Financiado por la Unión Europea
NextGenerationEU

Enhanced validation of tabular Synthetic Data: assessing Propensity Score Resemblance Metrics

Retreat GRBIO 2025

Nora Amama Ben Hassun

Daniel Fernández Martínez, Jordi Cortés Martínez
Institute for Research and Innovation in Health (IRIS)
Universitat Politècnica de Catalunya - BarcelonaTech (UPC)

17th July 2025

Outline

- Motivation
- Objectives
- Background
- Results: Simulation Study
- Conclusions & Future research

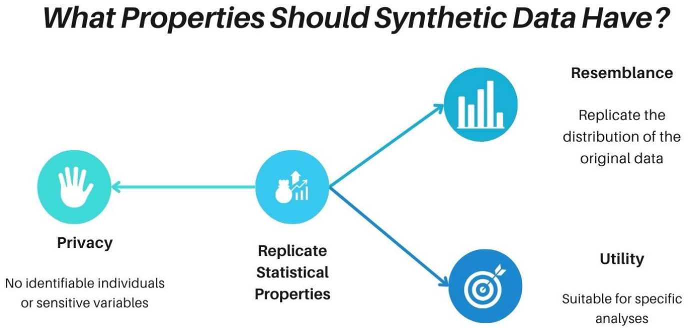
Introducing Synthetic Data (SD)

Synthetic data (SD): data artificially generated to **replicate the statistical properties** of real data while preserving confidentiality.

Nowok, B., et al. (2016). “Synthpop: Bespoke Creation of Synthetic Data in R”. *Journal of Statistical Software*

Introducing Synthetic Data (SD)

Synthetic data (SD): data artificially generated to **replicate the statistical properties** of real data while preserving confidentiality.

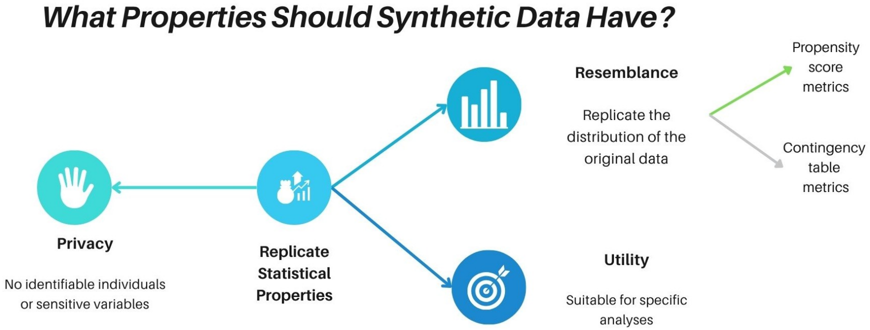


Nowok, B., et al. (2016). "Synthpop: Bespoke Creation of Synthetic Data in R". *Journal of Statistical Software*

Introducing Synthetic Data (SD)

Synthetic data (SD): data artificially generated to **replicate the statistical properties** of real data while preserving confidentiality.

synthpop



Nowok, B., et al. (2016). “Synthpop: Bespoke Creation of Synthetic Data in R”. *Journal of Statistical Software*

Objectives

1. **Present current propensity score metrics**

2. **Explore a new approach to compute propensity scores**

Example: Propensity score metrics

Original data (OD)

Id	Cost (€/kWh)	Region	Consumption (kWh)
ID-018	0.1249	South	448.4685
ID-902	0.2401	North	678.0603
ID-330	0.1963	South	1097.0372
ID-004	0.1697	West	920.6955
ID-705	0.0812	West	635.3353

Example: Propensity score metrics

Merging OD and SD

Id	Cost (€/kWh)	Region	Consumption (kWh)	$\mathbb{I}_{\{0,1\}}$
ID-018	0.1249	South	448.4685	0
ID-902	0.2401	North	678.0603	0
ID-330	0.1963	South	1097.0372	0
ID-004	0.1697	West	920.6955	0
ID-705	0.0812	West	635.3353	0
ID-085	0.0811	South	1262.5204	1
ID-402	0.0616	South	365.0383	1
ID-266	0.2232	South	655.0748	1
ID-197	0.1702	West	796.0875	1
ID-554	0.1916	West	966.5329	1

Example: Propensity score metrics

Propensity scores

Id	Cost (€/kWh)	Region	Consumption (kWh)	$\mathbb{I}_{\{0,1\}}$	$\widehat{p_i}$
ID-018	0.1249	South	448.4685	0	0.1749
ID-902	0.2401	North	678.0603	0	0.4441
ID-330	0.1963	South	1097.0372	0	0.9562
ID-004	0.1697	West	920.6955	0	0.8427
ID-705	0.0812	West	635.3353	0	0.4432
ID-085	0.0811	South	1262.5204	1	0.9863
ID-402	0.0616	South	365.0383	1	0.1049
ID-266	0.2232	South	655.0748	1	0.4804
ID-197	0.1702	West	796.0875	1	0.6874
ID-554	0.1916	West	966.5329	1	0.8814

Propensity Score Mean-Squared Error (pMSE)

Hypothesis Test

$$\begin{cases} H_0 : p(\mathrm{o} \mid X) = p(\mathrm{s} \mid X) \\ H_1 : p(\mathrm{o} \mid X) \neq p(\mathrm{s} \mid X) \end{cases}$$

pMSE Formula

$$\text{pMSE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - c)^2$$

pMSE under H_0

$$\text{pMSE} \overset{H_0}{\sim} \frac{\left(\frac{n_1}{N}\right) \frac{n_2}{N}}{N} \cdot \chi^2_{k-1}, \quad k = \text{no. glm parameters}$$

