Financiado por la Unión Europea
NextGenerationEU

Evaluation of Metrics for Assessing Synthetic Tabular Data Quality

Workshop on the occasion of the 10th Anniversary of the GRBIO

Nora Amama Ben Hassun

Daniel Fernández Martínez, Jordi Cortés Martínez

Universitat Politècnica de Catalunya - BarcelonaTech (UPC)

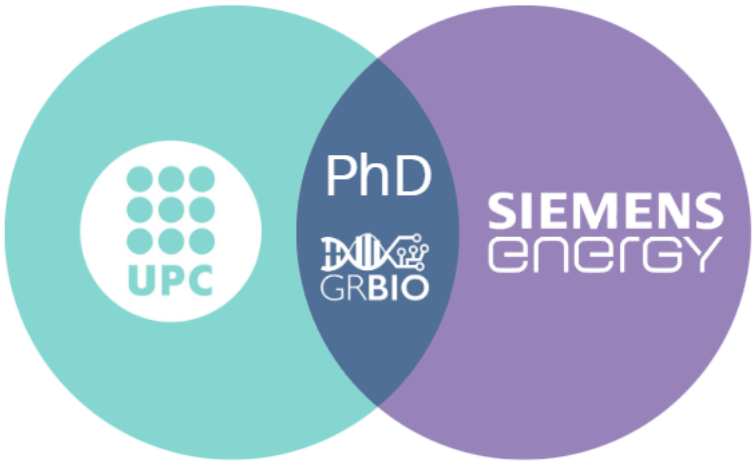
31st January 2025

Càtedra ENIA Siemens Energy AI Chair

Energy Sustainability for a Decarbonized Society 5.0

Càtedra ENIA Siemens Energy AI Chair

