AN AGENT BASED ESTIMATION METHOD OF HOUSEHOLD MICRO-DATA INCLUDING HOUSING INFORMATION FOR THE BASE YEAR IN LAND-USE MICROSIMULATION

Kazuaki Miyamoto, Faculty of Environmental and Information Studies, Tokyo City University, 3-3-1 Ushikubo-nishi, Tsuzuki, Yokohama, 224-8551 Japan, miyamoto@tcu.ac.jp

Nao Sugiki, Research & Planning Department, Docon Co., Ltd., 4-1, 5-chome, 1-jo, Atsubetsuchuo, Atsubetsu, Sapporo, 004-8585 Japan, ns1491@docon.jp

Noriko Otani, Faculty of Environmental and Information Studies, Tokyo City University, 3-3-1 Ushikubo-nishi, Tsuzuki, Yokohama, 224-8551 Japan, otani@tcu.ac.jp

Varameth Vichiensan, Faculty of Engineering, Kasetsart University, Phaholyothin Rd., Jatujak, Bangkok 10900, Thailand, fengvmv@ku.ac.th

ABSTRACT

Much development has been realized in land-use microsimulation models that are often used to forecast the changes in the micro-data, e.g., the detailed attributes and location of individual households. However, less attention has been paid in synthesizing the base year micro-dataset. The purpose of this study is to build a system to rationally estimate the micro-dataset of the base year for land-use microsimulation. This paper presents a result of continued development by incorporating location and housing type choice models into our earlier system, making the micro-data synthesizer comprehensive in that it produces details attributes of household such as its member and age composition, housing type, as well as its spatial location. This system, wherein a Monte Carlo simulation is used, deals with both continuous and discrete attribute variables by agent (i.e., a household in this study). It uses sample data that contain full information on the micro-data to establish the correlation between the attributes and the existing statistical data as the control total in each zone. To reproduce the correlation between continuous attribute variables, independent variables that can be obtained by the principal component analysis of the original variables are introduced and employed as intervening variables. The housing type and location choice are determined by means of Logit models. In addition, the system is composed of several iterative adjustment processes such that the data production satisfies the available control total by attribute. In order to develop the system in a rational and objective manner, an indicator is proposed to evaluate the goodness-of-fit between two micro-datasets. In other

words, by introducing the indicator, this study contributes to the development of not only an estimation system but also an approach to system development. Finally, a case study of the system application to a person-trip-survey dataset of the Sapporo metropolitan area is presented; the results validate the usefulness of the system and approach.

Keywords: Land Use Model, Micro-Simulation, Household Micro-data, Population Synthesis, Agent-based Approach

INTRODUCTION

Microsimulation is becoming a popular approach in land-use modelling or urban modeling, to describe the detailed changes in the land-use and transport system of a metropolis. In this study, a set of data of households required for the base year in the simulation of land-use microsimulation models, of which a residential location model is a typical example, is considered. A household is characterized by many attributes such as number of members, age of the members, location, and housing type. Every of these attributes of a household must, in principle, be defined for the base year in the microsimulation. However, this kind of data is usually not available, because the retrieval of individual data from administrative registers or census violates the right to privacy and is generally prohibited in most countries. Therefore, the data used for the microsimulation models are “synthetic populations” created from generally accessible aggregate data provided by the national census, with additional information obtained by conducting sample surveys. In most of the existing procedures used to generate synthetic populations, the number of households by type is estimated after setting the household types with the iterative proportional fitting (IPF) procedure. However, this approach has several difficulties when the model has to deal with many types of attributes of household, as described later in this paper. Another approach is to generate a set of individual synthetic households, each of which has its unique attributes; this is called “micro-data” in this study.

The purpose of the present study is to develop a consistent method for estimating or generating a set of synthetic household micro-data for a study area. In our previous study, a set of attributes of a member composed of member’s relationship with the household head and member’s gender and age is only considered (Miyamoto et al., 2010). In this paper, other household micro-data such as housing type and the location within the study area are also added to the previous estimation system. Although the system was originally developed for households, the method can be applied to “agents” in general. Since it is difficult to represent the system in a single mathematical form under different conditions in terms of the available data, such a system becomes ad hoc. Therefore, to develop the system in a rational and objective manner, this study has proposed an indicator to evaluate the goodness-of-fit between two micro-datasets. Together with the extension of the estimation system, the way of goodness-of fit evaluation is also extended (Otani et al., 2010).

The paper is structured as follows. First, state-of-the-art approaches to generating synthetic populations are described. After discussing the limitation of the existing methods, a system is presented that is built to generate a set of agent-based household micro-data. An indicator is then introduced to evaluate the goodness-of-fit between two micro-datasets. Finally, the usefulness of the system and approach is evaluated by applying this system to person-trip-survey data of the Sapporo metropolitan area.

POPULATION SYNTHESIS

IPF based Methods

In the past, travel demand microsimulation such as TRANSIM essentially uses the population synthesizer; the other applications include Guo and Bhat (2007), Ryan et al. (2010), Auld et al. (2010), etc. The Iterative Proportional Fitting (IPF) procedure is popularly used to generate synthetic populations. It was originally proposed by Deming and Stephan (1940) and Beckman et al. (1996) were the first to apply IPF to the problem of generating synthetic populations. Several variants of IPF application in population synthesis were proposed, e.g. Guo and Bhat (2007), Auld and Mohammadian (2010), etc. In land use microsimulation, household and population syntheses are also very essential due to very detail analysis of household attributes are concerned, e.g., Lee and Waddell (2010). Among the household synthesis in land use microsimulation analysis, Miyamoto et al. (1986) generated synthetic households in a metropolis by applying the simultaneous probability maximization principle with margins as constraints; their method is said to be equivalent to IPF in terms of the solving procedure.

