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

of the algorithm was determined through scenario analysis of the optimal combination of iteration and temperature on the cooling issue.

This paper consists of four sections. Section 1 draws the necessity of new population synthesizer. In section 2, algorithm of population synthesis is introduced based on the IPF and SA reviewed including to measure its reproducibility according to hill climbing and cooling schedule problems. In the section3, we create synthetic population at capital area using IPF and SA. And performance of algorithms is evaluated. In the section 4, the main findings are discussed.

2. Algorithm for creating Synthetic Population

2.1. Iterative Proportional Fitting: (IPF)

IPF is an algorithm introduced by Deming and Stephan in 1940 which stated that primary marginal distribution is identified by external resource if more than two populations are divided into the form of distribution. However, if it is not feasible to identify the number of population in each cell of joint distribution of more than two dimensions, the number of population in the marginal distribution and the number of samples in each cell are used to realize the number of population in each cell by specifically utilizing iterative algorithm. In other words, the basic idea is to proportionally adjust distribution of sample groups according to each dimension if the information of joint distribution between each of the variables in the sample group is given making it converge to the population for estimating joint distribution (Beckman, 1996, Bradley, 2003. etc.). Procedures of applying IPF if the number of variables to be considered is two are as follows.

【Step1】 Assumptions

; The number of each X and Y is i and j, respectively, as random variable. In other words, as for population with $X = x_1, x_2, \cdots x_I$, $Y = y_1, y_2, \cdots y_J$, the limit distribution of X and Y in the population is given with $p_i \coloneqq p(X = x_i), i = 1,2, \cdots I$, $p_j \coloneqq p(Y = y_j), j = 1,2, \cdots J$. In addition, information of N different samples $(X_k Y_k), k = 1,2, \cdots, n$ exists.

【Step2】 Calculate $\tilde{p}_{x_i}^{(0)} \coloneqq \sum 1(X_k = x_i)$and $\tilde{q}_{y_j}^{(0)} \coloneqq \sum 1(Y_k = y_j)$ of limit distribution of X and Y from the population

【Step3】 For all i, proportionally adjust them so that limit distribution of samples coincides with ones of population. In other words, $\tilde{p}_{x_i}^{(1)} \approx \tilde{r}_{x_i}^{(1)} \times \tilde{p}_{x_i}^{(0)}$ *(eq. a)*, $\tilde{r}_{x_i}^{(1)} = \frac{p_i}{\tilde{p}_{x_i}^{(0)}}, i = 1,2, \cdots I$

【Step4】 For all j, proportionally adjust them so that limit distribution of the sample coincides with the one of population as shown in the 3rd procedure.

【Step5】 For all i and j, stop applying aforementioned procedures if the difference of limit distribution between newly adjusted sample and the population is smaller than allowable errors. Otherwise, go back to the 2nd procedure and calculate $\tilde{p}_{x_i}^{(2)}, \tilde{p}_{y_j}^{(2)}$ again.

【Step6】 Repeat 2nd procedure ~5th procedure

IPF is referred to as ranking method. The reason why it is specifically referred to as raking method is that information of column and row is exchanged in turn reducing the error. It was derived from a phenomenon of levelling out the bumpy grounds to the left/right and up/down repetitively, in other words, removing the difference of height. In addition, the raking method represents the most important assumption of algorithm. In order to level out the bumpy ground, it is recommended to perform top-down raking one time and left-right raking for another time. Top-down raking influences on the row and left-right raking represents to influence on the column. The important point here is to vertically consider the influence on column and row in the IPF as shown in the vertical raking and in the assumption of independent processing. Therefore, the interaction information of column and row is not considered in the IPF (Beckman, 1996). This plays a crucial role when IPF is practically used that reference variable

requires highly systematic interrelationship with target variable. However, it represents how efficiency of weight is enhanced with low interrelationship between reference variables. Being vertical between variables represents independent relationship.

However, it is difficult for IPF to be applied on convergence for parts without statistics according to the result of sample survey. In other words, if the result of scope is 0 from the result of sample survey, there exists limitation on analysis since iterative IPF converges to zero. This is specifically called as zero-cell. Beckman (1996) has intended to solve zero cell problems by entering small random numbers to the zero cell (seed method) or applying random values of samples in the entire area for the information to be entered in the zero cell.

In this study, it is intended to use sample data of household travel diary survey establishing the representative data of complicated individual social and economic characteristics and to create synthetic population to be used in the activity-based model by using attribute-specific limit distribution in the census data. However, according to the result of analysis on sample survey, samples were not collected in the joint distribution of a significant number of individuals. This specifically represents that additional research methodology is required to create synthetic population through IPF in the procedure of establishing activity-based model.

2.2. Simulated Annealing (SA) Method

The origin of a word, Simulated Annealing (SA method), was derived from thermodynamics and metallurgy. Molten iron has the least energy structure when it is slowly cooled off and solidified. Annealing procedures describe local-exploring method.(Kirkpatrick et. al. ; 1983) In other words, the randomly selected solution among all the solutions that are adjacent to current solution is to be renewed. Afterwards, temperature is steadily reduced that new targets of solution are limitedly selected. In the last procedure, the one and only move improving cost function (f) is practically selected.