Energy Sustainability for a Decarbonized Society 5.0



Outline

- Context
- Motivation
- State of the art
- PhD Objectives
- Methodology

Context

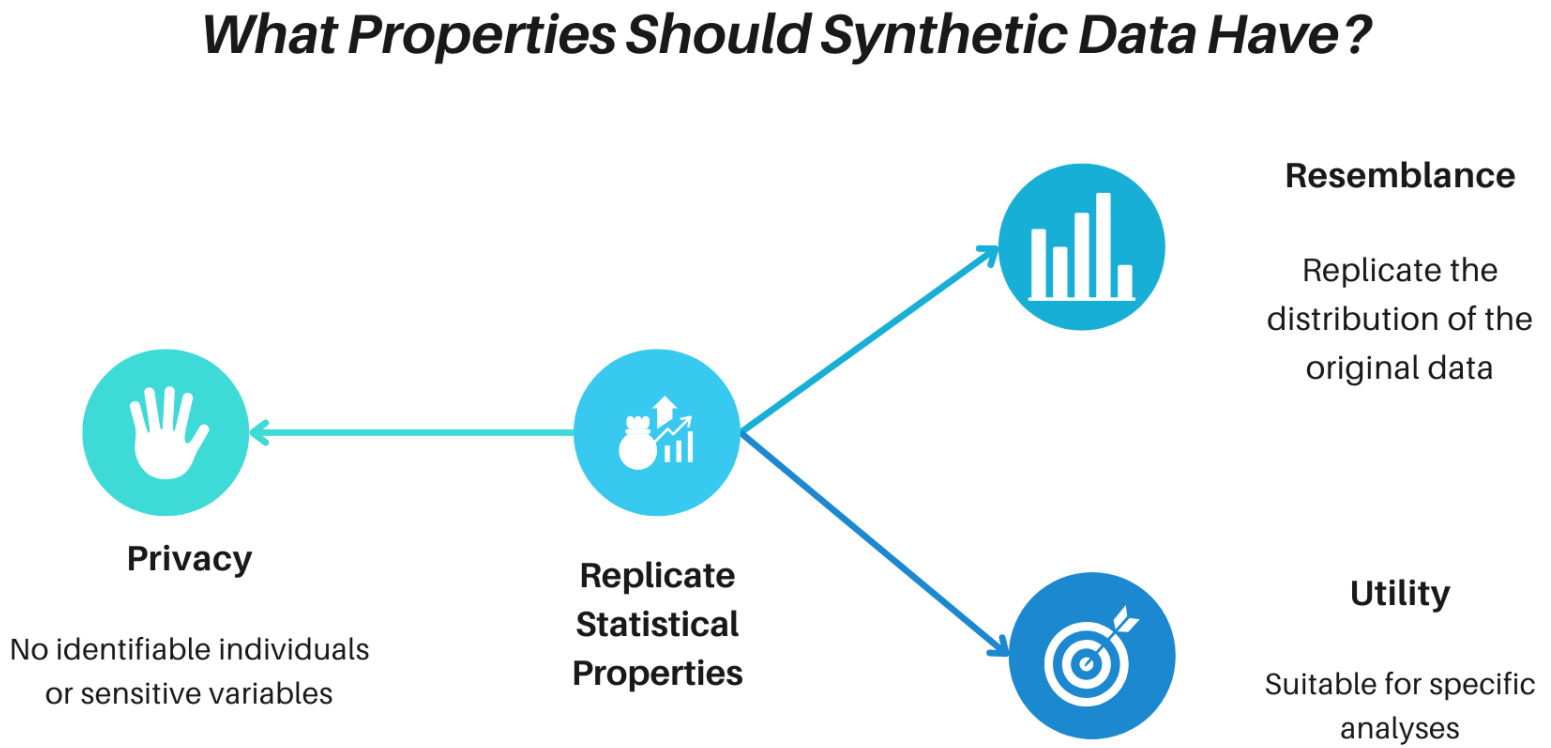
SD definition

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