Snoke, J., et al. (2018). "General and specific utility measures for synthetic data". *Journal of the Royal Statistical Society Series A: Statistics in Society*

Kolmogorov-Smirnov Statistic (SPECKS)

Hypothesis Test

$$\begin{cases} H_0: & F^{\mathrm{o}}(p) = F^{\mathrm{s}}(p) \quad \forall p \in [0,1] \\ H_1: & F^{\mathrm{o}}(p) \neq F^{\mathrm{s}}(p) \quad \exists p \in [0,1] \end{cases}$$

SPECKS Statistic

$$D = \sup_{\hat{p}_i} \left| \hat{F}^{\mathrm{o}}(\hat{p}_i) - \hat{F}^{\mathrm{s}}(\hat{p}_i) \right|$$

SPECKS under H_0

$$D \overset{H_0}{\sim} KS(n_o, n_s)$$

Percentage Over 50% (PO50)

Hypothesis Test

$$\begin{cases} H_0 : p(\mathrm{o} \mid X) = p(\mathrm{s} \mid X) \\ H_1 : p(\mathrm{o} \mid X) \neq p(\mathrm{s} \mid X) \end{cases}$$

PO50 Statistic

$$\mathrm{PO50} = 100\frac{m}{N} - 50$$

$m = \sum_{i=1}^{N} \mathbb{I}_{\{\hat{y}_i = y_i\}}$
where $y_i \in \{0, 1\}$ and $\hat{y}_i = \mathbb{I}\{\hat{p}_i > c\}$

PO50 under H_0

$$\mathrm{PO50} \overset{H_0}{\sim} N(100(p_0 - 1/2), \frac{100^2}{N} p_0(1 - p_0)), \quad p_0 = P(\hat{y}_i = y_i)$$

Results: Simulation Study

Which is the best resemblance metric in the different scenarios?

Results: Simulation Study

Which is the best resemblance metric in the different scenarios?

Objectives:

1. Control type I error (α)
2. Statistical power ($1 - \beta$)

Results: Simulation Study

Which is the best resemblance metric in the different scenarios?

Objectives:

1. Control type I error (α)
2. Statistical power ($1 - \beta$)

Method for generating data: Normal(0,1)

Results: Simulation Study

Which is the best resemblance metric in the different scenarios?

Objectives:

1. Control type I error (α)
2. Statistical power ($1 - \beta$)

Method for generating data: Normal(0,1)

Scenarios:

- Sample sizes (n): 50, 100, 250, 500, 1000, 5000, 10000
- Variables (p): 2, 5, 10, 25, 50, 100

Results: Simulation Study

Which is the best resemblance metric in the different scenarios?

Objectives:

1. Control type I error (α)
2. Statistical power ($1 - \beta$)

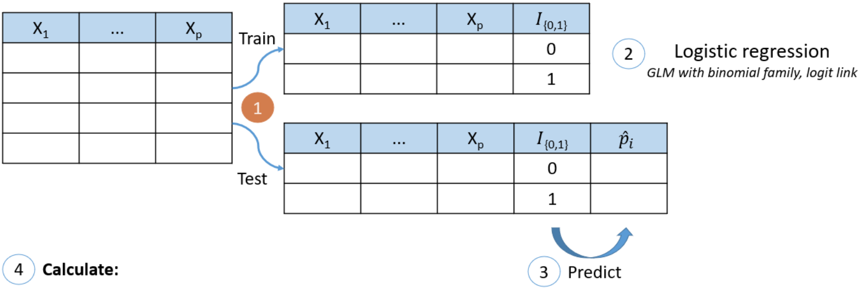
Method for generating data: Normal(0,1)

Scenarios:

- Sample sizes (n): 50, 100, 250, 500, 1000, 5000, 10000
- Variables (p): 2, 5, 10, 25, 50, 100

We did **NOT** use **SD** at any point in the simulation study.