The general flow of the simulated annealing is to change metals in a high temperature to low temperature increasing the intensity of metal as a procedure of development in solving for the optimal solution. Specifically, metal annealing adjusts the speed applied from high energy to low energy level according to how high temperature is changed to low temperature and hence increase the intensity of metal. The more it is rapidly changed from high temperature to low temperature, the lesser the intensity of metal is turned out to be. Then, instable structure of atom is formed to represent optimum condition. On the other hand, as the metal is slowly cooled off, the atom structure of metal is shaped with stable lattice form representing the optimal energy level. To be specific, particular solutions near current solution are identified by repetitively improving the solution that the value of function and control variable T (temperature) influence on each other. In addition, T (temperature) value is slowly decreased with the similar principle that is mentioned earlier. Therefore, a degree of solution to be changed is significant due to high (temperature). However, as T (temperature) becomes closer to 0, the range of change is reduced.

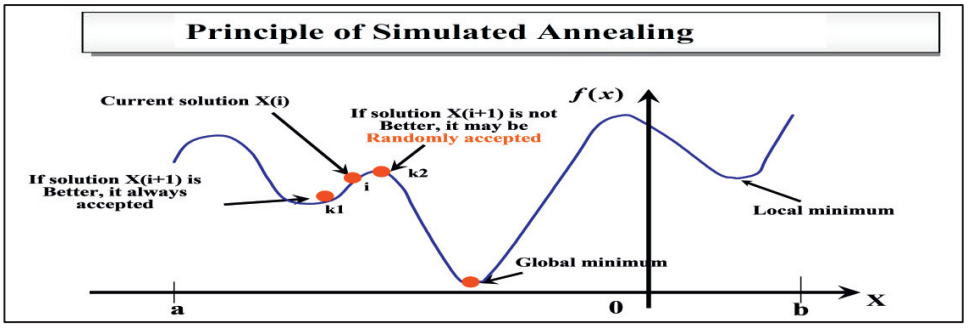

Fig. 1. General framework for Simulated Annealing.

Generalized procedures of finding optimal solution through simulated annealing are shown in the <Fig. 1>. Simulated annealing applied for creation of synthetic population in this study starts from initial solution that is randomly created. The neighbouring x^{l} is selected from the neighbourhood group $N(x)$ that is pre-established from

the current solution *(x)*. If the cost of next solution is lower than the one of current solution, the neighbouring solution x^1 is accepted as the next solution. If this is not the case, the chance is determined to be $e^{-(F(x^1)-F(x_n))/T}$ (this is frequently performed as the Monte Carlo Simulation method). Here, T represents "temperature" as a control variable. This procedure is repeated in a slow pace of *T* until ending conditions are satisfied.

In general, Newton Rapson method or Secant method is used for optimization problem. However, such methods are limited to find local minimum as the optimal solution. However, simulated annealing method performs exploration process through the entire scope of solution finding the global minimum. Due to this reason, it is reasonable to see that SA is designed to efficiently find solutions in optimization problem.

In this study, algorithm of satisfying joint distribution of multi-dimension in the synthetic population can be explained as follows by using SA. First of all, it is required to select neighbouring distribution x^l from the neighbourhood distribution group *N(x)* that is previously selected in the current distribution *(x)* in each of the procedures. At this time, if the error of the next distribution is smaller than the cost of current solution, neighbouring distribution is accepted as the next distribution. If this is not the case, neighbouring distribution tends to be accepted through probabilistic procedures. Here, probability is established according to the size of error of neighbouring distribution. In other words, the probability of being accepted is increased with smaller error to prevent hill climbing phenomenon in the procedure of finding optimal solution. In addition, a control variable of temperature *(T)* is established stably optimizing the solution by repeating the slow reduction of *T* until ending conditions are satisfied. This can be specifically explained as follows.

【Step1】 Initialization : Setup of default temperature: 10000 °C , Maximum Iteration setting

【Step2】 Setup of the total amount of columns and rows in the population and enter observed values of sample distribution(Diagonalization process, Yosef Sheffi, 1985)

【Step3】 Setup of sample distribution composed with random numbers that satisfy total amount restrictive conditions.(Before distribution)

【Step4】 Setup of sample distribution composed with random numbers that satisfy total amount restrictive conditions (After distribution)

【Step5】 Calculation of absolute error on the before/after distribution as well as observed data. (total absolute error :TAE)

$$TAE = \sum_{ijk} \left| O_{ij}^{(k)} - S_{ij}^{(k)} \right| \quad \text{(eq. 1)}$$

$O_{ij}^{(k)}$: Estimated values of *i* column and *j* row

$S_{ij}^{(k)}$: Estimated values of *i* column and *j* row in the *k*-th distribution

Compare the total amount of absolute error in the before and after distribution

$$\Delta E = TAE^{after} - TAE^{Before} \quad \text{(eq. 2)}$$

【Step6】 Calculation of selection probability

$$Pr(accept) = \begin{cases} 0 \ , & \Delta E > 0 \\ exp(\Delta E/T) , & \Delta E \le 0 \end{cases} \quad \text{(eq. 3)}$$