Context

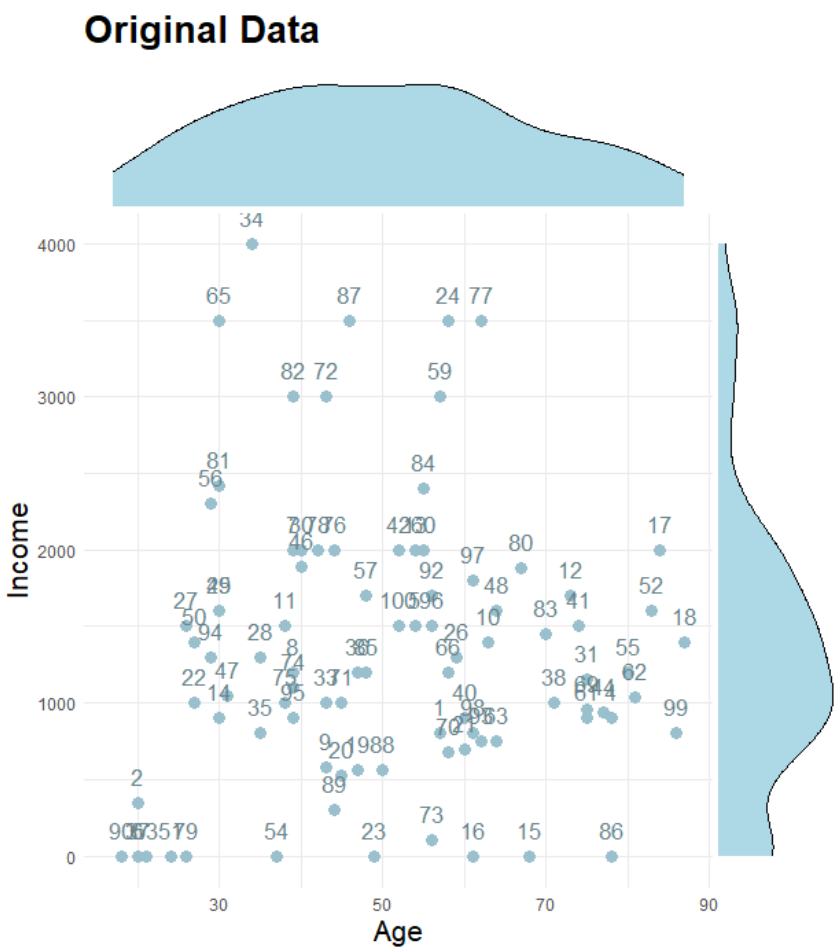
SD definition

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



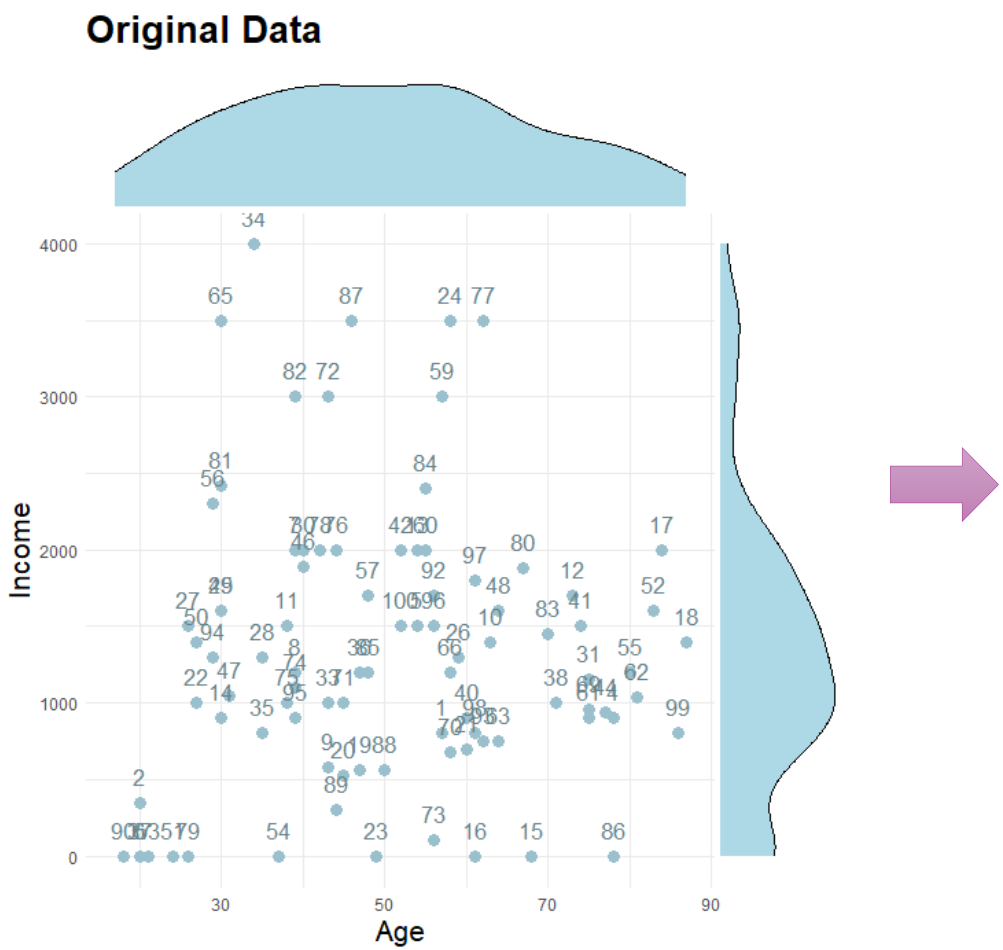
Context

Example



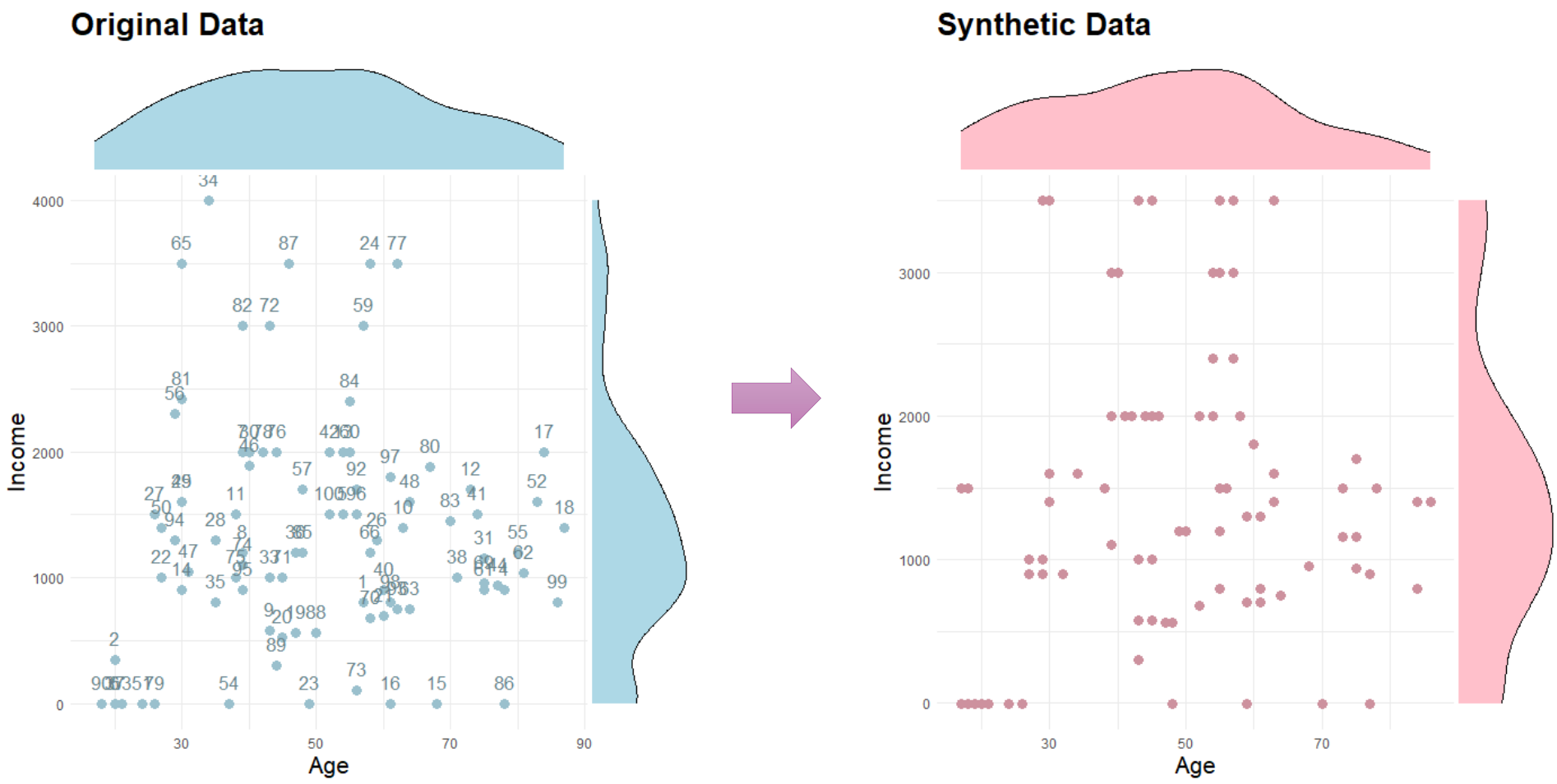
