Agent-Based Household Micro-Datasets: An Estimation Method composed of Generalized Attributes with Probabilistic Distributions from Sample Data and Available Control Totals by Attribute

Nao SUGIKI\textsuperscript{a}, Varameth VICHIENSAN\textsuperscript{b}, Noriko OTANI\textsuperscript{c}, Kazuaki MIYAMOTO\textsuperscript{d}

\textsuperscript{a} Research & Planning Department, Docon Co., Ltd., 4-1, 5-chome, 1-jo, Asubetsu-cho, Asubetsu-ku, Sapporo, Hokkaido, 004-8585 Japan; E-mail: ns1491@docon.jp
\textsuperscript{b} Faculty of Engineering, Kasetsart University, 50 Phahonyothin Rd, Ladyao, Jatujak, Bangkok, 10900 Thailand; E-mail: fengvmv@ku.ac.th
\textsuperscript{c} Faculty of Environmental and Information Studies, Tokyo City University, 3-3-1 Ushikubonishi, Tsuzuki-ku, Yokohama, 224-8551 Japan; E-mail: otani@tcu.ac.jp
\textsuperscript{d} Faculty of Environmental and Information Studies, Tokyo City University, 3-3-1 Ushikubonishi, Tsuzuki-ku, Yokohama, 224-8551 Japan; E-mail: miyamoto@tcu.ac.jp

Abstract: The purpose of this study is to build a system to rationally estimate the agent-based household micro-dataset of the base year for land-use microsimulation. Attributes of a household can be classified into a general set of categories. This system, wherein a Monte Carlo simulation is used, deals with those attributes, either continuous or discrete, in a generalized scheme. It uses sample data that contain full information on the micro-data to establish the correlation between the attributes and the available statistical data or census data. To reproduce the correlation between continuous attribute variables, independent variables that can be obtained based on the sample data are introduced and employed as intervening variables. Attributes of a whole household are probabilistically determined based on Logit and other models obtained from sample data. Finally, a case study of the system application to a person-trip-survey dataset of the Sapporo metropolitan area is presented.

Key Words: Land-use microsimulation, Agent-based household micro-dataset, Estimation system, Household attributes.

1. INTRODUCTION

Microsimulation is becoming a popular approach in land-use modeling 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 individual households required for the base year in residential microsimulation is considered. An individual household is characterized by many attributes such as member composition and the age of the members, housing type and the location, the number of car owned and income. Each 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 like pigeonholes 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. Another approach is to generate a set of individual synthetic households, each of which has its unique set of 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. In our earlier studies, a set of attributes of a member composed of the relationship with the household head and the gender and age only was considered (Miyamoto et al., 2010(a)) was made as a first phase. In the second phase, the other attributes such as household type and the locations were added to the attributes to be estimated (Miyamoto et al., 2010(b)). In this present paper, a general scheme to generate or estimate a set of individual households with comprehensive attributes is presented by classifying the attributes into general categories. Although the system was originally developed for households, the method can be applied to “agents” in general. Since the method is not a single mathematical framework with an objective function, our study has also proposed an indicator to evaluate the goodness-of-fit between two micro-datasets originally presented in Otani et al. (2010) and subsequently extended in Otani et al. (2011) in order to develop the system in a rational and objective manner.

The present 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 that is built to generate a set of agent-based household micro-data is presented. 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.

2. POPULATION SYNTHESIS

2.1 IPF Based Methods

In the past, travel demand microsimulation such as TRANSIM essentially uses the population synthesizer; the other applications include Guo et al. (2007), Ryan et al. (2010), Auld et al. (2010(a)), etc. The Iterative Proportional Fitting (IPF) procedure is popularly used to generate synthetic populations. It was originally proposed by Deming et al. (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 et al. (2007), Auld et al. (2010(b)), etc. In land use microsimulation, household and population syntheses are also very essential because highly detailed analysis of household attributes are concerned, e.g., Lee et al., (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 et al. (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 et al. (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.

2.2 Agent-Based Approaches
In contrast to the IPF based method, Moeckel et al. (2003) used an “agent”-based approach. Monte Carlo sampling by agent was employed so that as many features as required by a microsimulation model can be selected. In this case, the number of features is only limited by the possibility of determining reasonable relationships between the selected attributes. With the basic idea presented in Moeckel et al. (2003), this study has continuously developed a system to estimate a set of household micro-data based on probabilistic distributions, which can be obtained from sample data and control totals by attribute obtained from the census data (Miyamoto et al., 2010(a), 2010(b)). While our previous papers presented the system with a part of household attributes, this paper presents a system that covers a general set of attributes.

