Population Synthesis using Discrete Copulas

Peijun Ye[1], *IEEE Member*, Xiao Wang[2], *IEEE Member*

***Abstract*— Synthetic population is one of the most important foundations of disaggregated travel demand forecasting and agent-based traffic simulation. This paper proposes a new sample-based method for synthetic population generation, which can be viewed as an alternative of the traditional Iterative Proportional Fitting. The method introduces bootstrapping techniques to compute a discrete copula function. Based on the copula function, associations among different attributes can be estimated and the population structure can be recovered. Experiments using actual Chinese national population data indicate that the new method can achieve the same level of accuracy as Iterative Proportional Fitting while acquire better results of the partial joint distributions.**

***Index Terms*— Population Synthesis; Copula; Association Estimation; Agent-Based Traffic Simulation**

I. Introduction

Agent-Based simulation has become an indispensable approach to travel demand analysis and traffic management strategy evaluation[1]–[4]. Computational models of human travel behavior are introduced to build virtual agents, and the simulation system then generates systemic traffic patterns by interacting those agents with each other in an artificial environment. In such process, synthetic population is an important foundation, which integrates various travel behavior models and provides the simulation an initial state. Thus, the quality of synthetic population determines or seriously impacts the credibility and reliability of the simulation and traffic demand prediction results.

Current methods of population synthesis can be categorized into two types, according to the data sources it uses. The first type is called the sample-based methods, which uses cross-classification tables and disaggregate samples as its input. The cross-classification tables are officially (usually by the National Bureau of Statistics) published marginal or partial joint distributions of target population. Each table covers a small part (but not the whole) of investigated attributes and reveals total population number (also called frequency) under each value combinations. Disaggregate samples are usually a small proportion of original census records, with private information omitted such as name and accurate address. Each sample record covers the whole investigated attributes and provides complete information of an individual. As two representatives, synthetic reconstruction and combinatory optimization are the most extensively applied sample-based methods[5][6]. The second type of population synthesis is called the sample-free methods, which only treats the cross-classification tables as its input. This is because disaggregate samples are available only in a few countries like the Public Use Microdata Samples (PUMS) in the U.S. and the Sample of Anonymized Records (SAR) in the UK. Thus in the application that lack disaggregate samples, sample-free methods are the only choices. Currently, Gargiulo, Barthelemy, Ye, Farooq and other scholars have proposed several sample-free methods[7]–[9]. For comparative analysis between the two types, we refer the reader to literature [10] and [11].

In sample-based methods, synthetic reconstruction keeps the valid correlation structure from samples in synthetic process. Thus it is applied in many projects in the field such as Integrated Land Use, Transportation and Environment (ILUTE)[12][13] and TRansportation ANalysis SIMulation System (TRANSIMS)[14]. However, the Iterative Proportional Fitting (IPF) algorithm adopted by this method has some limitations. It usually suffers from the zero element problem. In addition, IPF requires the same marginal sums of each data source to guarantee its convergence, which is usually not satisfied by the inconsistent input data. With this motivation, this paper proposes an alternative method using discrete copulas to generate a population for partial joint distributions while keeping valid correlation structure from samples simultaneously. Comparative experiments are further conducted which indicates that the new method can acquire higher accuracy.

The remainder of this paper is organized as follows. Section II states the problem and briefly introduces IPF algorithm and copula theory. Section III elucidates our new copula-based method. Section IV presents the experiment data source and population evaluation results. And finally, Section V concludes this paper with additional discussions.

II. Problem Statement and Relative Work

A. Synthetic Reconstruction Method

As alluded before, the focus of this paper is limited to the sample-based methods, which mainly involves synthetic reconstruction and combinatorial optimization. It is because when the disaggregate sample available, this type of methods can exploit the correlations among different attributes provided by the sample, so that the final synthetic population usually contains a more realistic structure. For the scenario of combinatorial optimization, it is only able to

This work is supported by National Natural Science Foundation of China (No. 61603381 and No. 61702519) and the grant of China Scholarship Council (No. 201704910284).

[1]Peijun Ye is with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; Qingdao Academy of Intelligent Industries, Qingdao 266109, China; and a visiting scholar in Department of Cognitive Science, University of California San Diego (e-mail: peijun.ye@ia.ac.cn).