Context

Example



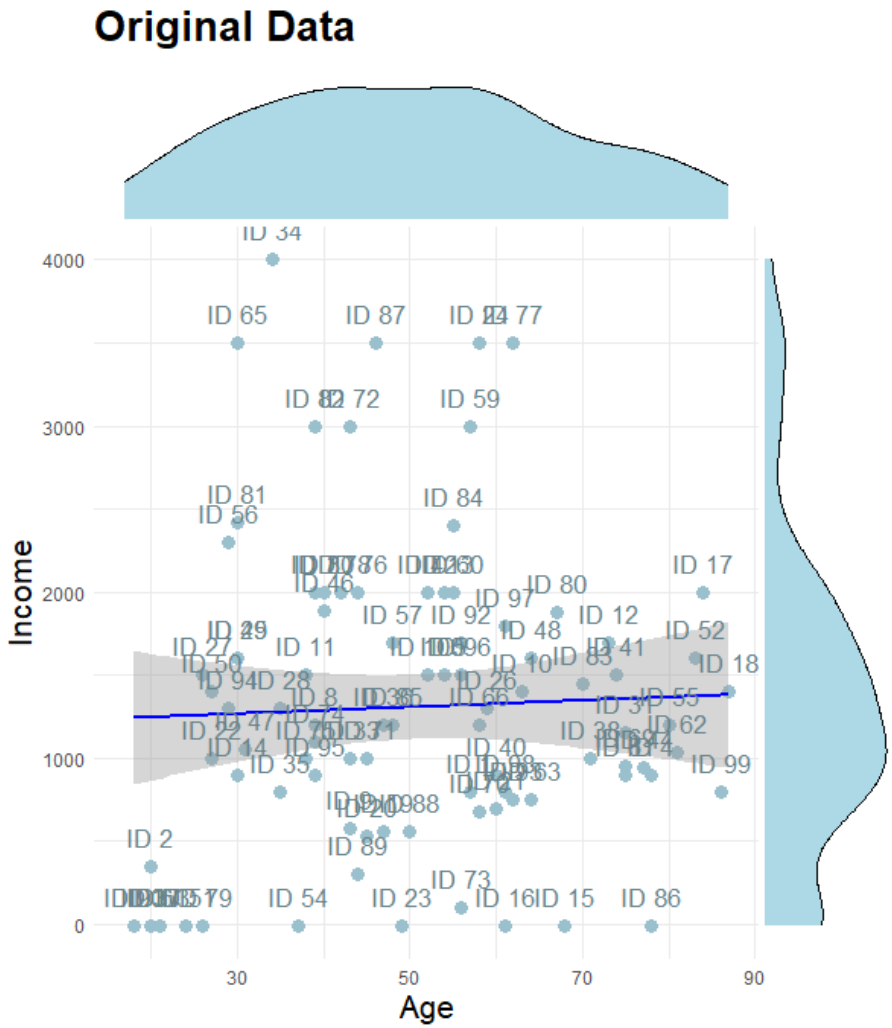
Context

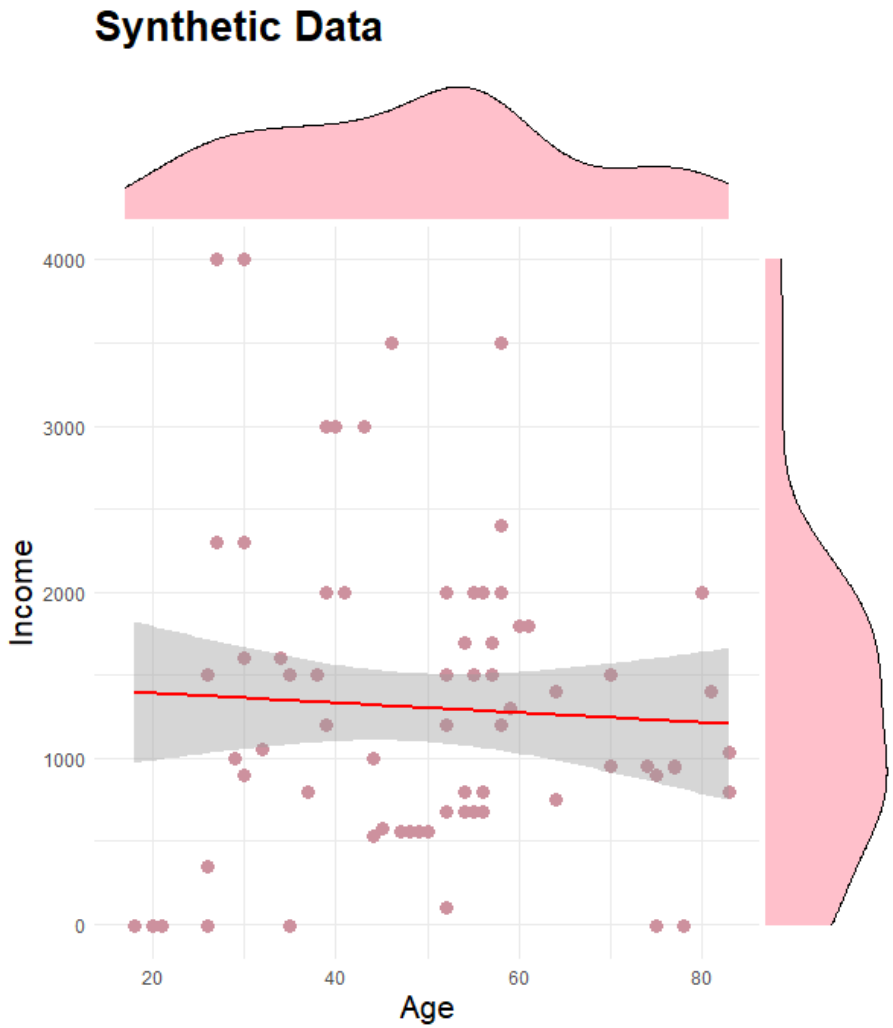
Example



Context

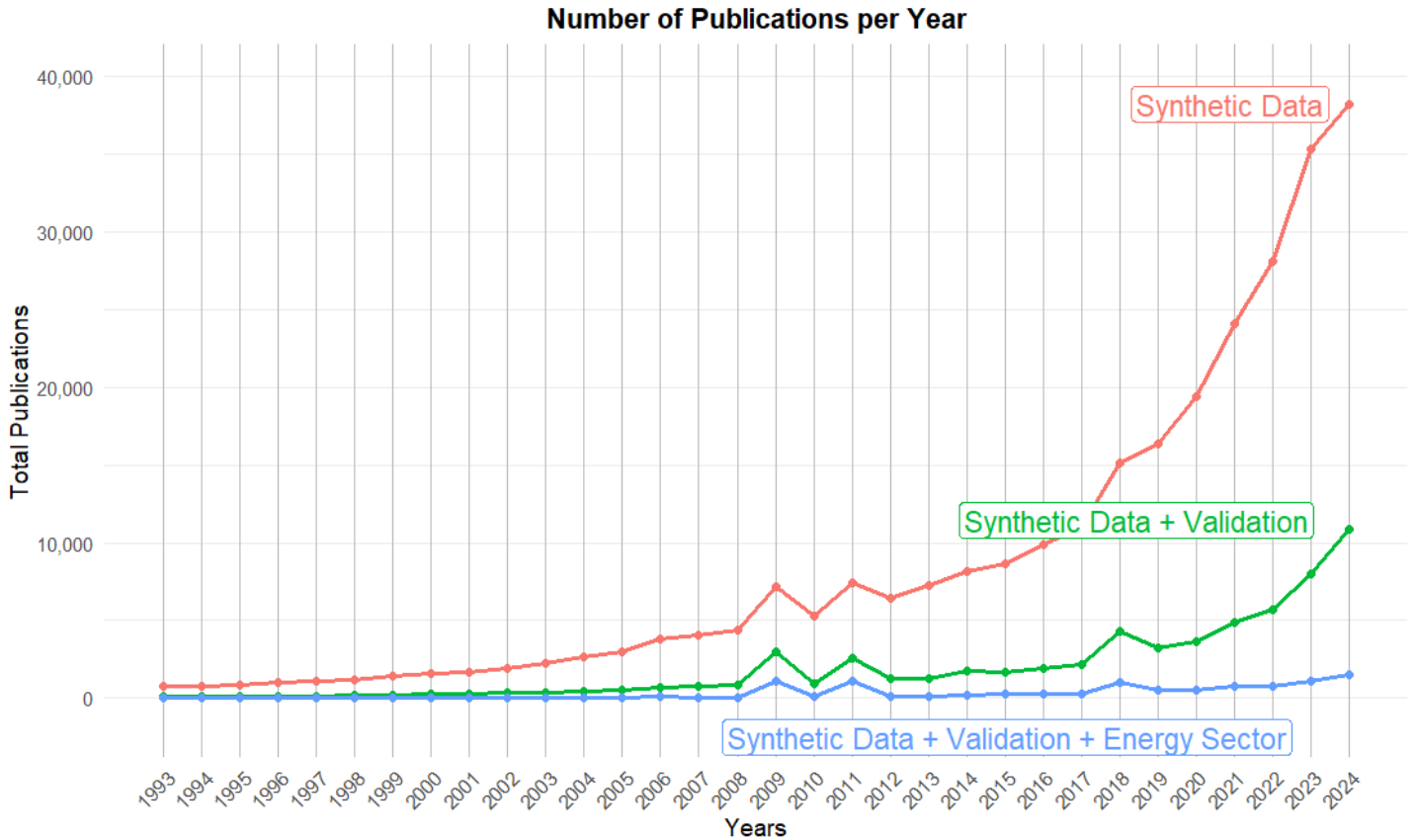
Example





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”).