2.3 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 method would produce many zero-value cells if the problem deals with many data categories. Practically a small value close to zero is assigned for these cells, e.g., 0.01 or 0.1. However such operation influences probabilities and is based on loose theoretical foundation (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. Moreover, it is, in reality, impossible for an IPF method to deal with more than three or four kinds of attributes in each of which a limited number of categories are set up. This study aims to synthesize general attributes of individual households with no ad hoc categories.

3. GENERAL FRAMEWORK

3.1 Household Attributes
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. This study is a series of extensions of the agent-based synthesis method proposed by Moeckel et al. (2003).
Based on our previous work (Miyamoto et al., 2010(a), 2010(b)), attributes of a household are represented in general as follows:

- Households are the target agents.
- A household is characterized by its structure or composition: members and their relationship with the household head.
- Each member is characterized by continuous and discrete attributes.
- The numbers of households or individuals by the structure or attribute have control totals in the study area or by given unit of analysis.
- A household has its attributes as a whole, some continuous and some discrete. Moreover, some attributes are summed up to the control totals of the study area or other spatial units of analysis. In general, the attributes may be classified into four categories which are the combinations of the above-mentioned two dimensions: individual member or household as a whole; and continuous or discrete, as illustrated in Figure 1.

![Figure 1. A general framework of a household attributes](image)

### 3.2 Estimation Concept

In the estimation of the initial household micro-dataset for the base year of the simulation, the system of estimation is built based on the following basic concepts.

- A microsimulation land-use model forecasts future state by Monte Carlo simulation with a given set of probabilities in the change of various agents. This system estimates the initial micro-dataset, in principle, by the same probabilistic procedure with given constraints of the total agents (households) or sub-agents (members of household), by attribute, if necessary.
- Since the attributes are considered to be deterministic in a hierarchical structure, the estimation procedure is also structured to follow the hierarchy, where the causal relationships between attributes are taken into consideration.
- The system consists of several iterative adjustment processes such that the data production satisfies the available control totals by attribute.
- It uses sample data that contain full information on the micro-data to establish the correlation between the attributes and the existing statistical or census data as the control total.
- To reproduce the correlation between continuous attribute variables, independent variables that can be obtained by the principal component analysis based on the original variables of sample data are introduced and employed as intervening variables.
Attributes of a whole household such as housing type are probabilistically determined based on Logit models or by means of multiple regression equations which can be calibrated with the sample data.

4. SYSTEM DEVELOPMENT

4.1 Micro-Data

As per the presuppositions, the micro-data of household structure and the attributes of the members are 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,

$$ h_{ms} = \{c_{ms}, x_{ms}\} $$

where $c_{ms}$: member composition of the household $s$ having $m$ members

$x_{ms}$: 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):

$$ A = \{a_i = (a_{i1}, a_{i2}, \cdots, a_{ik}) | 1 \leq i \leq N \} $$

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 $a_i = (45, 999, 999, 42, 15, 999, 999, \cdots)$. 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 ($j$), the location ($z$), the number of car owned ($nc$) and income ($inc$) so that $h_{ms} = \{c_{ms}, x_{ms}\}$ is extended to equation (3):

$$ h_{ms} = \{c_{ms}, x_{ms}, j, z, nc, inc\} $$

where $j$ is the type of housing, $z$ is the zone in which the household is located, $nc$ is the number of car owned, and $inc$ is the income.

4.2 Representative Household Attributes

Within the general framework described before, a representative set of household micro-data is presented as illustrated in the parentheses in FIGURE 1 and is based on the following
assumptions:
- The study area is divided into zones, in which the limited kinds of census data are available.
- The target micro-data of the household composition is structured by the relationships of the members with the household head followed by their genders and ages.
- The household-specific attributes are housing type, zone location, the number of car owned and income.
- The number of households by the number of members and the number of individual persons by five-year age bands are obtained from the census data.
- The number of housings by type is obtained from the census data.
- The number of households in each zone is obtained from the census data.
- A certain number of micro-data samples are available.

In order to describe the system concretely, this set of representative attributes of a household is used in developing the system. In case that another attribute is necessary for an application, it can be estimated by the same procedure as used for the attributes under the same category.

4.3 Estimation Procedure
The estimation procedures are structured into the following order: housing composition, members’ attributes, housing and the location, and income & car ownership, respectively; illustrated in Figure 2. The structure is assumed to reflect the actual causal relationship between those attributes.

To avoid misunderstanding, it is worth mentioning that the estimation of choice probability of housing type and location based on discrete choice models with the sample set is used for determining the housing type and location of each household in the base year only, but not used for behavioral forecasting in the microsimulation model.

