© Statistical Research and Training Center

مجلّه‌ی پژوهش‌های آماری ایران

پژوهشکده‌ی آمار

J. Statist. Res. Iran **16** (2019): 465–490

دوره‌ی ۱۶، شماره‌ی ۲، پاییز و زمستان ۱۳۹۸، صص ۴۶۵–۴۹۰

DOI: 10.52547/jsri.16.2.465

DOR: 20.1001.1.25385771.2020.16.2.8.8

Simulated Synthetic Population Projection Using an Extended Model

Mohammad Taghi Moeti, Hamidreza Navvabpour* and Farzad Eskandari

Allameh Tabataba'i University

Received: 2021/10/09 Approved: 2021/11/10

Abstract. Population projections of small areas have attracted the attention of many researchers in applied demographics for many years. According to the suggested population policies of Iran in recent years and considering the need of different governmental agencies for having enough information about population and individual characteristics in small areas, studying and presenting an appropriate model of population projections for small areas seems more necessary than ever. Given that today not only population projections include estimating the number of populations and identifying their specific characteristics, but also more projections are likely to project different required characteristics of organizations. The present study attempts to introduce a model for population projections in small areas. In this study, "city" is considered as a small area. For the purpose of surveying population projection between two censuses in Iran, 2006 and 2011, Mahallat, a central city in this country, has been selected among many cities since its geographical area has not been changed from 1996 to 2011. Hence, the present article projects simulated synthetic population in 2011 with distinctive characteristics of 2006 population by presenting an extended model and comparing it with projected population from the existing model.

Keywords. Sample-based; interactive proportional updating; simulation of population; small area; synthetic population.

* Corresponding author

MSC 2010: 62-07, 62P99.

1 Introduction

Population projections are mostly used in assessing national or regional needs, e.g., in the field of new jobs and urban settlements. Therefore, population projection gives a picture of what the future size and structure of the population by sex and age might look like. Knowing the size of population and its characteristics not recorded in any censuses (smoking by age groups or Household fuel consumption) and its changes in years between two consecutive population and housing censuses, has led demographers to invent and use new models for population projection. For example, knowing per capita energy consumption of households in small areas, according to their size and type considerably helps planners to provide the required funds to establish the necessary infrastructures for using clean and optimal energy consumption in the areas.

To access information on such populations with different characteristics gathered in the census, the need to simulate synthetic population in the year of the census implementation (the base year) and obtain multi-dimensional distributions of households and individual characteristics not found in census results, seems necessary. For example, no small area information can be found showing the joint distribution of electricity consumption table in terms of household sizes, while the distribution of electricity consumption and the distribution of household size are marginally achievable from the census data. At the moment, it seems necessary to simulate the population showing these joint distributions.

There exist two main category of methods for simulating populations: the synthetic reconstruction (SR) methods and the combinational optimization (CO) methods.

While SR methods (Wilson and Pownall, 1976) generally use the iterative proportional fitting (IPF) technique (Deming and Stephan, 1940) along with a sample of the target population to obtain joint distribution variables of interest, CO methods generate a synthetic population by small area using marginal of the characteristics of interest and a sample of the target population in each small area (Voas and Williamson, 2001; Huang and Williamson, 2002). To improve these methods, various methods have been proposed by

Beckmanand et al. (1996), Huang and Williamson (2002), Guo and Bhat (2007), and Arentze et al. (2007).

A sample-free synthetic reconstruction method developed by Barthelemy and Cornelis (2012) and Gargiulo et al. (2010). A sample-based iterative proportional updating (IPU) method developed by Ye et al. (2009) for generating a synthetic population. lenormand and Deffuant (2013) focus on the iterative proportional updating developed by Ye et al. (2009) and compared the sample-based method with the sample-free method on an example.

Recently, weaknesses of the sample-based method (for example zero-cell and weights less than one) to simulate synthetically the population of selected small area with different characteristics for the base year were modified by Moeti and Navvabpour (2015). Navvabpour et al. (2015) by using sample-based and sample-free methods simulated the population synthetically and projected the population in four selected cities of Iran with some characteristics between the two censuses, 2006 and 2011, through a demographic model. In this article, the sample-based simulation method is used to reconstruct the population of Mahallat city synthetically in the base year (2006). Mahallat is chosen as a small area overlapping with the work of Navvabpour et al. (2015) due to its lack of geographical changes in recent years. Moreover, the population of this city is projected by applying an extended model in the years between the two censuses of 2006 and 2011. The results are compared to those presented by Navvabpour et al. (2015).

The rest of this paper is organized as follows: In Section 2 we presents characteristics of the reference population and the method for constructing synthetic population. In Section 3 we introduce the model used by Navvabpour et al. (2015) for population projection. In Section 4 an extended model for population projection is presented. Section 5 of this work is concerned with the criteria for the performance evaluation of both models, and provides a comparison of the population projection through these models based on the presented criteria. In Section 6, an application of the synthetic population construction to the electricity industry is investigated according to the recent characteristics. Discussion and conclusion are discussed in Section 7.

2 Individual and Household Characteristics

Given the use of the following characteristics in Navvabpour et al. (2015), we also considered these characteristics to compare results of using the proposed model and the demographic model.

- Age interval: [0-1), [1-2), ..., [85-∞),
- Sex: male, female, and
- Settlement area: urban, rural.

2.1 Reference Population

In this article, Mahallat is considered as reference city. In the 2006 census, the city had 14163 households and a population of 49246. To simulate and project the population of this city along with individual and household joint characteristics, frequency distribution of individual sex, distribution of individual ages within households and distribution of settlement area for the population of Mahallat are needed. The frequency distributions are available from the data center of the Statistical Center of Iran. In the following subsection, more explanation will be given concerning the simulation of the synthetic population of Mahallat.

2.2 Synthetic Simulation of Population

There are several methods for simulating the population; sample-based and sample-free methods are two of them. The sample-free method is less data demanding but it requires more data pre-processing. Indeed, this approach requires to extract the distributions for affecting individual into household from data (for example distribution of individual by activity status according to the age and Joint-distribution of household by type and size). The sample-based method starts with a sample of population. The purpose is to define a weight associated with each individual and each household of the sample in order to match the total number of each type of individuals and households to reconstruct population. In this article, the sample-based method is used for simulation of the synthetic population of Mahallat.

