Distribution System for Japanese Synthetic Population Data with Protection Level

Tadahiko Murata
Faculty of Informatics
Kansai University
Takatsuki, Japan
murata@kansai-u.ac.jp

Susumu Date
Cybermedia Center
Osaka University
Ibaraki, Japan
date@ais.cmc.osaka-u.ac.jp

Yusuke Goto
Faculty of Software & Information Science
Iwate Prefectual University
Takizawa, Japan
y-goto@iwate-pu.ac.jp

Toshihiro Hanawa
Information Technology Center
The University of Tokyo
Kashiwa, Japan
hanawa@cc.u-tokyo.ac.jp

Takuya Harada
College of Science & Engineering
Aoyama Gakuin University
Sagamihara, Japan
harada@ise.aoyama.ac.jp

Manabu Ichikawa
College of Systems Engineering & Science
Shibaura Institute of Technology
Saitama, Japan
m-ichi@shibaura-it.ac.jp

Hao Lee
Faculty of Informatics
Shizuoka University
Hamamatsu, Japan
lee@inf.shizuoka.ac.jp

Masaharu Munetomo
Information Initiative Center
Hokkaido University
Sapporo, Japan
munetomo@iic.hokudai.ac.jp

Akiyoshi Sugiki
Information Initiative Center
Hokkaido University
Sapporo, Japan
sugiki@iic.hokudai.ac.jp

***Abstract*—In this paper, we introduce a distribution system of synthesized data of Japanese population using Interdisciplinary Large-scale Information Infrastructures in Japan. Synthetic population is synthesized based on the statistics of the census that are conducted by the government and publicly released. Therefore, the synthesized data have no privacy data. However, it is easy to estimate the compositions of households, working status in a certain area from the synthetic population. Therefore, we currently distribute the synthesized data only for public or academic purposes. For academic purposes, it is important to encourage scholars or researchers to use a large-scale data of households. we define protection levels for the attributes in the synthetic populations. According to the protection levels, we distribute the data with proper attributes to those who try to use them. We encourage researchers to use the synthetic populations to be familiar to large-scale data processing.**

Keywords—Japanese synthetic populations, protection level, real-scale social simulations, large-scale data processing.

I. INTRODUCTION

In this paper, we propose a distribution system for Japanese synthetic population with protection level that has been developed through Joint Usage / Research Project using Japan High Performance Computing and Networking plus Large-scale Data Analyzing and Information Systems (JHPCN). The "Synthetic population" is a population data synthesized from the statistics of the census that are conducted by the government and publicly released. The synthetic population is required to implement real-scale social simulations (RSSS) that are targeted on a specific community or society. We have developed a distribution system of the synthetic population data for researchers who try to implement RSSSs.

Recently social simulations attract researchers who try to investigate mechanisms in a human community or society. One of the most famous examples in social simulations is the segregation model proposed by Schelling [1]. In his model, agents in a virtual world repeat moving until they satisfy their surroundings where neighbors of the same group (i.e., the same race or beliefs) are living more than they expect. The Schelling's segregation model shows that the segregation may happen even agents in the virtual world do not expect the high ratio of their neighbors from the same group.

The implication from the Schelling's segregation model is important, though, we should apply the implication to the real world in order to see what will happen in a specific community or society. If we can implement real-scale social simulations (RSSSs) that are conducted for a specific community or society, they enable us to see what will happen in the community or society. In order to implement RSSSs, compositions of the population in the target community are required besides models of decision making or activities of citizens or residents in the target area. The local government who collects taxes grasps compositions and income of each family, but such information is protected as the privacy data even within the government in many countries. In Japan, the head of an administrative organ, a local public entity or any other executive committees should manage the information in an appropriate manner to protect any secret of individuals or juridical persons by Articles 39 to 43 in the Statistics Act of Japan. Therefore, we try to synthesize populations that have the same statistical characteristics according to the statistics publicly released by the government.

Fig. 1 shows the system we have developed using high performance computers in Cybermedia Center of Osaka University, Information Initiative Center of Hokkaido University, Information Technology Center of The University of Tokyo and RIKEN Center for Computational Science. Using the large-scale parallel computing system "OCTOPUS" in Osaka University we synthesize whole populations in Japan. We transfer the synthesized populations to "Intercloud System" in Hokkaido University to develop the synthetic population database. We make a backup of the database using "HPCI shared storage" in the University of Tokyo and RIKEN Center for Computational Science (HPCI stands for High Performance Computing Infrastructure). The authors of this paper are the project members of this HPCI-JHPCN Synthetic Population project headed by the first author.