In principle, IPF generates the number of agents N_{ijk} (households in this study) in cells of a multidimensional table (i,j,k), which is regarded as a limited number of household types defined by a set of attributes. IPF uses a sample dataset to establish the correlation between dimensions or attributes under the condition that the mapping of the summations of N_{ijk} into lower dimensions should fit the margins given by the census data. Although the data obtained from this approach is useful for constructing microsimulation models, it does not correspond to the micro-data of individual agents but to the number of agents by type (household type in this study). Guo and Bhat (2007) improved the IPF procedure by alleviating the problem of zero cell value and the inability to control the statistical distributions of both household- and individual-level attributes. Pritchard and Miller (2009) also improved the IPF procedure by adding a function to allow many more attributes per agent through a Monte Carlo simulation that is based on a sparse list-based data structure. Both groups have made very useful developments that are still within the scope of the “cell”-based approach, which is inevitable as long as the IPF procedure is used.

Agent-Based Approaches

On the other hand, Moeckel et al. (2003) used an "agent"-based approach. Monte Carlo sampling by agent was employed so that as many features can be selected, as required by a microsimulation model. In this case, the number of features is only limited by the possibility of determining reasonable relationships between the selected attributes.

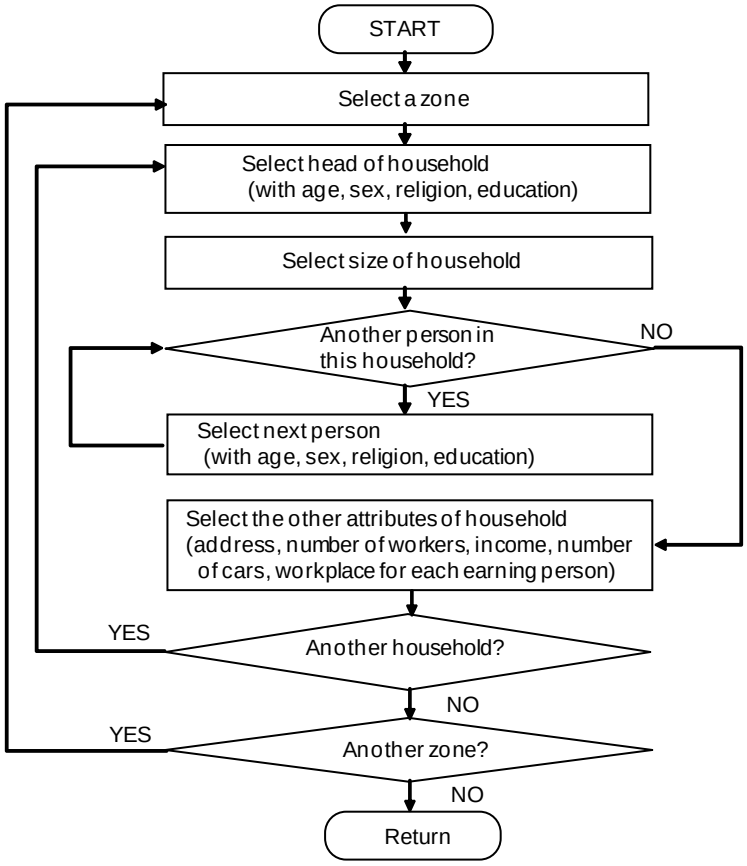


Figure 1 – Synthetic Population Generator proposed by Moeckel et al. (2003)

Difficulties in Conventional Population Synthesis

The IPF procedure has limitations when generating complex micro-datasets. The reliability of the initial input data for the cells of the matrix is required to be very high. IPF has to set zero-cells to 0.1 or 0.01. This condition influences probabilities and its theoretical foundation is weak Moeckel et al. (2003). Besides the issues pointed out there, the cell-based approach, which includes IPF and its extensions, has the following difficulties: neither attributes nor their categories can be set to universal values; the discrete setting of categories causes a similar problem to the modifiable area unit problem (MAUP) in zoning, because many attributes such as age and income are originally continuous; and additional attributes generate more zero-cell problems and reduce reliability.

IPF was originally devised to reduce the computation and memory burden of computers by setting a limited number of agent types. However, this approach does not necessarily alleviate this issue, because the number of combinations of attribute categories can increase to a large extent even though most cells have a zero value.

METHOD

Presuppositions

Considering the advances in computer capability, the agent-based approach has great potential for application to solving the problem of generating synthetic populations, because both the computation and memory burdens of computers have almost been overcome. Therefore, this study is a series of extensions of the agent-based synthesis method proposed by Moeckel et al. (Moeckel et al., 2003) Based on our previous work (Miyamoto et al., 2010), the following presuppositions are extended to cover housing type and the location as follows::

- Households are the target agents.
- Households are characterized by members, and the members are characterized by their relationship with the household head and their gender and age.
- The target micro-data of the households are the relationships of the members with the household head and their gender and age, housing type and the location in the zone.
- The number of households by the number of members and the number of individual persons by five-year age bands are available from the census data obtained in the study area.
- The number of housings by type in the study area is available from the census data.
- The study area is divided into zones in which limited kinds of census data is available.
- The number of households in each zone in the study area is available from the census data.
- A certain number of samples, for which the micro-data are known, are available.

Although the scope of household attributes is limited, this approach has the potential for expansion to deal with more household attributes, when necessary.