[2]Xiao Wang is with The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and Qingdao Academy of Intelligent Industries, Qingdao 266109, China.

generate the population in a small region using samples from larger areas. This is hardly satisfied in most cases, where only a small proportion of sample is accessible. For this reason, synthetic reconstruction is the most comprehensively adopted. Synthetic reconstruction (SR) is composed of two steps: joint distribution estimation of the target population and individual realization. The main operation is the former step, which adopts IPF algorithm. In a general case, the joint frequency distributions can be represented as

$$f = f\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\} \tag{1}$$

where $\{X_1, \cdots, X_n\}$ stands for n investigated variables, $\{x_1, \cdots, x_n\}$ means its related n values and f is the statistical population number under the specific variable combination. By using the sample frequencies of each f as initial distribution $f(0)$, IPF iteratively updates the frequencies. Specifically, in the k-th iteration, the algorithm computes

$$\begin{aligned}
f^{(1)}(k) &= \frac{f\{X_1 = x_1, \cdots, X_n = x_n\}(k-1)}{\sum_{x_1} f\{X_1 = x_1, \cdots, X_n = x_n\}(k-1)} \cdot N_1 \\
f^{(2)}(k) &= \frac{f^{(1)}\{X_1 = x_1, \cdots, X_n = x_n\}(k)}{\sum_{x_2} f^{(1)}\{X_1 = x_1, \cdots, X_n = x_n\}(k)} \cdot N_2 \\
&\vdots \\
f^{(n)}(k) &= \frac{f^{(n-1)}\{X_1 = x_1, \cdots, X_n = x_n\}(k)}{\sum_{x_n} f^{(n-1)}\{X_1 = x_1, \cdots, X_n = x_n\}(k)} \cdot N_n \\
f\{X_1 = x_1, \cdots, X_n = x_n\}(k) &= f^{(n)}(k)
\end{aligned} \tag{2}$$

Where N_i is the real marginal frequency of the i-th attribute from the total target population. Note that in Eq. (2), the denominator of the i-th formula is the sum of current fitted frequencies in dimension i. Thus in each iteration, the algorithm in essence sequentially fits the population number in proportion according to the real marginal frequency of one attribute. Given an error threshold in advance, the iteration will converge in several rounds generally. Once such joint frequency distribution obtained, we can generate the population according to the frequency under each attribute combination, or using Monte Carlo simulation to draw a smaller scale of similar structured population according to the probability converted by $\frac{f(X_1=x_1,\cdots,X_n=x_n)}{\sum f(X_1=x_1,\cdots,X_n=x_n)}$.

Obviously, operations of IPF are quite simple and easy to be implemented. However, its result is only consistent with 1-dimensional marginal. On the other hand, statistical cross-classification tables from census usually provide partial joint distributions simultaneously covering several variables. And such information is not exploited in IPF procedure. Therefore, it is essential to improve the algorithm or develop new methods to address this problem.

B. Copula Function Theory

Copula function is used to estimate associations among random variables. It is extensively applied in economics and statistics. Following Sklar[15], a joint distribution function F with marginal distribution functions $F_1, F_2, \cdots, F_n$ can be written as

$$F(x_1, x_2, \cdots, x_n) = \mathbb{C}(F_1(x_1), F_2(x_2), \cdots, F_n(x_n)) \tag{3}$$

where $\mathbb{C}$ is called copula function, which indicates the association among variables. When $F_1, \cdots, F_n$ are strictly monotonically increasing, so that the margins are continuous, $\mathbb{C}$ is known to be unique. Currently, several copula functions are proposed, such as Gaussian copula[16] and logit copula[17]. However, when one or more marginal distribution is discrete, this is no longer the case[18]. The copula function is not unique and it is more complicated than the continuous case. Nevertheless, there are still several candidates for applications like independence copula, minimum copula and so on[19]. In contrast with Person correlation coefficient that only represents linear associations among variables, copula function is applicable for any type of distributions. It is more general and widely used. Thus, in this paper, we will discuss how to create a synthetic population via copula functions.