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

$$p_i = \sum_k v_{ik} x_k$$  \hspace{1cm} (4)

or written in the matrix form,

$$p = Vx$$  \hspace{1cm} (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 3. From equations (4) and (5), the following equations are obtained.

$$x = Wp$$  \hspace{1cm} (6)

$$x_i = \sum_k W_{ik} p_k$$  \hspace{1cm} (7)

To generate a synthetic household, a random number $\text{ran}_{is}$ is generated for a member $i$ of household $s$; $p_{is}$ is obtained from Figure 3 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.

![Figure 3. Correlation determination using independent variables](image)

### 4.5 Computation Flowchart

With the abovementioned concepts and procedures, the entire estimation system may be summarized in the form of a flowchart, which is presented in Figures 4a to 4c. Figure 4a illustrates the procedure to estimate a household’s attributes, that for the housing type and location in Figure 4b, and car ownership and income in Figure 4c. In words, 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 $ran_s$ (0 ≤ $ran_s$ ≤ 1) is generated for household $s$. The household member composition $c_{ms}$ of $s$ is determined as that of sample number $j_{ms}$.

- (9)-(10): If the member composition $c_{ms}$ is rare in the sample dataset, the age composition $x_{ms}$ is determined to be the same as that of the sample number $j_{ms}$.

- (9), (11)-(14): If the member composition $c_{ms}$ is common, the age composition $x_{ms}$ 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 $x_{ms}$ belongs to a pair of gender $(g)$ and age band $(y)$ which satisfies the control total then the current household number $(m, s)$ is replaced by a new random sampling.

- (19), (7)-(13), (20)-(22): If any of the ages in $x_{ms}$ belongs to a pair of gender $(g)$ and age band $(y)$ where the current household number is larger than the observed number then re-estimation is carried out. $x_{ms}$ is replaced by the new member composition $c_{ms}$ and the gender $(g)$ and age $(y)$ of the members, which reduces the differences between the estimated and observed total numbers.

- (23): The adjustment iteration is continued until all marginal conditions are satisfied.

- (24)-(29): This series of steps is the same as that of (1)-(6).

- (30)-(31): Random number $ran_s$ (0 ≤ $ran_s$ ≤ 1) is generated for household $s$. The housing type $j$ of $s$ is determined by Logit model of housing choice calibrated with the sample data.

- (32)-(34): Random number $ran_s$ (0 ≤ $ran_s$ ≤ 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.

- (35)-(36): This tests if 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.

Figure 4a. Comprehensive population synthesis: household attributes
Figure 4a

$\text{prob}_j(j) = \frac{\exp(V_{js})}{\sum_k \exp(V_{sk})}$

$\text{prob}_j(1) \cdot \text{prob}_j(2) \cdot \ldots \cdot \text{prob}_j(K)$

$0 \leq \text{ran}_s \leq 1$

Figure 4b. Comprehensive population synthesis: housing type & location

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

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

$S_m$ : number of households which has $m$ members

$\text{prob}_j(q)$ : is the probability of that $s$ chooses $q$ given by the logit model

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

Figure 4c

$\text{prob}_z(z) = \frac{\exp(V_{sz})}{\sum_j \exp(V_{sj})}$

$\text{prob}_z(1) \cdot \text{prob}_z(2) \cdot \ldots \cdot \text{prob}_z(Z)$

$0 \leq \text{ran}_s \leq 1$
(43): The adjustment iteration should be continued until all marginal conditions by housing type and zone are satisfied.

(44)-(49): This series of steps is the same as that of (1)–(6).

(50)-(51): Random number \( \text{ran}_s \) (\(0 \leq \text{ran}_s \leq 1\)) is generated for household \(s\). The number of cars \(n_s\) of household \(s\) is determined by Logit model of car ownership choice, calibrated with the sample data.

(52)-(55): Random numbers \( \text{ran}_{1s} \) and \( \text{ran}_{2s} \) (\(0 \leq \text{ran}_{1s}, \text{ran}_{2s} \leq 1\)) is generated for household \(s\). The income \(\text{inc}\) of household \(s\) is determined by multiple regression equation, calibrated with the sample data, and its standard error considering normal random number generated by Box-Muller transform with two uniform random numbers. The number of cars and the income of household are then determined for all households.
4.6 Goodness-of-Fit

The basic concept of the proposed estimation system has a sound basis and consistent with the general framework of the microsimulation model. However, it cannot be evaluated quantitatively with the conventional evaluation methods, for it is difficult to represent the system in a single mathematical form under different conditions in terms of the available data. 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. The evaluation criterion in this case is defined as the minimum of the normalized sum of weighted distances of attributes of all households in the study area. The calculation cannot be carried out only with a conventional algorithm for micro-data of a typical size, because the number of calculations increases in proportion to the factorial (N!) of the number (N) of agents. Therefore, genetic algorithm, especially one employing symbiotic evolution, is developed to solve the problem. The detailed information on the goodness-of-fit indicator is described in Otani et al. (2010).