Micro-Data

As per the presuppositions, the micro-data for our previous study is defined as a set of attributes composed of member's relationship with the household head and member's gender and age. The dataset is represented by h_{ms} for a household s that has *m* members,

$$h_{ms} = \{\mathbf{c}_{ms}, \mathbf{x}_{ms}\} \quad (1)$$

where $\mathbf{c}_{ms}$: member composition of the household s having *m* members

$\mathbf{x}_{ms}$: age composition

The data can be represented as a vector of ages in the sequence "general household member types," which is represented by member's relationship with the household head and member's gender: e.g., {head (male), head (female), husband, wife, one child (male), three children (male), one child (female), three children (female), father, mother, one grandchild (male)… }. In a more general form, the dataset is given by equation (2):

$$A = \{\mathbf{a}_i = (a_{i1}, a_{i2}, \cdots, a_{iR}) \mid 1 \le i \le N\} \quad (2)$$

where R is the number of possible member types, N is the number of households in the study area, and a_{ik} is age of member k. If member k does not exist in household i, a dummy number is appropriated to a_{ik} to indicate the absence. For example, the vector of *i*th household with a father (= head, age: 45), mother (= wife, age: 42), and son (= child, age: 15) is $\mathbf{a}_i = (45, 999, 999, 42, 15, 999, 999, \cdots)$. Here, 999 is a dummy number. This dummy number is effective in increasing the distance for the vector of household when a member type does not exist.

In this study, the micro-data is extended to have the housing type and the location, $h_{ms} = \{\mathbf{c}_{ms}, \mathbf{x}_{ms}\}$ is extended to equation (3):

$$(hhl)_{ms} = \{\mathbf{c}_{ms}, \mathbf{x}_{ms}, j, z\} \quad (3)$$

where j is the type of housing and z is the zone in which the household is located.

Basic Estimation Principles

The basic estimation principles are as follows.

- Overall approach of the study is the most likelihood estimation based on the probability obtained from a sampled set with given control totals.
- Estimation of the population micro-data set for the base year is carried out based on the probabilities obtained from a sampled micro-dataset.
- Each estimation or synthesis and adjustment is probabilistic and uses the Monte Carlo approach.
- The relationships between attributes (e.g., household member composition and age of members) are considered. These relationships are derived from the sample set.
- The synthetic population should be adjusted to fit the marginal conditions (e.g., number of persons by five-year age bands).

In order to avoid misunderstanding, it should be explained the purpose of estimating choice probability of housing type and location based on discrete choice models with the sample set. They are used only for determining housing type and location of each household in the base year. They are not used for behavioural forecasting in microsimulation.

Correlation between Continuous Attributes

The innovation of this study is in dealing with the relationships between continuous attributes. In this case, the age of household members $\mathbf{x}$ is considered. This procedure is applicable only to households whose membership composition $\mathbf{c}$ is sufficiently common. However, when the member composition $\mathbf{c}$ is very rare, the age composition $\mathbf{x}$ is determined as that of the sample household.
First, the original attribute variables $\mathbf{x} = (x_1, x_2, \cdots, x_m)$—i.e., age of household members—for sample households composed of *m* members are converted into independent or non-correlated variables $\mathbf{p} = (p_1, p_2, \cdots, p_m)$ by carrying out a principal component analysis:

$$p_i = \sum_k^m v_{ik} x_k \tag{4}$$

or written in the matrix form $$\mathbf{p} = \mathbf{V}\mathbf{x} \tag{5}$$

On the basis of the sample values, the cumulative frequency curve of p_i is drawn for *m* principal component variables, as shown in Figure 2.

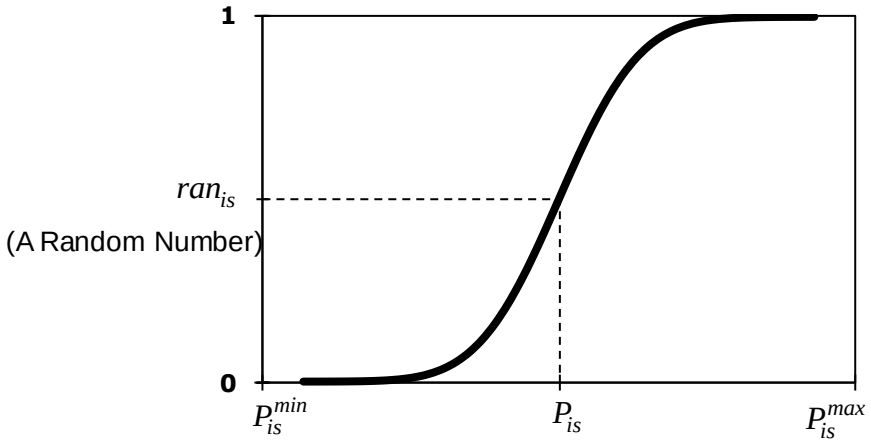


Figure 2 – Correlation Determination using Independent Variables

From equations (4) and (5), the following equations are obtained.

$$\mathbf{x} = \mathbf{W}\mathbf{p} \tag{6}$$

$$x_i = \sum_k^m W_{ik} p_k \tag{7}$$

To generate a synthetic household, a random number ran_{is} is generated for a member i of household s; p_{is} is obtained from Figure 2 for the principal component variable i. x_{is} (or age) is then obtained for member i of household s from equation (6). The procedure is

repeated to generate other synthetic households until the total number of households in the study area is attained.