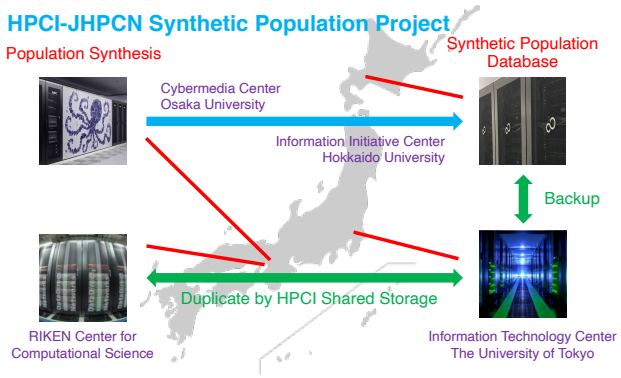

Fig. 1. HPCI-JHPCN Synthetic Population Project.

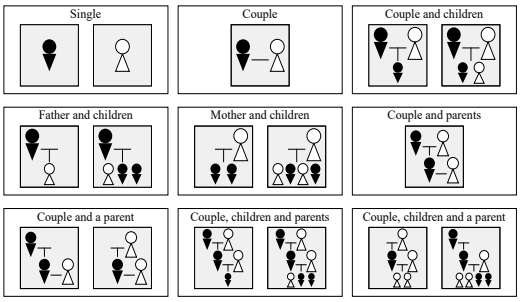

Fig. 2. Nine types of households synthesized by Murata et al. [2].

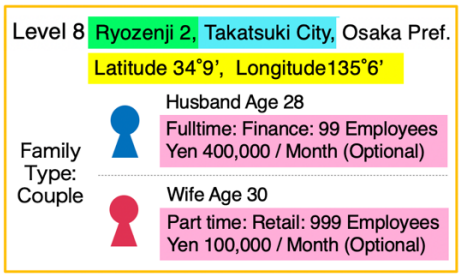

Fig. 3. Attributes of synthetic population.

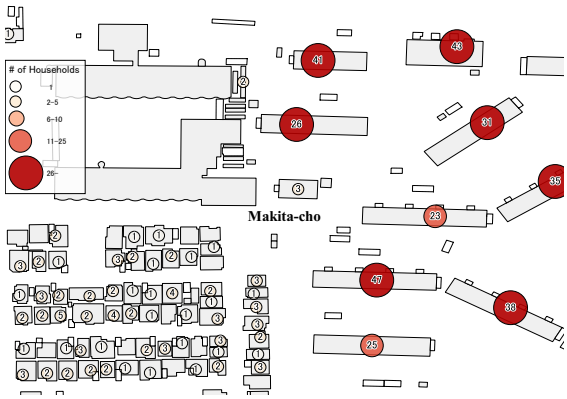

Fig. 4. The number of households assinged to each building (Map is depicted using the map of Geographical Information Agency of Japan).

In this paper, we briefly show how we synthesize population in Japan from the national census. Then we explain how much populations are synthesized in HPCI-JHPCN Synthetic Population Project. In order to encourage the usage of the synthetic population, we have set the protection level to distribute the synthetic population, that allows undergraduate students, graduate students and researchers to access to attributes of the synthetic population by each level. After the explanation about the protection level, we show several research examples by the protection level that employ the synthetic populations.

II. Synthetic Population Method

In order to implement RSSSs for a specific geographical community or society, we need to have compositions of a population of the target area. Some of the authors of this paper have already proposed a synthetic population method in their research [2-4]. They synthesize nine types of households shown in Fig. 2 [2]. These types of households cover 95% of Japanese Population. The attributes of the synthesized household are shown in Fig. 3. Each household has its family type (One of nine types of households in Fig. 2), attributes of its members such as role, sex, age, working status, and income.

As working status, they are categorized as a fulltime worker or part time worker. The type of industries and the size of the business they are working for are also synthesized [3]. The location of the household is shown in three levels. The first one is “prefecture”. In Japan, there are 47 prefectures. Each prefecture has one governor who is elected by an election. Tens of cities, towns and villages are included in a prefecture. There are 1,741 cities, towns or villages in Japan. Each city, town or village has its mayor who has also elected by an election. Each city, town or village has several areas or streets within its territory. “Ryozenji 2” in Fig. 3 is one of area names in Takatsuki City, Osaka Prefecture, Japan.