5. SYSTEM APPLICATION

5.1 Data

The data is obtained from the current person-trip-survey for the Sapporo metropolitan area; full-scale information is available for 19,394 households. In the case study, 10,000 households are randomly sampled from the original dataset to compose a virtual set of household population and used as a true population set A of the household micro-data being estimated. A virtual dataset A is generated as the observed household micro-dataset. Among the virtual population set A, 1,000 households are randomly sampled to compose a sample dataset B. The problem is to estimate the true population set A from the sample dataset B with the following given 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
- Number of car owned and income

On the basis of the sample dataset B, the following 20 “general household member” types are selected to compose , 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 compositions are selected for establishing the “Correlation between Continuous Attributes” by following the steps (11)-(18) in Figure 4a. This is because they have 10 or more degrees of freedom in the 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 4a. 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 4b.
5.2 Calibration
The housing type choice, location choice, and number of cars owned by a household are determined by multinomial Logit models, i.e., box (31) and (33) in Figure 4b, and box (51) in Figure 4c. The parameters are estimated from the sample data of the person-trip-survey in Sapporo. However, the person-trip-survey in Sapporo didn’t include the household income item since it would discourage the people to respond to the survey. Therefore, it was not possible to conduct the application to a real set of household data at this time.

Besides the estimation system proposed in this study, a naive estimation system has been created for the purpose of comparison in the stage of household composition. The naive estimation system uses a simple enlargement method for the sampled micro-data, which can be illustrated in similar to Figure 4a but without steps (12)-(18). In general, the proposed system has more degrees of freedom to represent the population dataset because it can generate the age compositions flexibly. The estimated datasets from the proposed and naive systems are denoted by $E_1$ and $E_2$, respectively.

5.3 Synthesized Data

Household Composition
Table 1a presents the summary of the synthesized data produced by the proposed and naive systems. Judging from the residuals from the observed data, $E_1$ appears to be slightly better than $E_2$ in the absolute residuals averaged by member compositions. However, with only this kind of comparison, it is difficult to determine which dataset is more similar to the observed one, i.e., which system is superior to the other. When plural estimation systems are available, an evaluation indicator may be required to select the one with the best performance. For the system development process, a goodness-of-fit indicator is also indispensable to assess the effect of modifications on the existing system. In the case study, the goodness-of-fits for $E_1$ and $E_2$ are calculated 10 times with $\text{DiffMax} = 99,999$. The average and standard deviations of the goodness-of-fits are shown in Table 1b. The goodness-of-fit of the proposed system is considerably smaller than that of the naive system. This result implies that the proposed system provides a more valid estimation of the micro-data.

Housing Type and Location
Table 2 summarized the synthesized data produced by the proposed and naive systems for the Own-Detached Housing Type. The average and standard deviation are the average of residuals and standard deviation due to different set of random numbers. Considering the average residual, $E_1$ seems to be slightly better than $E_2$. In terms of spatial location, the location distributions of the synthesized households with respect to the real data in the study area are shown in Figure 5. However, it is difficult to judge how the proposed system is superior to the naïve system only by visual inspection. This again demands an efficient evaluation indicator, as described earlier.

Car Ownership
The synthesis of car ownership by household with respect to the observed data is summarized in Table 3. Although we cannot say that the system represents the observed data very well, even such a simple model works to some extent at this stage of study. But with a more sophisticated modeling method, it is expected to represent the real data more satisfactorily.
Table 1a. Data synthesized by proposed and naive systems: Household member composition