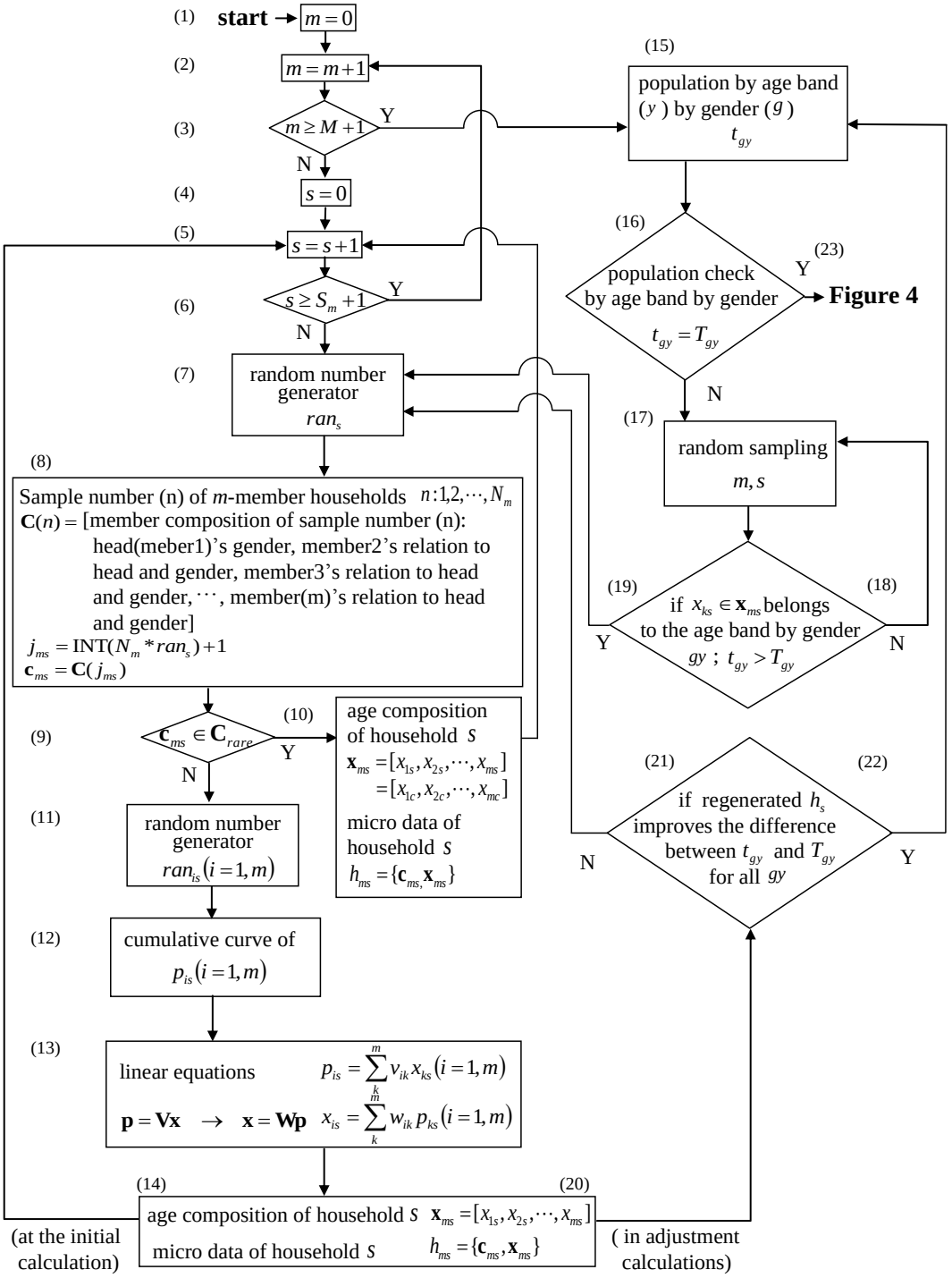
Therefore, by introducing the independent variables as intervening variables, the relationships between attributes are easily dealt with and the system becomes easy to operate.

Flowchart

With the abovementioned concepts and procedures, the entire estimation system may be summarized in the form of a flowchart shown in Figures 3 and 4. Figure 3 describes the part of estimating a household's attributes, while Figure 4, which follows Figure3, explains the part of estimating its housing type and location.

The steps in the flowchart are described as follows:

- (1)–(3): Estimation is carried out as the number of household members increases from 1 to *M*, which is determined from the sample dataset as the maximum number of household members.
- (4)–(6): Since the number of households that has *m* members, S_m, is usually available from the census, it is exogenously used as the control total of the study area. Estimation is carried out for 1 to S_m households.
- (7)–(8): Random number ran_s ($o \le ran_s \le 1$) is generated for household s. The household member composition $\mathbf{c}_{ms}$ of s is determined as that of sample number j_{ms} ($j_{ms} = INT(N_m * rsn_s) + 1$).
- (9)–(10): If the member composition $\mathbf{c}_{ms}$ is rare in the sample dataset, the age composition $\mathbf{x}_{ms}$ is determined to be the same as that of the sample number j_{ms}.
- (9), (11)–(14): If the member composition $\mathbf{c}_{ms}$ is common, the age composition $\mathbf{x}_{ms}$ is determined by the method described in the previous section.
- (15)–(16): The initial set of synthetic households does not satisfy the marginal conditions for the number of persons by five-year age bands.
- (17)–(18): A Monte Carlo approach is used to randomly select a household (*m*, *s*). If every age in $\mathbf{x}_{ms}$ belongs to a pair of gender (g) and age band (y) that satisfies the control total, the current household number (*m*, *s*) is replaced by a new random sampling.
- (19), (7)–(13), (20)–(22): If any of the ages in $\mathbf{x}_{ms}$ belongs to a pair of gender (g) and age band (y) where the current household number is larger than the observed number, re-estimation is carried out. $\mathbf{x}_{ms}$ is replaced by the new member composition $\mathbf{c}_{ms}$ and the gender (g) and age (y) of the members, which reduces the differences between the estimated and observed total numbers.