Using the projection technique proposed in [4], we projected each synthesized household on a building in a map. In the map of Geographical Information Authority of Japan, shapes of buildings are provided. We project each household on a detected building in the map, then we assign the attribute of the latitude and the longitude of the building to the assigned household. Fig. 4 shows an example of the number of households assigned to buildings in the map.

TABLE I shows the results of synthesizing populations of Japan using the statistics in the national census held in every 5 years, that is, 2000, 2005, 2010, and 2015. We have synthesized 100 sets of populations for each Census year according to the statistics of the national census. We have completed synthesizing populations for every year according to the statistics of each area in City, Town or Village. The fourth row shows the ratio of areas where we have projected households in that area. In 2015, we have projected synthesized households on buildings in 99.7% areas. The other 0.3% area in Japan does not have any buildings in its territory. Since the map created by Geographical Information Agency of Japan before July 30, 2014 has less information about buildings, we have assigned only 42.6% areas in 2010. We could not assign synthetic households before 2005 since no map with sufficient information is available.

As for the working status of each worker in the fifth row, we could synthesize it only in 2010 and 2015 because of the availability of the statistics of working status in Japan.

TABLE I. Synthesized Population by National Census

Year	2000	2005	2010	2015
City, Town & Village	O	O	O	O
Area	O	O	O	O
Buildings in Each Area	X	X	42.6%	99.7%
Working Status	X	X	O	O
File Size	20GB	20GB	28GB	28GB

TABLE II. PUBLICLY RELEASED SYNTHETIC POPULATIONS

Organization	Synthesized Area	Statistics
RTI International, USA	All states, USA Population: 300 million	2010 US Decennial Census 2007-2011 American Community Survey
CDRC: Consumer Data Research Center, UK	England & Wales, UK Population: (53 + 3) million	2011 UK Census
Kansai University, Japan	Japan Population: 120 million	2000 National Census 2005 National Census 2010 National Census 2015 National Census

TABLE III. DATA PROTECTION LEVEL

Level	Resolution	Working Status and Income※1	Users※2
1	Prefecture	×	Undergraduate
2	City, Town, Village	×	Undergraduate
3	Prefecture	○	Graduate
4	City, Town, Village※3	○	Graduate
5	Area	×	Graduate
6	Building Coordinate	×	Researcher
7	Area※3	○	Researcher
8	Building Coordinate※3	○	Researcher

※1 According to the research objective, income is provided as optional.

※2 According to the research objective, higher level can be considered.

※3 Population under 1,000 are excepted.

The last row in TABLE I shows the file size of each synthesized population. Therefore, the size of 100 sets of synthesized populations in each year become 2 TB to 2.8TB. Therefore, the size of all synthetic populations in 4 census years becomes about 10TB in total. We prepare the HPCI shared storage for the update of synthetic population using a new synthetic population method or the new synthetic population for the latest census which will be held in 2020.

We publicly released the synthetic populations of Japan to those who agree with the following conditions [5, 6].

1) The synthetic populations do not contain any data of the real households and individuals.
2) The synthetic populations contain only the same statistical characteristics of the real households and individuals.
3) The synthetic populations do not contain any statistical characteristics that are not used in the synthetic process.
4) The synthetic population will be updated when latest statistics or a modified synthetic method become available.
5) Simulations or analysis using the synthetic populations should be conducted on multiple sets of populations.
6) Outcomes of simulations and analysis should NOT be released any personal or private information that is relating to real households or individuals.

III. SYNTHETIC POPULATION DATABASE WITH PROTECTION LEVEL

Nation-wide synthetic populations are publicly released only in United States [7], United Kingdom [8] and Japan. TABLE II shows the target area (or country) each organization releases.

In order to implement real-scale social simulations (RSSSs), it is essential to use synthetic populations that have the same (or similar) statistical characteristics. However, so far, many RSSSs targeting a specific geographical area synthesize only attributes they need in their research. Thus, the quality of the synthesized populations cannot be considered in those researches. It is important to consider the quality of the synthesized populations and it should be improved if some problems are found in them.

Since we could synthesize many populations that have the same or similar statistical characteristics with the real statistics, we prepared 100 sets of synthesized populations as shown in the last section. Even though the synthesized population itself does not contain any private data of real households, it is easy for users to see which kind of persons are mainly living in a specific area in a city the users are considering. Therefore, we set the protection level that defines attributes users can access. TABLE III shows the protection level or access level for users of undergraduate students, graduate students and researchers.