A new approach for obtaining PS

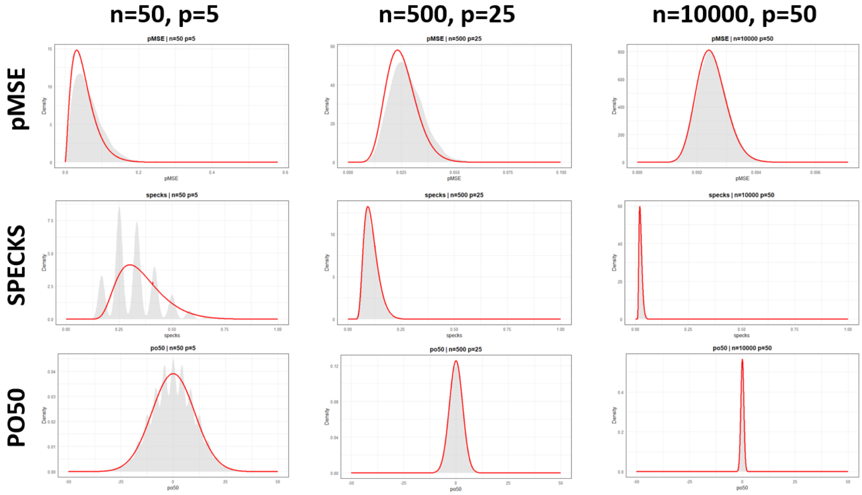


4 **Calculate:**

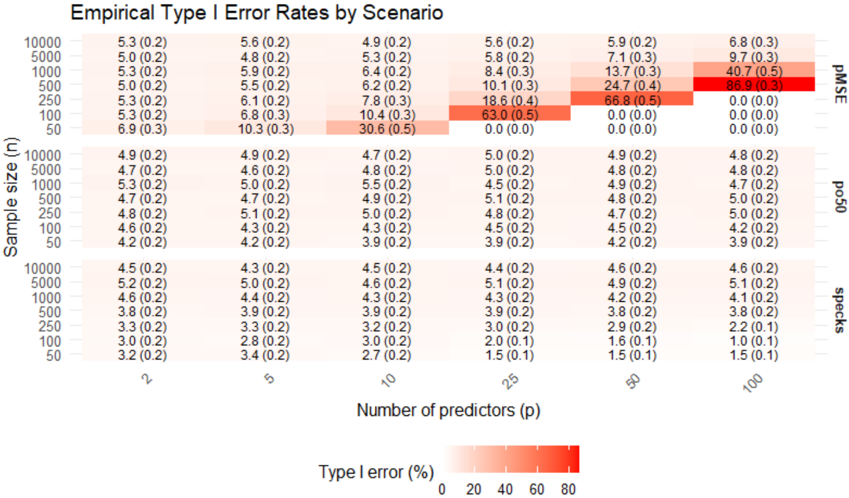
pMSE, SPECKS and PO50

Results

Empirical and theoretical metrics distribution



Probability of type I error (α)



Summary

With the current approach:

- ☑ Type I error controlled for pMSE
- ☑ SPECKS and PO50 fail because of overadjustment

With the new approach:

- ☑ Type I error controlled for all metrics
- ☑ Identified which scenarios are not good for the metrics.

Next steps and Future research

Assessing the statistical power

Scenario	Sample 1	Sample 2
1	Independent multivariate normal	Normal with different mean
2	Independent multivariate normal	Normal with different sd
3	Independent multivariate normal	Different distribution (Skew-Normal Distribution)
4	Independent multivariate normal	Correlated multivariate normal

Next steps and Future research

Assessing the statistical power

Scenario	Sample 1	Sample 2
1	Independent multivariate normal	Normal with different mean
2	Independent multivariate normal	Normal with different sd
3	Independent multivariate normal	Different distribution (Skew-Normal Distribution)
4	Independent multivariate normal	Correlated multivariate normal

Case study using existing datasets.

Next steps and Future research

Assessing the statistical power

Scenario	Sample 1	Sample 2
1	Independent multivariate normal	Normal with different mean
2	Independent multivariate normal	Normal with different sd
3	Independent multivariate normal	Different distribution (Skew-Normal Distribution)
4	Independent multivariate normal	Correlated multivariate normal

Case study using existing datasets.

Adapt Existing Metrics to SD

Best Resemblance Metric for Specific Analyses

Assess Missingness

TAKE-HOME MESSAGE

Now, we have
with a **type I error controlled**
multivariate resemblance metrics
to perform **synthetic data validation**.

Selected references

Hernandez M., Epelde G., Alberdi A., Cilla R., Rankin D. (2021). *Standardised Metrics and Methods for Synthetic Tabular Data Evaluation*. URL: `https://www.techrxiv.org/articles/preprint/Standardised_Metrics_and_Methods_for_Synthetic_Tabular_Data_Evaluation/16610896`, doi:10.36227/techrxiv.16610896.

Jordon J., Szpruch L., Houssiau F., Bottarelli M., Cherubin G., Maple C., Cohen S. N., Weller A. (2022). Synthetic Data – what, why and how? URL: `http://arxiv.org/abs/2205.03257`.

Nowok B., Raab G. M., Dibben C. (2016). Synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software 74*. doi:10.18637/jss.v074.i11.

Raab G. M. (2022). Utility and Disclosure Risk for Differentially Private Synthetic Categorical Data.

Raab G. M., Nowok B., Dibben C. (2021). Assessing, visualizing and improving the utility of synthetic data. URL: `http://arxiv.org/abs/2109.12717`.

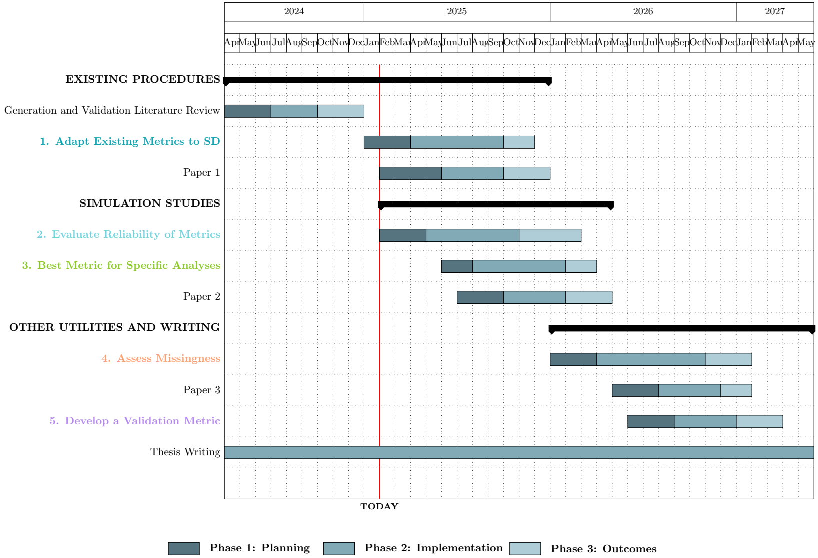
Snoke J., Raab G. M., Nowok B., Dibben C., Slavkovic A. (2018). General and Specific Utility Measures for Synthetic Data. *Journal of the Royal Statistical Society Series A: Statistics in Society 181*(3), 663–688. doi:10.1111/rssa.12358.

