Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 6 (2015) 50 – 63

Transportation Research Procedia

www.elsevier.com/locate/procedia

4th International Symposium of Transport Simulation-ISTS'14, 1-4 June 2014, Corsica, France

A Reproducibility Analysis of Synthetic Population Generation

Jooyoung Kim[a], Seungjae Lee[b] *

[a]*Research Professor, Integrated Urban Research Center, University of Seoul, Seoul 130-743, Korea*
[b]*Professor, Department of Transportation Eng, University of Seoul, Seoul 130-743, Korea*

Abstract

For the development of agent based traffic simulation model, population synthesis is critical to the accuracy of simulation outcomes. This paper attempts to develop the synthetic population generation based on the Simulated Annealing (SA) algorithm for the activity-based travel demand model. This algorithm leads to estimate the activity schedules according to the multi-dimensional characteristics of the synthetic populations. However, appropriate rules have not been established for the estimation of parameters in simulated annealing, and it requires a significant amount of time to find optimal solution. In order to apply SA into the synthetic population, hill climbing and cooling schedule should be considered. In this study, total absolute error was calculated to prevent hill climbing and used Metropolis- Hasting algorithm to determine whether to select or dismiss follow-up distribution. In addition, stability of the algorithm was determined through scenario analysis of the optimal combination of iteration and temperature "T" on the cooling schedule. Based on this result, the current condition of household travel diary survey and census data were used to compare the IPF(Iterative Proportional Fitting) of a previous methodology with the result of establishing suggested algorithm, performing procedures of creating synthetic population, and suggesting the validity of algorithm created with the synthetic population based on SA through statistical verification.

Selection and/or peer-review under responsibility of the Organizing Committee of ISTS'14

Keywords: Population Synthesis, Simulated Annealing, Metropolis-Hasting algorithm, Hill Climbing, Cooling Schedule

1. Introduction

Synthetic population represents an agent of the population of activity patterns with identical social and economic characteristics that are applied on the model for estimating traffic demand in the activity-based access. In addition, behavioural characteristics and individual features as a subject of activity are reflected accounting for estimation of

* Corresponding author. Tel.: +82-2-6490-2823; fax: +82-2-6490-2823.
E-mail address: sjlee@uos.ac.kr

Selection and/or peer-review under responsibility of the Organizing Committee of ISTS'14
doi:10.1016/j.trpro.2015.03.005

activity-based traffic demand with virtual population that is specifically designed to behave as or on behalf of human.(Bradley, 1999)

Studies related to the creation of synthetic population become more important due to development of simulation technology and are applied to diverse fields in foreign countries. According to the result of representative researches, Ryan et al. (2009) have performed a research in comparing IPF (Iterative Proportional Fitting) and CO (Combinatorial Optimization) that are being used the most in creation of synthetic population. According to the experiment performed on the sample size and a degree of elaboration of the chart to maximize accuracy of the result in measuring synthetic population, both methods have proved that the result in estimating the joint distribution of small-scaled group and realistic data is statistically significant. In Korea, a study dealing with creation of synthetic population for agent-based simulation has compared IPF and Copula in order to analyse the characteristics of methods in producing synthetic population by specifically utilizing distribution of ages and incomes as well as characteristics of individuals in the totalizing data. IPF is simply and efficiently calculated but has a limit of an independent structure, while Copula proves it feasible to maintain an independent structure between distinct attributes and be efficiently calculated.

Other than them, there are many cases that synthetic population is created for micro-simulation and is applied to simulation model. Harland et al.(2012) examine the performance of deterministic reweighting, conditional probability and simulating annealing over varying spatial scales. Beckman et al.(1996) has used TRANSIM in order to create synthetic population for the analysis of micro-traffic simulation in the area of Los Alamos. Bradley et al.(2001) have developed a model for creating synthetic population of San Francisco County Transit Authority (SFCTA) applying it to the activity-based travel simulation. Hensher et al.(2004) have developed TRESIS creating simulation model through generation of synthetic population in Sydney. TRESIS has estimated 3-dimensional distribution on the entire area. Hereupon, it was intended to completely utilize totalizing data performing a procedure of subdivision in a single cell on 1% of sample households and to enhance accuracy of the model.

As for synthetic population, multi-lateral social and economic characteristics shall be reflected. However, results of census also tend to entail a problem for not being able to identify individual characteristics that compose population as a totalizing data on individual or households. On the other hand, household travel diary survey is capable of analysing individual features and travel characteristics as a non-total data but might cause bias depending on the response rate and sampling method on survey targeting 2~3% of the sample of the total population in the areas to be analysed. As analysed earlier in the sample survey, middle-aged people consisted of the highest portion of 37.98% according to the population based on age, while the senior class with more than or equal to the age of 65 was turned out to be smallest portion as 5.38%. However, it is easy to identify how the senior class was turned out to be lower in contrast with the result of census conducted by National Statistical Office and also the value of 8.42% of the senior class of a portion in age-based population. Furthermore, this type of phenomenon tends to be diversely shown according to social and economic features. In this study, bias is specifically defined to a difference between the sample and expanded proportion. In this case, applying expanded proportion of collected data on the biased sample is inappropriate in reflecting characteristics of human behaviour. (Balmer ;2007, Ma ;1997, Bhat et al. ;2002, Bhat & Koppleman; 1999, Bowman and Ben Akiva ;2000).

It is identified that research is performed by using IPF or applying random proportion on the totalizing data on the creation of representative synthetic population. However, as for IPF, biased result of sampling survey causes direct influence on the result of creation for synthetic population. In addition, it also entails a problem for making it difficult to converge on the area without statistics according to the result of sampling survey. In other words, zero cell issue exists that clear solution has not been provided yet. According to the result of analysis on data, it was turned out that bias existed on the sample data of household travel diary survey. In addition, it was confirmed that statistical values did not exist on the number of sample, and the number of samples representing population was significantly insufficient. Therefore, this analysis was intended to review previous methodology of creating synthetic population along with theoretical consideration on it. Simulated Annealing (SA) is to be utilized for the case of zero cell issues or bias intending to develop algorithm of creating synthetic population in an improved form. However, appropriate rules have not been established for the estimation of parameters in simulated annealing, and it requires a significant amount of time to find optimal solution. In order to apply SA into the synthetic population, hill climbing and cooling schedule should be considered. In this study, total absolute error was calculated to prevent hill climbing and used metropolis standards to determine whether to select or dismiss follow-up distribution. In addition, stability