The characteristics of individuals and households are listed as below:

- u_1-u_{86}: urban men from the first age interval [0-1) to the age interval [85-∞),
- u_{87}-u_{172}: urban women from the first age interval [0-1) to the age interval [85-∞),

- u_{173}-u_{258}: rural men from the first age interval [0-1) to the age interval [85-∞), and
- u_{259}-u_{344}: rural women from the first age interval [0-1) to the age interval [85-∞).

where u_i (i=1,2, ...,344) represents the characteristics of individuals and households in an IPU table in sample-based method (lenormand and Deffuant, 2013; Moeti and Navvabpour, 2015).

3 Methods of Population Projections in Years Between Two Censuses

After simulating the population with desired characteristics synthetically, the same population with new characteristics will be projected for the years between two censuses (see figure 1). Information required for population projection is derived from 2% of sample data file of the 2006 Iran population and housing census. In addition to the available data from the 2006 census, information about the vital events such as birth and death during the years leading to the general population and housing census in 2011 is required. In this regard, data related to the vital events registered by the Civil Register Organization of Iran from 2006 to 2011 are used to project population in the small area, Mahallat, in the years between 2006 and 2011 censuses in Iran.

3.1 A Population Projection Model

A demographic model is commonly used to project populations. This model is based on the base year population. The population has decreased by death and emigration, and increased by birth and immigration. Navvabpour et al. (2015) used model I introduced in Equation 1. The model I is given as:

$$\begin{aligned} P_{a_i}^{Y} &= \beta\times P_{a_{i-1}}^{Y-1} - \beta\times P_{a_{i-1}}^{Y-1}\times D_{a_{i-1}}^{Y} \\ &+ \beta\times P_{a_{i-1}}^{Y-1}\times IM_{a_{i-1}}^{Y} - \beta\times P_{a_{i-1}}^{Y-1}\times EM_{a_{i-1}}^{Y} \\ &+ (1-\beta)\times P_{a_i}^{Y-1} - (1-\beta)\times P_{a_i}^{Y-1}\times D_{a_i}^{Y} \\ &+ (1-\beta)\times P_{a_i}^{Y-1}\times IM_{a_i}^{Y} - (1-\beta)\times P_{a_i}^{Y-1}\times EM_{a_i}^{Y} \\ &= \beta\times P_{a_{i-1}}^{Y-1}\times(1-D_{a_{i-1}}^{Y}+IM_{a_{i-1}}^{Y}-EM_{a_{i-1}}^{Y}) \\ &+ (1-\beta)\times P_{a_i}^{Y-1}\times(1-D_{a_i}^{Y}+IM_{a_i}^{Y}-EM_{a_i}^{Y}) \end{aligned} \qquad (1)$$

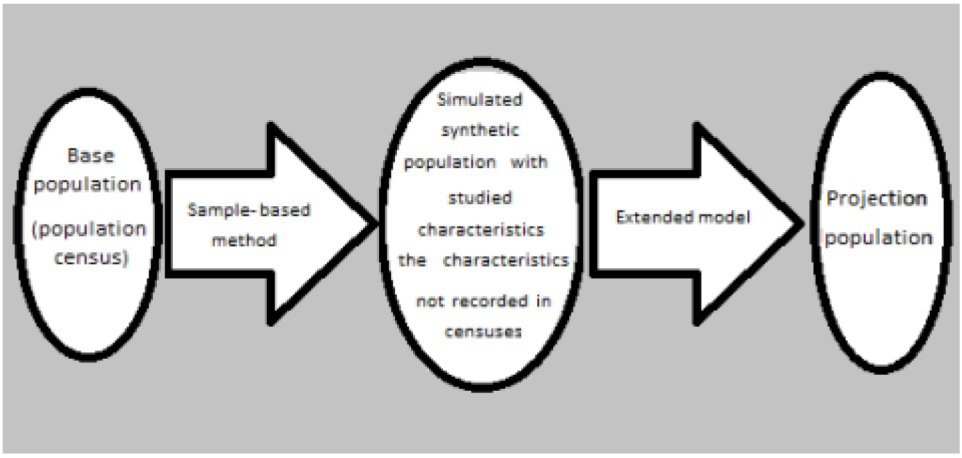
Base population (population census)
Sample- based method
Simulated synthetic population with studied characteristics the characteristics not recorded in censuses
Extended model
Projection population

Figure 1. The process of converting the base population to the projection population.

where a_i: i^{th} 5-year age interval, $i = 1, 2, ..., 18$,
$P_{a_i}^{Y}$: Population of age interval "i^{th}" in year Y,
$D_{a_i}^{Y}$: Mortality rate for age interval "i^{th}" in year Y,
$IM_{a_i}^{Y}$: immigration rate for age interval "i^{th}" in year Y,
$EM_{a_i}^{Y}$: emigration rate for age interval "i^{th}" in year Y, and
β : The percentage of people coming from one age interval to the next age interval.

In model I, by considering $\beta = 0.2$, the terms of the model can be rewritten as follows:

$0.2 \times P_{a_{i-1}}^{Y-1}$ represents 20% of $(i-1)^{th}$ age interval, $0.8 \times P_{a_i}^{Y-1}$ represents 80% of i^{th} age interval in the previous year, and the expressions in parentheses indicate the effect of the rate of vital events. Navvabpour et al. (2015) simulated the population of Mahallat by using the sample-based method synthetically and used model I to project the population between two censuses of 2006 and 2011.

Regarding the birth event, due to the rate distortion, the number of registered births was collected directly from the registration information of the Civil Registration Organization. Distribution of immigration and emigration to urban and rural areas can be extracted from the census using available information. Ofcourse, regarding the distribution of immigration, the pattern of immigration to rural areas has been modified (see Navvabpour et al. (2015)).

4 An Extended Model for Projecting Population

To project population of the coming years from the population of the base year (usually the census year), birth, mortality and migration rates are estimated and population of the coming years from 5 to 10 age interval is estimated through various models such as exponential, regression and time series models. In this article, by introducing an extended model, the population simulation results of Mahallat in 2006 are projected for the years between 2006 and 2011 Iran censuses. The final results are compared to those reported by Navvabpour et al. (2015).

