Synthetic population experiments with copula-based transferable models

Fabian Bastin[*], Cinzia Cirillo[†], Pascal Jutras-Dubé[‡]

Population synthesis refers to models that aim to create an artificial yet realistic population of observations from existing but limited disaggregated datasets. The information generated can be used for various purposes, including modeling, optimization, simulation, or, in general, generating new information for an application of interest [4]. In transportation research, these synthesizers have been extensively employed [19]. This capability is particularly valuable in agent-based models, such as transportation models based on micro-simulation, where understanding the spatial implications of policies is critical [14, 15].

A desirable property of any population synthesizer is to preserve the general characteristics of the population, which leads to different technical approaches. The most widely used, until recently, has probably been the Iterative Proportional Fitting (IPF) [5], popularized in transportation by Duguay et al. [6]. IPF selects households from the source sample while attempting to match given marginal totals, requiring a fitting stage and an allocation stage. In the fitting step, a contingency table is computed from the seed table (the source sample) and the marginal totals. In the allocation phase, households are randomly selected from the seed table to match the frequency given in the contingency table. However, the method presents several significant flaws, such as sampling zero issues Guo and Bhat [8].

Ye et al. [20] consider simultaneously fitting different types of agents and propose a heuristic approach called Iterative Proportional Updating (IPU), but they fail to accommodate the new synthetic information at multiple geographical resolutions simultaneously, leading to a loss of representativeness. Konduri et al. [12] extend their efforts by proposing an enhanced IPU algorithm that accounts for constraints at different levels of spatial resolution when generating a synthetic population. Farooq et al. [7] point out that fitting a contingency table to the available data may lead to errors if the information is incomplete or has been manipulated. They propose a Markov Chain Monte Carlo (MCMC) simulation-based approach (further explored, for instance, in Saadi et al. [17], Kukic et al. [13]) that uses partial views of the joint distribution of the real population obtained from the census to generate high-dimensional synthetic populations.

To properly reflect the properties of the population it aims to mimic, a synthetic population must share the same joint distribution of the variables it encapsulates. Various methods have been proposed to achieve this goal, but with a few notable exceptions [1], most rely on samples from a target population, such as census data or travel surveys, which can be costly to obtain. This often results in limited sample sizes, especially at smaller geographical scales. Additionally, some regions, such as those at the census tract level, may lack detailed data entirely, providing only marginal totals—which effectively represent the attribute distributions of these regions. Traditional methods like re-weighting estimate sampling weights from the attribute distributions in each district and simulate the population from the weighted samples [3, 16, 9]. However, re-weighting cannot produce attribute combinations not observed in the training samples but present in the actual population. More recent approaches, based on generative models, can generate new, out-of-sample attribute combinations due to their probabilistic nature but do not explicitly integrate the attribute distributions of the area under study [18, 2, 11]. Consequently, there is an increasing demand for population synthesis methods that can generate new samples while accurately matching the aggregate totals of the studied region.

[*]bastin@iro.umontreal.ca, Department of Computer Science and Operations Research, CIRRELT University of Montreal, Canada. https://orcid.org/0000-0003-1323-6787

[†]ccirillo@umd.edu, Department of Civil and Environmental Engineering, University of Maryland, USA. https://orcid.org/0000-0002-5167-0413

[‡]pjutrasd@purdue.edu, Department of Computer Science, Purdue University, 305 N. University Street, West Lafayette, IN 47907, USA. https://orcid.org/0009-0002-8706-4877

We consider the framework proposed by Jutras-Dubé et al. [10] capable of generating synthetic data for a target population by utilizing only known marginal totals, in combination with a sample from another population that exhibits similar structural relationships among variables. The approach integrates copula theory with machine learning (ML) generative modeling techniques to separate the learning of dependency structures from that of marginal distributions. This separation facilitates the framework's application across different populations with varying marginal distributions. A key advantage of this method is that it eliminates the need to select a specific copula family, allowing to choose the generative model that best fits the context of the population under study. It also gives the opportunity to fully leverage the capabilities of probabilistic generative models, which have been shown in recent literature to be highly effective in capturing complex dependencies between variables and generating diverse data [11].

