GENERATION OF SYNTHETIC POPULATION USING MARKOV CHAIN MONTE CARLO SIMULATION METHOD

Anu P. Alex[1], Prinsha T[2] and Manju V. S[3]

[1]Department of Civil Engineering, College of Engineering Trivandrum
[2]Department of Civil Engineering, Malabar Institute of Technology, Anjarakandy
[3]Department of Civil Engineering, College of Engineering Trivandrum

ABSTRACT

Activity based travel demand models are widely used in transportation planning to predict future demand of transportation. Disaggregate level data for the entire population is required as input to these models, which included household level and person level attributes for the entire study area. These data are usually collected by the population census, but are rarely available due to confidentiality reasons. Hence as a viable alternative, population synthesis techniques are used to supplement the microdata. An attempt has been made in this study to generate synthetic population using Markov Chain Monte Carlo Simulation method and to compare this with conventional method. Thiruvananthapuram Corporation in Kerala was selected as the study area and sample data were collected by household survey. The algorithm for population synthesis was coded in C++. The methodology was validated using 16 percentage of the collected data. Prediction accuracy of the method was compared with conventional method and was found better.

KEYWORDS

Microsimulation, Synthetic population, Beckman's method, MCMC method.

1. INTRODUCTION

Microsimulation is a mechanism for forecasting the state of a dynamic, complex system by simulating the behavior of the individuals in the system. The fundamental data required for microsimulation model are the details of individual and household attributes for the entire population of the study area. These details are usually collected in a population census but is not been made available to the public due to privacy and confidentiality reasons. Population synthesis techniques are commonly used as viable alternative to supplement the lack of availability of microdata. The process of population synthesis involves expanding a sample drawn from a population to a full set of synthetic population. These are algorithms that apply to sample data and its aggregated population data, in order to generate a synthetic population which is statistically representative of the actual population data. Different algorithms used for the process may generate synthetic population with different quality [1]. Since synthetic population is the input to activity based travel demand models, there had been a tremendous increase in the research works related to generation of synthetic population over the last fifteen years.

The conventional approach to synthesize base year population is based on a methodology originally developed by Beckman et al. [1]. This approach involves integrating aggregate data

DOI: 10.5121/ijci.2016.5117

from one source with disaggregate data from another source. Beckman's population synthesis approach uses the disaggregate data as "seeds" to create individual population records that are collectively consistent with the cross tabulations provided by the aggregate data. A new algorithm was proposed by Guo and Bhat [1] and they discussed about its data structure, operation and the step by step procedure. They generated synthetic population for Dallas /Fort-Worth area and census block groups were used as the aggregate data. Ryan et al. [2] discussed the synthetic reconstruction method based on Iterative Proportional Fitting (IPF) and Combinatorial Optimization method (CO) and the algorithms were tested. Programs to execute the CO and IPF methods were written in C++. Auld et al. [3] explained the population synthesis using Iterative Proportional Fitting Procedure. They used category reduction as a control for zero cell issue. The algorithm was designed to be used for any geographic area. The methodology was validated by generating the synthetic population for Southwestern Cook County. Ye et al. [4] tested the generation of synthetic population using Iterative Proportional Updating algorithm and matching the distributions of both household and person attributes in it. The algorithm involved iteratively adjusting and reallocating household weight. They included an example to illustrate the algorithm and its geometric interpretation. Ma [5] discussed about disaggregate travel demand models and activity based transportation planning models. Conceptual overview of the generation of synthetic population was described by the author.

Another method to population synthesis is Markov Chain Monte Carlo (MCMC) simulation based approach, which will overcome the limitations of IPFP [6]. Farooq et al. [6] described population synthesis using Markov chain Monte Carlo simulation. They have also explained about the simulation based approach and preparation of conditionals. The real population from Swiss census was used to compare the performance of simulation based synthesis with the standard IPF. Finally they have obtained better result for Monte Carlo Simulation method. Performance of synthesis procedures has been assessed using the Standardized Root Mean Square Error (SRMSE). The proposed methodology was implemented by Farooq et al. [7] for a real case study, where a synthetic population was generated for the base year of an integrated land use and transport model for the region of Brussels.