If $\Delta E \le 0$, determine acceptance rate through $exp(\Delta E/T)$. Determination of after-distribution acceptance rate by generating random number between 0 and 1 (Marcov Chain Monte Calro)

【Step7】 Repeat procedures in the step4~step6 and end calculation when TAE has the smallest value or satisfies ending conditions

In this study, conditions of sample data are assumed to be metal conditions in a high temperature, and two randomly measured distributions specifically indicate chilled metals. At this time, the level and difference of energy

are measured to be TAE(total absolute error), and repetitive renewal of the distribution with small error value makes it feasible to estimate the distribution with the lowest energy level, in other words, the distribution with the least absolute value of error. 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 addition, simulated annealing method does not always renew the best solution all the times that there is a chance for hill climbing to be occurred that more inappropriate solutions over current solution might be probabilistically selected. Hill climbing specifically represents limitation for how it might deviate from the optimal solution due to repetitive procedures even after the optimal solution is already found.

In this study, Metropolis-Hasting Algorithm (M-H algorithm) as a representative method of Marcov Chain Monte Carlo(Hastings ; 1970) was applied for the probability for the estimated neighbouring distribution to be accepted as the next distribution in order to prevent hill climbing phenomenon. The reason why M-H algorithm is used to prevent hill climbing for removing the procedures of finding optimal solution that probabilistically deviates from the best solution by adopting the concept of selection/dismissal probability. M-H algorithm is normally used when selecting the probability for neighbouring distribution x^1 to be accepted as the next distribution. In other words, it is an algorithm to be used for determining the transfer of an object to the next condition by considering the condition and energy of a certain object. M-H algorithm compares the difference of absolute errors obtained in the SA in the issue of creating synthetic population (eq. 2). Here, ΔE is a function converting restriction-minimizing function to the restriction-free minimizing problem. At this time, the probability of selecting or dismissing M-H algorithm should be selected (eq. 3).

When $\Delta E \geqslant 0$, newly estimated distribution is rejected, whereas if $\Delta E \leqslant 0$, it is not always the case that newly estimated distribution is to be accepted, but random sampling number is generated from the chance of $\exp(\Delta E/T)$ and in the range of {0,1} is compared to determine whether to accept neighbouring distribution or not. In other words, if it is Pr(accept)>R, new distribution is accepted renewing previous distribution. However, in case of R>Pr(accept), new distribution is removed. This is called as Metropolis Criteria. Another important element considered when applying simulated annealing is cooling schedule issue. The minimum value in the entire range is stably obtained when the temperature is lowered in an appropriate way. Therefore, the analysis was conducted on the requirements for finding the best solution on temperature and changes of the repetitive frequency to identify how cooling schedule influences the accuracy of algorithm.

3. 3. Creation of Synthetic Population in Capital Areas

3.1. Setup for Analysis Data

Areas of analysis were classified according to household travel diary survey data in 2010 and also categorized into 72 different activity patterns group to create synthetic population. 72 different numbers of cases were occurred depending on gender (2), age (6), and career (6), and it was also assumed that each of the case represented similar patterns of travel. As for values of the characteristics in measuring household travel diary survey, relationship with household, age, gender, whether to have driver license, career, employment type and career type were specifically identified. However, measuring all these joint distributions caused huge number of cases that ended up reducing efficiency of the analysis. Therefore, priority in the attributes in influencing the travel patterns of individuals was determined. Age was classified into 6 stages in an order of preschool children, juveniles, teenagers, middle-aged class, post middle-aged class, and the senior, and 9 items of the household travel diary survey were incorporated into the agriculture/fisheries industry/production/general labourers along with sales and service fields in total 6 types of career areas. Therefore, total 72 distinct capital populations were classified by separating 6 classes of age, 6 classes of career, and 2 classes of age. Totalizing data in each of the scopes on the age, gender, and career of the category of synthetic population were collected. In addition, data from 『Working Population of Current Residency, Work Place, and Areas (more than the age of 15)-City, District, and Province (2010) 』 provided by the National Statistics Office were used. Data of 『Population of Age and Gender - Eup, Myeon, Dong (2010) 』 were used and properly applied for the number of population according to the gender and age in each of the regions.

Table 1. Synthetic Population category settings.

Sex	Age		Carrer
Man	0~7	Preschool child	student
Woman	8~19	Youth	House maker/Preschool child
	20~29	20's	Specialized job/Technical post
	30~39	30's	Administrative job/ Office job
	40~65	elderly citizens	Sales/Service
	65	old age	Others

The following figure shows the setup for analysis data which is arranged according to age, sex and job and compares statistical results between sample and census data. For example, with respect to age, the census data comprises of almost 15.47% of population aged between 8 – 19 years whereas it is 21.98% for sample population. Same way, percentage of students in census population is 18.18% whereas the value is 26.77% for sample population.