Since a copula is a multivariate distribution on $[0,1]^d$, the first step is to cast the observations as vectors in $[0,1]^d$. This can be easily achieved by considering the empirical CDF (ECDF) of each feature X_i, $i = 1, \ldots, d$, defined as

$$\hat{F}_i(x) = \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbb{1}\left(x_i^{(j)} \leq x\right), \tag{1}$$

where $\mathbb{1}$ is the indicator function and $\left\{x_i^{(1)}, \ldots, x_i^{(m_i)}\right\}$ is the ordered set of distinct observations (i.e. $x_i^{(p)} < x_i^{(q)}$ if $p < q$) for the i-th marginal. However, as the marginal distribution functions of the source and target populations are typically discrete but with different ranges, we first construct the relaxed ECDF $\tilde{F}_i(\cdot)$ as the continuous piecewise linear function obtained by considering the linear interpolations between consecutive values $\hat{F}_i\left(x_i^{(k)}\right)$ and $\hat{F}_i\left(x_i^{(k+1)}\right)$, $k = 1, 2, \ldots, m_i - 1$. We then extend the copula C to the domain $[0,1]^d$ by setting $C\left(\tilde{F}_1(x_1), \ldots, \tilde{F}_d(x_d)\right)$ as the linear interpolation, component by component, from

$$C\left(\hat{F}_1\left(x_1^{(k_1)}\right), \ldots, \hat{F}_i\left(x_i^{(k_i)}\right), \ldots, \hat{F}_d\left(x_d^{(k_d)}\right)\right)$$

to

$$C\left(\tilde{F}_1\left(x_1^{(k_1)}\right), \ldots, \hat{F}_i\left(x_i^{(k_i+1)}\right), \ldots, \hat{F}_d\left(x_d^{(k_d)}\right)\right),$$

with $x_i^{(k_i)} \leq x_i \leq x_i^{(k_i+1)}$, $k = 1, 2, \ldots, m_i - 1$, $i = 1, \ldots, d$ (see [10] for more details). We then draw the number of desired observations by drawing from C and transforming the realizations in the required marginals using the pseudo-inverse distribution function, as summarized in Algorithm 1.

Algorithm 1: Synthetic population generation

Step 1 Normalize the source population data using the ECDFs $\hat{F}_i(\cdot)$, $i = 1, \ldots, d$.

Step 2 Train the model on the normalized data to learn a copula C.

Step 3 Generate a synthetic population of vectors in $[0,1]^d$ by sampling from C extended to $\tilde{F}_i(\cdot)$, $i = 1, \ldots, d$.

Step 4 Transform any generated vector $\boldsymbol{u} = (u_1, \ldots, u_d)$ in a vector $\boldsymbol{y}$ in the target population as

$$\boldsymbol{y} = \left((F_1^Y)^{-1}(u_1), \ldots, (F_d^Y)^{-1}(u_d)\right),$$

where $(F_i^Y)^{-1}(\cdot)$ is the pseudo-inverse distribution function of the i-th target marginal, defined as

$$(F_i^Y)^{-1}(u) = \min\left\{x_i^{(j)} : F_i^Y\left(x_i^{(j)}\right) \geq u\right\}.$$

Numerous choices can be made to select the model at Step 2, for instance Bayesian Networks, Variational Autoencoders, and Generative Adversarial Networks. We applied the proposed approach on data from the American Community Survey (ACS), and analyze the model's transferability across various geographical levels, including state, county, Public Use Micro Areas (PUMA), and census tract, as illustrated in Table 1, reporting standardized root mean square errors (SRMSE), where BN stands for Bayesian network, the best found model for this dataset.

ML models have the capacity to generate attribute combinations that are not observed in the population sample, particularly when the sample size is small and the probabilities of such combinations are low but not zero. Following

Method	SRMSE 1	SRMSE 2	SRMSE 3	SRMSE 4	SRMSE 5
Independent	0.3442	0.8723	1.6720	3.0777	5.9539
IPF	0.7020	1.5510	3.0708	6.2470	13.4378
BN	0.3334	0.7402	1.4119	2.6954	5.4334
BN Copula	0.0350	0.3221	0.8658	1.9979	4.5310

Table 1: Standardized root mean squared error (SRMSE) for the spatial transferability experiment from the county to PUMA levels

the methodology established by Kim and Bansal [11], who extensively discuss this desirable property, they propose computing the $F1$ score of the generated population with respect to the entire original population. A false positive is declared when an observation absent in the original population is generated, while a false negative is assigned to zero cells when the corresponding feature combination is present in the original population. However, when computing structural zeros and the $F1$ score, one faces the challenge of not having access to all feasible attribute combinations within the target population—a dataset we naturally do not possess. To overcome this limitation, we assume that our entire state sample encapsulates the complete population, as this enables us to estimate structural zeros by identifying combinations absent in this comprehensive sample, thus treating it as a complete representation of feasible combinations. The $F1$ score is then calculated within this framework to evaluate the balance between precision, which reflects the proportion of feasible synthetic data, and recall, which measures the coverage of these feasible combinations within the generated data, as illustrated in Table 2, taken from [10]. However, misclassification of sampling zeros as structural zeros can occur because rare but feasible combinations may be absent from the dataset due to their low occurrence probabilities. As a result, the number of structural zeros is typically overestimated, and the $F1$ score undervalued.