$m=[1,\cdots,M]$: number of household members

$s=[1,\cdots,S_m]$: household which has m members

S_m : number of households which has m members (Exogenous as the control total of the study area)

ran : random number $0 \leq ran \leq 1$

C_{rare} : rare household composition

$y=[1,\cdots,Y]$: age class

T_{gy} : total number of observed individuals belonging to g y (Exogenous as the control total of the study area)

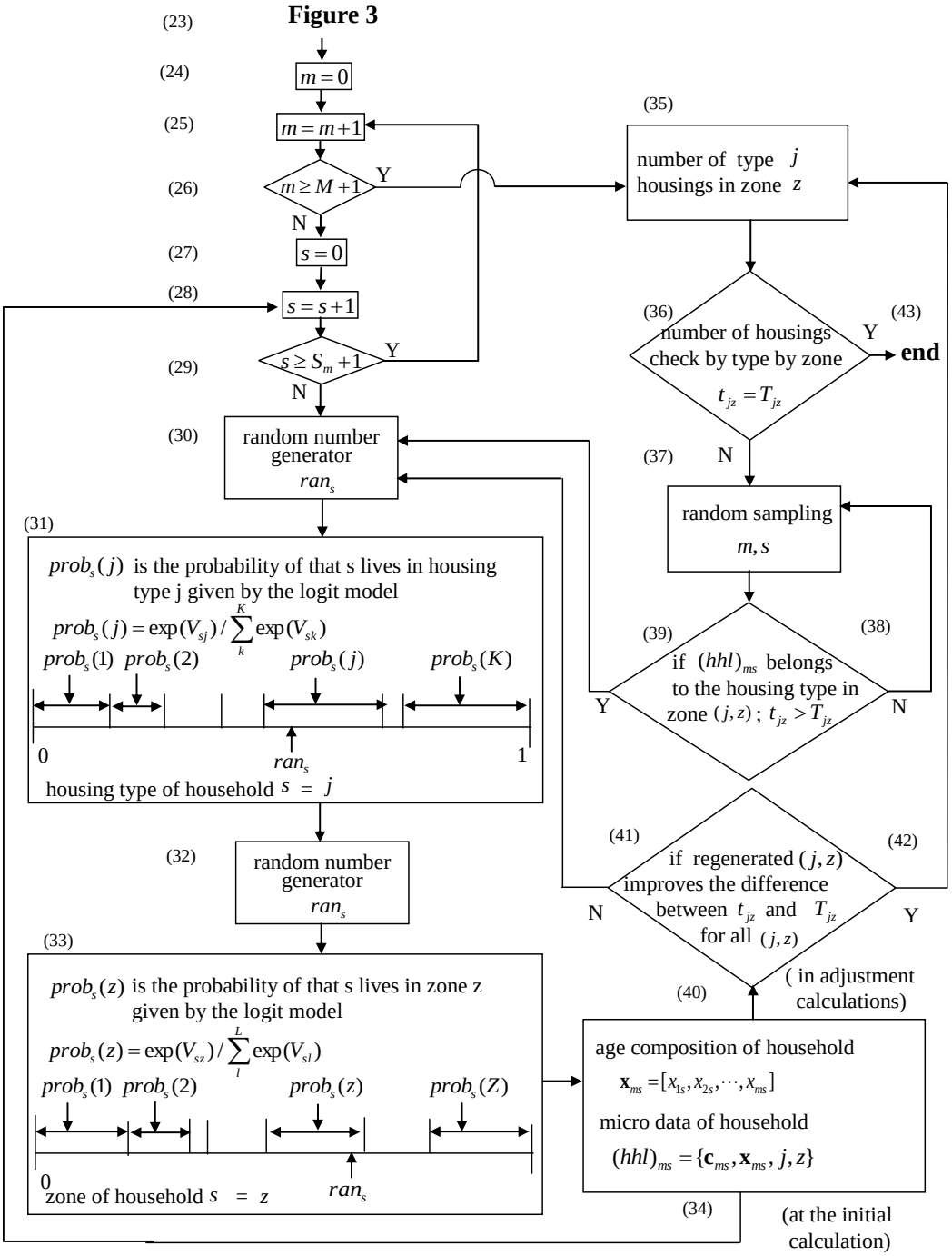
$\sum_m (m * S_m) = \sum_{gy} T_{gy}$

t_{gy} : total number of estimated individuals belonging to g y

Figure 3 –Comprehensive Population Synthesis: Part 1

12th WCTR, July 11-15, 2010 – Lisbon, Portugal

Figure 3



$m=[1,\cdots,M]$: number of household members

$s=[1,\cdots,S_m]$: household which has m members

S_m : number of households which has m members (Exogenous as the control total of the study area)

ran : random number $0 \leq ran \leq 1$

$prob_s(q)$: is the probability of that s chooses q given by the logit model

V_{sq} : utility of q for s in the logit model

$j=[1,\cdots,J]$: type of housing

$z=[1,\cdots,Z]$: zone in the area

T_{jz} : total number of type j housings in zone z

t_{gy} : total number of estimated type j housings in zone z

Figure 4 – Comprehensive Population Synthesis: Part 2

- (23): The adjustment iteration should be continued until all marginal conditions of household are satisfied.
- (24)–(29): This series of steps is the same as that of (1)–(6).
- (30)–(31): Random number ran_s ($o \le ran_s \le 1$) is generated for household s . The housing type j of s is determined by Logit model of housing choice which is obtained by the sample dataset.
- (32)–(34): Random number ran_s ($o \le ran_s \le 1$) is again generated for household s . The zone z in which s is located is determined by Logit model of location choice which is obtained also by the sample dataset. Then, initial housing type and zone is determined for all households which are estimated in Figure3.
- (35)–(36): The initial set of synthetic households does not satisfy the marginal conditions for the number of housing by type and by zone.
- (37)–(38): A Monte Carlo approach is used to randomly select a household (*m*, *s*). If (j, z) belongs to a pair of housing type (j) and zone (z) which satisfies the control total, a new random sampling (*m*, *s*) is generated.
- (39), (30)–(33), (40)–(42): If any of the ages in (j, z) belongs to a pair of housing type (j) and zone (z) where the current household number is larger than the observed control total, re-estimation is carried out. (j, z) is replaced by a new pair of housing type and zone, which reduces the differences between the estimated and observed total numbers.
- (43): The adjustment iteration should be continued until all marginal conditions by housing type and zone are satisfied.