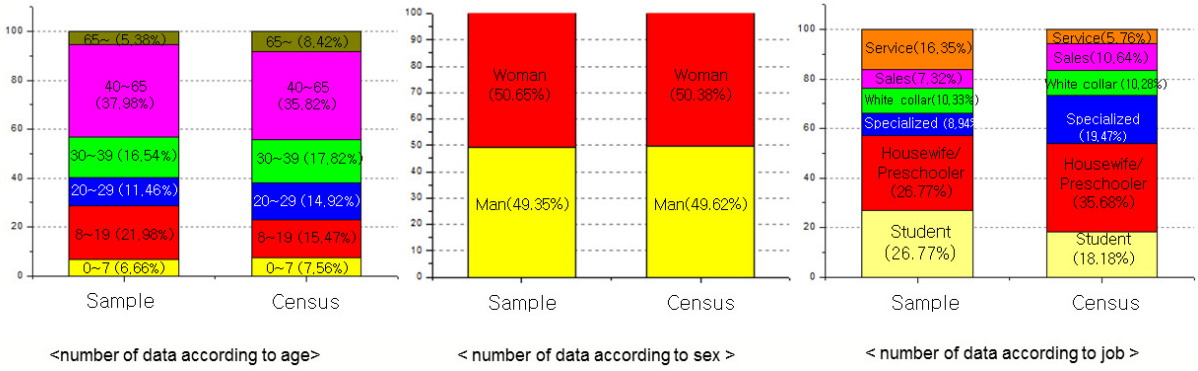

Fig. 2. Synthetic Population Category settings.

3.2. Creation of Synthetic Population using IPF

Synthetic population in each of the groups by activity types created in each of the areas for analysis represented a minor disparity in the result from census-based data. IPF is changed depending on the adjusting variables of column and row that it is theoretically impossible for each of the columns and rows to be completely coincide. However, it is very similar with distribution of census-based data. The results of synthetic population using IPF have been shown in Fig. 3.

In the figure, x-axis represents age of sample and generated population (six ranges), y-axis represents six job categories and z-axis represents the distribution of population for census and generated population. Age is categorized into 6 ranges i.e. 0-7, 8-19, 20-29, 30-39, 40-65, 65~. Similarly job is categorized into six categories i.e. student, Housewife/ Preschooler, Specialized, White collar (Administrative or office job), Housewife and others.

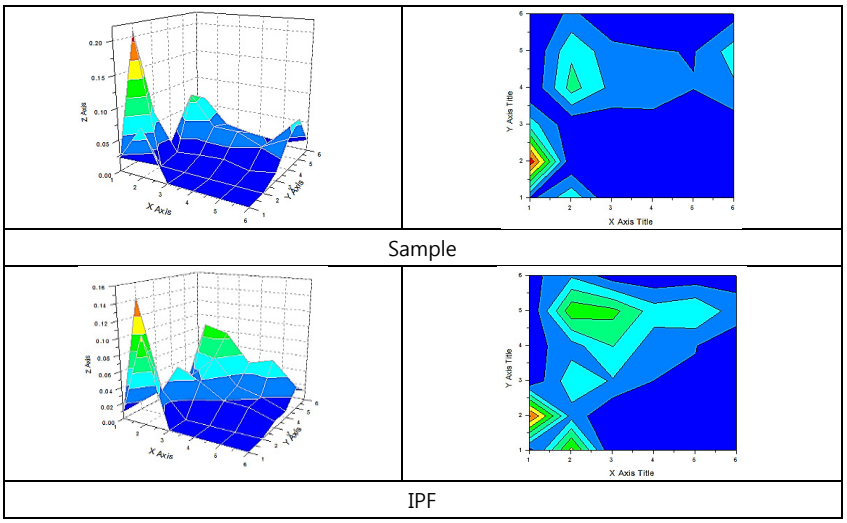

Fig. 3. Synthetic Population distribution results (up: Sample, low: IPF)

3.3. Creation of Synthetic Population by using Simulated Annealing (SA) algorithm

Algorithm was applied in the same manner as explained earlier, but Diagonalization process (Yosef Sheffi, 1985) was specifically used for sequentially setting variables in each of the categories to establish randomly selected sample distribution to satisfy the requirements of total amount of prescription. This is to finalize phased category distribution as an assumption to satisfy the requirements of total amount in population.

Hill climbing and cooling schedule issues should be considered for application of SA. First of all, this study calculated total absolute error (TAE) by utilizing M-H algorithm to prevent hill climbing and determined whether to select or dismiss follow-up distribution. In addition, stability was determined through optimal combination between iteration and temperature according to cooling schedule. As for hill climbing, metropolis criteria were applied according to M-H algorithm solving the issue of hill climbing. However, as for cooling schedule, the change of frequency of maximum iteration, an ending condition of initial temperature, was analysed by establishing a certain scenario.