State of the art - Software

Synthpop R package

(Nowok *et al.*, 2016)

State of the art - Software

Synthpop R package

(Nowok *et al.*, 2016)

State of the art - Software

Synthpop R package

synthpop

(Nowok *et al.*, 2016)

1. **Social Diagnosis 2011 (SD2011):**



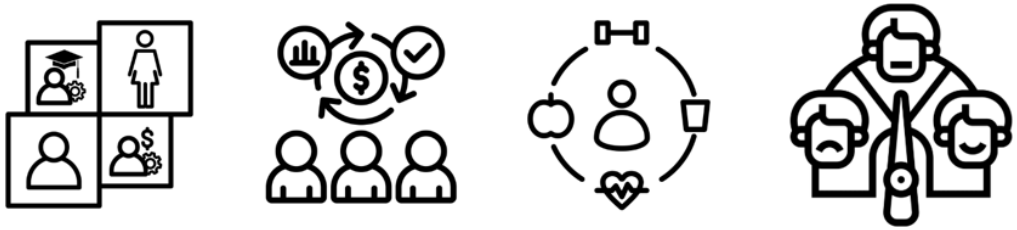

State of the art - Software

Synthpop R package

synthpop

(Nowok *et al.*, 2016)

1. **Social Diagnosis 2011 (SD2011):**



2. **Generation using `syn()`:**
   Original Data $\rightarrow$ Decision Tree

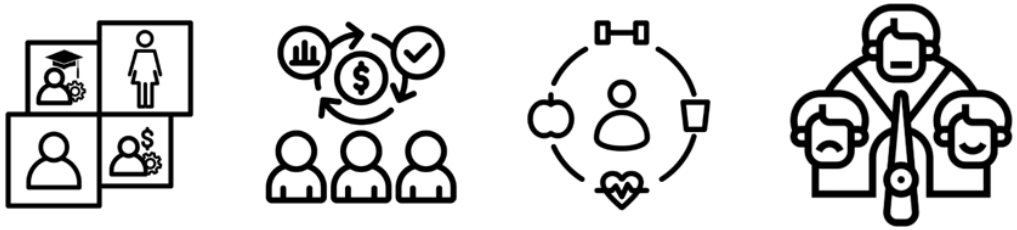
State of the art - Software

Synthpop R package

synthpop

(Nowok *et al.*, 2016)

1. **Social Diagnosis 2011 (SD2011):**



2. **Generation using `syn()`:**
   - Original Data → Decision Tree
   - Synthetic Data ← Randomization ↩

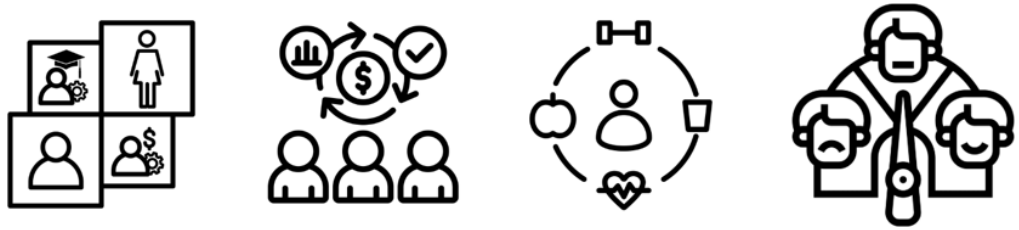
State of the art - Software

Synthpop R package

synthpop

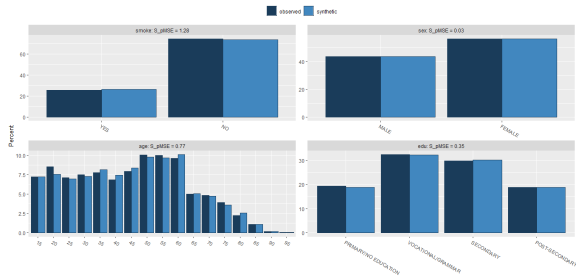
(Nowok *et al.*, 2016)

1. **Social Diagnosis 2011 (SD2011):**



2. **Generation using `syn()`:**
   - Original Data → Decision Tree
   - Synthetic Data ← Randomization ←

3. **Assessment (Snoke *et al.*, 2018):**



State of the art

Type of metrics

1. Resemblance (Raab *et al.*, 2021)

State of the art

Type of metrics

1. Resemblance (Raab *et al.*, 2021)
    - **Propensity score metrics**

State of the art

Type of metrics

1. Resemblance (Raab *et al.*, 2021)
    - **Propensity score metrics**
    - Contingency table metrics

State of the art

Type of metrics

1. Resemblance (Raab *et al.*, 2021)
    - **Propensity score metrics**
    - Contingency table metrics
2. Utility (Raab, 2022)

Propensity score metrics

Example: original data

Id	Cost	Region	Consumption
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

Propensity score metrics

Example: adding SD

Id	Cost	Region	Consumption	$\mathbb{I}_{\{0,1\}}$
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

Propensity score metrics

Example: propensity scores

Id	Cost	Region	Consumption	$\mathbb{I}_{\{0,1\}}$	$\hat{p}_i$
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 metrics

Kolmogorov-Smirnov Statistic (SPECKS)

