The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Synthetic and Distribution Method of Japanese Synthesized Population for Real-Scale Social Simulations

Tadahiko Murata[*1] Takuya Harada[*2]

[*1] Department of Informatics, Kansai University, Japan

[*2] Research Institute for Socionetwork Strategies, Kansai University, Japan

In this paper, we describe how synthesized populations are essential in real-scale social simulations (RSSS), and the current situation of the population synthesis for whole populations in Japan. RSSS is simulations using the real number of populations or households in social simulations. This paper describes how we have completed to synthesize multiple sets of populations based on the statistics of each local government in Japanese national census in 2000, 2005, 2010 and 2015. We have started to distribute those multiple sets of the synthesized populations for researchers of RSSSs in Japan. In distributing the synthesized populations, we should protect personal or private information in the synthesized populations. We show some scheme how to protect them using a cloud service or secure computations.

1. Introduction

In this paper, we try to develop a platform for Real-Scale Social Simulation (RSSS) by synthesizing whole households in Japan and providing the data of synthesized households for researchers who try to develop RSSS tools. RSSS is simulations using populations or households in the real scale.

Recently social simulations have attracted from many researchers to tackle with problems in our environments or communities. One of the most influential social simulations is the segregation model proposed by **Schelling (1971)**. In his model, he clearly shows how segregations happen due to the preference of residents to be a neighbor of the same race or group. His model shows that segregations can happen even if there is no hostility among races. His model is quite interesting and meaningful to give understanding of conflict and cooperation. He was awarded the 2005 Novel Memorial Prize in Economic Science.

Although Schelling's model is quite significant, interpretation is required to apply his model to real situations. If we are able to directly conduct simulations with real-scale environments and real-scale residents, it is easy to draw insight from simulation results. That is why RSSS has much attention from many researchers recently.

In order to conduct RSSS, real-scale populations are required. For example, when **Murata & Konishi (2013)** optimized the number of polling places with considering the voting rate and the number of polling places in a city using a scheme of RSSS, they should synthesize the population in the city and measure the distance of polling places from their homes. When **Murata & Du (2015a)** assessed effects of the pension program for each household in Japan, they should create and simulate demographic movement of all prefectures in Japan for 25 or over 100 years according to the statistics of Japanese census conducted in 2010.

2. Synthetic reconstruction methods

Tadahiko Murata; Address: Kansai University, 2-1-1, Ryozenji, Takatsuki, Osaka 569-1095, Japan; Phone: +81-72-690-2429, FAX; +81-72-690-2491; murata@kansai-u.ac.jp .

Since RSSS researchers should face to synthesize populations in the target area of their social simulation sooner or later, we have synthesized populations using the available statistics in each local government such as city, town and village in Japan according to the national census in 2000, 2005, 2010 and 2015. The number of cities, towns and villages in Japan is 1741 in the national census in 2015).

Methods synthesizing populations with individual attributes are known as Synthetic Reconstruction method (SR method) (**Wilson, 1976**). Originally an SR method employs real samples from the real statistics. That method increases the number of individuals from the samples in order to fit the real statistics. Here, we prefer using the term "synthesize" to the term "reconstruct" in this paper. Since a reconstruction method is expected to generate exactly the same attributes of each individual in the population, however, it is impossible to reconstruct the same attributes from a small number of statistics. Therefore, we can only synthesize a population that has the same statistical characteristics using SR methods. **Lenormand & Deffuan (2013)** compared SR methods that employ samples with a synthetic method without samples. They showed the synthetic method without samples is better than the former one.

We employ a synthetic method without samples in this paper. The basis of our method is a method proposed by **Ikeda et al. (2010)**. They proposed a method for synthesizing households of nine family types according to the nine real statistics using a simulated annealing method **(Davis, 1987)**. **Fig. 1** shows the nine