Anu et al. [8] conducted a study on generation of synthetic population for Thiruvananthapuram city by Beckman's method which is the conventional method. They have checked the prediction accuracy at both aggregate level and disaggregate level. They found that the prediction accuracy is above 90 percentage in the case of total number of households and total number of females and it is above 80 percentage in the case of total population and total number of males. The present paper attempts to develop synthetic population using MCMC method for Thiruvananthapuram city and to compare the results with conventional method.

2. Markov Chain Monte Carlo Method For Synthetic Population

The methodology involved in Markov Chain Monte Carlo (MCMC) simulation is to draw agents (households) directly from the joint distribution. For example, joint probability distribution of X, Y… etc. is a probability distribution that gives the probability that each of X, Y… etc. falls in any particular range or discrete set of values. Gibbs sampler is used to simulate the draw. Instead of joint distribution conditional distributions are used as input for Gibbs sampler. This method is applicable when joint distribution is unknown, but conditional distributions are known. The only information that is required to generate the next agent is the previous agents synthesized and the input conditionals. Conditional distribution of one variable is the distribution of one variable conditioned on all other variables. For example, X and Y are two jointly distributed random variables then the conditional distribution of Y given X is the probability distribution of Y when X is known to be a particular value. Gibbs sampler generates a Markov chain of samples each of which is correlated with the nearby samples. Steps involved in development of synthetic population using Monte Carlo simulation are shown in figure 1.

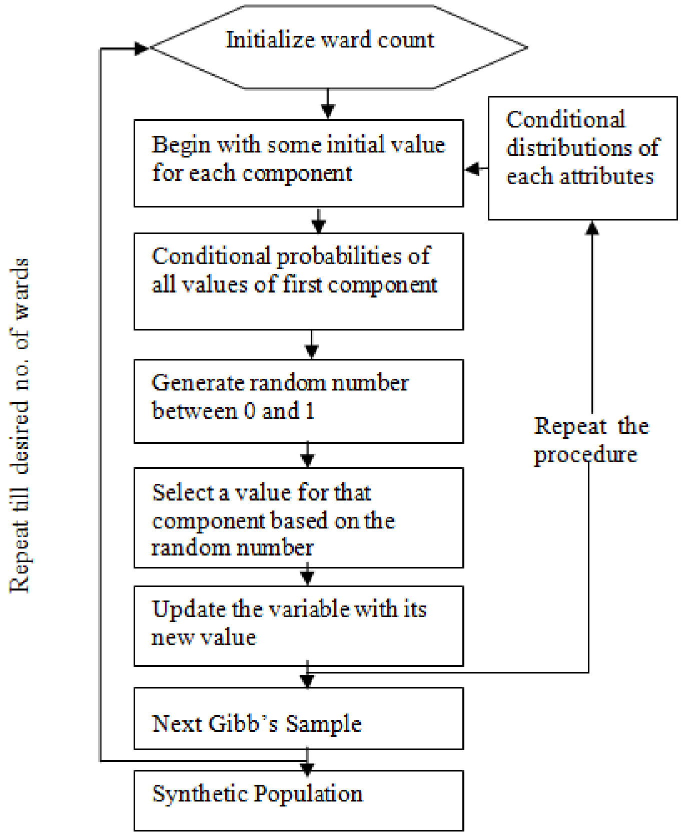

Initialize ward count
Begin with some initial value for each component
Conditional distributions of each attributes
Conditional probabilities of all values of first component
Generate random number between 0 and 1
Select a value for that component based on the random number
Update the variable with its new value
Next Gibb's Sample
Synthetic Population
Repeat the procedure
Repeat till desired no. of wards

Figure 1. Steps involved in Monte Carlo simulation

The method starts with inputting the conditional distributions of each attributes. First some initial value is chosen for each attributes of first household, which is the first Gibbs sample. Conditional probabilities for all possible values of the first attribute are determined. Next step is generation of random numbers between 0 and 1. Based on the range in which the random number falls, the attribute value is selected. This step is repeated for all the other attributes to get the next Gibbs sample. This whole procedure is repeated for required number of times to get synthetic households in a ward. Similarly synthetic population is generated for all the 87 wards of Thiruvananthapuram Corporation.