Hypothesis Test

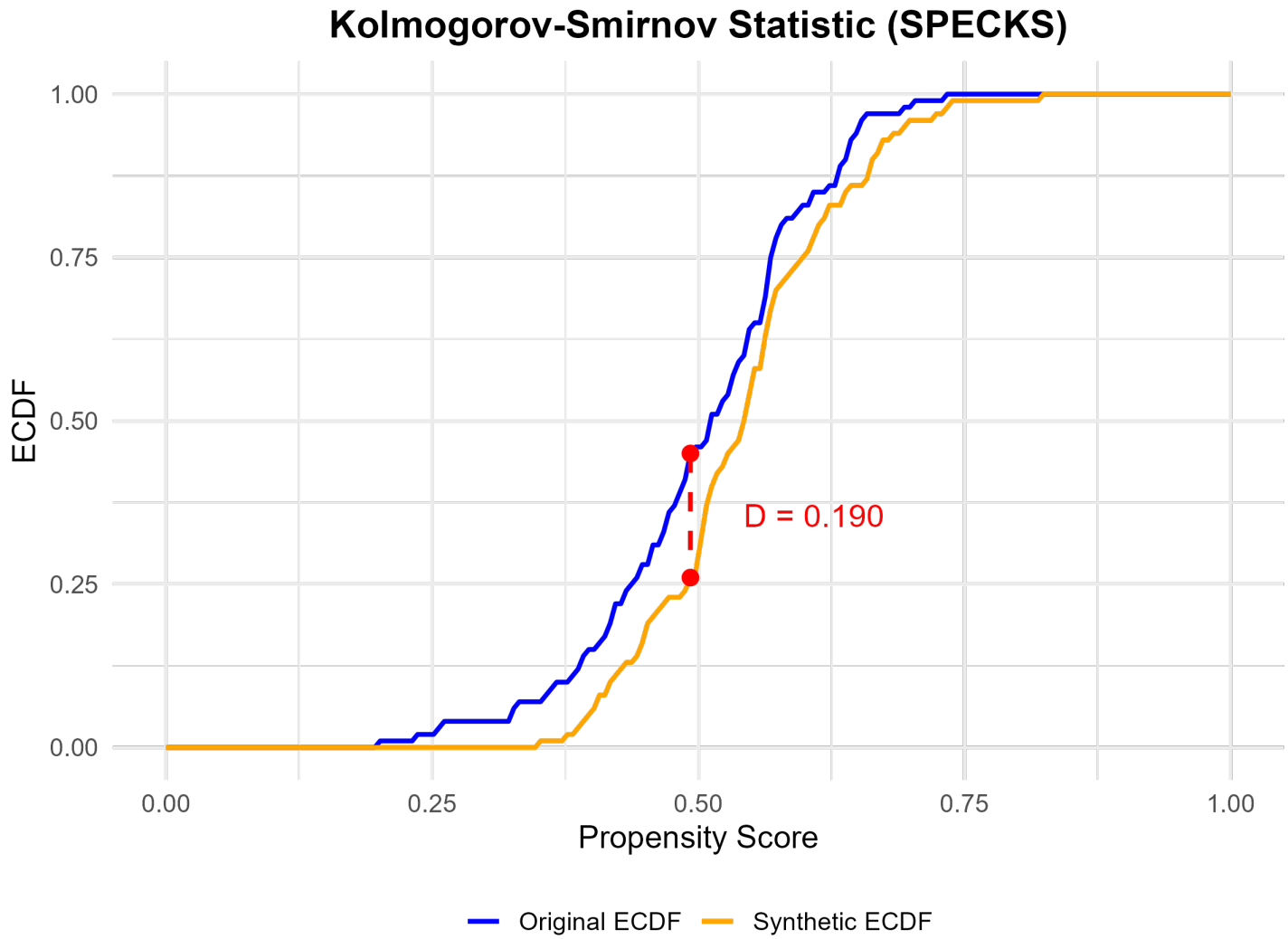
$$\begin{cases} H_0: & F^o(p) = F^s(p) \quad p \in [0,1] \\ H_1: & F^o(p) \neq F^s(p) \quad p \in [0,1] \end{cases}$$