Acknowledgements

- Esta tesis está financiada por la **Siemens Energy AI Chair. Energy Sustainability for a Decarbonized Society 5.0** (TSI-100930-2023-5), financiado por la **Secretaría de Estado de Digitalización e Inteligencia Artificial** dentro de la convocatoria **Cátedras ENIA 2022**. Además, cuenta con el apoyo de la **Unión Europea - Next Generation EU**.

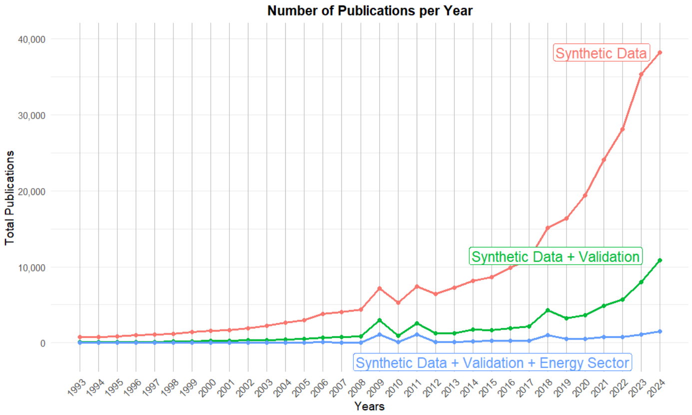
- This research was funded by the **MICIU** / **AEI** /10.13039/501100011033 (Spain) and by **FEDER (EU)** [PID2023-148033OB-C21] & and by **grant 2021 SGR 01421 (GRBIO)** administrated by the **Departament de Recerca i Universitats de la Generalitat de Catalunya (Spain)**.

Timeplan


2024
2025
2026
2027
Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May
EXISTING PROCEDURES
Generation and Validation Literature Review
1. Adapt Existing Metrics to SD
Paper 1
SIMULATION STUDIES
2. Evaluate Reliability of Metrics
3. Best Metric for Specific Analyses
Paper 2
OTHER UTILITIES AND WRITING
4. Assess Missingness
Paper 3
5. Develop a Validation Metric
Thesis Writing
TODAY
Phase 1: Planning
Phase 2: Implementation
Phase 3: Outcomes

Motivation

How synthetic data research is taking off?



1 **Source:** Dimensions.ai. **Keywords:** “Synthetic Data” AND (“Utility” OR “Resemblance”) AND (“Energy Sector” OR “Renewable Energy” OR “Electric Power” OR “Energy Industry”).

Context and Motivation

Current utilities and future challenges

Current utilities

- Enhancing data privacy
- Reducing bias
- Augmenting small datasets
- Accelerating training

Current challenges

- Minimizing the need for validation of real data
- Addressing generation dependency
- Validating synthetic data
- Capturing and replicating all extreme cases

Context and Motivation

Current utilities and future challenges

Current utilities

- Enhancing data privacy
- Reducing bias
- Augmenting small datasets
- Accelerating training

Current challenges

- Minimizing the need for validation of real data
- Addressing generation dependency
- Validating synthetic data
- Capturing and replicating all extreme cases

Context and Motivation

Current utilities and future challenges

Current utilities

- Enhancing data privacy
- Reducing bias
- Augmenting small datasets
- Accelerating training
- **Validation metrics framework**
- **Realism in specific scenarios**

Current challenges

- Minimizing the need for validation of real data
- Addressing generation dependency

Context and Motivation

Sustainable Development Goals

7 AFFORDABLE AND CLEAN ENERGY

Synthetic data supports energy optimization and renewable energy integration

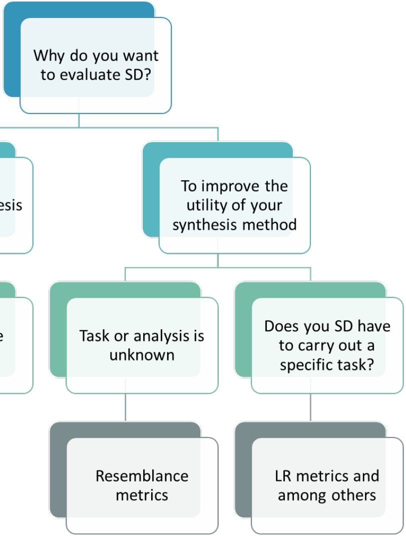
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

Synthetic data allows companies to safely innovate while maintaining confidentiality.