III. Population Synthesis Based on Discrete Copula

In population synthesis, the studied attributes can be discrete, enumerate, binary and even continuous. However, continuous values are often split into several intervals in order to reduce the computational complexity. Thus we assume that all the variables are discrete. For each frequency distribution represented by cross-classification table, it can be converted into a probabilistic distribution through dividing each cell by the total number of target population. Therefore, the basic problem is to estimate the joint probabilistic distribution. Given a set of marginal and partial joint distributions, our method starts by investigating the two with highest disaggregate level. Disaggregate level means the number of variables that a specific partial joint distribution contains. For example, the disaggregate levels of $ResidenceType \times ResidentialProvince \times EthnicGroup \times Gender$ and $AgeInterval \times EthnicGroup \times Gender$ are 4 and 3, respectively. If two partial joint distributions have the same disaggregate level, it is preferred to select the one that contains more attribute values. This is because more direct details from partial views will lead to a more accurate estimation. For example, when considering $Gender \times ResidenceType \times ResidentialProvince \times EthnicGroup$ (58 values) and $Gender \times ResidenceType \times ResidentialProvince \times AgeInterval$ (21 values), we should choose the former in priority. Another pre-operation is to fold the mutual dimensions of the two selected partial joint distributions. For example, if the most disaggregate distributions are $ResidenceType \times ResidentialProvince \times EthnicGroup \times Gender$ and $AgeInterval \times EthnicGroup \times Gender$, then we need to fold the mutual dimension $Gender$ in one of them, like

$$\begin{aligned}
&P(AgeInter, EthnicGroup) = \\
&\qquad \sum_{Gender} P(AgeInter, EthnicGroup, Gender)
\end{aligned} \tag{4}$$

This operation is able to guarantee that the two selected partial joint distributions do not include mutual variables. That is the two distributions can be represented by $F(X_1, \cdots, X_m)$ and $F(X_{m+1}, \cdots, X_{m+n})$, and $X_i \neq X_j$ ($i \neq j$ and $i, j \in \{1, \cdots, m+n\}$).

When the two partial distributions of target population determined, say $F(X_1, \cdots, X_m)$ and $F(X_{m+1}, \cdots, X_{m+n})$, we introduce a transform $(X_1, \cdots, X_m) \rightarrow X$ and $(X_{m+1}, \cdots, X_{m+n}) \rightarrow Y$ to convert multi-dimensional into 1-dimensional variables. Such transform can be easily implemented by mapping each value combination of $(X_1, \cdots, X_m)$ into an ordinal value. For convenience, the values of X and Y are denoted as $I_X = (x_1, \cdots, x_m)$ and $I_Y = (y_1, \cdots, y_n)$. Furthermore, without loss of generality, it can be assumed that $x_1 < x_2 < \cdots < x_m$ and $y_1 < y_2 < \cdots < y_n$.

The second step of our method is to estimate the association between X and Y. Let

$$F(x, y) := P(X \leq x, Y \leq y) \tag{5}$$

be the joint cumulative distribution function (cdf),

$$\begin{aligned} F_X(x) &:= F(x, +\infty) = P(X \leq x, Y \leq +\infty) \\ F_Y(y) &:= F(+\infty, y) = P(X \leq +\infty, Y \leq y) \end{aligned} \tag{6}$$

be the marginal cdfs. Following Sklar's theorem, there is a copula function that

$$F(x, y) = \mathbb{C}(F_X(x), F_Y(y)) \tag{7}$$

where $\mathbb{C}$ is not unique. Here, we propose to use empirical copula function, which is computed via disaggregate sample. Let $S = \{< x^{(1)}, y^{(1)} >, < x^{(2)}, y^{(2)} >, \cdots, < x^{(K)}, y^{(K)} >\}$ be the sample records. The empirical copula $\mathbb{C} : I_X \times I_Y \rightarrow \mathbb{R}$ is defined as

$$\mathbb{C}(x, y) = \begin{cases} 0 & \text{if } x < x_1 \text{ or } y < y_1; \\ \frac{\#\{< x^{(k)}, y^{(k)} > \in S | x^{(k)} \leq x, y^{(k)} \leq y\}}{K} & \text{otherwise.} \end{cases} \tag{8}$$

