# A benchmarking methodology to assess population synthesis algorithms: A case study for Ile-de-France (France).

Jean-Giono Zehounkpe*[1,2], Sebastian Hörl[2], and Latifa Oukhellou[1]

[1]*COSYS-GRETTIA, Université Gustave Eiffel, France*
[2]*IRT SystemX, France*

## Introduction

In the last twenty years, agent-based models (ABMs) have emerged as a method to replicate the complexity in social systems and have been applied to a diverse range of applications in transportation systems Bastarianto et al. (2023). The ABMs quality depends on that of the synthetic travel demand, which includes a synthetic population of households and persons with sociodemographic attributes and their daily activity patterns in time and space Hörl & Balac (2021).

Synthetic populations are artificial populations generated based on observed populations that aim to share statistical properties similar to those of the observed populations Garrido et al. (2020). Population synthesis, the process of generating a synthetic population, has been performed using both deterministic algorithms, such as Iterative Proportional Fitting (IPF) and Direct Sampling (DS), and probabilistic algorithms, including Monte Carlo Markov Chain (MCMC) and Bayesian Networks (BN) (Barthelemy & Cornelis, 2012; Farooq et al., 2013; Sun & Erath, 2015).

Although several strictly statistical metrics have been used to assess the performance of population synthesis algorithms, frameworks that compare those algorithms in terms of the synthetic population quality are rare in the existing literature. To our knowledge, Bigi et al. (2024) worked on that gap in the literature.

Following this work, this paper proposes a framework for benchmarking population synthesis algorithms, with a particular focus on how they behave when missing observations in the training data are present.

## Methodology

### *Problem statement*

Census data published by the statistical bureaus contain combinations of demographic and socioeconomic attributes values of a share of the population. We call *structural zeros* the attribute combinations that do not exist in reality and *sampling zeros* those that are not observed in the census data due to limited knowledge of the underlying population.

Since probabilistic population synthesis algorithms have the ability to generate attribute combinations that are either structural zeros or sampling zeros, we aim to quantify the tendency of a population synthesis algorithm to generate attribute combinations that are either sampling or structural zeros.

### *The approach*

It is difficult to discern whether missing attribute combinations are structural or due to the small sample size, usually ranging from 1% to 10% of the population depending on the country Sun & Erath (2015). In

France, however, the national statistical bureau (INSEE[1]) publishes a large, weighted census data set that is representative for about 30% of the national population. Since, the INSEE data are more detailed and richer than most data sets used in the literature, we can assume that sufficiently few *sampling zeros* exist in the data.

Given the detailed data set, we choose to replicate the classic workflow of population synthesis from the provision of (potentially sparse) census data to the generation of a synthetic population. In that scope, we consider the INSEE data as ***ground truth*** (GT) and generate differently sized ***samples*** (S) based on that ground truth. We then generate ***synthetic populations*** (SP) based on those samples. That way, we can see how the sample size affects the quality of the synthetic population.

Let us call the missing combinations of attribute values in GT *ground truth zeros*. Assessing whether a population synthesis algorithm has a tendency to generate attribute combinations that are sampling zeros or structural zeros then translates into studying whether an algorithm has a tendency to generate attribute combinations that are ground truth zeros or sampling zeros that have been generated by our process.

Our approach consists of three stages. First, we generate $N$ weighted samples based on GT and the persons' weights using the Truncate-Replicate-Sample (TRS) method with $N$ random seeds Lovelace & Ballas (2013). TRS generates persons from a count value by copying every observation as many times as the integer component of the person's weight. Another person is generated stochastically based on the fractional part of the weight. Second, we generate for each GT sample $M$ other samples using $M$ sampling rates. Third, we generate $T$ synthetic populations using the $M$ samples as training data. Thus, $N \times M \times T$ data sets are generated at the end of the process allowing us to understand stochasticity effects in the population synthesis process.

### *The zeros analysis framework*

The sampling and population synthesis stages generate a new kind of zeros while inheriting the zeros from the previous stage. While the synthetic population inherits the ground truth zeros and the sampling zeros from the second stage, some of them may be filled by synthetic persons. Looking at the trajectories of the zeros over the three stages leads us to define the framework made of the following aggregated cases:

- **Reproduction:** An attribute combination is either present or missing in the ground truth and still present or missing in the synthetic population, respectively.
- **Removal:** An initially existing attribute combination no longer exists in the synthetic population.
- **Recovery:** An initially existing attribute combination is removed during the sampling but again present in the synthetic population.
- **Hallucination:** An initially not existing attribute combination appears in the synthetic population.