The proposed model II (introduced in Equation 2) consists of three parts; the first part includes the first age interval ($i = 0$), the second part includes the other age intervals ($i = 1, 2, ..., \frac{80}{a}$) and the third part includes the last age interval ($i = \frac{80}{a} + 1$).

$$\begin{aligned} P^{Y+k}_{[ai,ai+a)} &= \sum_{j=1}^{a} \Big[P^{Y+(k-j+1)I(j\leq k)}_{[ai+j-k-1,ai+j-k)I(j-k\in\{2,3,4\})+[0,1)I(j-k\notin\{2,3,4\})} \times \\ &\quad \prod_{m=1}^{k} \left(C^{Y+m}_{[ai,ai+a)} \right)^{I(j+m>k+1)} \Big] \qquad i=0 \end{aligned} \tag{2}$$

$$\begin{aligned} P^{Y+k}_{[ai,ai+a)} &= \sum_{j=1}^{a} \Big[P^{Y}_{[ai+j-k-1,ai+j-k)} \times \\ &\quad \prod_{m=1}^{k} C^{Y+m}_{[a(i-I(j+m\leq k+1)),a(i-I(j+m\leq k+1))+a)} \Big] \qquad i = 1, 2, ... \frac{80}{a} \end{aligned}$$

$$\begin{aligned} P^{Y+k}_{[ai,\infty)} &= \sum_{j=1}^{\infty} \Big[P^{Y}_{[ai+j-k-1,ai+j-k)} \times \\ &\quad \prod_{m=1}^{k} C^{Y+m}_{[a(i-I(j+m\leq k+1)),a(i-I(j+m\leq k+1))+a)} \Big] \qquad i = \frac{80}{a} + 1 \end{aligned}$$

where:

$ai = a \times i$,

$C^{j}_{[l,u)} = 1 - D^{j}_{[l,u)} + IM^{j}_{[l,u)} - EM^{j}_{[l,u)}$,

$D^j_{[l,u)}$= Mortality rate of population in l to u age interval in j^{th} year,
$IM^j_{[l,u)}$= immigration rate in l to u age interval in j^{th} year,
$EM^j_{[l,u)}$= emigration rate in l to u age interval in j^th year,
$P^j_{[l,u)}$ = Population of l to u age interval in j^{th} year,
$P^j_{[0,1)}$ = Number of births in j^{th} year for $i = 0$ and $k = 1, 2, ...,$ and

$$I(x) = \begin{cases} 1 & x \text{ is True} \\ 0 & \text{otherwise} \end{cases}$$

In addition, Y represents the base year, $a(= 5$ or $10)$ represents age interval length, $k(= 1, 2, ...)$ represents the number of years in which projection will be done.

4.1 Model Features

We express two important features of Model II in Theorems 1 and 2.

***Theorem* 1.** *If in model II, in the base year (Y) and in each age interval, the ratios of one-year age in that age interval are equal, the population project for next year by Model II (That is, year $Y + 1$ or $k = 1$) is equal to that of Model I, where*

$$\hat{\beta} = \begin{cases} 0.2 & a = 5, \\ \\ 0.1 & a = 10. \end{cases}$$

Proof. See proof in Appendix 1. □

***Theorem* 2.** *Suppose Model I is rewritten as follows:*

$$P^{Y+k}_{[ai,ai+a)} = \beta_1 \times P^{Y+k-1}_{[ai-a,ai)} \times C^{Y+k}_{[ai-a,ai)} + (1 - \beta_2) \times P^{Y+k-1}_{[ai,ai+a)} \times C^{Y+k}_{[ai,ai+a)}$$

in the base year (Y) and in each age interval, the population project for next year by Model II is equal to that of Model I, where

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}1 \\ \hat{\beta}2 \end{bmatrix} = \begin{bmatrix} \frac{P^Y_{[ai-1,ai)}}{\sum_{n=1}^{a} P^Y_{[ai-a+n-1,ai-a+n)}} \\ \frac{P^Y_{[ai+a-1,ai+a)}}{\sum_{n=1}^{a} P^Y_{[ai+n-1,ai+n)}} \end{bmatrix} = \begin{bmatrix} \frac{P^Y_{[ai-1,ai)}}{P^Y_{[ai-a,ai)}} \\ \frac{P^Y_{[ai+a-1,ai+a)}}{P^Y_{[ai,ai+a)}} \end{bmatrix}$$

Proof. See proof in Appendix 2. □

Given that the base year is the census year of 2006. Age intervals include 5 years, and population projection years are from 2007 to 2011. Therefore, $Y = 2006$, $k = 1, 2, ..., 5$, $a = 5$, and model II can be rewritten as Equation 3:

$$\begin{aligned} P^{2006+k}_{[5i,5i+5)} &= \sum_{j=1}^{5} \Big[P^{2006+(k-j+1)I(j\leq k)}_{[5i+j-k-1,5i+j-k)I(j-k\in\{2,3,4\})+[0,1)I(j-k\notin\{2,3,4\})} \times \\ & \prod_{m=1}^{k} \left(C^{2006+m}_{[5i,5i+5)} \right)^{I(j+m>k+1)} \Big] \qquad i=0 \end{aligned} \tag{3}$$

$$\begin{aligned} P^{2006+k}_{[5i,5i+5)} &= \sum_{j=1}^{5} \Big[P^{2006}_{[5i+j-k-1,5i+j-k)} \times \\ & \prod_{m=1}^{k} C^{2006+m}_{[5(i-I(j+m\leq k+1)),5(i-I(j+m\leq k+1))+5)} \Big] \qquad i=1,2,...16 \end{aligned}$$

$$\begin{aligned} P^{2006+k}_{[5i,\infty)} &= \sum_{j=1}^{\infty} \Big[P^{2006}_{[5i+j-k-1,5i+j-k)} \times \\ & \prod_{m=1}^{k} C^{2006+m}_{[5(i-I(j+m\leq k+1)),5(i-I(j+m\leq k+1))+5)} \Big] \qquad i=17 \end{aligned}$$

