SynC: A Unified Framework for Generating Synthetic Population with Gaussian Copula

Colin Wan[1], Zheng Li[23], and Yue Zhao[4]

[1] Department of Statistical Sciences,
University of Toronto, Toronto, ON, Canada
colin.wan@mail.utoronto.ca
[2] Northeastern University - Toronto Campus,
Toronto, ON, Canada
[3] Arima Inc., Toronto, ON, Canada
winston@arimadata.com
[4] Department of Computer Science,
University of Toronto, Toronto, ON, Canada
yuezhao@cs.toronto.edu

Abstract. Synthetic population generation is the process of combining multiple socioeonomic and demographic datasets from various sources and at different granularity, and downscaling them to an individual level. Although it is a fundamental step for many data science tasks, an efficient and standard framework is absent. In this study, we propose a multi-stage framework called **SynC** (**Syn**thetic Population via Gaussian **C**opula) to fill the gap. SynC first removes potential outliers in the data and then fits the filtered data with a Gaussian copula model to correctly capture dependencies and marginal distributions of sampled survey data. Finally, SynC leverages neural networks to merge datasets into one and then scales them accordingly to match the marginal constraints. We make four key contributions in this work: 1) propose a novel framework for generating individual level data from aggregated data sources by combining state-of-the-art machine learning and statistical techniques, 2) design a metric for validating the accuracy of generated data when the ground truth is hard to obtain, 3) demonstrate its effectiveness with the Canada National Census data and presenting two real-world use cases where datasets of this nature can be leveraged by businesses, and 4) release an easy-to-use framework implementation for reproducibility.

Keywords: synthetic data · microsimulation · copula · Canadian census · outlier removal · neural networks · data science.

1 Introduction

Synthetic population is used to combine socioeconomic and demographic data from multiple sources, such as census and market research, and downscale them to individual level. Often for privacy reasons, such datasets are released at aggregated regional levels (e.g., only averages or percentages are released for a region

with multiple residents). However, practitioners often find individual level data far more appealing, as aggregated data lack information such as variances and distributions of residents within that region. For the downscaled synthetic population to be useful, it needs to be realistic and consistent. The first condition means that simulated data should mimic realistic distributions and correlations of the true population as closely as possible (e.g., if wealthier neighborhoods spend more money on vacations, a similar pattern needs to also be reflected on simulated individuals), and the second condition implies that when we aggregate downscaled samples, the results need to be consistent with the original data. A more rigorous and detailed explanation can be found in Section 3.

Recently, there is an increasing demand and interest in micro-simulated and synthesized population data, as it is useful in decision-making tasks such as travel demand estimation [2] and agent-based models [4]. A well-generated population, formed by a set of simulated individuals each described by attributes such as age, sex, education, and income, contains hidden dependency within, and reflects the diversity of different individuals and unique characteristics of the geographic area [9]. Due to privacy regulations such as the General Data Protection Regulation (GDPR), organizations across the world are forbidden to release personal level data. Therefore the most commonly available data, usually accessible through census surveys and syndicated market research campaigns, must be in an anonymous and aggregated fashion. Synthetic population generation, therefore, is a crucial step for practitioners to reconstruct the lost information in order to maximize their model performance.

Common techniques for synthetic population generation are synthetic reconstruction (SR) [2] and combinatorial optimization (CO) [7,18]. SR methods involve information from two separate data sources: regional level joint distribution data from census tables, and surveyed data representing the true population of interest. The algorithm uses techniques such as iterative proportional fitting (IPF) [6], or matrix ranking, to generate the desired simulation upon convergence. CO methods, although less popular, are less strict on the requirement of data. The principle behind CO methods is to first segment the population into distinct, mutually exclusive groups, representing smaller geographic regions, for which the marginal distributions for a set of desired attributes are retrievable. The algorithm then draws from the known microdata to form a combination that satisfies the marginal constraint while optimizing a fit function.

Let alone the underlying assumptions and requirements on the dataset, both approaches have limitations which usually cannot be easily resolved. For IPF to converge to a reasonable approximation, SR requires a surveyed sample to represent the true distribution of the population, which is often hard to obtain accurately. On the other hand, CO methods, although giving promising results, face severe challenges on the optimization side. Typical resolutions, such as simulated annealing [3,11], incur extremely high computational costs.

