Artificial Population: Synthesizing Population from Census Data

Rongqing Meng
College of Information Systems and Management
National University of Defense Technology
Changsha, China
meng.rongqing.realize@outlook.com

Xiaogang Qiu
College of Information Systems and Management
National University of Defense Technology
Changsha, China
michael.qiu@139.com

***Abstract*—Artificial Population is very useful in many application domains. In this paper, we use census data to synthesize artificial population. We define the structures of household and individual. The data patterns of census data are analyzed too. Based on these definitions and patterns, we design the algorithms to synthesize population. We use these algorithms to synthesize artificial population of Beijing. The result indicates that our algorithms are effective.**

Keywords—artificial population; census data; individual

I. Introduction

Artificial population is the infrastructure for computational experiments, and has been applied in multi domains such as disease transmission [1], economic [2], traffic [3] and so on. It is so widely used because artificial population provides the heterogeneous individuals and somehow has the same characteristic as the actual population. Especially it can be the initial data for Agent-Based or Individual-Based Simulation [4]and can be basic data for construct artificial society [5].

So synthesizing artificial population is very important. There have been a lot of scholars find ways to construct artificial population. For example, in TRANSIMS Beckman et.al use Two-Step IPF(Iterative Proportional Fitting)[6] to synthesize the city scale population [7], in Big Italy, [8] Stroud et.al use IPF to synthesize the population of Italy, Kirill et.al summarize the state of art of population synthesis [9]. As we can see IPF is the basic method to synthesize artificial population. To synthesize population, two kinds of data are needed, that are integrated data and non-integrated data [7]. Integrated data is statistical data which is from census or social statistic. The integrated data indicts the macro characteristic of population and it's a long distance from the individual information. non-integrated data is the sample individuals of population who are anonymous. It is the actual data from census table or social investigation. Each item in non-integrated data corresponds to a real person. In china, the integrated data is easy to get because each country will construct conducting a census of the population in 10 years. But the non-integrated data is hard to get because of legal reasons. So the IPF method isn't suitable for synthesizing Chinese population.

The aim of this article is to synthesize population from census data and the statistic of artificial population approaching the real population. We use the process of synthesizing the population of Beijing City based on the 2010 census data as a case study. Firstly the structures of household and individual are defined. From these structures we know what data are is required. Then the data patterns of census data are analyzed. We filter the original data and select the suitable data. Based on these analyses, we design the algorithms to synthesize population. Finally, we briefly present the artificial population of Beijing.

II. House Hold and Individual

Actually artificial population is composed by individuals and if part of individuals is collected together, household is constructed. There are so many characteristics of household and individual, and it's unrealistic to assign all the attributes, so we should choose the key attributes.

A. Household

Household is a group of individuals who stay together and it is the basic unit of census. It can be represented as (1).

$$HHold = \{hId, admId, type, memNet\} \qquad (1)$$

- hID: It is the unique identity number of household.
- admId: It is the unique identity of administrative area. Each household is contained by an administrative area, just as the actual government frame. So the households somehow are constrained in some area and the same with individual which is contained in the houshold.
- type: It is used to represent the type of household. According to the relationship of individuals who are in the same household, the household can be divided into two types: family household and. collective household. Family household means the members of the household has family relationship, such as father/mother and son. Collective household means the member of the household just live together, no family relationship, such as classmate.

- memNet: It is used to store the member and relationship of this household. According to the type of household, memNet is stored by two forms, one is network and the other is list. Network is used to present family relationship and list is used to present collective relationship. The family membership network as Fig. 1 shows.

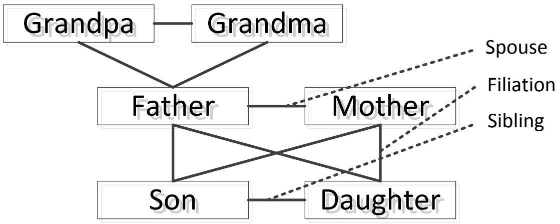

Fig. 1. Example of family member network

The basic family relation has three types, which are spouse, filiation and sibling. The other relationships can be infered from these three types. From the network, one can infer the how many generatoins of this family, such as two generation or three generaton.