The notation $\#$ means the number of samples that satisfy the conditions in the brace. The empirical copula above manifests the dependence between X and Y. However, it may probably not be compatible with $F_X(x)$ and $F_Y(y)$ when using original disaggregate samples. This is because in most cases, the samples are extracted randomly from original census data, which probably brings a bias to the achieved results. Thus they do not accurately reveal the correlation structure of the target population. To deal with this problem, we use bootstrap alternatively. Bootstrap is a re-sampling technique that is used in R-copula computation[20]. But it does not suffer restrictions from marginal cdfs there. In our application scenario, the bootstrap steps are as follows:

1. Initialize bootstrap sample S' as null;
2. Get a random individual record Ind from original samples and compute marginal distributions $\hat{F}_X(x)$ and $\hat{F}_Y(y)$ of $S' \cup Ind$;
3. If $\hat{F}_X(x)$ and $\hat{F}_Y(y)$ are 'closer' to $F_X(x)$ and $F_Y(y)$, then include Ind into S';
4. Repeat step 2 and 3 until convergence.

In essence, the obtained S' is a modified sample set consistent with marginal constraints from target population. It is an appropriate candidate to calculate empirical copula given in Eq. (8). The main algorithm for joint cdf estimation is shown in Alg. 1, where function $ComputeError$ calculates the fitness of S' (shown in Alg. 2) and function $ReplicateRandInd$ replicates a random individual from samples. The notation $Card(S)$ represents the number of elements in set S.

Algorithm 1 JointCDFEstimation($F_X(x)$, $F_Y(y)$, S)

Input:
$F_X(x)$, $F_Y(y)$, marginal cdfs;
S, original samples;
Output:
Joint cdf $\mathbb{C}(x, y)$.

$S' \leftarrow \{\ \}$;
$(e_x, e_y) \leftarrow$ ComputeError$(F_X(x), F_Y(y), S')$;
repeat
  $Ind \leftarrow$ ReplicateRandInd(S);
  $(\hat{e}_x, \hat{e}_y) \leftarrow$ ComputeError$(F_X(x), F_Y(y), S' \cup Ind)$;
  if $\hat{e}_x < e_x$ and $\hat{e}_y < e_y$ **then**
    $S' \leftarrow S' \cup Ind$;
    $(e_x, e_y) \leftarrow (\hat{e}_x, \hat{e}_y)$;
  end if
until $e_x + e_y < threshold$ or reaches maximum iteration
return $\mathbb{C}(x, y)$ computed via Eq. (8) using S'.

Algorithm 2 ComputeError($F_X(x)$, $F_Y(y)$, S)

Input:
$F_X(x)$, $F_Y(y)$, marginal cdfs;
S, samples;
Output:
Errors (e_x, e_y).

$(e_x, e_y) \leftarrow (0, 0)$;
for each (x, y) **do**
  if $Card(S) = 0$ **then**
    $(e_x, e_y) \leftarrow (e_x, e_y) + (|F_X(x)|, |F_Y(y)|)$;
  else
    $e_x \leftarrow e_x + \left| F_X(x) - \frac{\#\{< x^{(l)}, y^{(l)} > \in S | x^{(l)} \leq x\}}{Card(S)} \right|$;
    $e_y \leftarrow e_y + \left| F_Y(y) - \frac{\#\{< x^{(l)}, y^{(l)} > \in S | y^{(l)} \leq y\}}{Card(S)} \right|$;
  end if
end for
return (e_x, e_y).

As explained above, the obtained joint distribution only includes variables $(X_1, \cdots, X_n)$. If they do not cover all the studied attributes, then we select another marginal distribution with new attributes and repeat the process. Such repeat will bring expansion of at least one new attribute into our current variable set. Thus at last, we are able to

get the distribution of all attributes. The joint distribution acquired by empirical copula can be converted into frequency distribution by multiplying each probability with the total number of target population. Furthermore, an arbitrary total number is eligible instead of real target population number, so that any scale of population can be synthesized easily.

IV. Numerical Experiments

In this section, experiments of Chinese national population synthesis will be conducted to validate the proposed method. As a comparison, IPF is also applied as a benchmark. We will firstly introduce our data source and then present the results of the experiments below.