EVALUATION OF GOODNESS-OF-FIT

Since it is difficult to represent the system in a single mathematical form under different conditions in terms of the available data, the proposed estimation system becomes ad hoc. Therefore, to develop the system in a more rational and objective manner, an indicator is introduced to evaluate the goodness-of-fit between two micro-datasets. It is defined as the minimum of the normalized sum of three weighted distances, that in each member's age, each member's gender and member's relationship with the head of household; in housing type; and in residential zone. The detail information on the goodness-of-fit indicator is described in a parallel paper (Otani et al., 2010), which employs the genetic algorithm (GA) with the symbiotic evolution method.

SYSTEM APPLICATION

Data

The data used in the case study is obtained from the current person-trip-survey for the Sapporo metropolitan area; full-scale information is available for 19,394 households. 10,000 households are randomly sampled from the original dataset to compose a virtual set of household population and to be used as a true population set *A* of the household micro-data being estimated. These households are living in five different types of housing, shown in Table 1. Among their household members, 11,367 are male and 12,748 are female. The study area is divided into 8 zones, shown in Figure 5, having certain number of households, summarized in Table 2.

Table 1 – Housing Types of Virtual Households in Set A

Housing Type	Number of Households
Own-detached	5,233
Rent-detached	329
Own-apartment	1,395
Rent-apartment	2,889
Other	154
Total	10,000

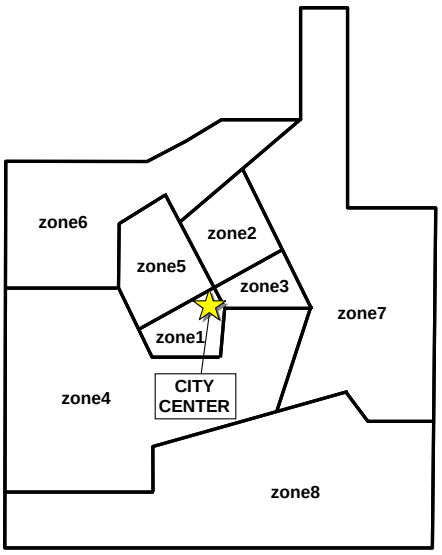


Figure 5 – Zone Map

Table 2 – Location of Virtual Households in Set A

Zone	Households
zone 1	1,095
zone 2	1,972
zone 3	1,254
zone 4	1,940
zone 5	1,551
zone 6	7,90
zone 7	5,70
zone 8	8,28
Total	10,000

Next, 1,000 households are randomly sampled from the virtual population set *A* to compose a sample dataset *B*. The problem is to estimate the true population set *A* from the sample dataset *B* with the control totals:

- Number of households with *m* members (*m* = 1, 2,…, 7)
- Number of individuals by gender and five-year age band
- Number of households by type of housing by zone

On the basis of the sample dataset *B*, the following 20 "general household member" types are selected to compose $\mathbf{a} = (a_1, a_2, \cdots, a_Z)$, which are represented by member's relationship with the household head and member's gender: head (male), wife, one child (male), two children (male), three children (male), grandchild (male), brother, father, other (male), two other members (male), head (female), one child (female), two children (female), three children (female), grandchild (female), sister, mother, child's wife, other (female), and two other members (female).

In addition, the following 10 membership composition types are selected for establishing the "Correlation between Continuous Attributes" by following the steps (11)–(18) in Figure 3. This is because they have 10 or more degrees of freedom in principal component analysis for the sample dataset *B*: Single (male), single (female), couple, head (female) + child (female), couple + child (male), couple + child (female), couple + mother, couple + two children (male), couple + child (male) + child (female), couple + two children (female). The other member composition types are treated as "rare" and are selected by following the steps (10) and (11) shown in Figure 3.

Allocation of a household into certain type of housing and location (zone) is conducted by employing the conventional multinomial Logit models, i.e., boxes (31) and (33) in Figure 4. The two models are then estimated; resulting in the parameters shown in Table 3 and Table 4 respectively.

Table 3 – Parameters for the Housing Type Choice Model

Explanatory variables	Parameters (t-values)			
	Own-detached	Rent-detached	Own-apartment	Rent-apartment
Number of household members	0.667(3.32)	0.645(2.72)	0.405(1.95)	0.503(2.47)
Age of head	0.166(4.28)	-0.116(-2.12)	0.106(2.64	0.128(3.24)
Sample size	1,000			
Log-likelihood	-1,118.6			
Likelihood ratio	0.304			

Table 4 – Parameters for the Location (Zone) Choice Model

Explanatory variables	Parameters (t-values)		
	zone 1	zone 2 - 5	zone 6 - 8
Distance from the city center	-0.010(-0.76)		
Number of household members	-	0.442(6.21)	0.672(7.23)
Age of head	-	-	-0.055(-2.45)
Dummy of Own-detached	-	1.553(6.21)	0.902(2.39)
Dummy of Own-apartment	-	-	-2.045(-4.77)
Dummy of Rent-apartment	-	1.079(5.14)	-0.639(-1.78)
Sample size	1,000		
Log-likelihood	-757.8		
Likelihood ratio	0.307		

