March 31, 2025.

Vianey **DARSEL**

Evaluation metrics for Population Synthesis

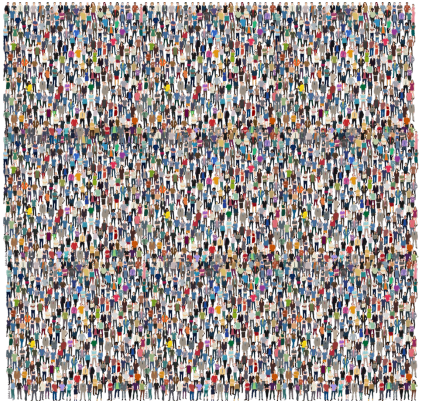
Definition of new metrics for a robust evaluation.

Université Gustave Eiffel

LABORATOIRE GRETTIA
GÉNIE DES RÉSEAUX
DE TRANSPORT TERRESTRES
ET INFORMATIQUE AVANCÉE

Introduction

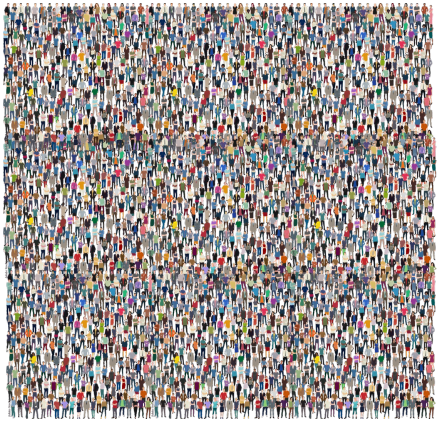
Population Synthesis in one scheme



Challenges in population synthesis

We cannot directly get a global synthetic population

Scalability issue



Privacy issue

How population synthesis is evaluated?

In the last 15 years, many algorithms have been applied to generate a synthetic population.

Problem

However, there is no consensus on the assessment of a generated synthetic population!

Often, only the global **distribution** of the generated population is evaluated.

It omits two other criteria: the **realism** and the **privacy**.

Interpretation in Machine Learning

Generating a synthetic population = generating a vector of mixed-type data (both continuous and categorical attributes)

This task is called **Tabular Data Synthesis**. We can inspire ourselves from contributions in this field to build relevant metrics for population synthesis.

Menu

Literature review on the evaluation

Evaluation criteria in the population synthesis and in Machine Learning

We can outline the criteria of evaluation depending on the field.

Criterion	PS	ML	Goal
Distribution	**X**	X	Comparing the distribution of the generated population with the true population.
Realism	X		Verifying that the generated samples are realistic
Privacy		X	Avoid any inference from the training data using the generated data
Diversity	X	X	Capacity of the generated data to cover all available data.
Performance on downstream tasks		**X**	Removing one variable from generated data and guess it from the other variables

Table: Description and definition of the different criteria in Population Synthesis and Machine Learning

X : most used criteria

Distribution evaluation: Population Synthesis

Preprocessing: Conversion into **categorical data**

Evaluation procedure: Compare part of the **multivariate** distribution.

Metrics:

	MAE	MARE	χ^2	SRMSE	R^2	NRMSE	Hellinger	JS divergence	Pearson	Cramer's V
Beckman et al. (1996)	2									
Ye et al. (2009)		1*	1*							
Farooq et al. (2013)				4*	4*					
Sun and Erath (2015)				5*						
Saadi et al. (2016)				1, 2, 3, 4						
Borysov et al. (2019)				1, 2, 3, 4	1, 2, 3				1, 2, 3	2
Kim and Bansal (2022)				1, 2						
Kukic et al. (2024)	1			1, 2, 3					1	
Bigi et al. (2024)						2, 3	2, 3	2, 3		
Kang et al. (2024)				1, 2						

Table: Metrics used in the distribution evaluation[1]

[1]The number indicates the number of variables that are taken into account to compute the multivariate frequencies. * indicates that the number corresponds to the total number of attributes (so it cannot be increased)

Distribution evaluation: Machine Learning

Distribution comparison

Preprocessing: Keep the **mixed data** format

Evaluation procedure: Compare the **marginals**.

Metrics:

- For categorical data: Total Variation Distance
- For numerical data: Kolmogorov-Sirnov Test

Limitation: No multivariate evaluation.