5 Performance Evaluation Criteria

To test the performance of the model II, individuals and households along with their characteristics in reference small area were reconstructed using the sample-based algorithm method. The mentioned algorithm was run by R software for the selected city. To use the sample-based algorithm we need a sample of filled households and marginal variables. In order to obtain these data, 2% of 2006 census households are selected and the dimensional marginals (age interval, sex and settlement area distributions) are extracted. Then, to obtain a weighted sample, we apply the sample-based algorithm from Ye et al. (2009) and lenormand and Deffuant (2013). With this sample we generate the synthetic population of Mahallat. Then, the reconstructed population (synthetic population) of the city is projected using the extended model II covering vital events for the years between the two censuses, 2006 and 2011. Table 1 shows the projected population of 2011. To evaluate performance of these models, we use the estimated proportion of good prediction ($P\hat{G}P$), proposed by lenormand and Deffuant (2013) as

$$P\hat{G}P = 1 - \frac{1}{2} \times \frac{\sum_{k=1}^{p} |O_k - \hat{E}_k|}{\sum_{k=1}^{p} E_k} \tag{4}$$

where O_k and $\hat{E}_k$ are the observed (census frequency) and estimated expected frequency (the population projection frequency) of household or individuals in k^{th} age interval, respectively. $P\hat{G}P$ criterion range from 0.5 to 1.

The second criterion used for comparing the results of these two models, is the percentage of relative absolute deviation of observed and estimated expected frequencies, given by

$$\hat{\delta}_k = \frac{|O_k - \hat{E}_k|}{E_k} \times 100 \qquad k = 1, 2, ..., p \tag{5}$$

where $\hat{\delta}_k$ represents the percentage of relative absolute deviation of observed and estimated expected frequencies for the k^{th} age interval. The third criterion used for comparing performance of the proposed model and model I is 95% nonparametric bootstrap confidence intervals for each age interval. Resampling with replacement has been carried out with a sample size equal to 2% of data file of the 2006 census. Synthetic population by the year 2006 is simulated using the sample-based method, and population from 2006 to 2011 is projected using model II. This process has been repeated 500 times. In each bootstrap sample, age distribution of the target population is estimated. Finally, sampling distribution estimator of the total population in every age interval is estimated. A percentage of 95 confidence intervals of nonparametric bootstrap for the estimator of the size of the population in each age interval $\hat{\theta}_n$ is calculated by the Equation 6 (see Wassermann (2006)).

$$(\hat{\theta}_n - z^*_{0.975}\hat{se}_{boot}, \hat{\theta}_n - z^*_{0.025}\hat{se}_{boot}) \tag{6}$$

where z^*_α is the α quantile of $z^*_{n,1}$, $z^*_{n,2}$, ..., $z^*_{n,b}$ and $z^*_{n,b} = \frac{\theta^*_{n,b} - \hat{\theta}_n}{\hat{se}_{boot}}$, $\hat{\theta}_n$ is the estimator of size of the population in each age interval from the initial sample size, $\theta^*_{n,b}$ is the bootstrap estimate of size of the population in each age interval in the b^{th} bootstrap sample, and $\hat{se}_{boot}$ is the estimated bootstrap standard error of $\theta^*_{n,b}$ calculated by Equation 7.

$$\hat{se}_{boot} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\theta^*_{n,b} - \bar{\theta}^*_n)^2}, \qquad \text{where} \qquad \bar{\theta}^*_n = \frac{1}{B}\sum_{b=1}^{B}\theta^*_{n,b} \qquad (7)$$

5.1 A Comparison Between the Existing and the Extended Models

Tables 1 to 4 summarize final results by applying Equation 3. To compare model I with model II, criteria 4, 5 and 6 are considered.

5.1.1 Comparing the $\hat{\delta}_k$ of Observed and Estimated Expected Frequencies

To compare population projections using models I and II, we use the percentage of relative absolute deviation of observed and estimated expected frequencies by applying Equation 5. Table 1 indicates that the population of age intervals projected by model II has a lower average $\hat{\delta}_k$. Table 2 indicates that the projected population using model II has a lower $\hat{\delta}_k$ in term of the type of characteristics related to the model I. This improvement in sum and average estimation of Mahallat population is evident in both Tables 1 and 2.

5.1.2 A Comparison of the Two Models Using $P\hat{G}P$

Table 3 shows the values of $P\hat{G}P$ criterion using Equation 4 for the population projection of Mahallat. As we can see $P\hat{G}P$ values using model II are greater than those for model I, this means the model II performs better than the model I in projecting the population of urban and rural for both men and women.

5.1.3 Bootstrap 95% Confidence Intervals

As Table 4 and Figure 2 show the model II, 95% bootstrap confidence intervals cover Mahallat population of all age intervals except for 0-4 and 80-84 age intervals. While the model I, 95% bootstrap confidence intervals cover Mahallat population in all age intervals except for 0-4, 30-34, 60-64, 75-79,

Table 1. Percentage of relative absolute deviation of observed and estimated expected frequencies to compare the two models for each age interval in Mahallat

Age interval	Population of 2011 Census	Population projection using model I	Population projection using model II	$\hat{\delta}_k$ for model I	$\hat{\delta}_k$ for model II
0-4	3691	3106	3127	15.85	15.28
5-9	3441	3421	3448	0.58	0.21
10-14	3545	3820	3652	7.76	3.02
15-19	4189	4681	4302	11.75	2.70
20-24	5823	6029	5640	3.54	3.14
25-29	5945	6026	6269	1.36	5.45
30-34	4588	5103	4645	11.22	1.23
35-39	3999	4211	4028	5.30	0.72
40-44	3758	3733	3679	0.67	2.10
45-49	3388	3307	3421	2.39	0.97
50-54	2840	2798	2835	1.48	0.16
55-59	2153	2113	2209	1.86	2.59
60-64	1406	1592	1368	13.23	2.67
65-69	1206	1323	1259	9.70	4.35
70-74	1161	1198	1109	3.19	4.46
75-79	1040	932	1075	10.38	3.35
80-84	626	600	570	4.15	8.87
85+	554	483	608	12.82	9.75
sum	53353	54476	53245	2.10	0.20
Average $\hat{\delta}_k$ of age intervals 0-4 to 85+				6.51	3.95

Table 2. Percentage of relative absolute deviation of observed and estimated expected frequencies to compare the two models for different characteristics of population in Mahallat