Results

A virtual dataset A is generated as the observed household micro-dataset (Z = 20, N = 10,000). To evaluate the estimation produced by our proposed system, a naive system is built to produce another set of estimation for comparison. It uses a simple enlargement method for the sampled micro-data, which is similar to the one shown in Figure 3 but without steps (12)–(18). The datasets estimated by the proposed and naive systems are denoted by E_1 and E_2, respectively. Notice that the proposed system has more degrees of freedom to represent population dataset because it is able to generate age compositions flexibly. Moreover, E_1 and E_2 are different due to the random numbers generated in Figures 3 and 4. In this case study, 8 different random numbers are generated. As a result, there are eight different estimated datasets produced by the two system; denoted by E_{11},…, $E_{18,}$ and E_{21},…, E_{28}, respectively.

Table 5 shows the estimation produced by the proposed and naive systems for Own-Detached Housing Type. The average and standard deviation are the average of residuals and standard deviation for those 8 different random numbers. By visual inspection, E_1 seems to be slightly better than E_2.

Table 5 – Estimation Produced by the Two Systems for Own-Detached Housing Type

Member Composition	Observed Samples	Residuals from the Observed data			
		Proposed System (E_1)		Naive System (E_2)	
		Average	Standard Deviation	Average	Standard Deviation
Single (male)	300	207.8	24.9	210.1	23.3
Single (female)	591	88.0	17.9	93.1	16.7
Couple	176	-36.4	5.7	-39.0	12.3
Head (female) + child (female)	1535	-131.4	24.0	-128.1	22.5
Two other members	120	-23.9	10.0	-22.1	6.7
Couple + child (male)	150	1.4	12.4	2.1	12.9
Couple + child (female)	456	-81.0	15.8	-96.4	14.5
Couple + mother	507	12.3	21.5	34.1	14.6
Three other members	83	25.8	11.3	20.8	7.9
Couple + two children (male)	132	-30.4	5.9	-37.0	13.4
Couple + child (male) + child (female)	193	28.9	10.2	27.4	4.9
Couple + two children (female)	394	11.3	23.3	3.4	23.3
four other members	192	-31.1	9.4	-26.0	12.9
Five members	322	-30.4	8.4	-30.8	8.6
Six members	70	-11.5	2.3	-11.0	4.9
Seven or more members	12	0.8	1.8	-0.6	2.4
Total	5,233				

In terms of spatial location, the distributions of the estimated households over zones in the study area are shown in Figure 6: observed data versus the ones produced by proposed and naive systems. However, it is difficult to judge how the proposed system is superior to the naïve system only by visual inspection. This calls for an efficient evaluation indicator, described earlier.

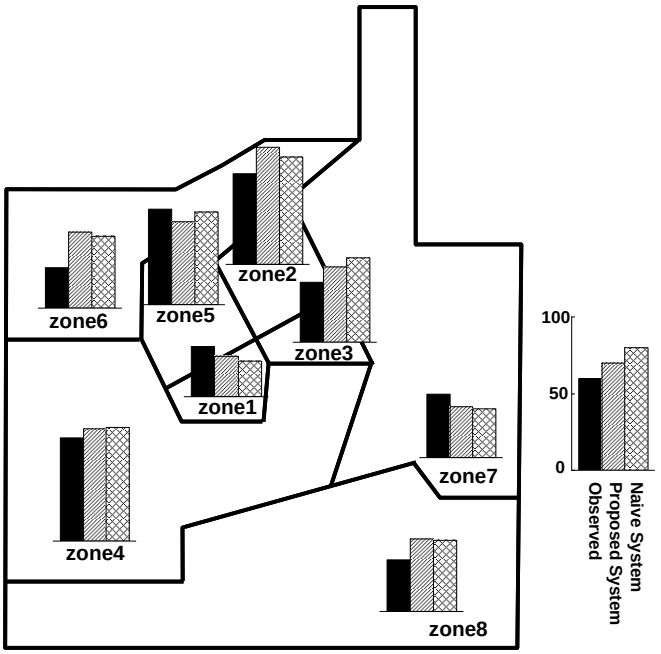


Figure 6 Households by Household Member Composition and Living Zone for the *Couple + child (male) + child (female)* Household

Goodness-of-Fit of the Estimation

The goodness-of-fit indicators of the estimations produced by the two systems are calculated. Due to its probabilistic nature under GA, 10 repetitions are realized with $DiffMax$ = 99,999, w_a = 1, w_h = 0.1, w_r = 0.1 for $E_{11},\ldots, E_{18}$ and $E_{21},\ldots, E_{28}$. The summary statistics of the goodness-of-fit indicators, i.e., the average and standard deviations, are shown in Table 6. To interpret, the lower average value, the better estimation is. Therefore it can be evaluated that the proposed system generally out performs the naïve system. However, this does not imply strict superiority of the proposed system in this application. Since both systems adopted the same agent based approach except the part of age composition, this made the comparison only valid for the difference caused by limited part. In addition, due to small number of samples (1,000), a limited number of extreme samples will greatly affect the results. In the other words, the goodness-of-fit evaluation represents relative evaluation with respect to the distance of the two datasets, i.e., the indicator would be zero if the two datasets are completely identical. On the other hand, the absolute evaluation of the goodness of fit is on-going in our future research.