Table 2 Set up cases for SA Algorithm parameter.

division	Temperature T	Maximum Iteration	division	Temperature T	Maximum Iteration
Case 1	10,000	25,000	Case 7	30,000	25,000
Case 2	10,000	50,000	Case 8	30,000	50,000
Case 3	10,000	75,000	Case 9	30,000	75,000
Case 4	20,000	25,000	Case 10	10,000	100,000
Case 5	20,000	50,000	Case 11	20,000	100,000
Case 6	20,000	75,000	Case 12	30,000	100,000

According to the results of applying the Simulated Annealing algorithm, the trends of the convergences were derived as follows. First of all, if the temperature was lower, different values were obtained depending on the iterative calculation frequency, and if the temperature was in the middle range, the same results were obtained in the mediocre level of iterative calculation. On the other hand, if the temperature was high, optimal values were obtained differently even if the iterative calculation frequency was increased. This was due to the reason that the entire-scope specific solution could not be found due to high speed of convergence of the Simulated Annealing algorithm if the

temperature was low making it stay in a particular range. If the temperature was high, iterative calculation was required in a certain degree in a repetitive manner to find the optimal solution due to long lasting process of convergence. However, it is not always recommended to increase initial temperature and iterative calculation frequency as an ideal parameter for the successful setup of algorithm. This is because the influence of temperature becomes lower as the iterative calculation frequency is lower, and it becomes more sensitive to temperature as the iterative calculation increases. Therefore, it causes inefficiency due to repetitive calculations along with costs in time for the analysis if temperature is high. In conclusion, if initial temperature is more than 20,000, and iterative calculation frequency is more than 100,000, it converges to one certain value for setup of temperature.

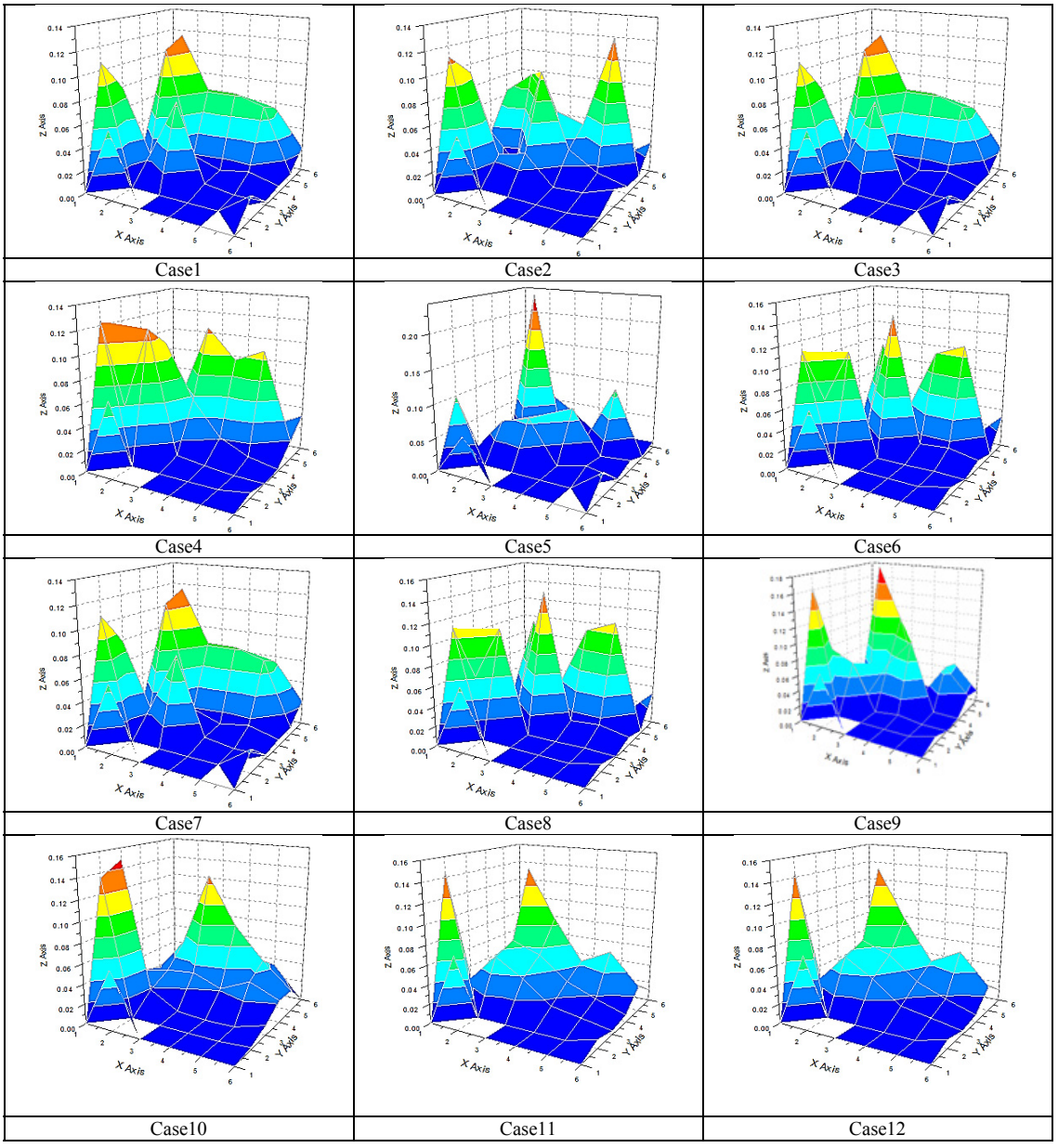

Fig. 4. The result of SA Algorithm based on different parameter cases.