A. Data Source

Like many countries, Chinese national census results are usually published in the form of cross-classification tables. Each table contains only a part of attributes, and provides individual number (frequency) under every attribute value of the whole investigated population. These tables are the most important data source for population synthesis because they reflect the structure of target population directly. Specifically, in our experiments, we use the data from the 5-th national census results[21]. Two types of tables can be exploited in population synthesis. One is called Short Table, which involves several basic characteristics and covers all of the investigated population. The other is called Long Table, which not only contains all the characteristics of Short Table but also includes additional detailed features like migration pattern, educational level, economic status, marriage and family, procreation, housing condition, etc.. Long Table is applied for particular individuals stochastically selected in advance, thus its result covers only a small part of population. According to the instructions published by National Bureau of Statistics, about 9.5% of population are investigated by Long Table. In our experiments, we use Short Table as the synthetic population constraints and Long Table as the evaluation criterion. That means after generating the total synthetic population, we stochastically extract 9.5% of them to compare with the Long Table results. Since the Long Table also comes from real population, such evaluation is much convincing.

Disaggregate sample is another data source for the experiments. Unlike the cross-classification tables that are available for public use, National Bureau of Statistics does not publish such sample officially. We have only collected 1,180,111 individual records, accounting for 0.95‰of the total population. These records all come from Long Table, each of which gives details attribute values of a particular individual with private information omitted. Even though the sample scale is quite small, it provides valuable correlation structure among different attributes thus helps us synthesize a valid population.

B. Experiment Results

According to the method proposed in Section III, we consider the attributes *Gender*, *Residential Province*, *Residence Type*, *Household Type*, *Age Interval* as the studied variables and select the most disaggregate distributions as the constraints. The attribute values, input and evaluation cross-classification tables are listed in Table I and Table II. Both the copula-based method and IPF are applied to conduct 5 independent experiments. Relative error of each attribute value is computed as

$$RE = \frac{\left|Count_{LT} - Count_{syn}\right|}{Count_{LT}} \tag{9}$$

where $Count_{LT}$ and $Count_{syn}$ stand for the frequencies from the Long Table and sampled synthetic population, respectively. Since all the frequencies in the Long Table are not zero, the formula above always makes sense. For every attribute combination, average RE of the 5 independent experiments are reported as follows.

TABLE I
Population Attributes and Values

Attributes	Values	Number of Values
Gender	Male, Female	2
Residential Province	Beijing, Tianjin, $\cdots$	31
Residence Type	City, Town, Rural	3
Household Type	Family, Collective Household	2
Age Interval	$\{0-5, \cdots, 96-100, \geq 100\}$	21

The first evaluation indicator is 1-dimentional marginal frequencies. We investigate each attribute respectively. Fig. 1 gives the results of the studied 5 variables. In the first sub-figure, copula-based method gets a larger deviation in the total number of population than IPF. But the total relative error is less than 1%, which is acceptable in most applications. In addition, proportions of different groups are recovered well by both methods. Specifically, copula-based method keeps a similar accuracy with IPF for *Gender* and *Residence Type*, and even performs slightly better than the latter for *Household Type*. The second and third sub-figures clearly illustrate that our method is able to reconstruct a more accurate population. However, differences between the two methods are not significant. Therefore, we are inclined to conclude that for 1-dimentional marginal frequencies, the copula-based method can achieve the same level of accuracy as IPF.

Our second evaluation indicator concentrates on partial joint frequency distributions. Fig. 2 shows the results of two distributions. Sub-figure (a) treats the first Long Table given in Table II as the benchmark (zero line). To illustrate more clearly, we draw frequency deviations for male and female separately. For each zero line, frequency deviations from IPF method are placed above the line whereas deviations from copula-based method are marked below the line. Each frequency deviation is an absolute value, computed as