2.1. Development of Algorithm

An algorithm for developing synthetic population based on MCMC method was coded in C++. 84 percentage of the collected data were used as input for the program and remaining 16 percentage of the data were used for validation.

2.1.1. Pseudo Coding

```
Pre Processor Directives
{
Variable declaration;
```

```
Read conditional probability distribution files of all attributes, household file, person file, ward total file;
For no. of wards
{
Read total no. of households from ward total file;
For total no. of households
{
For no. of attributes
{
Generate random number;
Check the range of conditional probability where the random no. falls;
Choose the attribute value;
Assign persons from person file;
List Household and person attributes;
}
End
}
```

3. Outcome Of The Study

Outcome of the study is synthetic population for 87 wards of Thiruvananthapuram Corporation, which included the household level and person level attributes of each household. The number of households as per 2011census is191446 for Thiruvananthapuram Corporation and number of households predicted in the study are 213051 and 191466 by Beckman's method [8] and by MCMC method respectively. This shows that synthetic households were predicted with an accuracy of 100 percentage in MCMC method and 89 percentage in Beckman's method. Total population as per 2011 census is 762535.The synthesized data are 705036 by Beckman's method and 785352 by MCMC method respectively. The population was predicted with an accuracy of 92 percentage in Beckman's method and 97 percentage in MCMC method.

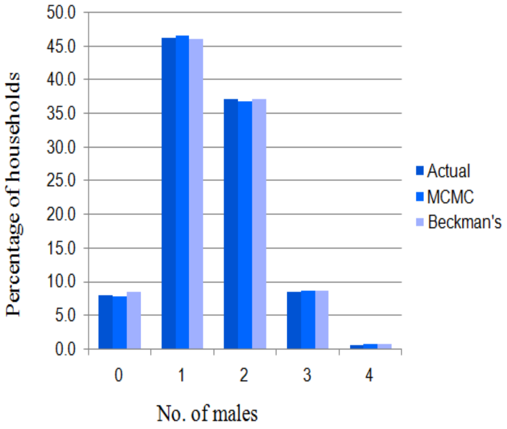
4. Validation

Validation of the method was done by generating synthetic population for ward 18 (Kannammoola). This comprises 16 percentage of the total data collected. Synthetic population was generated for Kannammoola ward and the results were compared with data obtained from survey and conventional method [8]. Prediction accuracy in aggregate level results are shown in Table 1. Synthetic population of Thiruvananthapuram Corporation generated using Monte Carlo simulation technique was found better in prediction accuracy than Beckman's method.

Both household level and person level attributes were also compared. Household level validation results are shown in figures 3 to 6. It shows that the actual and synthesized population is almost equal when compared with household level attributes like number of males, number of females, vehicle ownership and number of workers in case of MCMC method compared to Beckman's method. Percentage accuracy obtained for each case is shown in Table 2.

Table 1. Results of Validation at Aggregate Level

Attributes	Actual population from census data	Synthesized population by Beckman's method	Prediction accuracy (%)	Synthesized population by MCMC method	Prediction accuracy (%)
Total number of households	191446	213051	88.71	191446	100
Total population	762535	705036	92.46	785352	97
Total number of males	371037	343714	92.63	356550	95.93
Total number of females	391498	361322	92.29	428802	90.47