3.4. Verification of results in creating Synthetic Population at Metropolitan area

Evaluation of Validity and Accuracy: Assumption that sample represents the population

As for comparison and evaluation methods for the estimated distribution, a certain method exists with an assumption that sample represents the population. Four parameters are employed in this section to calculate measure of effectiveness (MOE). To be specific, statistics such as RMSE (root mean squared error), MAE (mean absolute error), Pearson correlation analysis, and x^2 (goodness of fit test) are available on the sample. However, the proportion of total amount should be determined as a criterion for evaluating accuracy since there is a huge difference between finally estimated result and the total amount. Also, for the evaluation in convergence of algorithm, we analyzed Sum of Squared Error (SSE), Relative Average Absolute Difference (RAAD).

RMSE (root mean squared error) and MAE (mean absolute error) are all used for calculating square and absolute values of errors in order to evaluate the difference between estimated values and observed values by the model as well as measured values. *N* means the number of variables, θ_i means joint distribution, and $\hat{\theta}_i$ indicates sample data.

$$RMSE(\theta_i, \hat{\theta}_i) = \sqrt{\frac{\sum_{i=1}^{n}(\theta_i, \hat{\theta}_i)}{N}} \quad \text{(eq. 4)}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|(\theta_i - \hat{\theta}_i)\right| \quad \text{(eq. 5)}$$

Correlation coefficient(ρ) is to identify patterns of estimated distribution. Correlation coefficient is between -1 and 1. As it becomes closer to +1, it means to be positively correlated. If the correlation coefficient is 0, it indicates that they are not linearly related.

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \quad \text{(eq. 6)}$$

Here, N indicates the number of sample, x_i, y_i means the *i*th number of *x* and *y*, $\bar{x}, \bar{y}$ are averages of *x* and *y*, and s_x, s_y indicate standard deviation of *x* and *y*.

Chi-square distribution is used to verify equivalence of multi-variables distribution. Verification statistics to be used for equivalence of multi-variables distribution are as follows.

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{h}\frac{(Y_{ij} - n_i\hat{p}_j)^2}{n_i\hat{p}_j} \quad \text{(eq. 7)}$$

The result of comparing creation of synthetic population by IPF and SA on four different cases is as follows. As for RMSE and MAE, IPF was turned out to be superior a little bit. As for correlation coefficient, SA was turned out to be less linearly correlated. As for chi-square and verification statistics, they were all belonged to the proper range, but a significance level of ρ was turned out to exceed 0.05 that could not negate the hypothesis in the confidence level of 95%. Therefore, both methods could not negate the original hypothesis that represented sample data and proportion on the identical distribution, and, hence, it is reasonable to say that data distribution estimated by both methods is different.

Assumption that sample does not represents the population

Verification is required when sample cannot represent population if parameter values are unknown in order to evaluate method of creating synthetic population. In other words, joint distribution derived by the household travel diary survey is needed for verification on procedures of creating synthetic population if it is not identical with joint distribution of real population. This is because it is not feasible to identify characteristics of data combined with features of population in the procedure of creation for synthetic population to be performed in this study. Hereupon,

virtual synthetic population is created according to samples of household travel diary survey. Therefore, it is determined that statistical verification is needed on an assumption that sample might not represent the population.

Table 3. Evaluation of algorithm of generating synthetic population.-a

division		RMSE	MAE	ρ	x^2
IPF		4.61931	3.89352	0.19180	54.5726
SA	Case 1	4.54888	3.86377	0.24131	50.553
	Case 2	4.54888	3.86377	0.24131	50.553
	Case 3	4.54888	3.86377	0.24131	50.553
	Case 4	5.19233	3.96242	0.19434	50.935
	Case 5	5.26769	4.36426	0.02959	61.14
	Case 6	5.21462	3.95517	0.13046	49.143
	Case 7	4.8364	3.92649	0.10199	55.414
	Case 8	4.7224	3.9806	0.10766	51.103
	Case 9	4.7224	3.9806	0.10766	51.103
	Case 10	4.4150	3.6108	0.21539	48.174
	Case 11	4.1535	4.1215	0.18956	55.879
	Case 12	4.1535	4.1215	0.18956	55.879

In such cases, statistical verification is feasible through confidence analysis. To be specific, it is feasible to evaluate confidence of joint distribution estimated through Cronbach Alpha Coefficient. Cronbach Alpha Coefficient, that is statistically used the most, is a coefficient for internal stability. If the internal consistency is high, it represents reliability of joint distribution on an assumption that detailed categories of joint distribution categorized in each area. Cronbach Alpha Coefficient value is between 0 and 1. If it is more than 0.6, the value is determined to be reliable. If it is more than 0.7, the value is highly reliable. In addition, if it is between 0.8 and 0.9, the value is determined to be very reliable. This represents how there is no clear solution in terms of % of significance level when performing general hypothesis verification even if subjective elements are included on the confidence coefficients, and, therefore, Cronbach Alpha can be placed in every range. Cronbach Alpha Coefficient is calculated through correlation coefficient as follows.

$$Cronbach\ \alpha = \frac{(n \times ave(cor))}{1+[(n-1) \times ave)cor)]} \quad \text{(eq. 8)}$$

Here, ave(cor) is the average of correlation coefficient, and n is the number of item.

