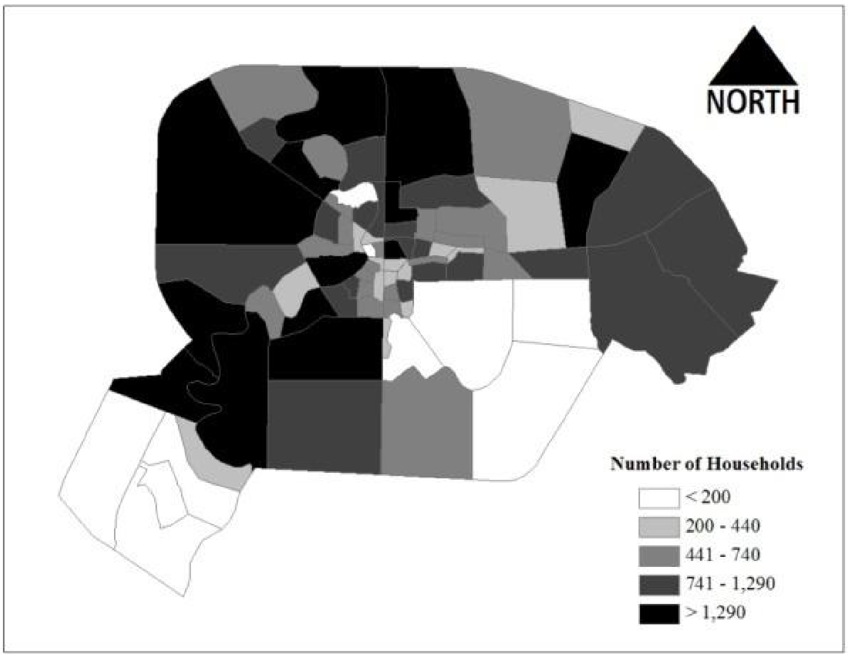

Figure 3 Study zones and distribution of total number of households

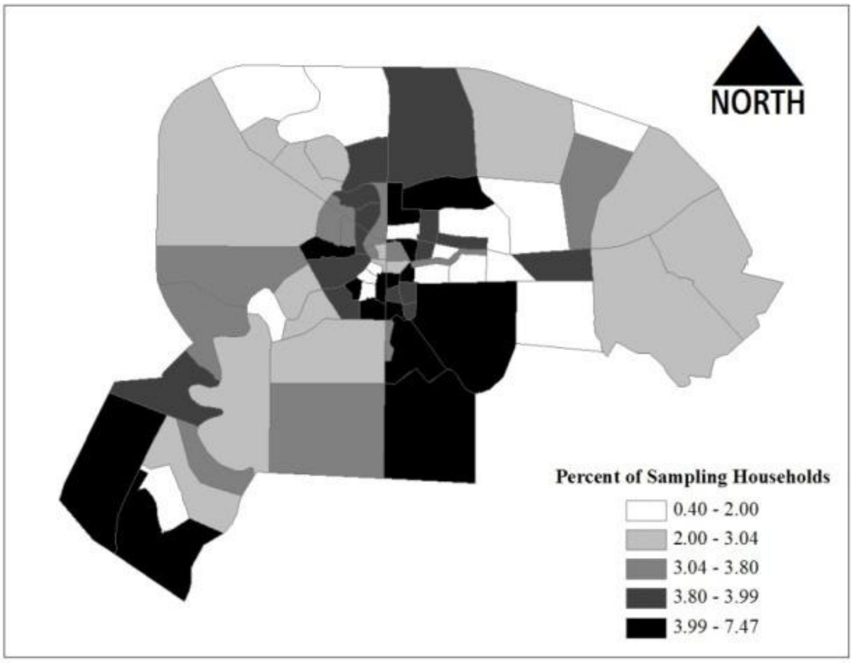

Figure 4 Percent of sampling households

Table 6 Population variables used for creating a synthetic population distribution in this paper

Variables	Class	Category	Description
hh_size (*i*)	Household	1-2	Size of the household 1: 1 or 2 persons 2: 3 or Greater
total_hhinc (*j*)		1-3	Total household incomes 1:low (< 15,000) 2:medium (15,000-30,000) 3:high (>30,000)
relation (*k*)	Individual	1-5	Relation to the householder 1: householder, 2:spouse, 3:son or daughter / son or daughter - inlaw, 4:father/mother, 5: others
sex (*l*)		1-2	Gender 1:male 2:female
age (*m*)		1-4	Age 1: 0-14, 2:15-29, 3:30-49, 4:50 or Greater

6.2.3 Model Formulation

As the results from section 3, the least squares optimization method with the objective function on the conditional probability is used for this case study. Equation (30) to (34) presents the model formulation for this case study.