To resolve these limitations, we propose a new framework called **SynC** (**Syn**thetic Population with Gaussian **C**opula) to simulate microdata by sampling features in batches. The concept is motivated by [9] and [10], which are

purely based on copula and distribution fitting. The rationale behind our framework is that features can be segmented into distinct batches based on their correlations, which reduces the high dimensional problem into several sub-problems in lower dimensions. Feature dependency in high dimensions is hard to evaluate via common methods due to its complexity and computation requirements. In this study, we use Gaussian copula functions, a family of multivariate distributions that can describe dependencies among random variables, to determine the underlying dependency structure among features.

To improve generation stability, SynC framework first removes potential outliers from the original data. The features of the filtered data are then segmented into batches, and distinct copula functions are fitted on each batch along with a set of core features. Then a predictive model is trained for each batch of features to generate realistic individuals with the given information. To finalize, the framework scales the data contained in each feature to satisfy the given aggregated information. In this study, we make the following technical contributions:

1. We propose a novel combination framework which, to the best of our knowledge, is the first published effort to combine state-of-the-art machine learning and statistical instruments (e.g., outlier detection, Gaussian copula, and neural network) to synthesize population data.
2. We design a metric for validating the accuracy of generated data in which the ground truth is hard to obtain.
3. We empirically demonstrate SynC's effectiveness on Canada National Census data. Additionally, we provide two real-world use cases to show how the generated datasets could be leveraged by businesses to gain extra insights.
4. To foster reproducibility and transparency, all code, figures and results are openly shared[1]. The implementation is also readily accessible to be adapted for similar use cases.

2 Related Works

2.1 Synthetic Reconstruction

Synthetic reconstruction (SR) is the most commonly used technique to generate synthetic data. Given the marginal data of the interested features and the surveyed sample data representing the true population, this approach attempts to reconstruct the desired distribution from the survey data while constrained by the marginal. First, a joint distribution, whose marginal distributions match the given data and preserves the correlation in the survey data, is estimated by an iterative process. Then a sample of individuals is selected from the joint distribution to form a synthetic population.

Different implementations may be applied to different situations depending on the context. Typical iterative procedures used to estimate the joint distribution are iterative proportional fitting (IPF) and matrix ranking.

[1] `https://github.com/winstonll/SynC`

The goal of IPF algorithm is to fit an n-dimensional contingency table base on sampled data and fixed marginal distributions. The inner cells are scaled one dimension at a time, such that they match the given marginal distribution. The process is repeated until the entries converge.

IPF has lots of advantages such as maximizing entropy, minimizing discrimination information [8], and resulting in maximum likelihood estimator of the true contingency table [12]. However, the algorithm requires the initial survey data to represent the true population well enough for the approximated data to be reasonably close. Moreover, IPF can only estimate categorical variables in a contingency table—continuous variables cannot be sampled using this method.

2.2 Combinatorial Optimization

Given a subset of individual level data, with features of interest, the fundamental idea behind combinatorial optimization (CO) is to find the best combination of individuals that satisfy the marginal distributions while optimizing a designed objective function[1]. For instance, the algorithm can be initialized with a random subset of individuals. Based on the fitness function, one individual is swapped with another in the pool if such action increases the fitness of the group. The process iterates until convergence, or a certain threshold of fitness is reached. Compared to SR approaches, CO methods can reach more accurate approximations. However, this often comes at the expense of exponential growth of computational power [19].

2.3 Copula-Based Population Generation

Copula is a statistical distribution that can be used to understand the dependency structure among different distributions (refer to Section 3 for details). It has been used in microsimulation tasks recently. Kao et al. simulate household samples by finding joint distributions through copula models [10]. However, downscaling is not possible, and the sampled data stay at the same level of granularity as the input. Jeong et al. discuss an enhanced version of IPF where the fitting process is done via copula functions [9]. Similar to IPF, this algorithm relies on the integrity of the input data, which, as discussed before, can be a problem for practitioners. Our method is less restrictive on the initial condition of the input dataset as it only requires an aggregated level data. Therefore the framework is more robust compared to previous approaches.

2.4 Outlier Removal in Microsimulation

Outliers or anomalies are the deviant samples from the general data distributions that may be a result of recording mistakes, incorrect responses (intentional or unintentional), or tabulation errors. The presence of outliers often leads to unpredictable results [22]. Microsimulation tasks are sensitive to outliers presented in the data [15], and therefore their removal is necessary to reduce the risk.