B. *Individual*

Individual is the virtual person and correspond the actual person. It is the basic item of population. It can be represented as (2).

$$Ind = \{iId, hId, gen, age, hRole, sRole\} \tag{2}$$

- iId: It is the unique identity number of each individual.
- hID: It is the unique identity number of household which the individual is in. Each individual can only be contained by one household.
- gender: It is used to indicate the individual's gender.
- age: It is used to indicate the individual's age.
- hRole: It is used to indicate the individual's role in the household. According to the type of the household, the hRoles are different. If the household type is family, the hRole can father, mother, grandpa, grandma and so on. This role is compatible with household's memNet.
- sRole: It is used to indicate the individual's role in the society. Actually, social roles are related to individual's age. As TABLE I shows, it's a demo table that represent the relationship between social role and age. At every age one individual may have different role. For example, at the age between 19 and 22 one may go to college or be on work, even be unemployed. The rates of different social roles depend on the dropout rate or the unemployment rate. These data can be obtained from census data or survey. The age group and role set can be defined according to the requirement.

TABLE I. RELATIONSHIP BETWEEN SOCIAL ROLE AND AGE

Role Age	Infant	Child	Pupil	Junior	Senior	undergraduate	Graduate	Worker	Elder	Unemployed
0-2	1.0	0	0	0	0	0	0	0	0	0
3-6	0	1.0	0	0	0	0	0	0	0	0
7-12	0	0	1.0	0	0	0	0	0	0	0
13-15	0	0	0	0.95	0	0	0	0	0	0.05
16-18	0	0	0	0	0.95	0	0	0	0	0.05
19-22	0	0	0	0	0	0.55	0	0.43	0	0.07
23-25	0	0	0	0	0	0	0.10	0.80	0	0.10
26-60	0	0	0	0	0	0	0	0.90	0	0.10
60-100	0	0	0	0	0	0	0	0	1.0	0

III. Census data

Census is carried out by a country at a specified time to reveal the nature, law and trend of population. From the census data we can learn the demographic characteristic. In china, the sixth national census was conducted in November 1, 2010 and the report of census is open on internet [10].

The minimal area of census data report is subdistrict. On each web site of district bureau of statistics we can find the communique of this district [11]. The communique contains the population size of each subdistrict, and no more information about subdistrict. So the subdistrict isn't enough to generate population.

Fortunately, the city bureau of statistics provides the detailed data of each district. The census data is made up of two parts. The first part is the statistical data of short table and the second part is the statistical data of long table. Here short table is the simple version census table which should be filled in by everybody and long table is the more complex table which should be filled by sampling in ten percent. Because the short table data comes from the entire population, so it's the ideal data source to synthesize and verify the artificial population. The short table data contains eight subparts, which are overview, race, age, education, family, death, household registry, and housing. We choose six tables as the source data to synthesize population, as TABLE II shows. A0101 contains include the number of family household and collective household, the number of males and females. This table constrains the number of population and household. A0107 provide the number of population of different age group. So from this table we can get the age distribution in different age group. A0110 provide the number distribution of family member, this is the link between household and individual. A0111 descripts the distribution of generation of the family according to the number of family member. This table can be used to generate the member relationship network. A0502 directly give the number of one person household. A0501 give the number of different size family hold from the city view.

CHOOSED TABLES

Label	Name
A0101	The number of households, population and gender ratio
A0107	The partition of population according to age and gender in different district
A0110	The number distribution of family member in different district
A0111	The type of family household
A0502	The one person household in different district
A0501	Families of different sizes in the city

IV. ALGORITHM

In the above we define the basic structures and select the data table. Here we will design the algorithms which use the data table to fill the structures.

Presently, the census lacks the detailed data about collective household, so here we only consider how to generate the family household.

A. Algorithm for generating family household sturcture

As Fig. 2 shows, it is the algorithm to generate the family household structure. We take the two generation family and three generation family as example. Household size distribution can be gotten from census data. Then Household size determines the number of generation of the family. The first individual of this family is generated randomly, and call the algorithm for generation individual's attributes. After the first person fulfills his/her attribute, based on these attributes, subsequent individuals are generated [12].