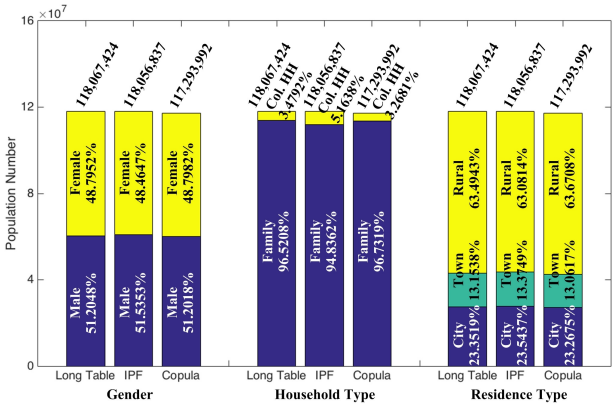
$$Dev = \left|Count_{LT} - Count_{syn}\right| \tag{10}$$

As can be seen, the data points from copula-based method are more clustered near zero lines in general, both for male and female. Quantitatively, copula-based method gets

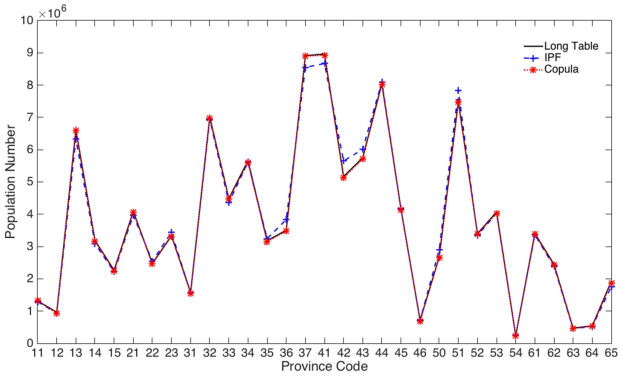
TABLE II

INPUT AND EVALUATION BENCHMARK CROSS-CLASSIFICATION TABLES

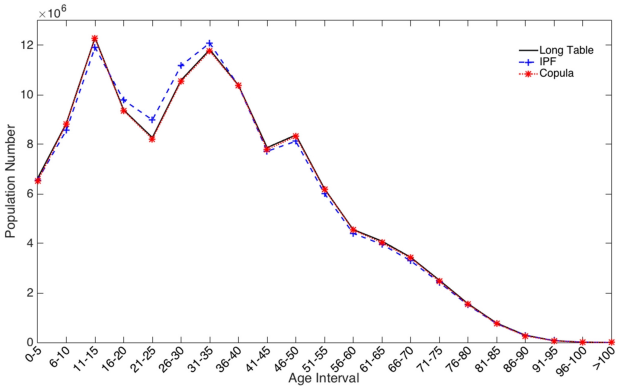
Input Marginal Distributions (Short Table)			Evaluation Benchmark (Long Table)		
No.	Attributes of Distribution	Table Codes	No.	Attributes of Distribution	Table Codes
1	$Gender \times ResidentialProvince \times ResidenceType \times HouseholdType$	$t0101a - t0101c$	1	$Gender \times ResidentialProvince \times ResidenceType \times HouseholdType$	$l0101a - l0101c$
2	$Gender \times ResidentialProvince \times ResidenceType \times AgeInterval$	$t0107a - t0107c$	2	$Gender \times ResidenceType \times AgeInterval$	$l0102a - l0102c$



(a) Gender, Household Type, Residence Type



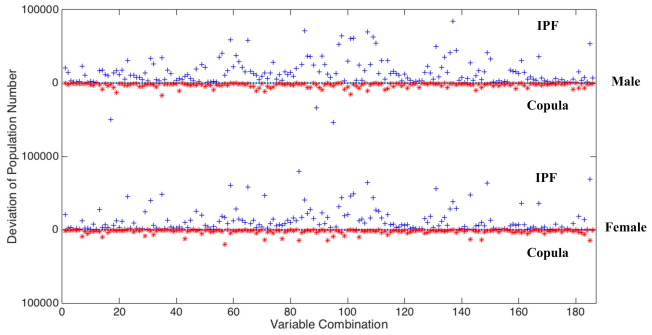
(b) Residential Province



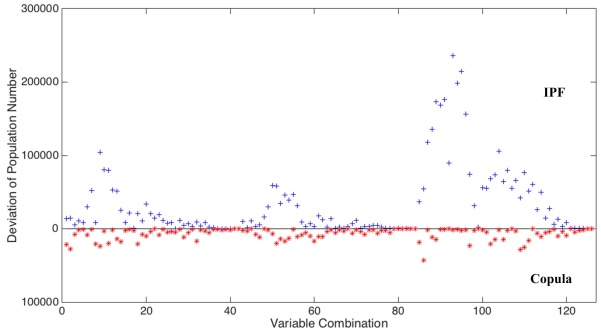
(c) Age Interval

Fig. 1. 1-dimensional Marginal Frequencies for 5 Attributes.