Age interval	Population of 2011 Census	Population projection using model I	Population projection using model II	$\hat{\delta}_k$ for model I	$\hat{\delta}_k$ for model II
Urban males	23278	24022	23214	3.2	0.3
Urban females	23132	23398	23020	1.1	0.5
rural males	3779	3547	3800	6.1	0.6
rural females	3164	3509	3210	10.9	1.5
sum	53353	54476	53245	2.1	0.2

Table 3. $P\hat{G}P$ statistic for model I and model II in Mahallat by gender

	Urban males	Urban females	Rural males	Rural females	total
$P\hat{G}P$ for model I	0.961	0.970	0.927	0.934	0.971
$P\hat{G}P$ for model II	0.981	0.980	0.968	0.971	0.983

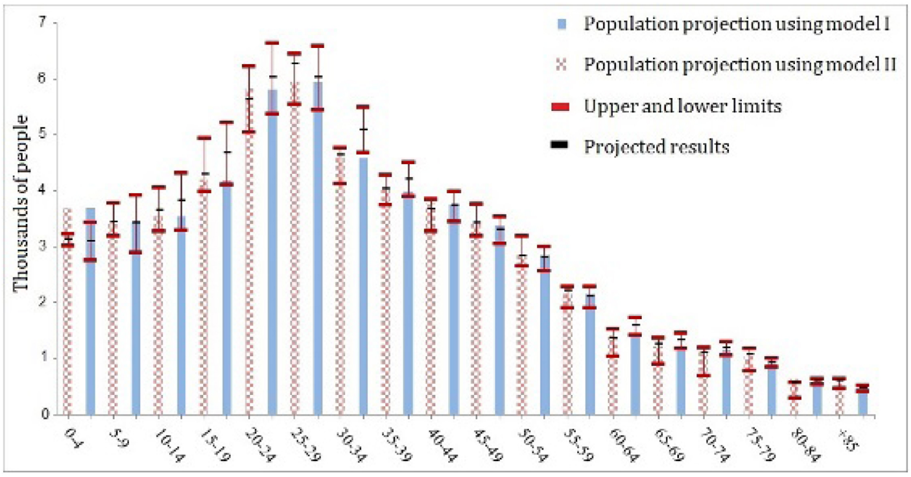

Figure 2. Bootstrap 95% confidence intervals for projected 2011 population by models I and II.

and 85 years old and more. Thus, 95% bootstrap confidence intervals improve the number of individuals of age interval in the model II, also the average length of confidence intervals is 567 for model II and 593 for model I.

6 An Application

Assume that there are two datasets in which the data are interconnected in some ways. A new dataset is to be driven out of the two aforementioned datasets whose data are not found in any other datasets. One of those initial datasets includes household of information as identified through census, and the other one consists information on electricity industry subscribers' energy consumption per capita. The purpose is to make a dataset of the characteristics of both datasets. Moreover, suppose that this dataset is constructed from a synthetic population based on characteristics per capita household energy consumption, size and type of household in Mahallat city. This dataset considerably helps planners to provide the required the funds and the infrastructures for clean and optimum energy consumption in the city. Aiming at this objective, 480 households in Mahallat are selected randomly. The required household information for the year 2006 and the households' electricity bill subscription code are then gathered. It should be noted that since

Table 4. Age distribution in 2011 census and 95% bootstrap confidence interval for each age interval in Mahallat

		Model I			Model II		
Age interval	population	Lower limit	Upper limit	Population projection using model I	Lower limit	Upper limit	Population projection using model II
0-4	3691	2773	3439	3106	3022	3246	3127
5-9	3441	2915	3929	3422	3195	3798	3448
10-14	3545	3316	4326	3821	3295	4063	3652
15-19	4189	4125	5238	4682	4004	4952	4302
20-24	5823	5396	6661	6028	5056	6231	5640
25-29	5945	5454	6598	6026	5557	6467	6269
30-34	4588	4694	5511	5103	4141	4784	4645
35-39	3999	3909	4515	4212	3767	4294	4028
40-44	3758	3471	3996	3734	3298	3863	3679
45-49	3388	3064	3551	3307	3210	3768	3421
50-54	2840	2581	3016	2799	2664	3211	2835
55-59	2153	1919	2308	2113	1926	2299	2209
60-64	1406	1433	1750	1592	1046	1530	1368
65-69	1206	1184	1462	1323	917	1375	1259
70-74	1161	1078	1319	1199	705	1213	1109
75-79	1040	855	1010	933	786	1189	1075
80-84	626	551	649	600	305	585	570
85+	554	429	540	484	481	661	608

Table 5. IPU Table

	Annual household electricity consumption			Household size			Type of household						
Number of households	Less than 300 kWh	...	More than 2400 kWh	1	...	6 and more than 6	Single	...	Other households	W_1	...	W_{480}	W_{final}
1													
2													
...													
480													
WS													
E													
δ_b													
WS_1													
WS_2													
...													
WS_{480}													
δ_a													

the census data of each household characteristic is available for the year 2006, the data related to the random sample are also collected for that same year. Then, the electricity consumption of each customer within that year is accumulated separately via their electricity bill subscription codes provided by Electricity Distribution Department. Characteristics of this new dataset are summarized as follows:

- Annual household electricity consumption: less than 300 kilowatt hours (kWh), 300-600, 600-900, 900-1200, 1200-1500, 1500-1800, 1800-2100, 2100-2400, and more than 2400 kilowatt hours,
- Household size: 1, 2, 3, 4, 5, 6 and more than 6, and
- Type of household: single, couples without children, couples with children and other households.

Data collected for the 480 households of the sample is summarized in the IPU table, Table 5, out of which a new dataset is created using R software; characteristics of the new dataset are summarized in Tables 6, 7, and 8.