Table 6 – Goodness of Fit of the Estimations

Proposed System			Native System		
	Average	Standard deviation		Average	Standard deviation
E_{11}	0.17115	0.00065	E_{21}	0.18456	0.00099
E_{12}	0.17721	0.00198	E_{22}	0.18506	0.00203
E_{13}	0.17738	0.00067	E_{23}	0.18522	0.00149
E_{14}	0.18378	0.00329	E_{24}	0.19013	0.00169
E_{15}	0.19247	0.00202	E_{25}	0.19262	0.00277
E_{16}	0.19405	0.00206	E_{26}	0.20019	0.00347
E_{17}	0.19985	0.00102	E_{27}	0.20103	0.00406
E_{18}	0.20570	0.00136	E_{28}	0.20117	0.00405
Average	0.18770	0.00163	Average	0.19250	0.00257

CONCLUDING REMARKS

A system is proposed in this study to rationally estimate a micro-dataset of the base year for land-use microsimulation. The system, which uses Monte Carlo simulations, deals with both continuous and discrete attribute variables by agent (i.e., household in this study) including housing type and location. It uses sample data, which contains full micro-data information, to establish the relationships between attributes and uses existing statistical data as the control total in each zone. Continuing our previous development by incorporating location and housing type choice models, this paper presents a comprehensive micro-data synthesizer that produces details attributes of household such as its member and age composition, housing type, as well as its spatial location. In addition, to develop the system in a rational and objective manner, an indicator is introduced to evaluate the goodness-of-fit between two micro-datasets. By introducing the proposed indicator, this study has developed not only an estimation system but also an approach to system development itself. Detail description on the goodness-of-fit indicator can be found in a parallel paper (Otani et al., 2010). The application of the system to person-trip-survey data for the Sapporo metropolitan area has proved the usefulness of the system and the approach. Finally, it is worth mentioning that many attributes of micro-data items are not presently included yet, e.g., income, job, workplace, study place, car ownership, etc. Production of these attributes in the complete micro-dataset is still on-going and will be reported in a subsequent paper.

ACKNOWLEDGEMENTS

This study was supported by a Grant-in-Aid for Scientific Research (20360232) from the Japan Society for the Promotion of Science.

REFERENCES

Auld, J., and Mohammadian, A. (2010) "Efficient Methodology for Generating Synthetic Populations with Multiple Control Levels." *The 89th TRB Meeting Compendium of Papers DVD*, Washington, D.C.

Auld, J., Rashidi, T. H., and Mohammadian, A. (2010) "Evaluating Transportation Impacts of Forecast Demographic Scenarios Using Population Synthesis and Data Transferability." *The 89th TRB Meeting Compendium of Papers DVD*, Washington, D.C.

Beckman, R. J., Baggerly, K. A., and McKay, M. D. (1996). " Creating Synthetic Baseline Populations." *Transportation Research A*, 30(6), 415-435.

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." *Annals of Mathematical Statistics*, 11, 427-444.

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

Lee, B. H. Y., and Waddell, P. (2010)"Residential Mobility and Location Choice: Nested-Logit Model with Sampling of Alternatives." *The 89th TRB Meeting Compendium of Papers DVD*, Washington, D.C.

Miyamoto, K., Ando, J., and Shimizu, E. (1986). "A Housing Demand Model based on Disaggregate Behavioral Analysis." *Journal of Japan Society of Civil Engineers*, 365(4(4)), 79-88.

Miyamoto, K., Sugiki, N., Otani, N., and Vichiensan, V. (2010)"Agent-Based Estimation Method of Household Microdata for Base Year in Land Use Microsimulation." *The 89th TRB Meeting Compendium of Papers DVD*, Washington, D.C.

Miyamoto, K., Sugiki, N., and Vichiensan, V. (2009). "An Estimation Method of Household Micro-Data for the Base Year in Land-Use Micro Simulation." Proceedings of CUPUM'09 CD-Rom, Hong Kong.

Moeckel, R., Spiekermann, K., and Wegener, M. (2003). "Creating a Synthetic Population." Proceedings of CUPUM'03 CD-Rom, Sendai.

Moriarty, D. E., and Miikkulainen, R. "Efficient Learning from Delayed Rewards through Symbiotic Evolution." *Proceedings 12th International Conference on Machine Learning*, 396-404.

Otani, N., Miyamoto, K., and Sugiki, N. (2009a). "Goodness-of-Fit Evaluation Method between Micro-Data Sets which contains Discrete Attributes in Land-Use Micro-Simulation." Proceedings of Infrastructure Planning, Japan Society of Civil Engineers (in Japanese).

Otani, N., Miyamoto, K., and Sugiki, N. (2009b). "Goodness-of-Fit Evaluation Method between Observed and Estimated Sets of Micro-Data in Land-Use Micro-Simulation " Proceedings of CUPUM'09 CD-Rom, Hong Kong.

Otani, N., Sugiki, N., and Miyamoto, K. (2009c) "Goodness-of-Fit Evaluation Method between Micro-Data Sets which contains Discrete Attributes in Land-Use Micro-Simulation." *Proceedings of Infrastructure Planning, Japan Society of Civil Engineers*, (in Japanese).

Otani, N., Sugiki, N., and Miyamoto, K. (2010) "Goodness-of-Fit Evaluation Method between Two Sets of Household Micro-Data for Land-Use Microsimulation Model",12th WCTR, July 11-15, 2010 – Lisbon, Portugal

Pritchard, D. R., and Miller, E. J. (2009). "Advances in Agent Population Synthesis and Application in an Integrated Land Use / Transportation Model." *The 88th TRB Meeting Compendium of Papers DVD*, Washington, D.C.

Ryan, J., Maoh, H., and Kanaroglou, P. S. (2010)"Population Synthesis for Microsimulating Urban Residential Mobility." *The 89th TRB Meeting Compendium of Papers DVD*, Washington, D.C.