13 CLIMATE ACTION

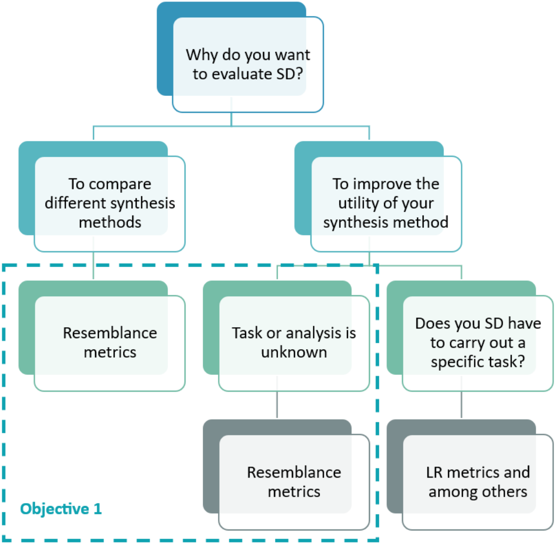
Synthetic data supports the modeling of energy efficiency strategies.

General classification



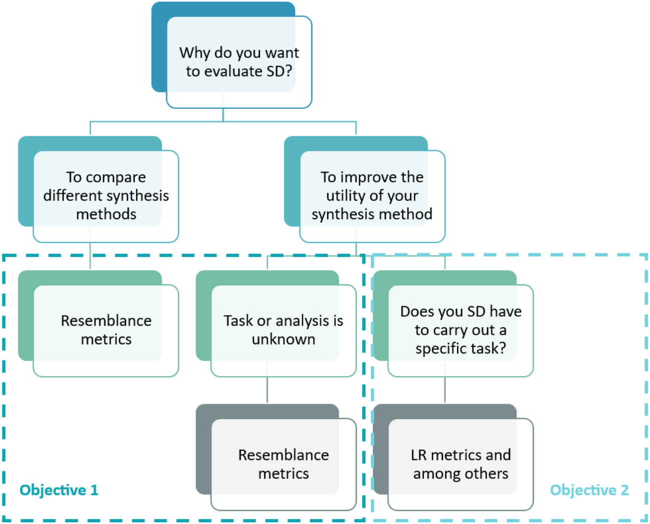
Source: United Nations Economic Commission for Europe. *Synthetic Data for Official Statistics: A Starter Guide*. 2023.

General classification



Source: United Nations Economic Commission for Europe. *Synthetic Data for Official Statistics: A Starter Guide*. 2023.

General classification



Source: United Nations Economic Commission for Europe. *Synthetic Data for Official Statistics: A Starter Guide*. 2023.

Simulation study process

X_1	...	X_p	$I_{\{0,1\}}$	$\hat{p}_i$
			0	
			0	
			1	
			1	

OD

SD

1 Logistic regression

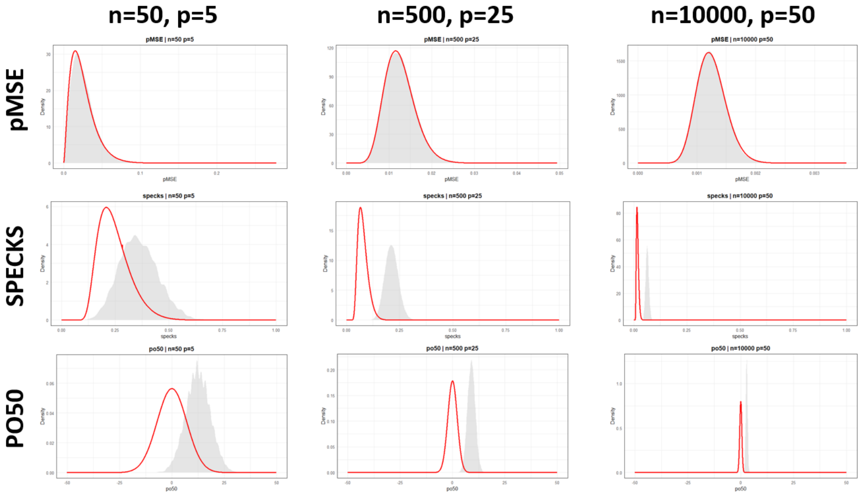
GLM with binomial family, logit link

2 **Calculate:**

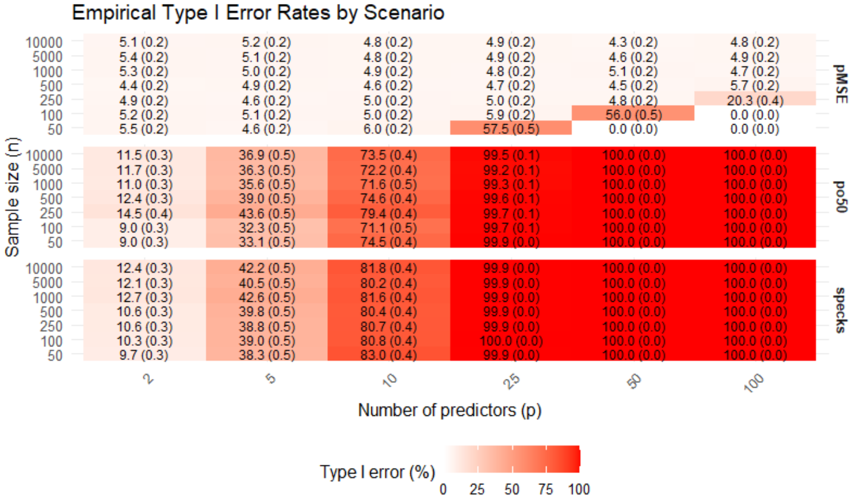
pMSE, SPECKS and PO50

Results

Empirical and theoretical metrics distribution



Probability of type I error (α)



Types of Synthetic Data

Types of data synthesis	Quality
Real non-public datasets	High
Real public data	High, although there are limitations (e.g. aggregated data)
Simulation engine	Will depend on the fidelity of the existing generating model
Generated from generic assumptions	Will likely be low

El Emam, K., et al. (2020). “Practical synthetic data generation: balancing privacy and the broad availability of data”. *O’Reilly Media*

State of the art

Current Tools

SynthRO: a dashboard to evaluate and benchmark synthetic data

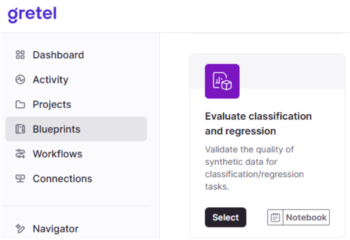
?