$$Minimize\ \left\{\left(\sum a^*_{ijklm} - \sum a_{ijklm}\right)^2 + \left(\sum b^*_{ijklm} - \sum b_{ijklm}\right)^2 + \cdots + \left(\sum e^*_{ijklm} - \sum e_{ijklm}\right)^2\right\} \quad (30)$$

where,

$a_{ijklm}, b_{ijklm}, \ldots., e_{ijklm}$ = estimated conditional probability

$a^*_{ijklm}, b^*_{ijklm}, \ldots., e^*_{ijklm}$ = conditional probability from sampling data

Subject to:

$$\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ijklm} = n_{...m}, \ \forall m, \qquad \text{/* total population by age group */} \quad (31)$$

$$\sum_{m=1}^{M}\sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ijklm} = n_{..k..}, \ \forall k, \qquad \text{/* total population by gender */} \quad (32)$$

where,

i=1, 2 /* size of the household */
j = 1-4 /* total household income groups */
k = 1-6 /* relation to the householder */
l = 1-2 /* gender */
m = 1-4 /* age groups */

6.3 Summary of Results and Discussions

As the results, total 267,148 populations are successfully distributed over the 182 groups of population characteristics using the least squares optimization procedure. Table 7 summarizes these results based on each variables selected for the case study. This estimated entire area population distribution can produce several category population distributions which are required in order to generate the synthetic population data in the area. Hence, this result provides the key to the next step in creating the synthetic baseline population data. Figure 5 presents examples of baseline synthetic population distribution in the selected categories. Finally, table 8 compares the conditional probabilities of sampling data and the estimated conditional probabilities, as well as square errors.

7. SUMMARY AND CONCLUSIONS

This paper presents the application of least squares procedure for estimating baseline population distribution in the area where only partial distribution data are available. From the results of numerical example, it was found that the least squares optimization procedure which uses the objective function on conditional distribution gives the better approximate distribution than both the method of the least squares optimization procedure which uses the objective function based on the total probability distribution and the method on direct expansion using the sampling data. Further, the method was also applied in the real case study, the results show that it successfully estimated the population distribution in the study area using the partial information given by NSO and the population survey data.

Both the numerical example and the case study present in this paper were evaluated under the spread sheet program which has its limitation of capabilities in holding the constraints. Hence, the applications of proposed methodology presented in this paper are still limited to the problem with simpler dimensional contingency table. The evaluation of this methodology with higher dimensional contingency table (more control variables involved) is needed. Finally, the algorithm which is used to solve the problems in this paper is basically based on the Newton's method. Therefore, other algorithms should be also suggested and evaluated in order to improve the estimated distribution results.

Table 7 Results of synthetic population

Variables	*Category*					*Total*
	1	2	3	4	5	
hh_size	86,185	180,963				
total_hhinc	117,568	87,442	62,138			
relation	76,781	47,970	62,662	31,979	47,755	267,148
sex	128,508	138,640				
age	48,990	66,616	89,649	61,893		

Note: see table 6 for category and variable descriptions

Table 8 Comparisons of conditional probability distribution between the sampling data and the estimated data

Variables	*Category*				
	1	2	3	4	5
	Conditional Probability Distribution from Sampling Data				
hh_size	0.247	0.753			
total_hhinc	0.479	0.340	0.181		
relation	0.362	0.223	0.268	0.028	0.118
sex	0.484	0.516			
age	0.093	0.239	0.413	0.255	
	Estimated Conditional Probability Distribution				
hh_size	0.323	0.677			
total_hhinc	0.440	0.327	0.233		
relation	0.287	0.180	0.235	0.120	0.179
sex	0.481	0.519			
age	0.183	0.249	0.336	0.232	
	Square Errors				
hh_size	0.006	0.006			
total_hhinc	0.002	0.000	0.003		
relation	0.006	0.002	0.001	0.008	0.004
sex	0.000	0.000			
age	0.008	0.000	0.006	0.001	