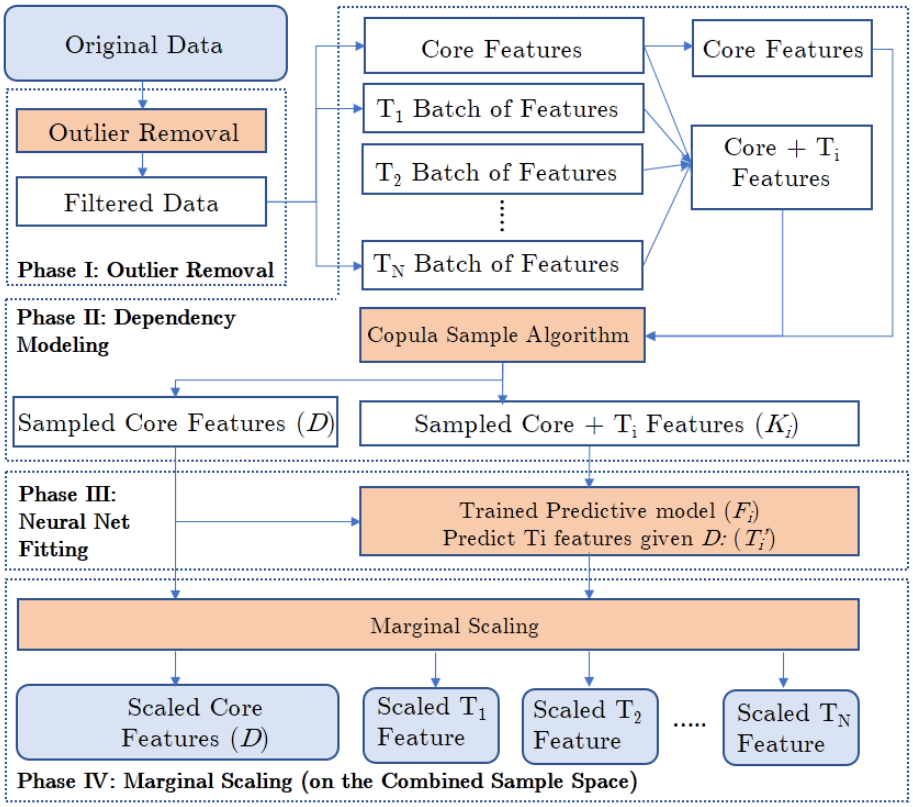

Fig. 1. Flowchart of the SynC framework

Many unsupervised outlier detection algorithms have been proposed recently by leveraging the latest deep learning techniques [23] and outlier ensembles [21]. In this study, outlier removal is included as part of the proposed framework, and its effectiveness is demonstrated through comparative studies.

3 Proposed Framework

The proposed framework is designed by ensuring the generated multivariate data samples of individuals satisfy the following three criteria [13]:

(i) Individual components satisfy the marginal distributions from the original data.
(ii) Correlations among features need to be logical and directionally consistent with the input data.
(iii) Aggregating simulated data must agree with the original postal code level data.

To illustrate the importance of these criteria, suppose that the variable of interest is income. The first criterion implies that the assumed underlying distribution of sampled individual incomes must align with the true distribution of income (which is commonly modelled by lognormal distributions). The second criterion implies that sampled individual incomes should generally be reflected by other features that correlate highly with income, such as spendings on vacations, luxury goods, and financial investments. Finally, the third criterion implies that if we aggregate incomes (or any variable) of all sampled individuals in a particular region, the result should match the average income from the original data. Our philosophy is to simulate individuals by generating features in batches base on core characteristics and scaling the result to maintain hidden dependency and marginal constraints.

As illustrated in Fig. 1, the proposed framework contains four phases that generated synthetic microdata with only aggregated level data described above. In **Outlier Removal Phase**, aggregated data are passed through outlier detection algorithms so that outliers are eliminated to avoid skewness. In **Dependency Modeling Phase**, a copula model is fitted to the core (systematically selected) variables of the filtered data. Then, in **Neural Network Fitting Phase** we train a predictive model using core variables as input and the remaining features as output. Finally, to match the marginal constraints from the input data, simulated data is scaled in the **Marginal Scaling Phase** to produce the final output.