Correlation comparison

Preprocessing: Convert into **categorical** data if needed

Evaluation procedure: Compare the **correlation** between each couple of variables.

Metrics:

- For categorical data: Contingency Similarity
- For numerical data: Pearson Correlation Coefficient

Realism evaluation: Population Synthesis

Interpretation: Generated individuals that do not exist should not have been generated → concept of "Structural Zeros"

Metric: rate of "Structural Zeros"

$$\frac{\sum_{x \in X_{gen}} \mathbb{1}_{x \notin B \cup C}}{|X_{gen}|}$$

Limitation: Curse of dimensionality

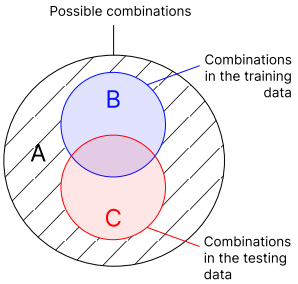


Figure: Venn diagram for structural zeros

Realism evaluation: Population Synthesis

Interpretation: Generated individuals that do not exist should not have been generated → concept of "Structural Zeros"

Metric: rate of "Structural Zeros"

$$\frac{\sum_{x \in X_{gen}} \mathbb{1}_{x \notin B \cup C}}{|X_{gen}|}$$

Limitation: Curse of dimensionality

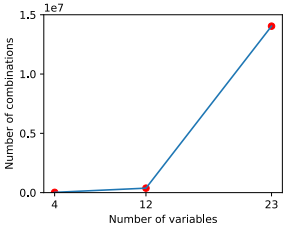
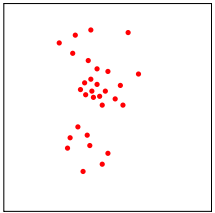
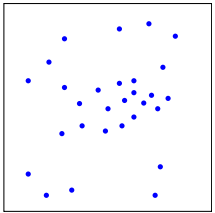


Figure: Evolution of the number of combinations with the number of variables

Realism evaluation: Machine Learning



(a) Real data



(b) Generated data

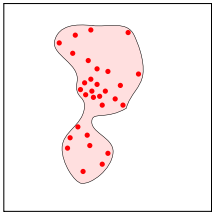
Interpretation: Generated data should belong to the original data support

Preamble: γ-support $\mathcal{S}^{\gamma}$ of a distribution is the minimum volume that supports a probability mass of γ

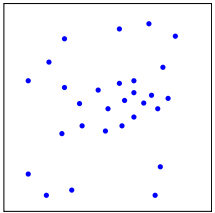
Metric: α-Precision: rate of generated samples that belongs to $\mathcal{S}^{\alpha}_{real}$

Limitation: Complex to evaluate the metrics, and requires enough data to have a good approximation of the distribution.

Realism evaluation: Machine Learning



(a) Support of real data



(b) Generated data

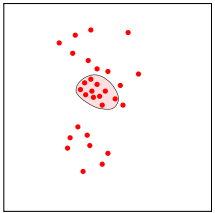
Interpretation: Generated data should belong to the original data support

Preamble: γ-support $\mathcal{S}^{\gamma}$ of a distribution is the minimum volume that supports a probability mass of γ

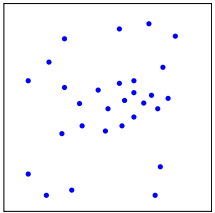
Metric: α-Precision: rate of generated samples that belongs to $\mathcal{S}^{\alpha}_{real}$

Limitation: Complex to evaluate the metrics, and requires enough data to have a good approximation of the distribution.

Realism evaluation: Machine Learning



(a) Alpha-support of real data



(b) Generated data

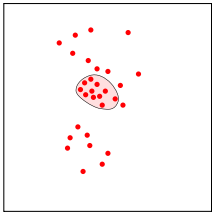
Interpretation: Generated data should belong to the original data support

Preamble: γ-support $\mathcal{S}^{\gamma}$ of a distribution is the minimum volume that supports a probability mass of γ

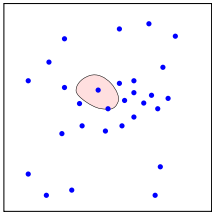
Metric: α-Precision: rate of generated samples that belongs to $\mathcal{S}^{\alpha}_{real}$

Limitation: Complex to evaluate the metrics, and requires enough data to have a good approximation of the distribution.