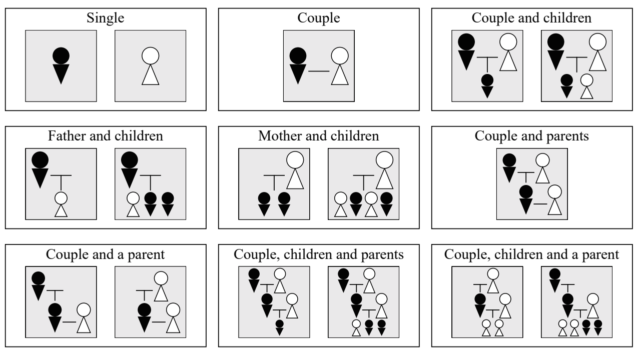
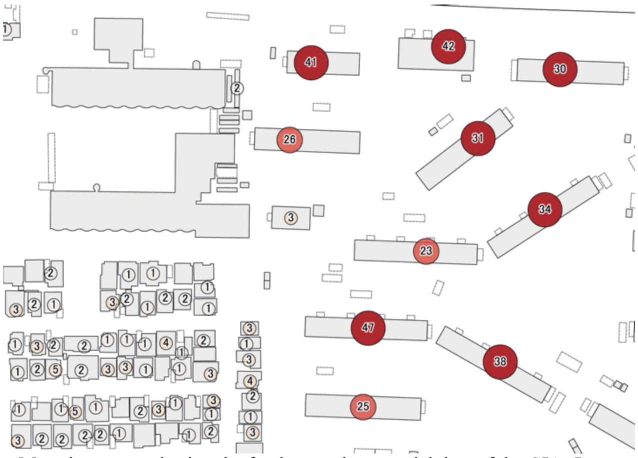


Fig. 1 Nine Family Types.



Maps is generated using the fundamental geospatial data of the GIA, Japan

Fig. 2 Households Projection on buildings.

family types they synthesized. 95% households in Japan come from these family types. Each family member has attributes of sex, age, kinship in its household. **Murata & Masui (2014, 2015b)** modified the objective function and a transition method in their simulated annealing method. Although their method (Ikeda, 2010; Murata, 2014, 2015b) can synthesize a population that has the same statistical characteristics with the real statistics, their method tried to synthesize a reduced population with only 500 or 1,000 households. The synthesized population is too small for social simulations in a real city, town or village.

In order to cope with the problems arisen in the reduced number of populations, we tried to synthesize exactly the same number of individuals in a target area such as states, counties, and prefectures using statistics of prefectures **(Murata, 2016, 2017a)**. We first increased the number of real statistics for each family type and modified a transition method (Age-Changing method) in their SA method by considering role in a family (2016). We then proposed another transition method (Age-Swap method) in their SA method that keeps the distribution of the initial population that is fit to the real statistics (2017a). Age-changing method has a better performance in reducing the error when the number of transitions in an SA method is relatively small. On the other hand, Age-swap method can reduce better than Age-changing method when the number of transitions in an SA method is relatively large.

When we increase other attributes such as geographical characteristics **(Harada, 2017)** or occupation and income **(Murata, 2017b)** to the synthesized population, populations by local governments such as city, town or village are required **(Murata, 2018)**. There are finer statistics that are statistics for each "basic unit block." The number of "basic unit blocks" in Japan is around 1.9 millions. A population synthesis method using statistics of "basic unit block" is proposed by Harada & Murata (2018). We have conducted population synthesis using the above algorithms with high performance computers in Osaka University. **Fig. 2** shows an example of a household projection on buildings in a map of Japan. Each figure in a circle shows the number of households residing in the corresponding building.

Table 1 Distributed 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

3. Synthesized Population Distribution

Using the above synthetic methods, we have generated synthesized populations for whole Japan. We are trying to prepare the database of synthesized populations using database. There are only two organizations that distributes synthetic populations in the national level in the world. **Table 1** shows the distributed synthesized populations. Those organizations distribute nation-wide populations of their country.

Although they are distributing only one set of synthesized populations, we have synthesized 10 sets of populations now. Since any methods synthesizes populations based on the limited number of statistics, there is no guarantee that the synthesized population is exactly the same as the real population. Therefore, RSSS should be conducted on several sets of populations and find common outcome from the simulations, or a unique outcome among them. When we find a common result, it seems to be obtained from any populations with the same statistical characteristics of the real population. When we find some unique result, we should carefully see how the obtained result is caused. In order to conduct such multiple simulations, we distribute several sets of populations.

We are distributing synthetic populations with the following notations.

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 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.

Although synthesized populations are not real populations, residents may consider that their privacy is offended by releasing their personal information such as their occupations, income or educational back grounds. Therefore, we require researchers to conduct their simulations or analysis using multiple sets of synthesized populations in Item 5). We also require researchers not to release outcomes of their simulations or analysis in any forms that enables others to identify or estimate a private information in a certain household.

4. Further Challenges for Data Protection

In order to protect the personal or private information in the synthesized populations, we are planning to employ a cloud service that enables simulations using the synthesized populations. By employing a cloud-style service, we do not have to distribute the synthesized data themselves to researchers but allow them to access the synthesized data in their RSSSs. In order to realize such an interface for accessing the synthesized populations, we should develop online programming tools for utilizing the synthetic populations in simulations or analysis.