Average of correlation coefficients is the average of correlation coefficient of pairs in the 2way dimension distribution. Here, if the calculated confidence coefficient is greater, it represents strong correlation between all the pairs in the detailed categories. However, if the calculated confidence coefficient is smaller, it represents inconsistency that optimal combination is in instable condition and is also sensitive to the change. In other words, high correlation coefficients specifically mean to have stable form of joint distribution and to be closer to optimal solution as an indirect inference.

According to the result of confidence analysis with IPF, confidence was analysed to be stable. IPF, that was proved to be statistically superior by many scholars, is confirmed to be statistically superior and have stable confidence in this study. According to the result of analysing SA in each of the cases in terms of confidence, it was analysed that low temperature was entailed with low confidence. In addition, an increase in iterative calculation frequency to a

certain degree improves confidence over when the iterative calculation frequency is too small in the interim level of temperature. On the other hand, as the iterative calculation frequency increases in the interim range of temperature, it becomes more reliable. However, it was not confidential at all in the iterative calculation frequency of 75,000 after it became more reliable up to a certain point. When the temperature was set to be high, it was overall reliable. Therefore, a significant conclusion was reached that creation of synthetic population by simulated annealing established in this study is to set the initial temperature to be high, and the criterion of iterative calculation frequency should be established in a proper range.

Table 4. Evaluation of algorithm of generating synthetic population.-b

Division		Cronbach Alpha Coefficient	note
IPF		0.74179	“highly reliable”
SA	Case 1	0.5187	“low reliable”
	Case 2	0.5187	“low reliable”
	Case 3	0.5187	“low reliable”
	Case 4	0.61408	“reliable”
	Case 5	0.70944	“highly reliable”
	Case 6	0.56713	“low reliable”
	Case 7	0.62392	“reliable”
	Case 8	0.7171	“highly reliable”
	Case 9	0.7171	“highly reliable”
	Case 10	0.7117	“highly reliable”
	Case 11	0.7776	“highly reliable”
	Case 12	0.7776	“highly reliable”

Convergence of algorithm

The Sum of Squared Error(SSE) and Relative Average Absolute Difference(RAAD) are used measure of effect for the evaluation in convergence of algorithm.

$$SSE(\theta_i, \hat{\theta}_i) = \sum_{i=1}^{n} (\theta_i - \hat{\theta}_i)^2 \quad \text{(eq. 9)}$$

$$RAAD = \frac{1}{n} \sum_{n=1}^{n} \left| \frac{\hat{\theta}_i - \hat{\theta}_{i-1}}{\hat{\theta}_{i-1}} \right| \quad \text{(eq. 10)}$$

Here, θ_i is the observed distribution proportion at iteration$i$, $\hat{\theta}_i$ is estimated distribution proportion at iteration$i$. Variation of Sum of Squared Error (SSE) and Relative Average Absolute Difference (RAAD) according to iterations represent SA algorithm is good at astringency than IPF. During initial iterations, SA shows greater sum of squared error but later both algorithms converge near 100th iteration. Whereas SA converges more accurately if analyzed by RAAD at 100th iteration.

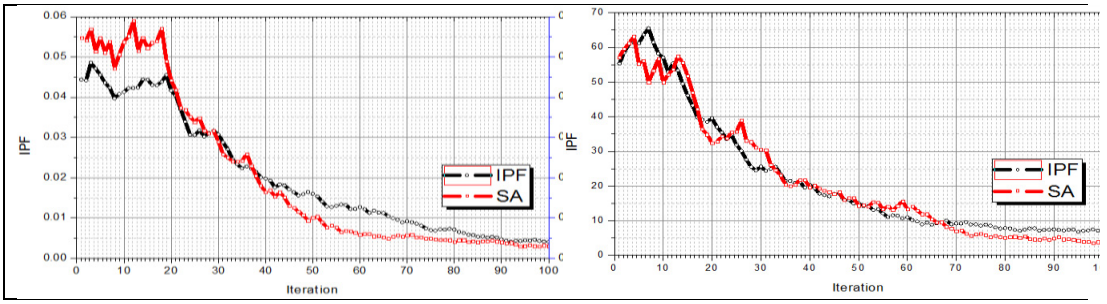

Fig. 5. Variation of MOE (measure of effectiveness) according to the number of iterations (left: SSE, right: RAAD).

4. Conclusion

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. The synthetic populations have been evaluated to measure the reproducibility of the algorithm.