The color of cells in TABLE III corresponds to the color in Fig. 3. Fig. 3 shows all attributes that are provided with the users who can access to Level 8. In the data of Level 1, only the family type, the age of husband, and the age of wife are provided as such a household is in anywhere in Osaka Prefecture. In Level 2, users can see that household in anywhere in Takatsuki City. In Level 3, users can see the working status of that household but only in Osaka Prefecture. In Level 4, users can see the working status in Takatsuki City. In Level 5, users can access to the area information in Takatsuki City, but they do not have any working status. In Level 6, users can access to the building coordinate in Takatsuki City without any working status. Level 7 provides the area information with working status. Finally, Level 8 provides all the attributes for the household with working status and the building coordinate in a specific area of Takatsuki City, Osaka Prefecture.

In the following section, we provide some research examples that employ synthetic populations by the protection level. We provide six examples from Protection Levels 3 to 8.

IV. RESEARCH EXAMPLES USING SYNTHETIC POPULATION WITH PROTECTION LEVEL

In order to employ synthetic populations in real-scale social simulations, it should be considered how to combine the attributes of synthetic populations with attributes of data the users have. In this section, we provide research examples that we have already employed synthetic populations to implement synthetic populations in a target community or society.

A. Level 3: Public Pension Simulation in Japan [9]

In this research, Du and Murata [9] employed synthetic populations for whole Japan by prefecture. They try to see that the value of pension received by pensioners varies where they live since commodity prices vary by prefecture. The wage increasing rate varies in prefectures with high population density and prefectures with low density. They apply the probability of job hunting and job leaving by age to members of households of the synthetic population. They also employ the wage increasing rate by prefecture and calculate the income replacement rate of pension payment to the average income of working people at that time. They found that 50%

income replacement rate (that is a governmental target) can be attained only in certain prefectures and only by couples both of them had continuously worked until they retired. Their research shows that some financial supporting scheme will be required for those of single households or couples without fully paid from the pension program due to their insufficient payments of pension premiums.

B. *Level 4: Labor Market Simulation in Iwate Pref. [10,11]*

In this research, Goto [10] employed synthetic populations of Ofunato City, Iwate Prefecture (See Fig. 6). He tries to see the effect of the cash-for-work (CFW) program in Ofunato City in the reconstruction process from the Great East Japan Earthquake and Tsunami in 2011. As shown in Fig. 7, CFW programs provide jobs to refugees but it aims to train and prepare them to get new skills for new jobs. He employed the attribute of age, sex, working status including industry type and income of synthetic populations to simulate job hunting activities in Ofunato City. To do that, he defined a job-skill matrix that relates required skills with a certain job, and acquired skills with jobs when refugees are engaged with that job.

He also conducted the same simulations in neighboring cities and towns such as Rikuzentakata, Sumita, Kamaishi, and Otsuchi (See Fig. 6) [11]. They found that the type of CFW programs are different by cities in order to reduce the unemployment rate. Their simulation results indicate that the CFW program should be tailored according to acquired skills of residents in target areas.

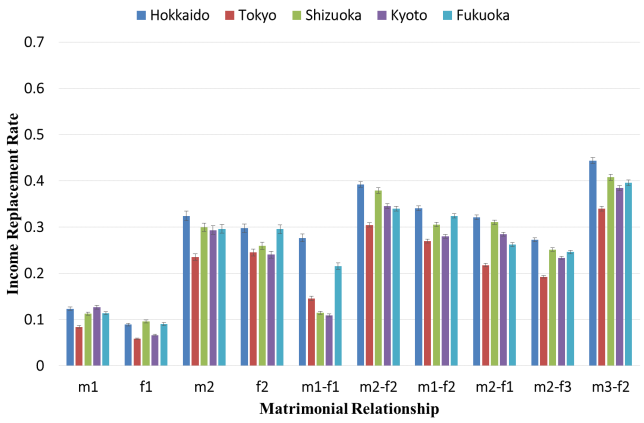

Fig. 5. Income Replacement Rate by Household Type and by Prefecture in 2050 [9].

Fig. 6. Ofunato and Neighboring Cities and Towns in Iwate Prefecture.

C. *Level 5: Voting Simulations in Takatsuki, Osaka Pref. [12]*

In this research, Murata and Konishi [12] employed households attributes by area in Takatsuki City, Osaka Prefecture. They try to see the balance between the voter turnout and the number of polling places in the city. Their simulation results show that there are several options to increase the voter turnout or reduce the voting costs while keeping voter turnout (See Fig. 8).