State of the art

Current Tools

SynthRO: a dashboard to evaluate and benchmark synthetic data

?



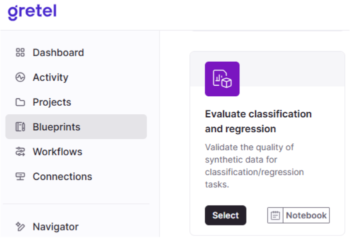
Gretel.ai

State of the art

Current Tools

SynthRO: a dashboard to evaluate and benchmark synthetic data

?



Gretel.ai

Ydata

State of the art - Metrics

Validation

1. Resemblance
   - Propensity score metrics (Raab *et al.*, 2021)
   - Contingency table metrics (not shown)
2. **Utility (Raab, 2022)**

Utility metrics

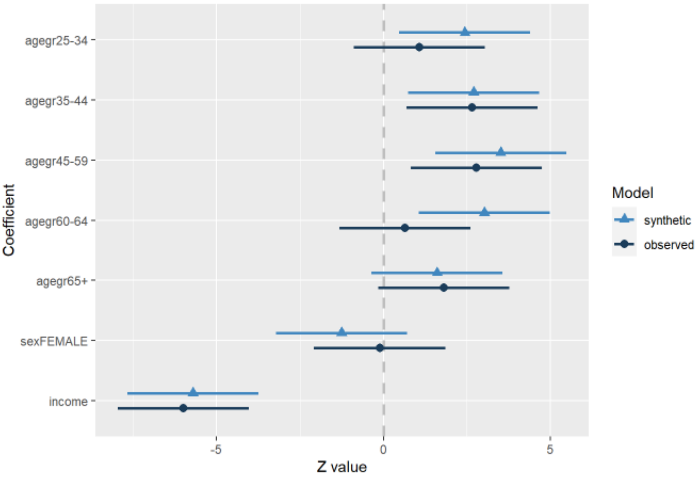
Confidence Interval Overlap (CIO)

Hypothesis Test

$$\begin{cases} H_0 : \beta_j^{\mathrm{o}} = \beta_j^{\mathrm{s}} \\ H_1 : \beta_j^{\mathrm{o}} \neq \beta_j^{\mathrm{s}} \end{cases}$$

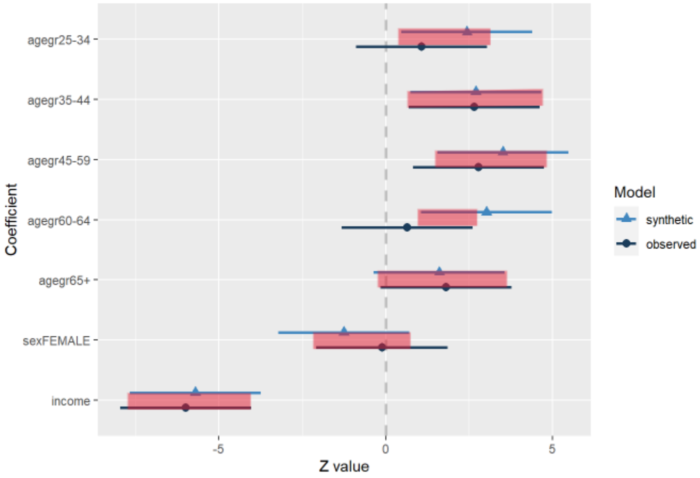
Utility metrics

Confidence Interval Overlap (CIO)



Utility metrics

Confidence Interval Overlap (CIO)



Utility metrics

Confidence Interval Overlap (CIO)

Hypothesis Test

$$\begin{cases} H_0 : \beta_j^{\mathrm{o}} = \beta_j^{\mathrm{s}} \\ H_1 : \beta_j^{\mathrm{o}} \neq \beta_j^{\mathrm{s}} \end{cases}$$

CIO Statistic

$$\mathrm{CIO} = \frac{1}{2}\left(\frac{\text{Overlap Length}}{\text{CI Length (Original)}} + \frac{\text{Overlap Length}}{\text{CI Length (Synthetic)}}\right) \overset{H_0}{\sim} N(0, 1)$$

Kolmogorov-Smirnov Statistic (SPECKS)

Synthetic Probability Error Comparison using Kolmogorov-Smirnov Statistic (SPECKS)

Kolmogorov-Smirnov Distribution

Under H_0, the statistic D follows a Kolmogorov-Smirnov distribution, whose cumulative distribution function (CDF) is given by:

$$P\left(\sqrt{n}D \leq x\right) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2}$$

where D is the Kolmogorov-Smirnov statistic, and n is the sample size (**?**).