Realism evaluation: Machine Learning



(a) Alpha-support of real data



(b) Alpha-support on generated data

Interpretation: Generated data should belong to the original data support

Preamble: γ-support $\mathcal{S}^{\gamma}$ of a distribution is the minimum volume that supports a probability mass of γ

Metric: α-Precision: rate of generated samples that belongs to $\mathcal{S}^{\alpha}_{real}$

Limitation: Complex to evaluate the metrics, and requires enough data to have a good approximation of the distribution.

Privacy: Machine Learning

Interpretation: Generated data should not be too close from real data.

Metrics: For each generated sample: Distance to Closest Record (DCR) in the real data.

Post-processing: median, quantile, shortest between training and testing data...

Limitation: Output of the post-processing could be improved (optimal value, deeper comparison between training and testing distances).

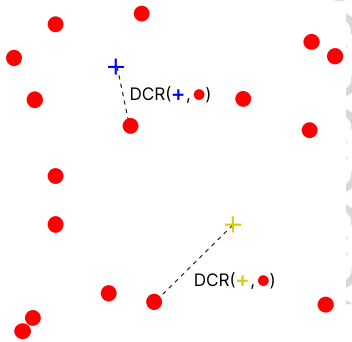


Figure: Illustration of the notion of DCR

Diversity evaluation: Population Synthesis

Interpretation: Generate individuals that exist, but were not present in the training data → concept of "Sampling Zeros"

Metric: rate of "Sampling Zeros"

$$\frac{\sum_{x \in X_{gen}} \mathbb{1}_{x \in C \setminus B}}{|X_{gen}|}$$

Limitation: What is the optimal value for this metric ?

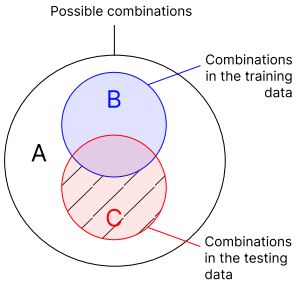
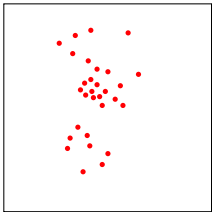
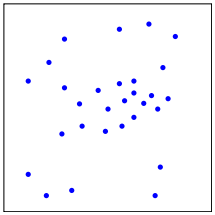


Figure: Venn diagram for sampling zeros

Diversity evaluation: Machine Learning



(a) Real data



(b) Generated data

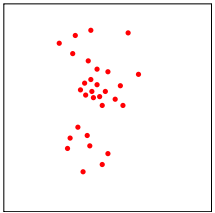
Interpretation: Real data should belong to the generated data support

Preamble: γ-support $\mathcal{S}^{\gamma}$ of a distribution is the minimum volume that supports a probability mass of γ

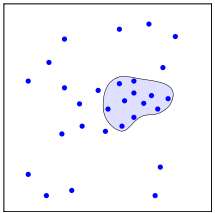
Metric: β-Recall: rate of real samples that belongs to $\mathcal{S}^{\beta}_{gen}$

Limitation: Complex to evaluate the metrics, and requires enough data to have a good approximation of the distribution.

Diversity evaluation: Machine Learning



(a) Alpha-support of real data



(b) Beta-support on generated data

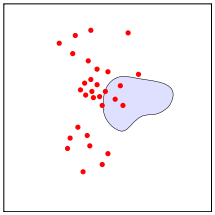
Interpretation: Real data should belong to the generated data support

Preamble: γ-support $\mathcal{S}^{\gamma}$ of a distribution is the minimum volume that supports a probability mass of γ

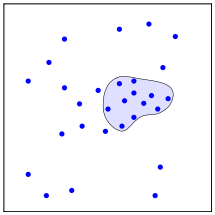
Metric: β-Recall: rate of real samples that belongs to $\mathcal{S}^{\beta}_{gen}$

Limitation: Complex to evaluate the metrics, and requires enough data to have a good approximation of the distribution.