Following this trajectory analysis, we draw three metrics that help us strengthen our comparative analysis. Those metrics are based on the attribute combinations and not their number of observations. We will focus on the observations in future work. The metrics names are : ***hallucination rate***, ***removal rate***, ***recovery rate***. The recovery and hallucination rates represent the tendency of a population synthesis algorithm to generate (wanted) observations for sampling zeros and (unwanted) observations for structural zeros, respectively. While we would expect the hallucination and removal rates to be closer to 0%, we would also assume the recovery rate to be closer to 100% for a high quality algorithm.

## Results

We present the results of our methodology applied to census data of people living in the region Ile-de-France in 2019 collected by INSEE. For our test, we have selected the following list of attributes described as follows: the category of age (13 modalities), the occupation category (8 modalities), the higher education (2

[1]Institut National de la Statistique et des Etudes Economiques

modalities), the gender (2 modalities), the structure of the household (17 modalities) and the number of cars in the household (4 modalities). In future work, we will examine the most frequently used attributes in the literature.

For benchmarking purposes, we choose four population synthesis algorithms: **Direct Sampling** (DS), **Iterative Proportional Fitting** (IPF), **Markov-Chain Monte-Carlo** (MCMC) and **Bayesian network** (BN). For robustness purposes, we set the values of N and M to 100 and 4.