3.1 Phase I: Outlier Removal

Outlier detection refers to the identification of rare items, events or observations which differ from the general distribution of a population [22]. Applying outlier detection to socioeconomical data before any analysis is particularly important, as a radical individual with extreme behaviours could easily skew the regional averages. As an unsupervised task, outlier removal algorithms are often selected by underlying assumptions of the data. When the ground truth is absent, we recommend to use the emerging outlier ensembling methods like SELECT [16] and LSCP [21]. Otherwise, supervised outlier ensemble frameworks such as XGBOD [20] could be effective in removing anomalies.

3.2 Phase II: Dependency Modeling

We propose using the copula model to address criteria i) and ii) since copulas, with its wide industrial applications, have always been a popular multivariate modeling methodology especially when the underlying dependency is essential. First introduced by Sklar in 1959 [17], a copula is a multivariate probability distribution where the marginal probability distribution of each variable is uniform. Let $X = (x_1, x_2...x_D)$ be a random vector in R^D, and the marginal cumulative distribution function be $P_i(x) = Pr[x_i < x]$, define U_i such that

$$U = (u_1, u_2, ..., u_D) = (P_1(x_1), P_2(x_2), ..., P_D(x_D)) \tag{1}$$

Algorithm 1: Gaussian copula sampling

Data: Postal Code Level Data
Result: Simulated Individual Level Data
initialization;
M = number of postal codes in a given country or region
D = dimension of input data size
Γ = correlation matrix of input data
for *j in* 1...M **do**
  Draw $Z_j = Z_{1,j}, \cdots, Z_{D,j} \sim N(0, \Gamma)$, where $N(\mu, \Gamma)$ denotes a d-dimensional Normal distribution with mean μ and correlation matrix Γ
  for *i in* 1...D **do**
    Set $u_{i,j} = \Phi(Z_{i,j})$ Set $y_{i,j} = F_{i,j}^{-1}(u_{i,j})$ where $F_{i,j}^{-1}$ is the inverse CDF of the marginal distribution of the i^{th} component in the j^{th} postal region. This implies that $Y_{i,j}$ follows the desired distribution
  end
end

A copula of the random vector X is defined as the joint CDF of a random uniform vector U:

$$C(u'_1, u'_2, ...u'_D) = Pr(u_1 < u'_1, u_2 < u'_2, ..., u_D < u'_D) \tag{2}$$

In other words, we can describe the joint distribution of a random vector X using its marginal distributions and some copula function. Additionally, Sklar's Theorem states that for any set of random variables with continuous CDFs, there exists a unique copula as described above. It allows us to isolate the modeling of marginal distributions from their underlying dependencies. Sampling from copulas is also widely used by empiricists to produced deserved multivariate samples based on a given correlation matrix and the desired marginal distributions of each of the components. Nelsen [14] outlines a simpler version of Algorithm 1 for bivariate sampling, which can easily be generalized to multivariate cases.

In order to properly specify F_i^{-1}, we make a reasonable assumption that the population of each postal code region is significant and diverse enough to ensure constant variance across the nation. We also assume that if the desired output is a positive continuous variable (such as income), the marginal distribution, $F_{Y_{i,j}}(y)$, follows a lognormal distribution with mean $\mu_{i,j}$ and standard deviation $\sigma_{i,j}$. In the case of categorical variables, $F_{Y_{i,j}}(y)$ follows a beta distribution with parameters α and β such that $\frac{\alpha}{\alpha+\beta} = \mu_{i,j}$ and $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = {\sigma_{i,j}}^2$. We further specify $\mu_{i,j}$ and $\sigma_{i,j}$ in the following way to represent postal code level mean and standard deviation for variable i.

$$\mu_{i,j} = \textit{mean of feature } i \textit{ in postal code } j \tag{3}$$

$$\sigma_{i,j} = \sigma_i * \sqrt{M} * \sqrt{p_j} \tag{4}$$

Algorithm 2: Batch sampling via Gaussian copula

Data: Postal Code Level Data
Result: Simulated Individual Level Data
initialization;
X = cleaned census table
N = total number of batches of features
S = predefined set of core features
T_i = ith batch of features
D = Initial sampled individual with only core features
for i *in* $1...N$ **do**
  $D' = X[S\&T_i]$
  K_i = Sampled data by applying *Algorithm* 1 to D'
  $F(T_i \mid S = s)$ = an approximated distribution function trained on K_i
  $T_i' = argmax_Y\ F(Y \mid S = D)$
  $D = D + T_i'$
end

Note that μ_{ij} is observed from the data as we are given postal code averages for each variable, and σ_i is the national level standard deviation for variable i and p_j is the population of postal code j.