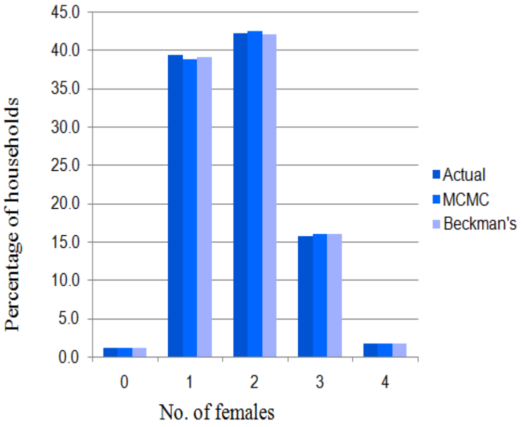
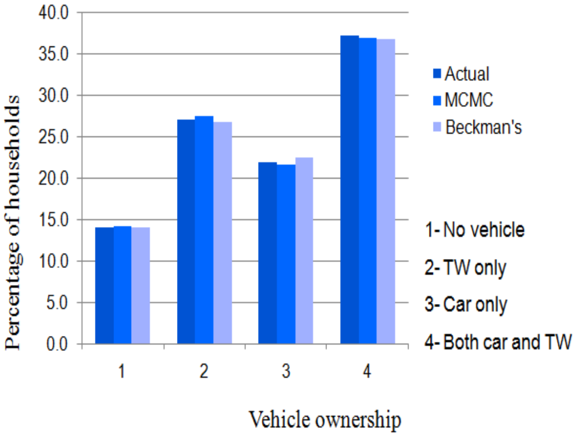


Figure 2. Comparison of HH based on no. of males

Figure 3. Comparison of HH based on no.of females



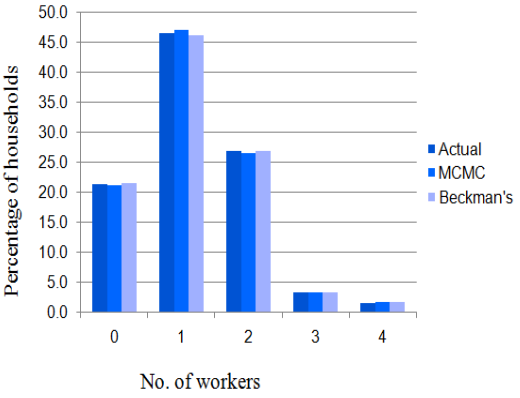


Figure 4. Comparison of HH based on Vehicle ownership

Figure 5. Comparison of HH based on number workers

Table 2. Summary of results HH level

Sl. No.	HH level Attributes	Accuracy (%) Beckman’s	Accuracy (%) MCMC
1	Number of males	97.04	98.77
2	Number of females	98.57	99.09
3	Vehicle ownership	98.97	99.02
4	Number of workers	98.48	99.38

It is found that household level attributes like number of males, number of females, vehicle ownership and number of workers have an accuracy above 97 percentage in Beckman’s method and about 99 percentage in MCMC method Person level validation results are shown in figure 7 to 11, which gives the comparison of percentage of individuals based on gender, age group, marital status, education level and employment status. It can be inferred that actual and synthesized population are closer in the case of gender and marital status. Percentage accuracy obtained for each case is shown in Table 3.