a 13.20% average relative error while IPF receives 50.03%. Also, similar deviation computation is applied to the second Long Table, which led to sub-figure (b). In this sub-figure, error differences seem more obvious. But the average relative errors are 8.71% (copula) and 6.78% (IPF), which is quite surprised. This indicates IPF may bring large deviations in some attribute combinations. In contrast, copula-based method performs much more stably. In summary, it can be concluded that our copula-based method is able to recover partial joint population structure better than IPF.



(a) $Gender \times Res.Province \times Res.Type \times HHType$



(b) $Gender \times Res.Type \times AgeInterval$

Fig. 2. Partial Joint Frequency Distributions for Long Tables.

For further evaluation of the partial joint distributions, we consider FreemanTukey statistic (FT^2)[22]. The FT^2 indicator is calculated as

$$FT^2 = 4 \cdot \sum_k \sum_i (\sqrt{O_{ki}} - \sqrt{E_{ki}})^2) \qquad (11)$$

where O_{ki} is the generated count corresponds to the i-th cell of the k-th table; E_{ki} is the given (known) count for the i-th cell of the k-th table. In our evaluation, there are only two Long Tables for benchmark, thus $k = 1, 2$. After computation, the total FT^2 values are 1,773,721 (IPF) and 189,386 (copula). Clearly, the error of our method only

accounts for 10.68% of IPF. This also proves that the copula-based method brings much lower variations in partial joint distributions.

V. Conclusions and Discussions

Copula function is widely used in economics and statistics. This paper introduces discrete copula into population synthesis as an alternative for traditional sample-based algorithm. Bootstrapping is adopted to calculate empirical copula function so that the dependence among different individual attributes can be estimated. The proposed method is tested in Chinese national population. Experiment results indicate that our new method can achieve the same level of accuracy as IPF in 1-dimensional marginal distributions, and performs significantly better in partial joint distributions.

The bootstrapping algorithm in Section III gives a re-sampling strategy by replicating one person at a time. This may cause some problems however, when the scale of synthetic population is large. On the one hand, error decrease brought by a suitable person may be 'overwhelmed' due to the computational truncation. In such case, the error will always stay at the same level and prevent the algorithm from convergence. Therefore, the computation stops only when its iteration number reaches the maximum. Unfortunately, the joint distribution obtained in this way may probably not be consistent with marginal constraints and will lead to an unreliable synthetic population. On the other hand, even if the error decrease is detected and the computation is terminated by the smaller error than threshold, the low efficient convergent process may take a long time. Several approaches can be considered to solve these potential problems. First, instead of updating one person each time, we can test a group of people. This operation will increase the error deviation so that the algorithm is able to select suitable persons in a right direction. Second, parallel computing is suitable for acceleration by partitioning the target population into smaller subsets. Each subset can be synthesized by a separate thread and then merged to get the final result. Another potential problem is the minor group sampling. When the bootstrapping operates on original samples, the type of individuals with low frequencies is rarely extracted. It is very likely to cause the lack of those minor groups in the final synthetic population. To deal with this problem, hierarchical sampling techniques may be introduced to keep the heterogeneity of original samples. In summary, the proposed method needs to be further improved. Also, experiments in this paper only consider two partial joint distributions, which seems not enough. Thus in the near future, more various applications need to be introduced to further test the accuracy of the method as well as its computational performance.

Acknowledgment

The authors would like to thank Prof. Angela Yu in Department of Cognitive Science, University of California San Diego, for her support of this work.

References

[1] H. Zhao, S. Tang and Y. Lv. Generating Artificial Population for Traffic Microsimulation. IEEE Intelligent Transportation Systems Magazine, 2009, 1(3): 22–28.