Diversity evaluation: Machine Learning



(a) Beta-support of real data



(b) Beta-support on generated data

Interpretation: Real data should belong to the generated data support

Preamble: γ-support $\mathcal{S}^{\gamma}$ of a distribution is the minimum volume that supports a probability mass of γ

Metric: β-Recall: rate of real samples that belongs to $\mathcal{S}^{\beta}_{gen}$

Limitation: Complex to evaluate the metrics, and requires enough data to have a good approximation of the distribution.

Performance on downstream tasks: Machine Learning

Main task: Machine Learning efficiency.

Idea: How efficient is a model that is trained on generated data to guess a generated variable.

Method: A model is trained to guess a variable from the generated data, and then evaluated on the (real) test data.

Limitation: It does not correspond to the objectives in our use case.

Conclusion on literature review

Throughout the literature, we notice:

- No consensus exists to evaluate the distribution
- For other criteria, current evaluation metrics have some limitations
- Some criteria are not explored in Population Synthesis

Derivation of 3 new metrics

Complementarity of the metrics

Criterion	Relevance	But not sufficient
Faithful distribution	Ensure looks like population	Recopying training data would lead to an almost perfect score
Realistic individuals	Verify each individual is plausible	Globally, individuals may not represent the population
Privacy protection	Ensure data privacy for training individuals	If the model learns nothing, the privacy is respected

Table: Explanation of the necessity of several metrics for the evaluation

Distribution metric: $\overline{SRMSE}$

We propose using the mean of the SRMSE on the distributions of all possible combinations of three variables.

$$SRMSE_{ijk}(X_{gen}, X_{test}) = \sqrt{\sum_{x^{ijk}} \left(f_{gen}(x^{ijk}) - f_{test}(x^{ijk})\right)^2 \times |\Omega_i| \times |\Omega_j| \times |\Omega_k|}$$
$$\overline{SRMSE}_3(X_{gen}, X_{test}) = \frac{1}{\binom{n}{3}} \sum_{(i,j,k) \in \binom{\{1,\dots,n\}}{3}} SRMSE_{ijk}(X_{gen}, X_{test}) \quad (1)$$

Arguments:

- Most widely used metric
- Considering the trivariate distributions allows grabbing marginals, and bivariate distributions without exploding the computation time
- Balanced metric on all combinations

Realism metric: *SSCIOT*

Unrealistic individual: at least one couple of its attributes is absent from both training and testing sets.

We propose using the Share of Samples with a Couple of Instances that is Out of Testing data (*SSCIOT*).

$$SSCIOT(X_{gen}, X_{test}, X_{train}) = \frac{\sum_{x \in X_{gen}} \left(1 - \prod_{(i,j) \in \binom{\{1,\dots,n\}}{2}} \mathbb{1}_{x_{ij} \in (B \cup C)_{ij}}\right)}{|X_{gen}|} \quad (2)$$

Arguments:

- This metric that does not suffer from the curse of dimensionality

$(B \cup C)_{ij}$ is the restriction of $B \cup C$ to the variables i and j

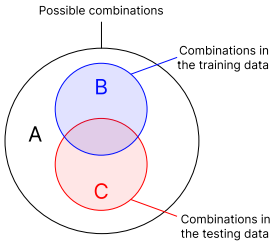


Figure: Venn diagram

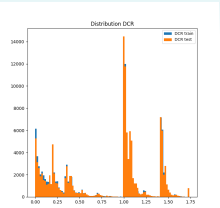
Privacy metric: Wasserstein-DCR

Goal: A generated individual should be in probability at the same distance from a training sample and from a testing sample.

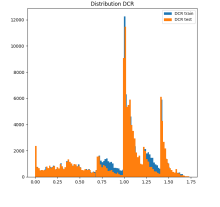
Design of the metric:

- Compute the DCR for each set to get an approximation of the distribution
- Compare these distributions with the Wasserstein distance

Note: this metric takes positively into account sampling zeros



(a) Privacy respected



(b) Privacy not respected

Equations for Wasserstein-DCR (WDCR)

Distance for mixed-type data: $d(x, y) = \sqrt{\sum_{i \in num}(q_{x_i} - q_{x_j})^2 + \sum_{i \in cat} \mathbb{1}_{x_j \neq y_j}}$

DCR between a set S and a sample x: $DCR(S, x) = \inf_{y \in S} d(x, y)$

Wasserstein Distance: $W_2(\nu_{DCR_{test}}, \nu_{DCR_{train}}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(DCR_{test\,i} - DCR_{train\,i}\right)^2}$

where $DCR_{set1} \leq DCR_{set2} \leq ... \leq DCR_{setn}$.