Method	Sampling Zeros	Structural Zeros	Precision	Recall	$F1$ Score
Independent	200	21289	0.2306	0.3148	0.2662
IPF	0	0	1.0000	0.0889	0.1633
BN	245	2367	0.7941	0.4505	0.5748
BN Copula	236	2871	0.7583	0.4446	0.5605

Table 2: Diversity and feasibility of the synthesised population at the state level

While encouraging, the $F1$ scores indeed remain low and are lower than those reported by Kim and Bansal [11] in their experiments. This was, however, expected due to the size of the original datasets, which do not capture all possible combinations, thereby limiting the validation of the proposed approach. To address this issue, we examine large synthetic datasets, both without and with model transfer, and demonstrate that knowledge of the true population distribution allows for obtaining $F1$ scores closer to one.

We also investigate the effect of synthetic population mis-specifications on a simple logit model. Specifically, we show that market shares are not properly recovered when the SRMSE is large, along with a significant number of misidentified zero cells, as seen in the case of IPF, where the multivariate distribution governing the population of interest is not properly captured. On the other hand, the proposed copula-based method mitigates these undesirable effects, particularly when considering model transfer, allowing for a better estimation of the market shares of the choice options.

References

[1] J. Barthelemy and P. L. Toint. Synthetic population generation without a sample. *Transportation Science*, 47(2): 266–279, 2013.

[2] S. S. Borysov, J. Rich, and F. C. Pereira. How to generate micro-agents? A deep generative modeling approach to population synthesis. *Transportation Research Part C*, 106:73–97, 2019.

[3] J. Castiglione, M. Bradley, and J. Gliebe. *Activity-Based Travel Demand Models: A Primer*. Transportation Research Board, Washington, D.C., 2014.

[4] K. Chapuis, P. Taillandier, and A. Drogoul. Generation of synthetic populations in social simulations: A review of methods and practices. *Journal of Artificial Societies and Social Simulation*, 25(2):6, 2022. ISSN 1460-7425.

[5] W. E. Deming and F. F. Stephan. 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.

[6] G. Duguay, W. Jung, and D. L. McFadden. SYNSAM: A methodology for synthesizing household transportation survey data. Working paper 7618, Institute of Transportation Studies, University of California, Berkeley, CA, USA, 1976.

[7] B. Farooq, M. Bierlaire, R. Hurtubia, and G. Flötteröd. Simulation based population synthesis. *Transportation Research Part B*, 58:243–263, 2013.

[8] J. Y. Guo and C. R. Bhat. Population synthesis for microsimulating travel behavior. *Transportation Research Record*, 2014:92–101, 2007.

[9] S. Hörl and M. Balac. Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. *Transportation Research Part C*, 130:103291, 2021.

[10] P. Jutras-Dubé, M. B. Al-Khasawneh, Z. Yang, J. Bas, F. Bastin, and C. Cirillo. Copula-based transferable models for synthetic population generation. *Transportation Research Part C*, 169:104830, 2024. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2024.104830.

[11] E.-J. Kim and P. Bansal. A deep generative model for feasible and diverse population synthesis. *Transportation Research Part C*, 148:104053, 2023.

[12] K. Konduri, D. You, V. Garikapati, and R. Pendyala. Enhanced synthetic population generator that accommodates control variables at multiple geographic resolutions. *Transportation Research Record*, 2563:40–50, 2016.

[13] M. Kukic, X. Li, and M. Bierlaire. One-step gibbs sampling for the generation of synthetic households. *Transportation Research Part C: Emerging Technologies*, 166:104770, 2024. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2024.104770.

[14] K. Müller and K. W. Axhausen. Population synthesis for microsimulation: State of the art. In *TRB 90th Annual Meeting Compendium of Papers DVD*, number 11-1789, Washington DC, USA, Jan. 2011. Transportation Research Board.

[15] D. R. Pritchard and E. J. Miller. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, 39(3):685–704, 2012.

[16] J. Rich. Large-scale spatial population synthesis for denmark. *European Transport Research Review*, 10(2), 2018. doi: 10.1186/s12544-018-0336-2. URL `https://doi.org/10.1186/s12544-018-0336-2`.

[17] I. Saadi, A. Mustafa, J. Teller, B. Farooq, and M. Cools. Hidden Markov model-based population synthesis. *Transportation Research Part B*, 90:1–21, 2016.

[18] L. Sun and A. Erath. A Bayesian network approach for population synthesis. *Transportation Research Part C*, 61:49–62, 2015.

[19] B. F. Yaméogo, P. Gastineau, P. Hankach, and P.-O. Vandanjon. Comparing methods for generating a two-layered synthetic population. *Transportation Research Record*, 2675(1):136–147, 2021.

[20] X. Ye, K. Konduri, R. Pendyala, B. Sana, and P. Waddell. Methodology to match distributions of both household and person attributes in generation of synthetic populations. In *TRB 88th Annual Meeting Compendium of Papers DVD*, number 09-2096, Washington DC, USA, Jan. 2009. Transportation Research Board.