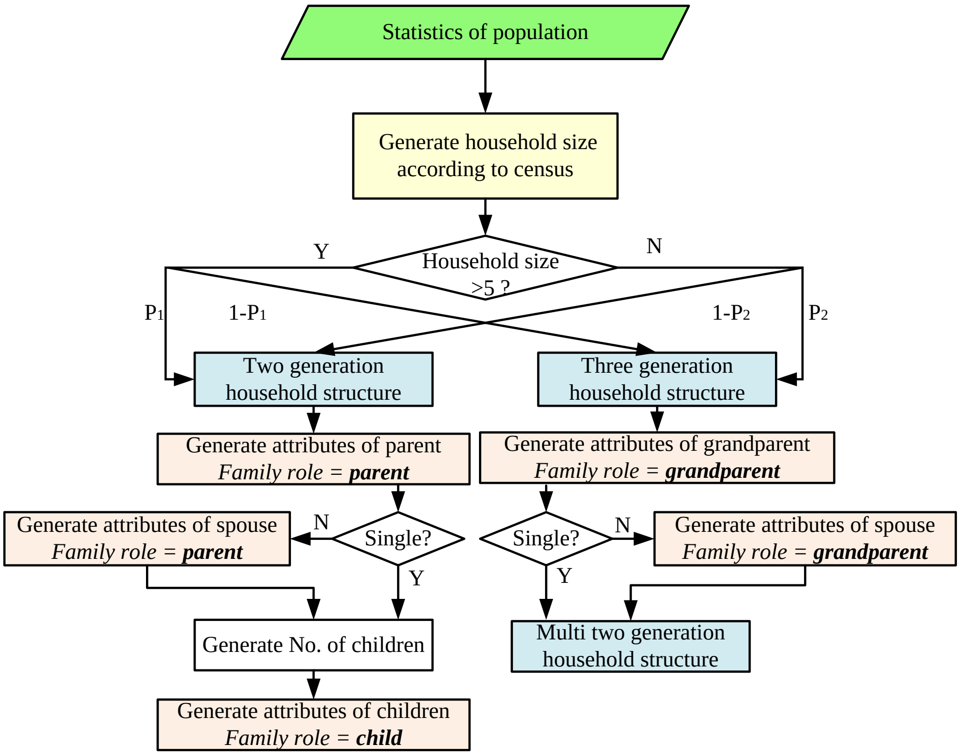
Statistics of population
Generate household size according to census
Household size >5 ?
Y
N
P1
1-P1
1-P2
P2
Two generation household structure
Three generation household structure
Generate attributes of parent Family role = parent
Generate attributes of grandparent Family role = grandparent
Generate attributes of spouse Family role = parent
N
Single?
Y
Single?
N
Y
Generate attributes of spouse Family role = grandparent
Generate No. of children
Multi two generation household structure
Generate attributes of children Family role = child

Fig. 2. Algorithm for generating family household sturcture

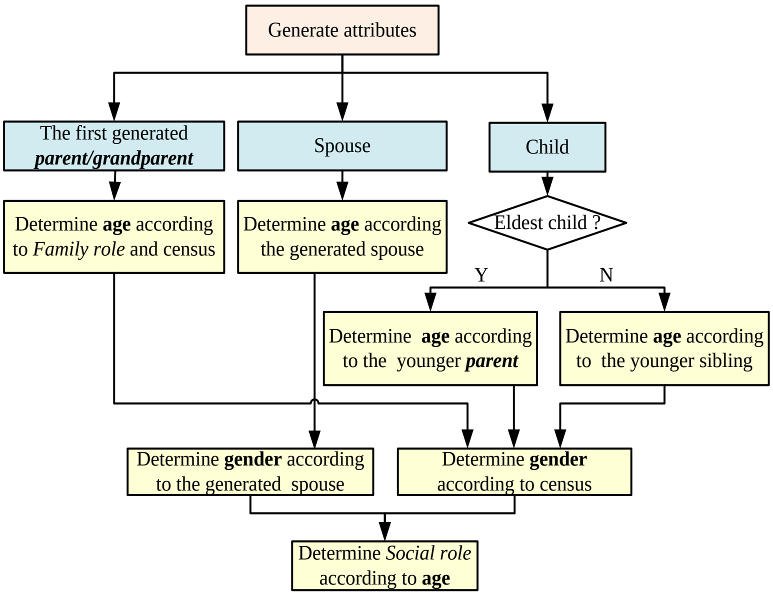
Generate attributes
The first generated *parent/grandparent*
Spouse
Child
Determine **age** according to *Family role* and census
Determine **age** according the generated spouse
Eldest child ?
Y
N
Determine **age** according to the younger *parent*
Determine **age** according to the younger sibling
Determine **gender** according to the generated spouse
Determine **gender** according to census
Determine *Social role* according to **age**