Table 6. Population size by households electricity consumption and household size in Mahallat

Household size	Less than 300 kWh	300-600	600-900	900-1200	1200-1500	1500-1800	1800-2100	2100-2400	More than 2400 kWh
1	422	211	138	18	149	131	61	0	0
2	614	695	585	292	352	284	37	0	0
3	874	349	400	303	326	1243	0	0	0
4	504	906	723	309	501	415	98	0	0
5	101	100	85	136	382	777	51	154	116
6+	38	52	65	33	173	345	160	61	131

Table 7. Population size by households electricity consumption and household type in Mahallat

The type household	Less than 300 kWh	300-600	600-900	900-1200	1200-1500	1500-1800	1800-2100	2100-2400	More than 2400 kWh
Single households	422	211	138	18	149	131	61	0	0
Couple households without children	528	559	490	234	329	246	0	0	0
Couple households with children	1425	548	754	677	1135	3452	313	199	235
Other households	179	246	248	291	171	412	106	16	12

Table 6 shows that in 2006, three-person households have had the highest electricity consumption among all households within the range 1500-1800 kWh. Moreover, households with two and four members, have had the highest consumption within the range 300-600 kWh. Also, the highest electricity consumption in single households has been within a range of less than 300 kWh and the highest electricity consumption in five and more-person households have belonged to the range 1500-1800 kWh. In addition, Table 6 shows that the electricity consumption has been proportionate across households with one, two, four, and six-and-more-than-six members. Thus, the electricity consumption of Mahallat can be improved upon having the electricity consumption for households of three and five members well-managed.

Table 8. Population size by households electricity consumption and household size and type in Mahallat

The type and size household	Less than 300 kWh	300-600	600-900	900-1200	1200-1500	1500-1800	1800-2100	2100-2400	More than 2400 kWh
Single households	422	211	138	18	149	131	61	0	0
Couple households without children	528	559	490	234	329	246	0	0	0
Two person households with one parent	12	3	3	23	38	105	42	13	9
Three person households	27	49	61	11	135	240	119	49	123
Three person households with one parent	7	32	14	41	40	64	13	3	3
Four person households	94	68	71	96	342	714	38	150	113
Four person households with one parent	16	29	32	62	40	112	14	0	0
Five person households	488	127	325	376	362	1350	157	0	0
Five person households with one parent	58	46	103	108	31	94	0	0	0
Six person and more than households and more	816	304	297	195	295	1148	0	0	0
Six person and more than households with one parent	86	136	95	58	23	37	37	0	0

Table 7 shows that couple households with children have had the highest electricity consumption among all households within the range 1500-1800 kWh. Also, among households without children, the highest electricity consumption has been within the range 300-600 kWh and within the range 1500-1800 kWh for other household types. In addition, Table 7 shows that the electricity consumption for the single households, couple households without children, and other-households type can be ignored due to their small number. Thus, the electricity consumption of Mahallat can be improved upon having the electricity consumption for couple households with children well-managed through investing on the existing consumption-reducing methods.

Table 8 shows that five-person households have had the highest electricity consumption among all households within the range 1500-1800 kWh. Moreover, this table shows the electricity consumption more than 2100 kWh have been only among two-person households with one parent, three-person households, and four-person households. In addition, Table 8 shows that a great deal of electricity consumption can be dealt with upon having the electricity consumption of four-person households, and six-person-and-more-than-six households well-managed.

7 Discussion and Conclusion

As noted before, the small area of Mahallat which was selected for this research overlaps with the work by Moeti and Navvabpour (2015) and Navvabpour et al. (2015). The sample used for synthetic population simulation in the sample-based method has been selected from 2% census data file disseminated by the Statistical Center of Iran in 2006. According to the comparative results of the two models, it can be concluded that the performance of the extended model II with correct results in point and distance estimation, is better than the performance of model I. Furthermore, the application employed in this work can be taken by any organization to the simulation of any new datasets constructed following the careful selection of various characteristics and variables of interest; results from these new datasets will greatly influence planning and development in the affiliated organizations. Finally, as a novel idea, model II can be used to project any simulated new dataset of various organization characteristics for the upcoming years.

References

Arentze, T., Timmermans, H., and Hofman, F. (2007). Creating Synthetic Household Populations: Problems and Approach. *Transportation Research Record*, 2014(1), 85–91. doi:10.3141/2014-11 URL http://dx.doi.org/10.3141/2014-11

Barthelemy, J., and Cornelis, E. (2012). Synthetic Populations: Reviewof the Different Approaches. *Tech. rep., LISER*

Beckman, R. J., Baggerly, K. A., and McKay, M. D. (1996). Creating Synthetic Baseline Populations. *Transportation Research Part A: Policy and Practice*, **30**, 415-429.

Deming, W. E., and Stephan, F. F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *Ann. Math. Statist.*, **11**, 427–444. doi:10.1214/aoms/1177731829 URL http://dx.doi.org/10.1214/aoms/1177731829

Gargiulo, F., Ternes, S., Huet, S., and Deffuant, G. (2010). An Iterative Approach for Generating Statistically Realistic Populations of Households. *PloS one*, **5**, e8828.

Guo, J., and Bhat, C. (2007).Population Synthesis for Microsimulating Travel Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, 2014, 92–101. doi:10.3141/2014-12 URL http://dx.doi.org/10.3141/2014-12

Huang, Z., and Williamson, P. (2001). A Comparison of Synthetic Reconstruction and Combinatorial Optimization Approaches to the Creation of Small-Area Micro Data. *working paper*2001/02

Lenormand, M., and Deffuant, G. (2013). Generating a Synthetic Population of Individuals in Households: Sample-Free vs Sample-Based Methods. *Journal of Artificial Societies and Social Simulation*, **16**, 12. doi: 10.18564/jasss.2319 URL http://dx.doi.org/10.18564/jasss.2319

Moeti, M. T., and Navvabpour, H. (2015). Population Simulations of Small Areas in the Base Year. *Journal of Population Association of Iran*, **10**, 89–107.

Navvabpour, H., Safakish, M., Moeti, M. T., Nourallahi, T., and Khondabi, B. (2015). Population Simulation in the Years Between Two Censuses in the Selected Cities. *Tech.* rep., Statistical Research and Training Center.

Voas, D., and Williamson, P. (2001). Evaluating Goodness-of-Fit Measures for Synthetic Microdata. *Geographical and Environmental Modelling*, **5**, 177–200. doi:10.1080/13615930120086078

Wasserman, L. (2010). All of Nonparametric Statistics. *Springer Publishing Company*, Incorporated, 1st edn.

Wilson, A. G., and Pownall, C. E. (1976). A New Representation of the Urban System for Modelling and for the Study of Micro-Level Interdependence. *Area*, (pp. 246–254)

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

Appendix

In this section we prove theorems 1 and 2.

Appendix 1

Proof of theorem 1:

The theorem is proved in different cases.