Methodology

2. Evaluate reliability of metrics

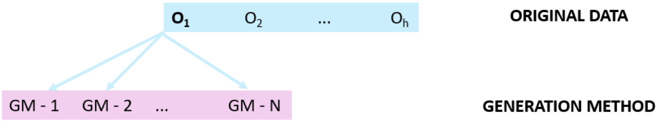
$\mathbf{O_1}$	O_2	...	O_h	**ORIGINAL DATA**

SIMULATION STUDY

Generation algorithm
Randomness level

Methodology

2. Evaluate reliability of metrics

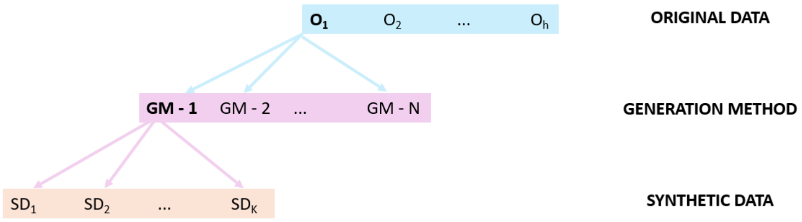


SIMULATION STUDY

Generation algorithm
Randomness level

Methodology

2. Evaluate reliability of metrics

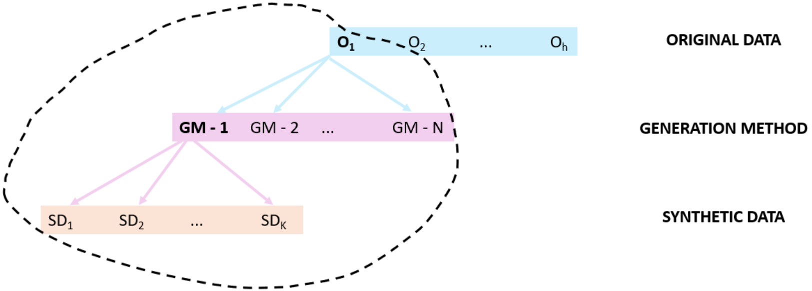


SIMULATION STUDY

Generation algorithm
Randomness level

Methodology

2. Evaluate reliability of metrics

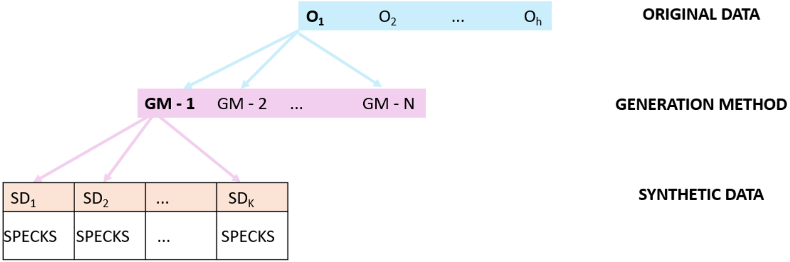


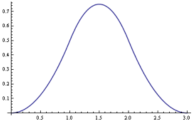
SIMULATION STUDY

Generation algorithm
Randomness level

Methodology

2. Evaluate reliability of metrics





SIMULATION STUDY

Generation algorithm
Randomness level

Methodology

3. The most suitable metric for specific statistical analyses

ORIGINAL DATA

SYNTHETIC DATA

Do we get the same result?

Methodology

3. The most suitable metric for specific statistical analyses

$\mathbf{O_1}$	O_2	...	O_h

ORIGINAL DATA

SIMULATION STUDY

Proportion of data
Sample size
Variable type
Outliers
Missings

Methodology

3. The most suitable metric for specific statistical analyses

O_1 O_2 … O_h

ORIGINAL DATA

SD_1	SD_2	…	SD_K
ARI	ARI	…	ARI
pMSE	pMSE	…	pMSE
SPECKS	SPECKS	…	SPECKS

SYNTHETIC DATA

SIMULATION STUDY

Proportion of data
Sample size
Variable type
Outliers
Missings

Methodology

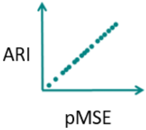
3. The most suitable metric for specific statistical analyses