Algorithm 1 incorporates the above assumptions with Gaussian copula. Data sampled using this algorithm satisfy criteria i) and ii).

3.3 Phase III: Neural Network Fitting

In theory, one could construct a complex enough correlation matrix or deep learning model to include all variables, and then fit a giant copula model using Algorithm 1. However, two reasons make this approach nearly impossible: 1) the curse of dimensionality makes it difficult to find computational resources to process all data at once, and 2) parts of the aggregated data may be updated, and it would be inefficient to train a new model each time a few variables are changed. In this section, we introduce an alternative method, Algorithm 2, to resolve this issue with minimal computation power required.

First, from the cleaned Census Table, X, select a set of core variables, S, containing variables such as age, gender, ethnicity, religion, family composition. We believe that a core set of features, such as age, education, ethnicity and family composition, universally affects all other behaviours of an individual. Therefore, with our domain expertise, we manually select variables to be a part of this core set. Next, we sample from the core variables using Algorithm 1 and only accept a sampled individual if it contributes towards matching the marginal distribution. For instance, if the number of 20-24 years old male in the current sampled pool already exceeded the marginal count, we reject any further sampled individual who is a 20-24 years old male. The resulting data set, D, should satisfy the first two criteria mentioned in Section 3.2.

Secondly, divide the non-core variables into a total of N batches, $T_1 ... T_N$, based on their correlations. Practically, we aim to have batches of around 200 features. Features in the same batch should be highly correlated while features among different batches should only be weakly correlated. For example, vehicle make and price are highly correlated and hence grouped into one batch, along with other associated features such as vehicle insurance cost and mileage. However, features such as coffee consumption or health care spending would have limited or no influence with the vehicle variables, condition on the core features (which in this case are the confounding variables), and as a result are grouped into different batches. Features are then merged with the core variables one batch at a time using Algorithm 1. The resulting sampled data set, K_i, contains features from S and T_i.

However, a problem immediately presents itself: the sampled individuals in K_i do not match D which has been meticulously sampled to satisfy the marginal constraints. Furthermore, individuals sampled in each K_i do not match the sampled individuals in K_j (for $j \neq i$). To overcome this problem, we can train a neural network on K_i to approximate the distribution of $P(T_i \mid S = s)$, and use the model to predict the values of features K_i given their values of the core features from D.

Finally, we merge the predicted values with the original data set D and iterate the process until all features are covered.

3.4 Phase IV: Marginal Scaling

The final step is to address criterion iii), which is to ensure the aggregated values of our sampled set agree with the original data.

For a continuous variable, Y, we have sampled a value in Phase III from distribution $F_{Y_{j,k}}$ for individual k in postal region j. We need to ensure the average of the simulated population, μ', agrees with the given average, μ. To achieve this, we multiply each sampled value by the scaling factor $\frac{\mu}{\mu'}$. The adjusted set of Y_k in postal region j preserves the diversity of the region from the original set and agrees with the given average.

For a categorical feature, X, with n classes, we first note that the predicted values from Phase III for individual k in postal region j, represented by $\mathbf{p_{j,k}} = \{p^i_{j,k}\}^n_{i=0}$, is a probability distribution. Hence it is natural to assume $X \sim Multi(1, \mathbf{p_{j,k}})$. To determine the exact class of individual k, we generate a random sample from the distribution. After initial sampling, the percentage of each category may not match the marginal constraint. To resolve this, SynC first randomly removes individuals from the categories that are over-sampled until their marginal constraints are satisfied. The removed individuals are then resampled using the normalized probability of the under-sampled categories. It is noted that the above processes are iterated until the desired result is achieved.

Table 1. An excerpt of three variables from Canadian census data

Postal	# Population	Avg Age	% with mortgage	% Speaks two languages
M5S3G2	467	35.1	0.32	0.69
M4Y1G3	269	37.2	0.35	0.67
L5M6V9	41	49.1	0.67	0.43

4 Results and Discussion

4.1 Introducing Question of Interest

To practically test the previously outlined framework, we generate a synthesized data using the latest Canadian National Census Data, which is collected by Statistics Canada and compiled once every five years. Our raw dataset is aggregated at the postal code level and made available to the general public. There are 793,815 postal codes across Canada (in the format A1A1A1, where A is a letter and 1 is a digit), with an average of 47 residents per postal code. The dataset contains more than 4,000 variables ranging from demographics, spending habits, financial assets, and psychographics. Table 1 illustrates a subset of with 3 postal codes and 4 variables.