D. *Level 6: AED Location Analysis in Tokorozawa, Sasitama Pref. [13]*

In this research, Ichikawa employed the location information (i.e., the building coordinate of each household) in the synthetic population. He applies the risk of heart attack to each household based on age of household members. The location analysis about AED (Automated External Defibrillator) indicate that only 55.7% of residents live within 300m from AED. That means 44.3% of residents have difficulties to get AED within ten minutes after a heart attack incident (See TABLE IV). His analysis indicates that more AED is needed to cover more people in the city from heart attack.

E. *Level 7: Flu simulation in Izu-Oshima Island, Tokyo [14]*

In this research, Ichikawa et al. [14] employed household attributes with working status (without income) by area. He assigned working places for workers and schools for young people. With the model of flu infection, they found that the transition of infected patients from area to area in the island as shown in Fig. 8. This simulation results indicate the importance of real-scale simulation to see how the flu spreads in specific areas.

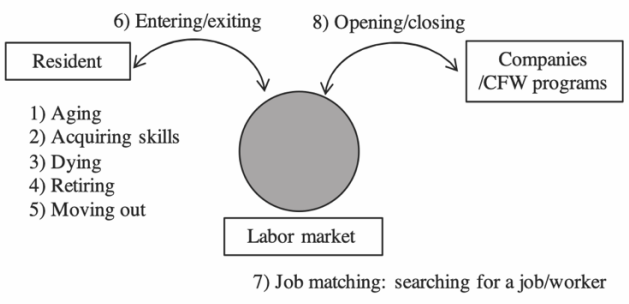

Fig. 7. Labor Market Dynamics in Ofunato City and Neighbor Cities [10].

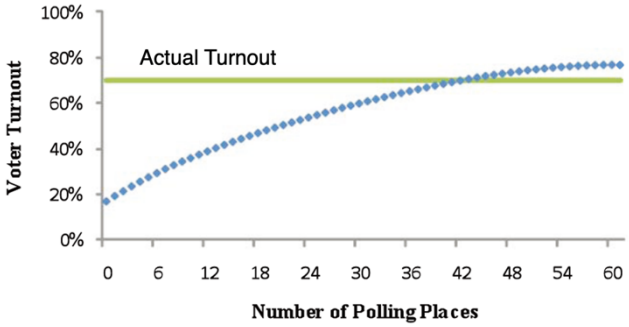

Fig. 8. Voter turnout in 2005 Lower House Election in Japan [12].

TABLE IV. THE NUMBER OF HOUSEHOLDS AND POPULATION WITHIN 300M FROM AED [13].

	Total	Within 300m	Rate
Households	134,985	77,417	57.4%
Population	319,294	177,803	55.7%

Fig. 9. The number of infected people by area in Izu-Oshima Island [14].

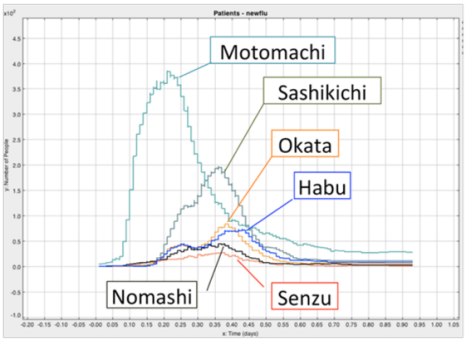

Fig. 10. The number of infected people by area in Izu-Oshima Island [14].

F. Level 8: COVID-19 simulation in Nagano Pref. [15]

In this research, Kurahashi and Nagai [15] employed household attributes with working status (without income) and building coordinate in a town, Nagano Prefecture. After projecting households in the town, they employ their models to simulate the spreading COVID-19 diseases in the town. Their simulation results indicate that the number of patients in the danger list can be reduced by applying PCR screening to all workers in shops or restaurants for visitors every day, and if some workers become positive in the screening, they are quarantined for two weeks.

V. Conclusion

In this paper, we introduced the high-performance computing systems in Japan for synthesizing and distributing data of populations based on the publicly released statistics for researchers to implement real-scale social simulations (RSSSs). We also introduced the protection level for synthetic populations and show some examples of usage of the synthetic populations by each level. The simulation results of these examples help decision makers to see the effectiveness of introduced measures or the challenges in the specific cities or prefectures. This is one of the important purposes of RSSSs. We have already publicly called for users of synthetic populations and received positive responses from researchers in Japan. We continue to improve the quality of synthetic populations and increase attributes that helps researchers to implement their RSSSs.