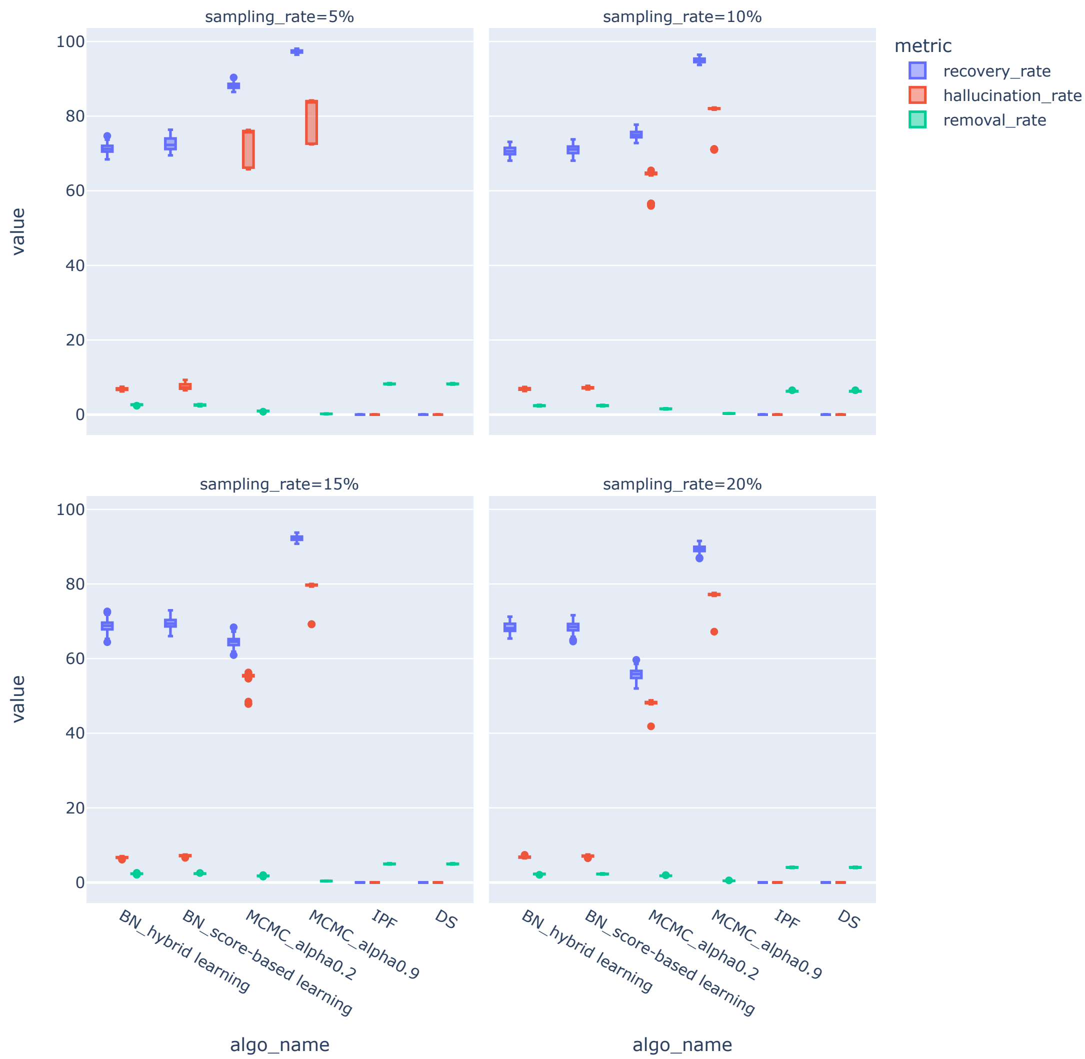


Figure 1: Population synthesis trajectory analysis metrics

In figure 1, we see that, the removal rate decreases as the training data get larger for all algorithms except MCMC for which it increases. This rate never goes above 8%. Also, the larger the training data the lower the recovery and hallucination rates, except for IPF and DS where those rates are equal to zero. There is gap of almost 70 percentage points between the two configurations for BN and MCMC for the metrics related to hallucination and recovery trajectories. While the recovery rates for BN (from 72% to 68%) and MCMC (from 97% to 89%) are high, the hallucination rates are lower for BN (7%) than for MCMC (from 83% to 77%). This means, both BN and MCMC have the tendency to generate attribute combinations that are sampling zeros (generated by our process) while MCMC has also the tendency to generate attribute combinations that are ground truth zeros.

## Conclusions

We introduce a framework to analyze population synthesis trajectories. While our goal is to benchmark population synthesis algorithms and see how they react in the context of missing observations in training data, we also provide three metrics derived from the trajectory analysis to evaluate the tendency of these algorithms to generate attribute combinations that are either sampling zeros or structural zeros.

The results of our use case suggest both BN and MCMC have a tendency to generate attribute combinations that are sampling zeros while MCMC tends to generate attribute combinations that are structural zeros in almost the same proportion as the sampling zeros. We have further the comparative analysis by tuning the algorithms parameters to see how the changes affect our metrics. While the changes had little impact for BN, they had considerably changed the shapes of the trajectories for MCMC. Therefore, our framework is designed to be used not only as an assessment tool but also as a calibration tool for population synthesis algorithms.

Future research will include more algorithms such as Generative Adversarial Network and Variational Auto-Encoder as they gained attention for their performances in recent literature Garrido et al. (2020). Further, given that they handle large data sets, we will expand the scope of the attributes and include all the 12 millions people living in the Ile-de-France region and look at the population at the household level.

## Acknowledgements

This research work is part of the PEPR MOBIDEC program managed by IFP Energies Nouvelles (IFPEN) and Université Gustave Eiffel (UGE) for the French government, funded itself by ”France 2030”.

## References

Barthelemy, J., & Cornelis, E. (2012). Synthetic populations: review of the different approaches.

Bastarianto, F. F., Hancock, T. O., Choudhury, C. F., & Manley, E. (2023, June). Agent-based models in urban transportation: review, challenges, and opportunities. *European Transport Research Review*, *15*(1), 19. doi: 10.1186/s12544-023-00590-5

Bigi, F., Rashidi, T. H., & Viti, F. (2024, April). Synthetic Population: A Reliable Framework for Analysis for Agent-Based Modeling in Mobility. *Transportation Research Record*, 03611981241239656. (Publisher: SAGE Publications Inc) doi: 10.1177/03611981241239656

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013, December). Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, 243–263. doi: 10.1016/j.trb.2013.09.012

Garrido, S., Borysov, S. S., Pereira, F. C., & Rich, J. (2020, November). Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C: Emerging Technologies*, *120*, 102787. doi: 10.1016/j.trc.2020.102787

Hörl, S., & Balac, M. (2021, September). Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. *Transportation Research Part C: Emerging Technologies*, *130*, 103291. doi: 10.1016/j.trc.2021.103291

Lovelace, R., & Ballas, D. (2013, September). 'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, *41*, 1–11. doi: 10.1016/j.compenvurbsys.2013.03.004

Sun, L., & Erath, A. (2015, December). A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, *61*, 49–62. doi: 10.1016/j.trc.2015.10.010