Fig. 3. Algorithm for generation individual's attribute

B. *Algorithm for generation individual's attributes*

As Fig.3 shows, the individual's attributes are assigned. Family role determines age. Then age determine social role [12].

V. CASE STUDY

We use the census data to sythesize the population of Beijing. The population is 19.619 million, and the number of household is 4.91 million. Fig. 4 shows the household size distribution of the generated population and the statistical data. The result indicates that out algorithms can help to synthesize artificial population.

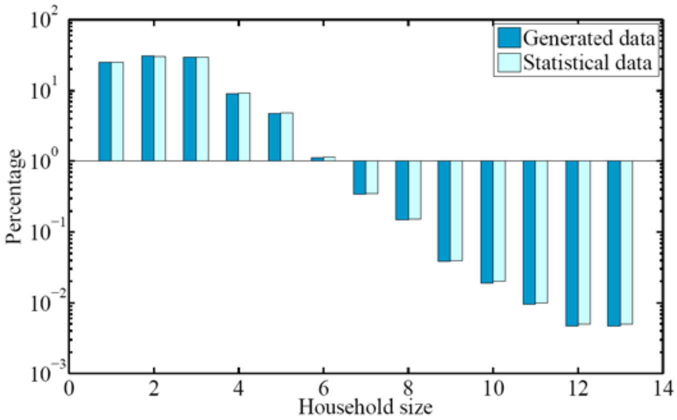

Fig. 4. Household size distribution of the generated population and the statistical data

Acknowledgment

The authors would like to thank the National Natural Science Foundation of China under Grant Nos.91024030, 9724008, 61503402, 6903402 and 4701544.

References

[1] Tang M, Mao X, Zhou H. Simulations of Influenza Viruses Transmission in Multiple Social Networks with an Artificial Society Model[J]. Journal of computational information Systems. 2014, 10(6): 2629-2637.

[2] Simulation in Economics: Evidence on Diffusion and Communication[J]. Journal of Artificial Societies and Social Simulation. 2006, 9(2).

[3] Lloyd M Smith R J B K. TRANSIMS: TRANSPORTATION ANALYSIS AND SIMULATION SYSTEM[J]. 1995.

[4] Charles M Macal C M M M. Tutorial on agent-based modelling and simulation[J]. Journal of Simulation. 2010, 4(3): 151.

[5] Epstein J M, Axtell R. Growing Artificial Societies: Social Science from the Bottom Up[M]. Washington, D.C.: The MIT Press, 1996.

[6] Deming B W E, Stephan F F. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known[J]. The Annals of Mathematical Statistics. 1940.

[7] Beckman R J, Baggerly K A, Mckay M D. Creating synthetic baseline populations[J]. Transportation Research Part A-Policy And Practice. 1996, 30(6): 415-429.

[8] Stroud P, Valle S D, Sydoriak S, et al. Spatial Dynamics of Pandemic Influenza in a Massive Artificial Society[J]. Journal of Artificial Societies & Social Simulation. 2007, 10(4): 9.

[9] M U Ller K, Axhausen K W. Population Synthesis for Microsimulation: State of the Art[C]. 2011.

[10] Information on http://www.bjstats.gov.cn/tjnj/rkpc-2010/indexch.htm

[11] Information on http://www.chystats.gov.cn/item/2013-10-22/100059169.html

[12] Yuanzheng Ge, Rongqing Meng, Zhidong Cao, Xiaogang Qiu and Kedi Huang (2014). Virtual city-an individual-based digital environment for human mobility and interactive behavior, Simulation: Transactions of the Society for the Modeling and Simulation International, 2014.