Acknowledgment

This work is supported by “Joint Usage / Research Center for Interdisciplinary Large-scale Information Infrastructures” and “High Performance Computing Infrastructure” in Japan (Project ID: jh190056-MDH, jh200022-DAH), and used HPCI Shared Storage (Project ID: hp190215), JSPS KAKENHI (Grant Number 17K03669, 20K10362), and the Kansai University Fund for Supporting Outlay Research Centers, 2020.

References

[1] T. Schelling, “Dynamic models of segregation,” J. of Mathematical Sociology, vol. 1, pp. 143–186, 1971.

[2] T. Murata, T. Harada, and D. Masui, “Comparing Transition Procedures in Modified Simulated-Annealing-Based Synthetic Reconstruction Method Without Samples,” SICE J. of Control, Measurement, and System Integration, vol. 10, no. 6, pp. 513-519, 2017.

[3] T. Harada and T. Murata, “Projecting Households of Synthetic Population on Buildings Using Fundamental Geospatial Data, SICE J. of Control, Measurement, and System Integration, vol. 10, no. 6, pp. 505-512, 2017.

[4] T. Murata, S. Sugiura, and T. Harada, “Income Allocation to Each Worker in Synthetic Populations Using Basic Survey on Wage Structure,” Proc. of 2017 IEEE Symposium Series on Computational Intelligence, (Hawaii, USA, Nov.27–Dec. 1, 2017), pp. 471-476, 2017.

[5] T. Murata and T. Harada, “Synthetic and Distribution Method of Japanese Synthesized Population for Real-Scale Social Simulations,” Proc. of 33rd Annual Conf. of the Japanese Society for Artificial Intelligence (Niigata, Japan, June 4-7, 2019), 3B4-E-2-05, 3 pages, 2019.

[6] T. Murata, T. Harada, M. Ichikawa, Y. Goto, L. Hao, S. Date, M. Munetomo, A. Sugiki, “Distribution of Synthetic Populations of Japan for Social Scientists and Social Simulation Researchers,” Proc. of Int’l Conf. on Machine Learning and Cybernetics (Kobe, Japan, July 7-10, 2019), 5 pages, 2019.

[7] W.D. Wheaton, J.C. Cajka, B.M. Chasteen, D.K. Wagener, P.C. Cooley, L. Ganapathi, D.J. Roberts, and J.L. Allpress, “Synthesized Population Databases: A US Geospatial Database for Agent-Based Models,” No. MR-0010-0905 (RTI Press), 12 pages, 2009.

[8] J. Robards, C.G. Gale, and D. Martin, “Creating a Synthetic Spatial Microdataset for Zone Design Experiments,” National Centre for Research Methods Working Paper 17/5, pp.1-40, 2017.

[9] N. Du and T. Murata, “Comparing Income Replacement Rate by Prefecture in Japanese Pension System,” Advances Social Simulation 2015 (Advances in Intelligent Systems and Computing, Springer), vol.528, pp.95-108, 2017.

[10] Y. Goto, “Stylized Fact Analysis of Cash-For-Work Programs in the Disaster Reconstruction Process,” Proc. of IEEE International Conference on Systems, Man, and Cybernetics 2018 (Miyazaki, Japan, Oct. 7-10, 2018), pp. 1144-1149, 2018.

[11] S. Abe and Y. Goto, “Simlation Analysis of Effective CFW Based on Regional Characteristics in the Reconstruction Process,” Proc. of the 22nd Conference of Division of Social Systems (Ishigaki, Japan, Mar. 15-17, 2020), pp.22-29, 2020 (in Japanese).

[12] T. Murata and K. Konishi, “Making a Practical Policy Proposal for Polling Place Assignment Using Voting Simulation Tool,” SICE J. of Control, Measurment, and System Integration, vol. 6, no. 2, pp. 124-130, 2013.

[13] M. Ichikawa, “Scenario Analysis of Night Emergency Transportation Using Social Simulation Approach,” Journal of the Society of Instrument and Control Engineers, vol.57, no.6, pp.407-412, 2018 (in Japanese).

[14] M. Ichikawa, H. Tanuma, and H. Deguchi, “Infectious Disease Simulation Model for Estimation of Spreading,” Development in Business Simulation and Experimental Learning, vol.38, pp.358-366, 2011.

[15] S. Kurahashi and H. Nagai, “Preventing Measures for 2019 Novel Coronavirus Diseases (COVID-19) in a Tourist Site,” http://www.u. tsukuba.ac.jp/~kurahashi.setsuya.gf/doc/slide-2020-n06-r2.pdf, 2020 (In Japanese).