Another way to protect personal information is to employ secure computation (Chida, 2014). The secure computation enables users to utilize sensitive data without allowing them to see exact values of them.

5. Conclusion

In this paper, we show the current status of population synthesis of whole populations in Japan. We have developed multiple sets of synthesized populations with the same statistical characteristics of the real populations in Japan. In synthesizing the populations, we utilized the statistics conducted in 2000, 2005, 2010, and 2015. After synthesizing the populations with sex, age, kinship in their household, we are increasing attributes of each individual such as geospatial data, occupation, and income. We hope enriching such synthesized population will help researchers who try to develop real-scale social simulations or analyze micro data to see characteristics of our communities or environments.

Acknowledgement

Part of this research is funded by Foundation for the Fusion of Science and Technology in 2017, JSPS KAKENHI 17K03669 in 2017, and Tateishi Science and Technology Foundation in 2018. Synthesized populations are generated using the large-scale computing systems (VCC) of Cybermedia Center, Osaka University.

References

[Chida, 2014] K. Chida, G. Morohashi, H. Fuji, F. Magata, A. Fujimura, K. Hamada, D. Ikarashi, R. Yamamoto, Implementation and evaluation of an efficient secure computation system using 'R' for healthcare statistics, J. of the American Medical Informatics Association, Vol. 21, Is. e2, pp. 326-331, 2014.

[Davis, 1987] L. Davies: Genetic algorithms and simulated annealing; Research Notes in Artificial Intelligence, Los Altos, CA: Morgan Kaufumann, 1987.

[Harada, 2017] T. Harada, T. Murata, Projecting household of synthetic population on buildings using fundamental geospatial data, SICE Journal of Control, Measurement, and System Integration, Vol. 10, No. 6, pp. 505-512, 2017.

[Harada, 2018] T. Harada, T. Murata, Geospatial data additional method using basic unit blocks, Proc. of SICE Symposium on Systems and Information 2018, 6 pages, 2018 (in Japanese).

[Ikeda, 2010] K. Ikeda, H. Kita, M. Susukita, Estimation method of individual data or regional demographic simulations, Proc. of SICE 43rd Technical Com. on System Engineer, pp. 11-14, 2010 (in Japanese).

[Lenormand, 2013] M. Lenormand, G. Deffuant, Generating a synthetic population of individuals in households: Sample free vs sample-based methods, Journal of Artificial Societies and Social Simulation, vol. 16, no. 4, pp. 1-9, 2013.

[Murata, 2013] T. Murata, K. Konishi, Making a Practical Policy Proposal for Polling Place Assignment Using Voting Simulation Tool, SICE Journal of Control, Measurement, and System Integration, Vol. 6, No. 2, pp. 124-130, 2013.

[Murata, 2014] T. Murata and D. Masui, "Estimating agents' attributes using simulated annealing from statistics to realize social awareness", Proc. of 2014 IEEE Int'l Conf. on System, Man & Cybernetics, pp. 717-722, 2014.

[Murata, 2015a] T. Murata, N. Du, Comparing income replacement rate by prefecture in Japanese pension system, Advances in Social Simulation, pp. 95-108, 2015.

[Murata, 2015b] T. Murata and D. Masui, "A two-fold simulated annealing to reconstruct household composition from statistics", Proc. of 2015 IEEE Int'l Conf. on System, Man & Cybernetics, pp. 1133-1138, 2015.

[Murata, 2016] T. Murata, T. Harada, D. Masui, Modified SA-based household reconstruction from statistics for agent-based social simulations", Proc. of 2016 IEEE Int'l Conf. on Systems, Man, & Cybernetics, pp. 3600-3605, 2016.

[Murata, 2017a] T. Murata, T. Harada, D. Masui, Comparing transition procedures in modified simulated-annealing-based synthetic reconstruction method without samples, SICE Journal of Control, Measurement, and System Integration, vol. 10, no. 6, pp. 513-519, 2017.

[Murata, 2017b] T. Murata, S. Sugiura, 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, pp. 471-476, 2017.

[Murata, 2018] T. Murata, T. Harada, Synthetic method for population of a prefecture using statistics of local governments, Proc. of 2018 IEEE Int'l Conf. on Systems, Man, & Cybernetics, pp. 1171-1176, 2018.

[Schelling, 1971] T. Schelling, "Dynamic models of segregation, Journal of Mathematical Sociology, Vol. 1, pp. 143-186, 1971.

[Wilson, 1976] A. G. Wilson, C. E. Pownall, A new representation of the urban system for modeling and for the study of micro-level interdependence, Area, vol.8, no. 4, pp. 246-254, 1976.