All variables except postal codes are expressed as either a percentage, a count, or an average (median is also given in some cases). Survey questions with binary responses are aggregated as a percentage of the area level total population (e.g., in postal code M5S3G2, 32% of the 467 residents own a mortgage), and numeric responses are aggregated as an average (e.g., the average age of the 41 residents in L5M6V9 is 49.1 years old).

This section presents the results of applying SynC to downscale Canadian census data to individual level. In order to evaluate the performance of SynC, we survey 30 individuals and recorded their demographic information. Our focus is on the postal regions that contains surveyed individuals.

4.2 Proposed Evaluation Metric

This particular experiment is hard to assess due to lack of true individual level data to benchmark. After all, if we had individual level data then a generation of synthetic population would not be needed in the first place. Therefore it is not possible for us to accurately measure the performance of our framework over the whole Canadian population. Instead, our proposed metric focuses on the assessing the practicality of the simulated populations relative to a locally surveyed sample. The evaluation function requires 1) a set of surveyed response representing real samples from the population and 2) simulated populations from the postal region of the surveyed individuals. For a surveyed dataset containing K demographic features from T individuals across M postal regions (each contains N_m surveyed individuals), M simulated populations each with T_m (the population of mth postal region) individuals are generated using the proposed

framework. The evaluation function is defined as the following:

$$\frac{1}{M}\Sigma_{m=1}^{M}\frac{1}{N_m}\Sigma_{i=1}^{N_m}\frac{1}{K}\Sigma_{k=1}^{K}\mathbb{1}\{x_{m,i}^{k}=y_{m,i}^{k}\} \tag{5}$$

where $x_{m,i}^{k}$ represents the kth feature of ith surveyed individual in mth postal region, and

$$\{y_{m,i}\}_1^{N_m}=\underset{\{y_{m,i}\}_1^{N_m}}{\arg\max}\,\Sigma_{j=1}^{N_m}\Sigma_{k=1}^{K}\mathbb{1}\{x_{M,j}^{k}=y_{M,j}^{k}\}. \tag{6}$$

where each $y_{m,i}$ is a simulated individual in mth postal region. In other words, the above evaluation function measures the average similarity across all surveyed postal regions between the person from true population and the set of simulated individuals in each region that resembles them the most.

4.3 Results of Experiment

We collect responses, as well as their residential postal codes, from 30 individuals in Toronto, Canada. For simplicity, the surveyed questions only contain five core features (Age, Sex, Ethnicity, Education, Income), seven additional personal features (Immigration Status, Marital Status, Family Size, Profession etc.) and four spending behaviour feature (Favourite Store for Cloth/Food/Grocery/Furniture). The five demographics features are selected as core variables and sampled using Phase II of the framework. The other features can be grouped into two batches, as spending behavior features clearly separates itself from the rest. Using Phase III of the framework, we generate a pseudo-population to train a 2-layered neural network. Table 2 shows an excerpt of the sampled pseudo-population with selected features for one of the regions after it has been scaled to fit the marginal constraints.

By quickly iterating over the sampled population, we identify the set of individuals that resemble the surveyed individuals the most. Table 3 is a comparison between a few surveyed and simulated individuals. Our performance score measured by the proposed metric, equation (5), is 82%. On a closer look it's easy to see that the generated individuals are actually more realistic than what the score can reflect. The metric function uses an indicator function for each feature to evaluate the accuracy, which does not capture the potential distance among categories (e.g. $Income_{50k-59k}$ should be "closer" to $Income_{60k-69k}$ compare to $Income_{150k+}$). Take the third individual for example: his ethnicity is predicted as Filipino when he is in fact Chinese. Although incorrect, the two countries are in close proximity with each other, and if the generated dataset is used for other analysis work, this difference only results in an immaterial error as in some real-world applications, both countries may be considered the same category (Asia or East Asia).

4.4 Real Applications

From a practitioner's perspective, synthetic population allows data-driven decision makers to make well informed decisions based on limited resources. In this section, we present two examples of how companies can leverage this technique.