Note: see table 6 for category and variable descriptions

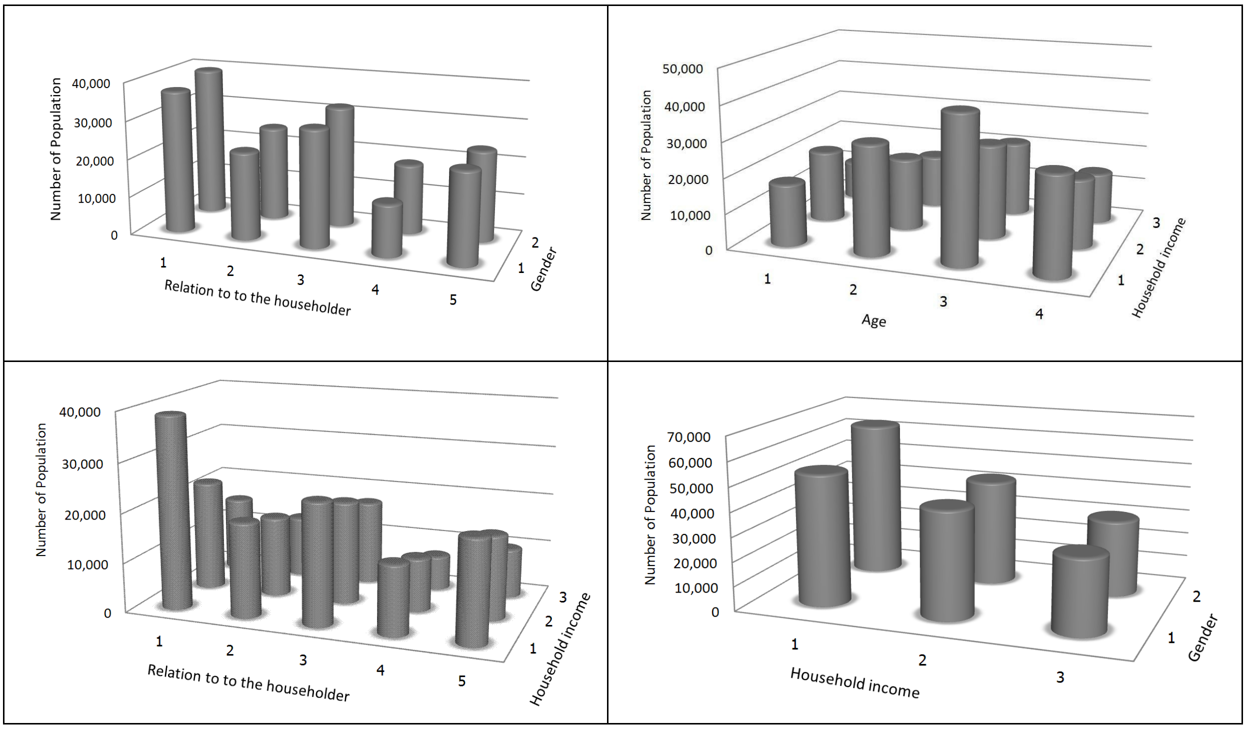

Figure 5 Examples of baseline synthetic population distribution
(**Note:** see table 6 for category and variable description)

ACKNOWLEDGEMENTS

The authors acknowledge the data support on the case study from the TDML project under the supervision of the Office of Transport and Traffic Policy and Planning (OTP). Further, staffs members from OTP and the PCBK Co. Ltd. are thanked for their assistance, encouragement, and guidance.

REFERENCES

Arentze, T., Timmermans, H., and Hofman, F. (2007) Creating Synthetic Household Population, **Transportation Research Record, No. 2014,** 85-91.

Beckman, R. J., Baggerly, K. A., and McKAY, M. D. (1996) Creating Synthetic Baseline Populations, **Transportation Research Part A, Vol. 30, No.6,** 415-429.

Bhat, C.R., Guo, J., Srinivasan, S., and Sivakumar, A. (2003) **Activity-Based Travel Demand Modeling for Metropolitan Areas in Texas: Software-related Processes and Mechanisms for the Activity-Travel Pattern Generation Micro-Simulation, Report No. 4028-5**, Texas Department of Transportation.

Birkin, M., Turner, A., and Wu, B. (2006) A Synthetic Demographic Model of the UK Population: Methods, Progress and Problems, Presented at 2nd International Conference on e-Social Science.

Bradley, M. Outwater, M. L., Jonnalagadda, N., and Ruiter, E. R. (2001) Estimation of Activity-Based Microsimulation Model for San Francisco, Presented at 80th Annual Meeting of the Transportation Research Board, Washington, D.C., January 2001.

Frick, M. and Axhausen, K. W. (2004) Generating Synthetic Populations using IPF and Monte Carlo Technique: Some New Results, Presented at 4th Swiss Transport Research

Conference**,** Ascona, March 2004.

Guo, J. Y. and Bhat, C. R. (2007) Population Synthesis for Microsimulating Travel Behavior, **Transportation Research Record, No. 2014,** 92-101.

Los Alamos National Laboratories (LANL) (2003) **TRANSIMS-Version 3.0 Vol 3-Modules: Chapter 2 Population Synthesizer, Report No. LA-UR-00-1725**, Los Alamos National Laboratories, Los Alamos, N.Mex.

Miller, E. J. (1996) Microsimulation and Activity-Based Forecasting, Presented at TMIP Conference, June 1996.

National Center for Health Statistics (2002) Imputing Missing Values in Two-Way Contingency Tables using Linear Programming and Markov Chain Monte Carlo, Presented at Conference of European Statisticians, Statistical Commission and Economic Commission for Europe, Finland, May 2002.

Slavkovic, A. B. and Fienberg, S. E. (2004) Bounds for Cell Entries in Two-Way Tables Given Conditional Relative Frequencies, **Lacture Notes in Computer Science, Vol. 3050**, 30-43.

Williamson, P., Birkin, M., and Rees, P.H. (1998) The Estimation of Population Microdata by using Data from Small Area Statistics and Samples of Anonymised Records, **Environment and Planning A, Vol. 30, No. 5**, 785-816