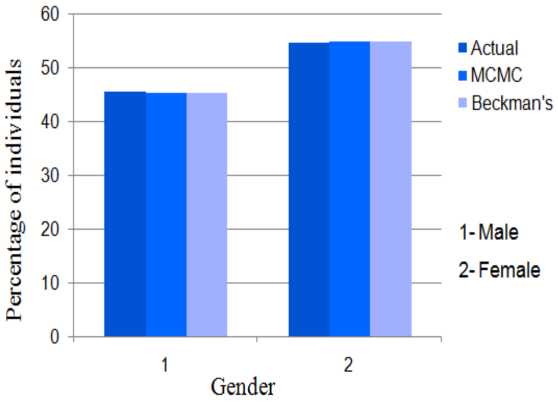


Figure 6. Comparison of persons based on gender

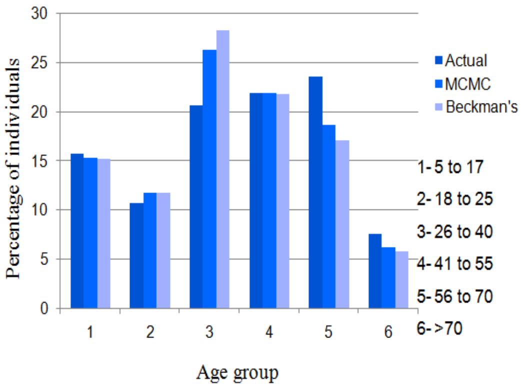


Figure 7. Comparison of persons based on age group

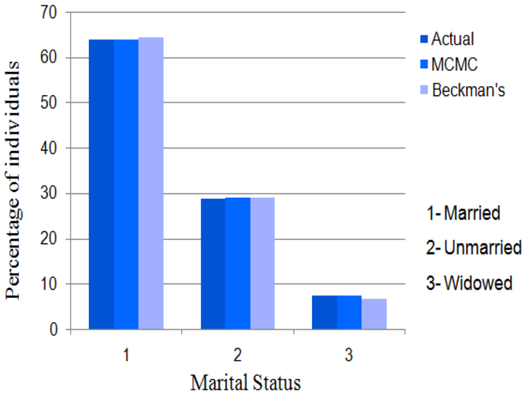


Figure 8. Comparison of persons based on marital status

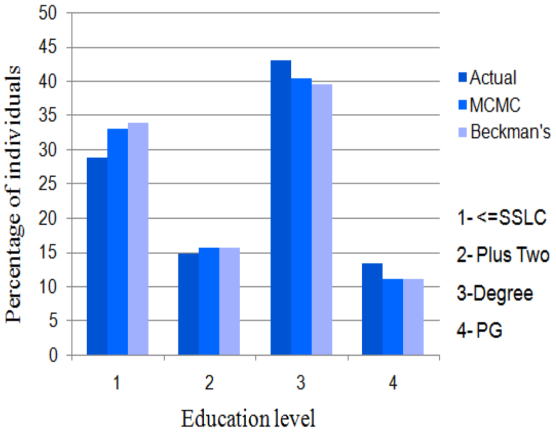


Figure 9. Comparison of persons based on education level

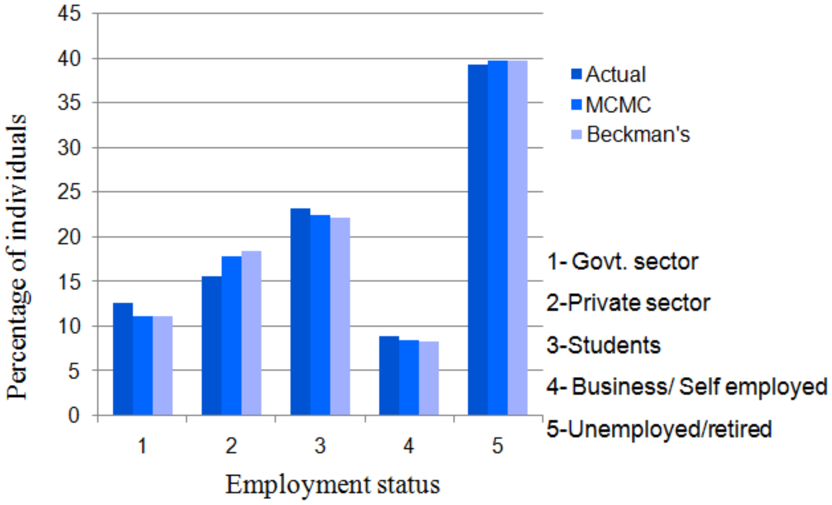


Figure 10. Comparison of persons based on employment level

Table 3. Summary of results person level

Sl. No.	Person level attributes	Accuracy (%) Beckman's	Accuracy (%) MCMC
1	Gender	99.67	99.79
2	Age group	83.5	87.11
3	Marital status	96.32	99.27
4	Employment	91.4	93.26
5	Education level	87.87	89.72

It is found that accuracy level is above 90 percentage in the case of gender, marital status and employment status and above 80 percentage in the case of age group and education level in Beckman's method. The accuracy level is above 99 percentage in gender and marital status and above 87 percentage in marital status, employment level and education level in MCMC method.