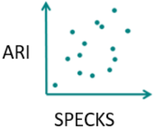
ORIGINAL DATA

O_1	O_2	...	O_h

SYNTHETIC DATA

SD_1	SD_2	...	SD_K
ARI	**ARI**	**...**	**ARI**
pMSE	pMSE	...	pMSE
SPECKS	SPECKS	...	SPECKS





SIMULATION STUDY

Proportion of data
Sample size
Variable type
Outliers
Missings

Methodology

4. Handling Missing Data

Methodology

4. Handling Missing Data

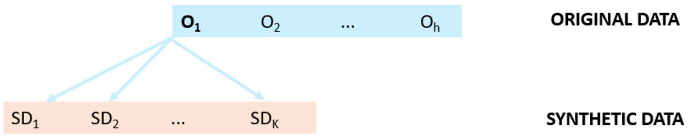
O_1	O_2	...	O_h	**ORIGINAL DATA**

SIMULATION STUDY

% missings
Generation algorithm
Imputation method

Methodology

4. Handling Missing Data

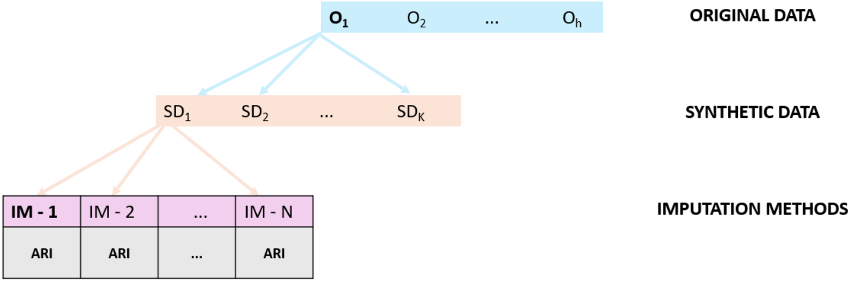


SIMULATION STUDY

% missings
Generation algorithm
Imputation method

Methodology

4. Handling Missing Data

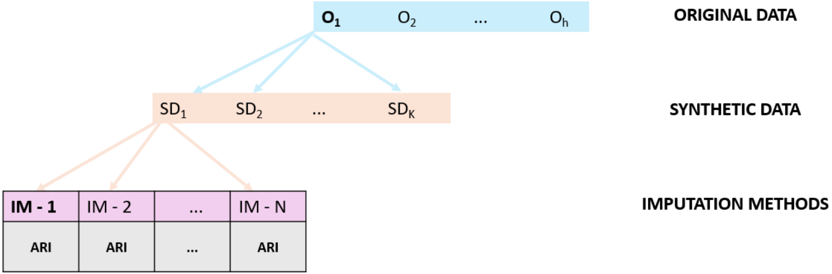


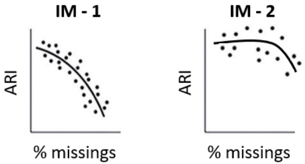
SIMULATION STUDY

% missings
Generation algorithm
Imputation method

Methodology

4. Handling Missing Data





SIMULATION STUDY

% missings
Generation algorithm
Imputation method

Methodology

5. Developing a Validation Metric

Proposed Validation Metric

Inspired by ?, we propose a weighted metric to evaluate synthetic data based only on **real and synthetic datasets**.

- **PCA-based structure**: Captures high-dimensional relationships.
- **Resemblance**: Measures similarity between synthetic and real data.
- **Privacy**: Ensures no leakage of sensitive information.

Pooled metric:

$$M = w_1 M_{PCA} + w_2 M_{Resemblance} + w_3 M_{Privacy}$$

Contingency Table Metrics

Data organization

Observed Data Frequency Table

Age Group	**Low Income**	**High Income**	**Total**
Young	100	260	360
Adult	90	50	140
Total	190	310	500

Synthetic Data Frequency Table

Age Group	**Low Income**	**High Income**	**Total**
Young	120	240	360
Adult	80	60	140
Total	200	300	500

Contingency Table Metrics

Process

1. **Frequency Tables**
2. **Application of Statistical Tests**

Pearson Statistic

$$\chi^2 = \sum_{j=1}^{k} \frac{(s_j - o_j)^2}{o_j},$$

where:

- o_j: Frequency for category j in the original data.
- s_j: Frequency for category j in the synthetic data.
- k: Number of categories.

Voas-Williamson Utility Measure (VW)

Hypothesis Test

$$\begin{cases} H_0: & o_j = s_j \quad \forall j \\ H_1: & o_j \neq s_j \quad \text{for some } j \end{cases}$$

Voas-Williamson Utility Measure (VW)

Hypothesis Test

$$\begin{cases} H_0: & o_j = s_j \quad \forall j \\ H_1: & o_j \neq s_j \quad \text{for some } j \end{cases}$$

VW Formula

Adjusts for the relative size of original (n_1) and synthetic (n_2) data.

$$VW = \sum_{j=1}^{k} \frac{\left(s_j - o_j \cdot \frac{c}{1-c}\right)^2}{c \cdot (o_j + s_j)}$$

Adjusted Rand Index (ARI)

Adjusted Rand Index (ARI)

Measures the similarity between two clustering results, adjusting for chance. It is given by:

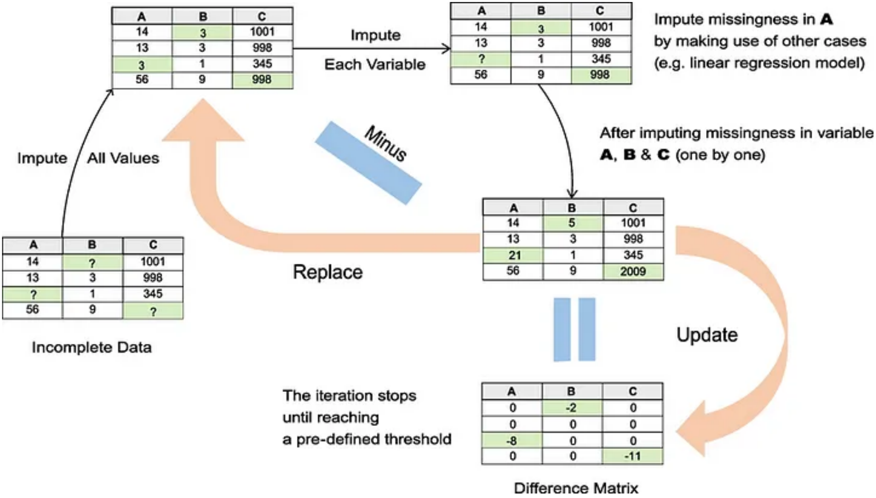
$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

Interpretation

- **1:** Perfect match of clusters.
- **0:** Random clustering.

Multiple Imputation by Chained Equations

MICE is an iterative method to handle missing data by imputing values **one variable at a time**.



Imputation methods