Interst on evaluating population synthesis

With this new rigorous framework, several contributions are possible:

- Do a benchmark on the current literature
- Perform model selection
- Test new models

Evaluating the robustness of the new metrics

Goal: Verifying the robustness (convergence and estimation of the minimal data) of the new metrics

Experience on census data from Ile-de-France (2015):

- Select most relevant variables in the data
- Train various models on two different data sizes (0.03% and 1%)
- Generate a synthetic population for each model
- Evaluate the generated population with several testing sets from several sizes (between 0.01% and 23% of the total population)

Note: by construction, the testing size for Wasserstein-DCR is the same as the training size, so the privacy metric is out of the study

Robustness of the distribution and the realism metrics

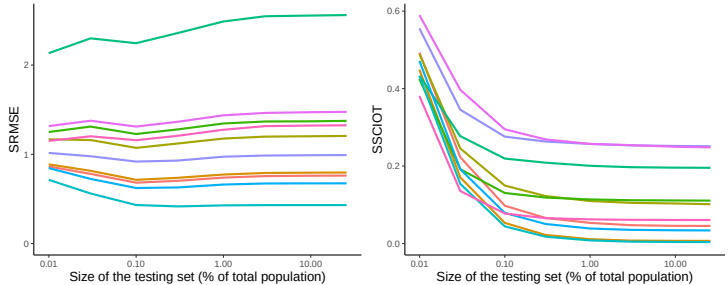


Figure: Evolution of the measures with the size of the testing data

- $\overline{SRMSE}$ can be well estimated with a reduced dataset.
- *SSCIOT* requires at least 0.1% of the total population in the testing data for a good estimation. This metric is based on the structure of the data and not on statistical properties.

Guidelines on the metrics on data management for evaluation

Recommendations in the evaluation:

- Evaluate your model with $\overline{SRMSE}$, *SSCIOT* and *WDCR*
- Split your data in 50/50 for training/testing
  At least half of data in testing is required for WDCR
- $\overline{SRMSE}$ should be evaluated by considering only the testing set
  Good approximation even with low data
- For *SSCIOT*, we recommand to use both the training and testing sets for the evaluation
  No methodology issue

Conclusion

- Population synthesis misses a systematic evaluation methodology
- Current metrics do not cover the goals of population synthesis
- We define three new metrics that evaluate: the distribution, the realism and the privacy
- These metrics have been tested, and guidelines on practical aspects given

A paper is being finalised on the findings of this work

Beckman, R. J., Baggerly, K. A., and McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6):415–429.

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

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

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

Kang, J., Kim, Y., Imran, M. M., Jung, G.-s., and Kim, Y. B. (2024). Generating Population Synthesis Using a Diffusion Model. In *Proceedings of the Winter Simulation Conference*, WSC '23, pages 2944–2955, <conf-loc>, <city>San Antonio</city>, <state>Texas</state>, <country>USA</country>, </conf-loc>. IEEE Press.

Kim, E.-J. and Bansal, P. (2022). A Deep Generative Model for Feasible and Diverse Population Synthesis. arXiv:2208.01403 [cs, stat].

Kukic, M., Li, X., and Michel Bierlaire (2024). One-step Gibbs sampling for the generation of synthetic households. *Transportation Research Part C: Emerging Technologies*, 166:104770.

Saadi, I., Mustafa, A., Teller, J., Farooq, B., and Cools, M. (2016). Hidden Markov Model-based population synthesis. *Transportation Research Part B: Methodological*, 90:1–21.

Sun, L. and Erath, A. (2015). A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, 61:49–62.

Ye, X., Konduri, K., Pendyala, R., Sana, B., and Waddell, P. (2009). Methodology to match distributions of both household and person attributes in generation of synthetic populations.

Thank you for your attention.

Vianey Darsel
COSYS - GRETTIA
vianey.darsel@univ-eiffel.fr

Tél. +33(0)7 81 64 83 74
https://grettia.univ-gustave-eiffel.fr
https://www.darsel.fr