<table>
<thead>
<tr>
<th>Member Composition</th>
<th>Observed</th>
<th>Proposed System ($E_1$)</th>
<th>Naive System ($E_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Estimated</td>
<td>Residual</td>
</tr>
<tr>
<td>Single (male)</td>
<td>1133</td>
<td>1237</td>
<td>104</td>
</tr>
<tr>
<td>Single (female)</td>
<td>1602</td>
<td>1498</td>
<td>-104</td>
</tr>
<tr>
<td>Couple</td>
<td>2560</td>
<td>2669</td>
<td>109</td>
</tr>
<tr>
<td>Head (female) + child (female)</td>
<td>243</td>
<td>201</td>
<td>-42</td>
</tr>
<tr>
<td>Two other members</td>
<td>354</td>
<td>287</td>
<td>-67</td>
</tr>
<tr>
<td>Couple + child (male)</td>
<td>795</td>
<td>703</td>
<td>-92</td>
</tr>
<tr>
<td>Couple + child (female)</td>
<td>832</td>
<td>911</td>
<td>79</td>
</tr>
<tr>
<td>Couple + mother</td>
<td>104</td>
<td>167</td>
<td>63</td>
</tr>
<tr>
<td>Three other members</td>
<td>315</td>
<td>265</td>
<td>-50</td>
</tr>
<tr>
<td>Couple + two children (male)</td>
<td>332</td>
<td>373</td>
<td>41</td>
</tr>
<tr>
<td>Couple + child (male) + child (female)</td>
<td>678</td>
<td>677</td>
<td>-1</td>
</tr>
<tr>
<td>Couple + two children (female)</td>
<td>309</td>
<td>280</td>
<td>-29</td>
</tr>
<tr>
<td>Four other members</td>
<td>182</td>
<td>171</td>
<td>-11</td>
</tr>
<tr>
<td>Five members</td>
<td>458</td>
<td>458</td>
<td>0</td>
</tr>
<tr>
<td>Six members</td>
<td>87</td>
<td>87</td>
<td>0</td>
</tr>
<tr>
<td>Seven or more members</td>
<td>16</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>10,000</td>
<td>10,000</td>
<td>0</td>
</tr>
<tr>
<td>Average residual</td>
<td></td>
<td>49.5</td>
<td>56.3</td>
</tr>
</tbody>
</table>

Table 1b. Goodness-of-fit: Household member composition

<table>
<thead>
<tr>
<th>Estimation System</th>
<th>Average</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed ($E_1$)</td>
<td>1 110 320 021</td>
<td>1 279 140</td>
</tr>
<tr>
<td>Naive ($E_2$)</td>
<td>1 179 062 599</td>
<td>8 472 622</td>
</tr>
</tbody>
</table>

Table 2. Data synthesized by proposed and naive systems: Own-detached housing type

<table>
<thead>
<tr>
<th>Member Composition</th>
<th>Observed Samples</th>
<th>Residuals from the Observed data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single (male)</td>
<td>300</td>
<td>207.8</td>
</tr>
<tr>
<td>Single (female)</td>
<td>591</td>
<td>88.0</td>
</tr>
<tr>
<td>Couple</td>
<td>176</td>
<td>-36.4</td>
</tr>
<tr>
<td>Head (female) + child (female)</td>
<td>1535</td>
<td>-131.4</td>
</tr>
<tr>
<td>Two other members</td>
<td>120</td>
<td>-23.9</td>
</tr>
<tr>
<td>Couple + child (male)</td>
<td>150</td>
<td>1.4</td>
</tr>
<tr>
<td>Couple + child (female)</td>
<td>456</td>
<td>-81.0</td>
</tr>
<tr>
<td>Couple + mother</td>
<td>507</td>
<td>12.3</td>
</tr>
<tr>
<td>Three other members</td>
<td>83</td>
<td>25.8</td>
</tr>
<tr>
<td>Couple + two children (male)</td>
<td>132</td>
<td>-30.4</td>
</tr>
<tr>
<td>Couple + child (male) + child (female)</td>
<td>193</td>
<td>28.9</td>
</tr>
<tr>
<td>Couple + two children (female)</td>
<td>394</td>
<td>11.3</td>
</tr>
<tr>
<td>four other members</td>
<td>192</td>
<td>-31.1</td>
</tr>
<tr>
<td>Five members</td>
<td>322</td>
<td>-30.4</td>
</tr>
<tr>
<td>Six members</td>
<td>70</td>
<td>-11.5</td>
</tr>
<tr>
<td>Seven or more members</td>
<td>12</td>
<td>0.8</td>
</tr>
<tr>
<td>Total</td>
<td>5,233</td>
<td></td>
</tr>
<tr>
<td>Average residual</td>
<td>47.0</td>
<td>48.9</td>
</tr>
</tbody>
</table>
7. 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, was built to deal with general attributes of a household. In the application of the system, attributes are composed of household composition and the members’ age and gender, housing type, the location, and the number of cars owned, which represent general types of attributes. 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 extending the scope from specific to general attributes, this paper presents a comprehensive micro-data synthesizer that produces detailed attributes of household. 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. 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.

ACKNOWLEDGEMENTS

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