[2] Y. Ou, S. Tang and F.-Y. Wang. Computational Experiments for Studying Impacts of Land Use on Traffic Systems. IEEE International Conference on Intelligent Transportation Systems (ITSC), St. Louis, USA, 2010:1813–1818.

[3] P. Ye and D. Wen. A Study of Destination Selection Model Based on Link Flows. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(1):428–437.

[4] P. Ye, X. Wang, C. Chen, Y. Lin and F. Wang. Hybrid Agent Modeling in Population Simulation: Current Approaches and Future Directions. Journal of Artificial Societies and Social Simulation, 2016, 19(1): 12.

[5] A. G. Wilson and C. E. Pownall. A New Representation of The Urban System for Modeling And for The Study of Micro-Level Interdependence. Area, 1976: 246–254. URL: http://www.jstor.org/stable/20001134.

[6] P. Williamson, M. Birkin and P. Rees. The Estimation of Population Microdata by Using Data From Small Area Statistics And Samples of Anonymised Records. Environment and Planning, 1998, 30: 785–816.

[7] F. Gargiulo, S. Ternes, S. Huet and G. Deffuant. An Iterative Approach for Generating Statistically Realistic Populations of Households. PloS one, 2010, 5(1): e8828.

[8] J. Barthelemy and P. L. Toint. Synthetic Population Generation Without A Sample. Transportation Science, 2013, 47(2): 266–279.

[9] B. Farooq, M. Bierlaire, R. Hurtubia and G. Flotterod. Simulation Based Population Synthesis. Transportation Research Part B: Methodological, 2013, 58: 243–263.

[10] M. Lenormand and G. Deffuan. Generating a Synthetic Population of Individuals in Households: Sample-Free Vs Sample-Based Methods. Journal of Artificial Societies and Social Simulation, 2013, 16 (4): 12. URL: http://jasss.soc.surrey.ac.uk/16/4/12.html.

[11] P. Ye, X. Hu, Y. Yuan and F.-Y. Wang. Population Synthesis Based on Joint Distribution Inference Without Disaggregate Samples. Journal of Artificial Societies and Social Simulation, 2017, 20 (4): 16. URL: http://jasss.soc.surrey.ac.uk/20/4/16.html.

[12] P. Salvini and E. J. Miller. ILUTE: An Operational Prototype of a Comprehensive Microsimulation Model of Urban Systems. Networks and Spatial Economics, 2005, 5(2): 217–234.

[13] D. R. Pritchard. Synthesizing Agents and Relationships for Land Use/Transportation Modeling. Master's Thesis, Department of Civil Engineering, University of Toronto, 2008.

[14] L. Smith, R. Beckman and K. Baggerly. TRANSIMS: TRansportation ANalysis and SIMulation System. Los Alamos National Laboratory Report, LA-UR-95-1641, 1995.

[15] A. Sklar. Fonctions de repartition a n dimensions et leurs marges. Publications de l'Institut de Statistique de L'Universite de Paris, 1959, 8: 229–231.

[16] I. Zezula. On Multivariate Gaussian Copulas. Journal of Statistical Planning and Inference, 2009, 139(11): 3942–3946.

[17] A. Nikoloulopoulos and D. Karlis. Multivariate Logit Copula Model With An Application To Dental Data. Statistics in Medicine, 2008, 27: 6393–6406.

[18] J. Neslehova. On Rank Correlation Measures For Non-Continuous Random Variables. Journal of Multivariate Analysis, 2007, 98: 544–567.

[19] R. Schefzik. Multivariate Discrete Copulas, with Applications in Probabilistic Weather Forecasting. Publications de l'Institut de Statistique de l'Universite de Paris, 2015, 59(1-2): 87–116. URL: https://arxiv.org/abs/1512.05629.

[20] O. P. Faugeras. Probabilistic Constructions of Discrete Copulas. Working paper, 2012. URL: https://hal.archives-ouvertes.fr/hal-00751393/file/copuladiscrete-submitted12-11-12.pdf.

[21] National Bureau of Statistics of the People's Republic of China. The 5-th National Census Data. URL: http://www.stats.gov.cn/tjsj/pcsj/rkpc/5rp/index.htm.

[22] J. Ryan, H. Maoh and P. Kanaroglou. Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms. Geographical Analysis, 2009, 41(2): 181–203.