One of the most important challenges companies face is how to grow their customer base. Most companies would have profiles of their current customers, but lack information on where similar targets might be. For example, an automotive company has historical purchase data that would suggest that their typical customers are highly educated young working professionals, single parents with three or more children, and retirees with low income, but may not know where are the best places to advertise based on these customer profiles. While traditional census data may provide some insights, it is difficult to conclude the distribution of demographics variables based on regional averages. Knowing that a neighborhood has a larger percentage of senior residents and a low average income does not necessarily conclude that seniors in this community earn less, as it is also possible that the remaining residents have meager income while the senior citizens are living comfortably. Synthetic population allows companies to understand the distributions of customer traits and hence allow them to make better decisions.

While being a powerful customer discovery tool, synthetic population can also be used to enhance a company's internal data. For example, a telecommunication company would like to know which customers are likely to purchase a new phone in the next 12 months, however, they may have limited data about their customers. With synthetic population, the telecommunication company can match their existing database of customer information with the synthetic population using non-personally-identifiable keys such as postal code, age, gender and ethnicity, and thus increasing the number of features that can be included in the telecommunication company's propensity models.

5 Conclusion and Future Directions

In this work, we propose a novel framework, SynC, for generating individual level data from aggregated data sources, using state-of-the-art machine learning and statistical methods. Additionally, we design a metric for validating the accuracy of generated data in which substantial fieldwork is required to collect the ground truth to validate the outcome using traditional methods. To show the proposed framework's effectiveness and boost reproducibility, we provide the code and data on the Canada National Census example described in Section 4. Finally, we present two real-world business use cases where datasets of this nature can be leveraged by businesses.

As a first attempt to formalize the problem, we see three areas where future works can improve upon. First of all, our method relies on Gaussian copulas and this can be further extended by leveraging other families of copulas to better model the underlying dependency structures. Secondly, we use beta and log-normal distributions to approximate marginal distributions for categorical and

Table 2. An expert of simulated data for one postal region

Postal	Sex	Age	Ethnicity	Immigration Status	Education	Profession	Marital Status	Family Size	Income	Favorite Store
M2M4L9	F	19	Latin	Immigrants	No degree	Ed services	Married	5+	<$10k	Banana
M2M4L9	F	65+	Chinese	Immigrants	No degree	Food services	Widowed	3	$10k to $19k	Banana
M2M4L9	M	51	Korean	Immigrants	College	Waste management	Separated	1	<$10k	Old Navy
M2M4L9	F	65+	South Asian	Immigrants	No degree	Edu services	Widowed	3	$20k to $29k	Zara
M2M4L9	F	58	Chinese	Immigrants	Bachelor	Scientific/Technical	Married	4	$40k to $49k	Gap
M2M4L9	M	60	Korean	Immigrants	Master	Finance	Married	2	<$10k	H&M
M2M4L9	F	58	Other	Non-immigrants	College	Health care	Married	3	$50k to $59k	Hot Renfrew
M2M4L9	M	26	Chinese	Immigrants	HighSchool	Retail	Never married	3	<$10k	Old Navy
M2M4L9	M	<14	Mid Eastern	Immigrants	N/A	N/A	N/A	N/A	N/A	N/A
M2M4L9	F	21	Chinese	Immigrants	Bachelor	Retail	Never married	3	¡$10k	Gap
M2M4L9	M	51	Chinese	Immigrants	No degree	Other services	Never married	2	$10k to $19k	Gap

Table 3. An comparison between simulated samples and surveyed samples

Type	Postal	Age	Sex	Education	Ethnicity	Family Size	Immigration Status	Marital Status	Income	Profession	Favorite Store
Simulated	M2M4L9	47	F	Bachelor	Chinese	2	Immigrants	Married	$40k to $49k	Scientific/Technical	Gap
Surveyed	M2M4L9	49	F	Bachelor	Chinese	3	Immigrants	Married	$5k to $59k	Scientific/Technical	H&M
Simulated	M2M4L9	46	M	Bachelor	Chinese	2	Immigrants	Married	<10k	Scientific/Technical	Gap
Surveyed	M2M4L9	51	M	Bachelor	Chinese	3	Immigrants	Married	$5k to $59k	Scientific/Technical	Gap
Simulated	M4Y1G3	22	M	Bachelor	Filipino	3	Immigrants	Married	$70k to $79k	Scientific/Technical	Banana Rep
Surveyed	M4Y1G3	23	M	Bachelor	Chinese	2	Immigrants	Never married	$5k to $59k	Finance	Banana Rep
Simulated	M5G2R3	28	M	Bachelor	Chinese	3	Immigrants	Married	$80k to $89k	Wholesale trade	Zara
Surveyed	M5G2R3	22	M	Bachelor	Chinese	2	Immigrants	Never married	$5k to $59k	Finance	Zara
Simulated	M5G2R3	22	F	Bachelor	Chinese	1	Immigrants	Never Married	<$10k	Finance	Holt Renfrew
Surveyed	M5G2R3	21	F	Bachelor	Chinese	2	Immigrants	Never married	$5k to $59k	Finance	Holt Renfrew