The Simulated Annealing (SA) was derived from thermodynamics and metallurgy. First of all, it is required to select neighbouring distribution xl from the neighbourhood distribution group N(x) that is previously selected in the current distribution (x) in each of the procedures. At this time, if the error of the next distribution is smaller than the cost of current solution, neighbouring distribution is accepted as the next distribution. If this is not the case, neighbouring distribution tends to be accepted through probabilistic procedures. Metropolis-Hasting Algorithm as a representative method of Markov Chain Monte Carlo was applied to calculate the probability for the estimated neighbouring distribution to be accepted as the next distribution in order to prevent hill climbing phenomenon (Hastings, 1970). The reason why M-H algorithm is used to prevent hill climbing for removing the procedures of finding optimal solution that probabilistically deviates from the best solution by adopting the concept of selection/rejection probability. M-H algorithm is normally used when selecting the probability for neighbouring distribution xl to be accepted as the next distribution. In other words, it is an algorithm to be used for determining the transfer of an object to the next condition by considering the condition and energy of a certain object.

First, we have analysed 599,790 persons and 1,365,071 trip data from the household travel survey for the activity based model. The sample data of the household travel survey was classified according to individuals' multi-dimensional socio-economic characteristics. When the sample data of the household travel survey have been applied to the activity-based model, we have investigate if the sampling bias might cause problems.

M-H algorithm compares the difference of absolute errors obtained in the SA in the issue of creating synthetic population. Here, ΔE is a function converting restriction-minimizing function to the restriction-free minimizing problem. At this time, the probability of selecting or dismissing M-H algorithm should be selected. The iterative proportion fitting (IPF), the existing method to generate synthetic population, is a simple and efficient calculation method. There are constraints in the use when there is sampling bias, namely when the part without statistics in sampling results, the problem of zero cells, occurs. Therefore, the simulated annealing (SA) technique which is applicable to the case with the problem of zero cells or sampling bias was utilized to develop the improved algorithm to generate the synthetic populations.

Second, a Simulated Annealing (SA) algorithm has been developed considering hill climbing and cooling schedule problems. In order to apply SA into the synthetic population, hill combing 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 of the algorithm was determined through scenario analysis of the optimal combination of iteration and temperature T on the cooling issue. Based on this result, the current condition of household travel diary survey and census data were used comparing the IPF 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.

This paper concludes that simulated annealing (SA) is a powerful algorithm that can be used to generate population synthesis. Results show that traditional algorithm i.e. IPF has zero cell problems or sample biased problems but SA algorithm can not only overcome aforementioned problems but also addresses issues like hill climbing and cooling schedule in an effective manner. We suggest that SA algorithm is more flexible than IPF in creating population synthesis.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRK) grant funded by the Korean Government (MEST), NRF-2014R1A2A2A01005155.

References

Balmer, M., 2007. Travel Demand Mmodeling for MULTI-AGENT TRANSPORT SIMULATIONS: Algorithms and Systems, Ph.D. Thesis, Uni. of ETH Zurich.

Beckman, R. J., K.A. Baggerly, M.D. McKay., 1996. Creating synthetic baseline populations. Transportation Research Part A, 30, 6, 415-429.

Bhat, C.R., and F. S. Koppelman., 1999. A Retrospective and Prospective Survey of Time-Use Research. Transportation(Kluwer Academic Publisher) 26, 119-139.

Bhat, C.R., Srinivasan, S., and Guo, J., 2002. Activity Based Travel Demand Analysis for Metropolitan Areas in Texas: Data Sources, Sample Formation and Estimation Results, Report 4080-3, prepared for the Texas Department of Transportation.

Bowman, J.L. and Ben-Akiva, M.E., 2000. Activity Based Travel Forecasting. Activity-Based Travel Forecating Conference.

Bradley, M., 1999. Methodology and Results of Generating a Prototypical Population. working paper, July 8.

Bradley, M., 2003. A Discussion of the Population Synthesis Approach for Atlanta. working paper, July 28.

Bradley M., M. Outwater, N. Jonnalagadda and E. Ruiter., 2001. Estimation of an Activity-Based Micro-Simulation Model for San Francisco. Paper presented at the 80th Annual Meeting of the Transportation Research Board.

Deming W.E., and Stephan F., 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Annals of Mathematical Statistics.

Harland K., Heppenstall A., Smith D. and Birkin M., 2012. Creating Realistic Synthetic Populations at Varying Spatial Scales : A Comparative Critique of Population Synthesis Techniques, Journal of Artificial Societies and Social Simulation, 15(1)

Hastings, W. K., 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika.

Hensher, D.A., P.R. Stopher, P. Bullock and T. Ton., 2004. TRESIS (Transport and Environmental Strategy Impact Simulator): Application to a Case Study in NE Sydney. presentation at the 83rd Annual Meeting, Transportation Research Board.

Kirkpatrick, S., Gelatt Jr, C. D., Vecchi, M. P., 1983. Optimization by Simulated Annealing. Science

Ma, J., 1997. An Activity-Based Approach and Micro-simulated Travel Forecasting System : A Pragmatic Synthetic Scheduling Approach. Thesis, The Pennsylvania State University.

Ryan, J., H. Maoh, P. Kanaroglou., 2009. Population Synthesis: Comparing the Major Techiques Using a Small Complete Population of Firms. McMaster Universoty, working paper series, CSpA 026.

Yosef Sheffi., 1985. Urban Transportation Networks. Prentice-Hall