- $a = 5$, $i = 1, 2, ..., 16$ (Part 2 Model II).

  Suppose Model I is rewritten as follows:

$$P^{Y+k}_{[ai,ai+a)} = \beta \times P^{Y+k-1}_{[ai-a,ai)} \times C^{Y+k}_{[ai-a,ai)} + (1-\beta) \times P^{Y+k-1}_{[ai,ai+a)} \times C^{Y+k}_{[ai,ai+a)}$$

According to the problem hypotheses ($a = 5$, $k = 1$ and $\beta = 0.2$), we have

$$\begin{aligned} P^{Y+1}_{[5i,5i+5)} =& 0.2\times P^{Y}_{[5i-5,5i)} \times C^{Y+1}_{[5i-5,5i)} \\ &+ (1-0.2)\times P^{Y}_{[5i,5i+5)} \times C^{Y+1}_{[5i,5i+5)} \qquad (*) \end{aligned}$$

Now in this case we prove that part 2 model II is equal to (*).

According to the assumption, for $[5i-5,5i)$ and $[5i,5i+5)$, we have

$$\frac{P^{Y}_{[5i-5,5i-4)}}{P^{Y}_{[5i-5,5i)}} = \frac{P^{Y}_{[5i-4,5i-3)}}{P^{Y}_{[5i-5,5i)}} = ... = \frac{P^{Y}_{[5i-1,5i)}}{P^{Y}_{[5i-5,5i)}}$$

and

$$\sum_{n=1}^{5} P^{Y}_{[5i-5+n-1,5i+n-5)} = P^{Y}_{[5i-5,5i)}$$

So, we have it easily

$$P^{Y}_{[5i-1,5i)} = \frac{1}{5} \times P^{Y}_{[5i-5,5i)} \qquad (1)$$

and

$$\frac{P^{Y}_{[5i,5i+1)}}{P^{Y}_{[5i,5i+5)}} = \frac{P^{Y}_{[5i+1,5i+2)}}{P^{Y}_{[5i,5i+5)}} = ... = \frac{P^{Y}_{[5i+4,5i+5)}}{P^{Y}_{[5i,5i+5)}}$$

and

$$\sum_{n=1}^{5} P^{Y}_{[5i+n-1,5i+n)} = P^{Y}_{[5i,5i+5)}$$

So, we have it easily

$$\begin{aligned} \sum_{n=1}^{4} P^{Y}_{[5i+n-1,5i+n)} &= P^{Y}_{[5i,5i+1)} + P^{Y}_{[5i+1,5i+2)} + ... + P^{Y}_{[5i+3,5i+4)} \\ &= \frac{4}{5} \times P^{Y}_{[5i,5i+5)} \qquad (2) \end{aligned}$$

According to the assumption and part 2 model II

$$\begin{aligned}
P_{[5i,5i+5)}^{Y+1} =& \sum_{j=1}^{5}\left[P_{[5i+j-2,5i+j-1)}^{Y} \times C_{[5(i-I(j+1\leq 2)),5(i-I(j+1\leq 2))+5)}^{Y+1}\right]\\
=& P_{[5i-1,5i)}^{Y} \times C_{[5i-5,5i)}^{Y+1}\\
&+ P_{[5i,5i+1)}^{Y} \times C_{[5i,5i+5)}^{Y+1}\\
&+ P_{[5i+1,5i+2)}^{Y} \times C_{[5i,5i+5)}^{Y+1}\\
&+ P_{[5i+2,5i+3)}^{Y} \times C_{[5i,5i+5)}^{Y+1}\\
&+ P_{[5i+3,5i+4)}^{Y} \times C_{[5i,5i+5)}^{Y+1}\\
=& P_{[5i-1,5i)}^{Y} \times C_{[5i-5,5i)}^{Y+1} + \sum_{n=1}^{4} P_{[5i+n-1,5i+n)}^{Y} \times C_{[5i,5i+5)}^{Y+1}
\end{aligned}$$

According to (1) and (2):

$$P_{[5i,5i+5)}^{Y+1} = \frac{1}{5} \times P_{[5i-5,5i)}^{Y} \times C_{[5i-5,5i)}^{Y+1} + \frac{4}{5} \times P_{[5i,5i+5)}^{Y} \times C_{[5i,5i+5)}^{Y+1}$$

which is the same as (*) and the proof is completed.

- $a = 5$, $i = 17$ (part 3 model II).

  Suppose Model I is rewritten as follows:

  It is clear $C_{[85,90)}^{Y} = C_{[85,\infty)}^{Y}$, we have

$$P_{[85,\infty)}^{Y+1} \quad = \quad 0.2 \times P_{[80,85)}^{Y} \times C_{[80,85)}^{Y+1} + P_{[85,\infty)}^{Y} \times C_{[85,\infty)}^{Y+1} \qquad (**)$$

Now in this case we prove that part 3 model II is equal to (**).

According to the assumption and for $[80, 85)$, we have

$$\frac{P_{[80,81)}^{Y}}{P_{[80,85)}^{Y}} = \frac{P_{[81,82)}^{Y}}{P_{[80,85)}^{Y}} = ... = \frac{P_{[84,85)}^{Y}}{P_{[80,85)}^{Y}}$$

and

$$\sum_{n=1}^{5} P_{[80+n-1,80+n)}^{Y} = P_{[80,85)}^{Y}$$

So, we have it easily

$$P^{Y}_{[84,85)} = \frac{1}{5} \times P^{Y}_{[80,85)} \qquad (3)$$

and

$$P^{Y}_{[85,86)} + P^{Y}_{[86,87)} + ... = P^{Y}_{[85,\infty)} \qquad (4)$$

According to the assumption and part 3 model II

$$\begin{aligned} P^{Y+1}_{[85,\infty)} &= \sum_{j=1}^{\infty} \left[P^{Y}_{[85+j-2,85+j-1)} \times C^{Y+1}_{[5(17-I(j+1\leq 2)),5(17-I(j+1\leq 2))+5)} \right] \\ &= P^{Y}_{[84,85)} \times C^{Y+1}_{[80,85)} + P^{Y}_{[85,86)} \times C^{Y+1}_{[85,90)} + P^{Y}_{[86,87)} \times C^{Y+1}_{[85,90)} + ... \\ &= P^{Y}_{[84,85)} \times C^{Y+1}_{[80,85)} + P^{Y}_{[85,\infty)} \times C^{Y+1}_{[85,\infty)} \end{aligned}$$

According to (3) and (4):

$$P^{Y+1}_{[5i,5i+5)} = \frac{1}{5} \times P^{Y}_{[80,85)} \times C^{Y+1}_{[80,85)} + P^{Y}_{[85,\infty)} \times C^{Y+1}_{[85,\infty)}$$

Which is the same as (**) and the proof is completed.