continuous variables, respectively, and other families of distributions could be considered. In particular, the κ-generalized model [5] could be a better candidate for money related variables. Lastly, including more geographical features (such as city-level features, population, GDP) can potentially increase the accuracy of predicted individual level traits.

References

1. Barthelemy, J., Toint, P.L.: Synthetic population generation without a sample. Transportation Science **47**(2), 266–279 (2013)
2. Beckman, R.J., Baggerly, K.A., McKay, M.D.: Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice **30**(6), 415–429 (1996)
3. Černỳ, V.: Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. Journal of optimization theory and applications **45**(1), 41–51 (1985)
4. Cetin, N., Burri, A., Nagel, K.: A large-scale agent-based traffic microsimulation based on queue model. In: In proceedings of swiss transport research conference (strc), monte verita, ch. Citeseer (2003)
5. Clementi, F., Gallegati, M., Kaniadakis, G., Landini, S.: κ-generalized models of income and wealth distributions. Review Network Economics **225**, 1959–1984 (2016)
6. Deming, W.E., Stephan, F.F.: On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics **11**(4), 427–444 (1940)
7. Huang, Z., Williamson, P.: A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata. Department of Geography, University of Liverpool (2001)
8. Ireland, C.T., Kullback, S.: Contingency tables with given marginals. Biometrika **55**(1), 179–188 (1968)
9. Jeong, B., Lee, W., Kim, D.S., Shin, H.: Copula-based approach to synthetic population generation. PloS one **11**(8), e0159496 (2016)
10. Kao, S.C., Kim, H.K., Liu, C., Cui, X., Bhaduri, B.L.: Dependence-preserving approach to synthesizing household characteristics. Transportation Research Record **2302**(1), 192–200 (2012)
11. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. science **220**(4598), 671–680 (1983)
12. Little, R.J., Wu, M.M.: Models for contingency tables with known margins when target and sampled populations differ. Journal of the American Statistical Association **86**(413), 87–95 (1991)
13. Münnich, R., Schürle, J.: On the simulation of complex universes in the case of applying the german microcensus. DACSEIS research paper series No. 4 (2003)
14. Nelsen, R.B.: An introduction to copulas. Springer Science & Business Media (2007)
15. Passow, B.N., Elizondo, D., Chiclana, F., Witheridge, S., Goodyer, E.: Adapting traffic simulation for traffic management: A neural network approach. In: 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013). pp. 1402–1407. IEEE (2013)
16. Rayana, S., Akoglu, L.: Less is more: Building selective anomaly ensembles. ACM Transactions on Knowledge Discovery from Data (TKDD) **10**(4), 42 (2016)

17. Sklar, M.: Fonctions de repartition an dimensions et leurs marges. Publ. inst. statist. univ. Paris **8**, 229–231 (1959)
18. Voas, D., Williamson, P.: An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. International Journal of Population Geography **6**(5), 349–366 (2000)
19. Wong, W., Wang, X., Guo, Z.: Optimizing marker planning in apparel production using evolutionary strategies and neural networks. Optimizing decision making in the apparel supply chain using artificial intelligence (AI): form production to retail. Woodhead Publishing Series in Textiles pp. 106–131 (2013)
20. Zhao, Y., Hryniewicki, M.K.: XGBOD: improving supervised outlier detection with unsupervised representation learning. In: 2018 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2018)
21. Zhao, Y., Nasrullah, Z., Hryniewicki, M.K., Li, Z.: LSCP: Locally selective combination in parallel outlier ensembles. In: SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, Calgary, Canada (May 2019)
22. Zhao, Y., Nasrullah, Z., Li, Z.: PyOD: A python toolbox for scalable outlier detection. arXiv preprint arXiv:1901.01588 (2019)
23. Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: International Conference on Learning Representations (2018)