5. CONCLUSION

Travel demand models are essential tool in transportation planning for predicting the future demand of transportation. There are two approaches, trip based and activity based modeling out of which, activity based models provide more realistic representation of travel behavior. Hence now a day's activity based models are widely used. Disaggregate level population data are essential for activity based models which are usually collected by population census. But due to confidentiality reasons these data are not available for the public. Hence population synthesis techniques are used to generate synthetic population as input to activity based travel demand models. An attempt has been made in this paper to generate synthetic population for Thiruvananthapuram city which can be used as an input for activity based travel demand model for the city. An algorithm to generate synthetic population based on MCMC method has been presented in this paper.

Algorithm for the development of synthetic population was coded in C++. 84 percentage of the collected data were used as input for the program. The method was validated in the disaggregate level by applying it to ward number 18 of Thiruvananthapuram Corporation, which included 16 percentage of the collected data. Validation at aggregate level was done using census data. The following conclusions were made from the study;

- The prediction accuracy is more in the case of Markov Chain Monte Carlo simulation method compared to the Beckman's method both in household level and person level attributes
- Prediction accuracy of synthetic households was greater for MCMC method than Beckman's method by 12%
- Prediction accuracy of total population was greater for MCMC method than Beckman's method by 5%.
- Prediction accuracy of total number of males was greater for MCMC than Beckman's method by 14%.

Prediction accuracy of total number of females was greater for MCMC than Beckman's method by 2%.

ACKNOWLEDGEMENT

Authors express their gratitude to Kerala State Council for Science, Technology and Environment (KSCSTE) for funding the project.

REFERENCES

[1] J.Y. Guo, and C. R. Bhat, (2007). "Population synthesis for microsimulating travel behaviour", *Transportation Research Record,* 2014 (12) 92–101.

[2] J. Ryan, H. Maoh, and P. Kanaroglou, (2007), "Population synthesis: comparing major techniques using a small complete population of firms" working paper of *Mc Master university*

[3] J. Auld, A.K Mohammadian, and K. Wies, (2008). "Population synthesis with control category optimization", paper presented at the 10th *International Conference on Application of Advanced Technologies in Transportation,* Athens, Greece, May 2008.

[4] X. K. Ye, K.C. Pendyala, B. Sana, and P. Waddell, (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.

[5] Lu Ma., (2011) "Generating disaggregate population characteristics for input to Travel-demand models" A dissertation presented to the *University of Florida*

[6] B. Farooq, M. Bierlaire, and G. Flotterod, (2013), "Simulation based population synthesis", *Transportation research board*

[7] B. Farooq, M. Bierlaire, and G. Flotterod, (2013), "Simulation based generation of synthetic population for Brussels case study", *Transportation research board*

[8] P. A. Anu, T. Prinsha, and V.S. Manju, (2015) "Generation of synthetic population using Beckman's method" The 16th *National Conference on Technological Trends,* Trivandrum, September 2015

AUTHORS

Mrs. Anu P. Alex, is working as Assistant Professor in Civil Engineering, College of Engineering Trivandrum, Kerala. She obtained B.Tech in Civil Engineering from Kerala University and M.E. in Transportation Engineering and Management from NIT, Trichy. She has more than 15 research papers in National and International Journals and Conferences.

Prinsha T., is working as Assistant Professor in Civil Engineering Malabar Institute Of Technology, Kannur, Kerala. She obtained B.Tech in Civil Engineering from CUSAT and M.Tech (Traffic and Transportation Engineering) from College of Engineering, Trivandrum.

Dr. Manju V. S., is working as Associate Professor in Civil Engineering at College of Engineering, Trivandrum. She graduated in B.Tech (Civil Engineering) under University of Kerala and M. Tech (Transportation Planning) from National Institute of Technology, Calicut. She did her doctoral programme under University of Kerala. She has published more than 20 research papers in National and International Journals and Conferences.