- $a = 5$, $i = 0$ (part 1 model II).

Suppose Model I is rewritten as follows:

we know $P^{Y}_{[0,1)}$ number of births in Y year, so

$$P^{Y+1}_{[0,5)} = P^{Y+1}_{[0,1)} + (1-0.2) \times P^{Y}_{[0,5)} \times C^{Y+1}_{[0,5)} \qquad (***)$$

Now in this case we prove that part 1 model II is equal to (***).

According to the assumption, for $[0,5)$ we have

$$\frac{P^{Y}_{[0,1)}}{P^{Y}_{[0,5)}} = \frac{P^{Y}_{[1,2)}}{P^{Y}_{[0,5)}} = ... = \frac{P^{Y}_{[4,5)}}{P^{Y}_{[0,5)}}$$

and

$$\sum_{n=1}^{5} P^{Y}_{[n-1,n)} = P^{Y}_{[0,5)}$$

So, we have it easily

$$\sum_{n=1}^{4} P^{Y}_{[n-1,n)} = P^{Y}_{[0,1)} + P^{Y}_{[1,2)} + ... + P^{Y}_{[3,4)} = \frac{4}{5} \times P^{Y}_{[0,5)} \qquad (5)$$

According to the assumption and part 1 model II

$$\begin{aligned} P^{Y+1}_{[0,5)} =& \sum_{j=1}^{5} \left[P^{Y+(2-j)I(j\leq 1)}_{[j-2,j-1)I(j-1\in\{2,3,4\})+[0,1)I(j-1\notin\{2,3,4\})} \times \left(C^{Y+1}_{[0,5)} \right)^{I(j+1>2)} \right] \\ =& P^{Y+1}_{[0,1)} + P^{Y}_{[0,1)} \times C^{Y+1}_{[0,5)} + P^{Y}_{[1,2)} \times C^{Y+1}_{[0,5)} \\ & + P^{Y}_{[2,3)} \times C^{Y+1}_{[0,5)} + P^{Y}_{[3,4)} \times C^{Y+1}_{[0,5)} \\ =& P^{Y+1}_{[0,1)} + \sum_{n=1}^{4} P^{Y}_{[n-1,n)} \times C^{Y+1}_{[0,5)} \end{aligned}$$

According to (5):

$$P^{Y+1}_{[0,5)} = P^{Y+1}_{[0,1)} + \frac{4}{5} \times P^{Y}_{[0,5)} \times C^{Y+1}_{[0,5)}$$

which is the same as (***) and the proof is completed.

For the case $a = 10$, it is proved in a similar way.

Appendix 2

Proof of theorem 2:

The theorem is proved in following case:

$a = 5$, $i = 1, 2, ..., 16$ (part 2 model II).

By substituting values β in model I, we have

$$
\begin{aligned}
P^{Y+1}_{[ai,ai+a)} =& \frac{P^{Y}_{[ai-1,ai)}}{P^{Y}_{[ai-a,ai)}} \times P^{Y}_{[ai-a,ai)} \times C^{Y}_{[ai-a,ai)} \\
&+ (1 - \frac{P^{Y}_{[ai+a-1,ai+a)}}{P^{Y}_{[ai,ai+a)}}) \times P^{Y}_{[ai,ai+a)} \times C^{Y}_{[ai,ai+a)} \\
=& P^{Y}_{[ai-1,ai)} \times C^{Y}_{[ai-a,ai)} + (P^{Y}_{[ai,ai+a)} - P^{Y}_{[ai+a-1,ai+a)}) \times C^{Y}_{[ai,ai+a)} \\
=& P^{Y}_{[ai-1,ai)} \times C^{Y}_{[ai-a,ai)} + P_{[ai,ai+a-1)} \times C^{Y}_{[ai,ai+a)} \qquad (*)
\end{aligned}
$$

Now in this case we prove that part 2 model II is equal to (*).

$$
\begin{aligned}
P^{Y+1}_{[ai,ai+a)} =& \sum_{j=1}^{a} \left[P^{Y}_{[ai+j-2,ai+j-1)} \times C^{Y+1}_{[a(i-I(j+1\leq 2)),a(i-I(j+1\leq 2))+5)} \right] \\
=& P^{Y}_{[ai-1,ai)} \times C^{Y+1}_{[ai-a,ai)} \\
&+ P^{Y}_{[ai,ai+1)} \times C^{Y+1}_{[ai,ai+a)} \\
&+ P^{Y}_{[ai+1,ai+2)} \times C^{Y+1}_{[ai,ai+a)} \\
&+ ... \\
&+ P^{Y}_{[ai+a-2,ai+a-1)} \times C^{Y+1}_{[ai,ai+a)} \\
=& P^{Y}_{[ai-1,ai)} \times C^{Y+1}_{[ai-a,ai)} + \sum_{n=1}^{a-1} P^{Y}_{[ai+n-1,ai+n)} \times C^{Y+1}_{[ai,ai+a)} \\
=& P^{Y}_{[ai-1,ai)} \times C^{Y+1}_{[ai-a,ai)} + P^{Y}_{[ai,ai+a-1)} \times C^{Y+1}_{[ai,ai+a)}
\end{aligned}
$$

which is the same as (*) and the proof is completed.

For the other cases, it is proved in a similar way.

Mohammad Taghi Moeti
Department of Statistics,
Allameh Tabataba'i University,
Tehran, Iran.
email: *mt_ moeti@yahoo.com*

Hamidreza Navvabpour
Department of Statistics,
Allameh Tabataba'i University,
Tehran, Iran.
email: *hnavvabpour@atu.ac.ir*

Farzad Eskandari
Department of Statistics,
Allameh Tabataba'i University,
Tehran, Iran.
email: *askandari@atu.ac.ir*