SPECKS Statistic

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

Propensity score metrics

Example: SPECKS



PhD Objectives

PhD Objectives

1. **Adapt existing metrics to SD**

PhD Objectives

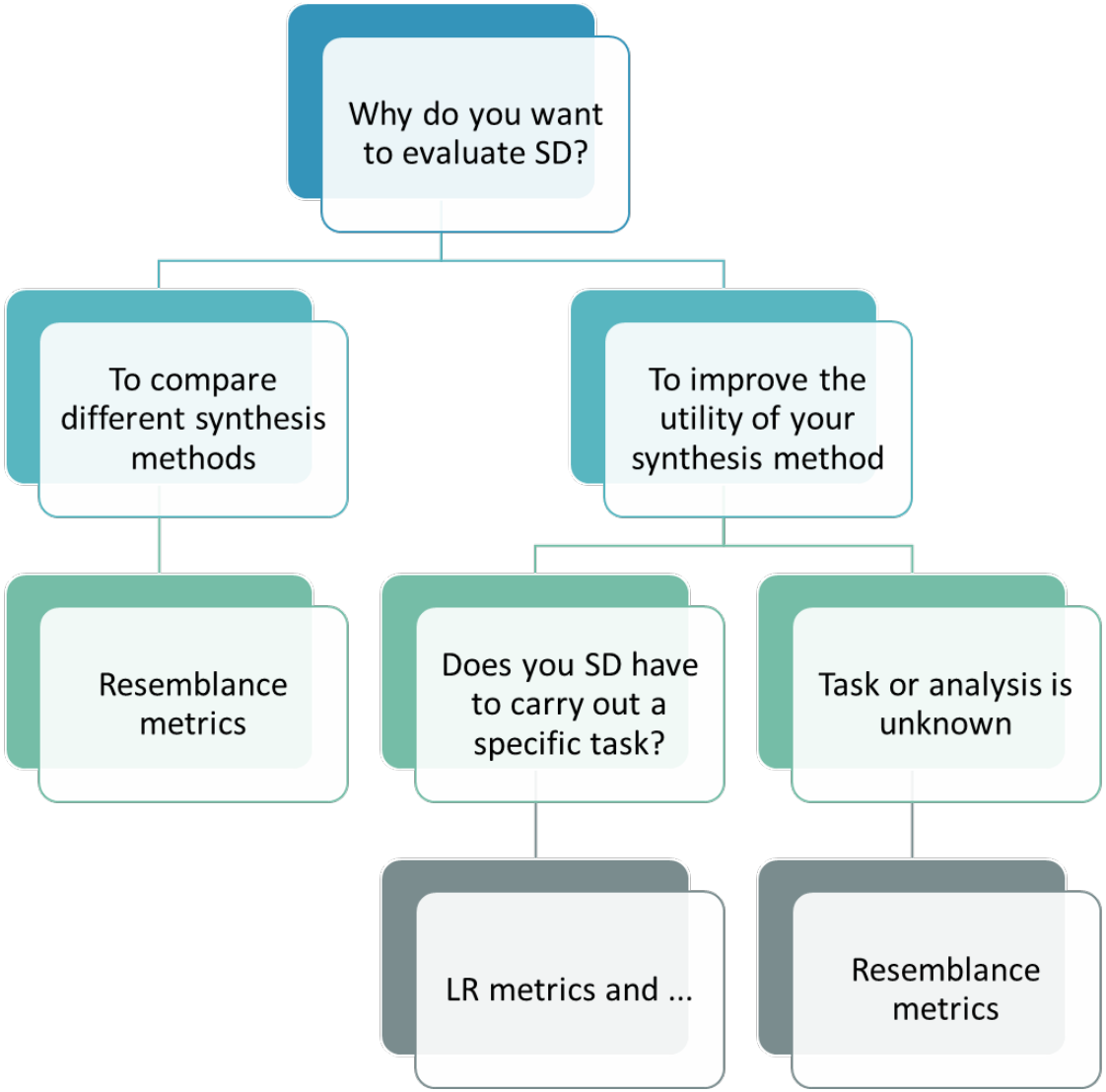
1. **Adapt existing metrics to SD**
2. **Evaluate reliability of metrics**

PhD Objectives

1. **Adapt existing metrics to SD**
2. **Evaluate reliability of metrics**
3. **Given an specific analysis, evaluate the resemblance-metrics' performance**

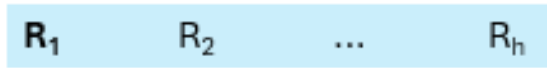
Methodology

1. Adapt existing metrics to SD



Methodology

2. Evaluate reliability of metrics



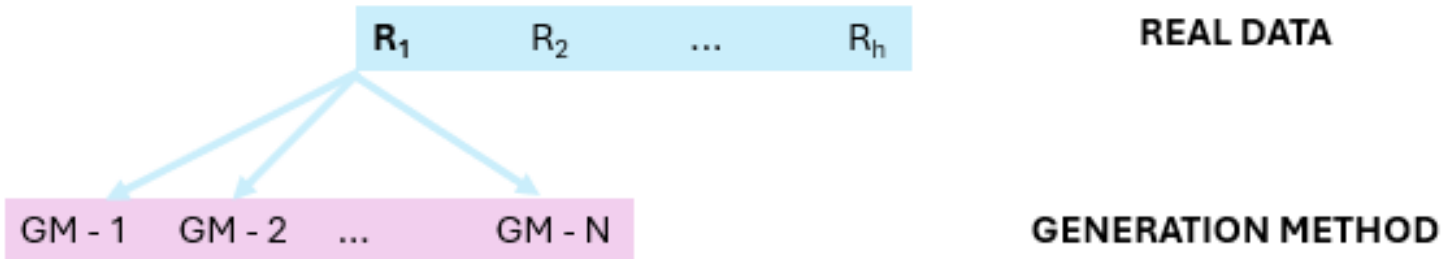
REAL DATA

SIMULATION STUDY

Generation algorithm
Randomness level

Methodology

2. Evaluate reliability of metrics

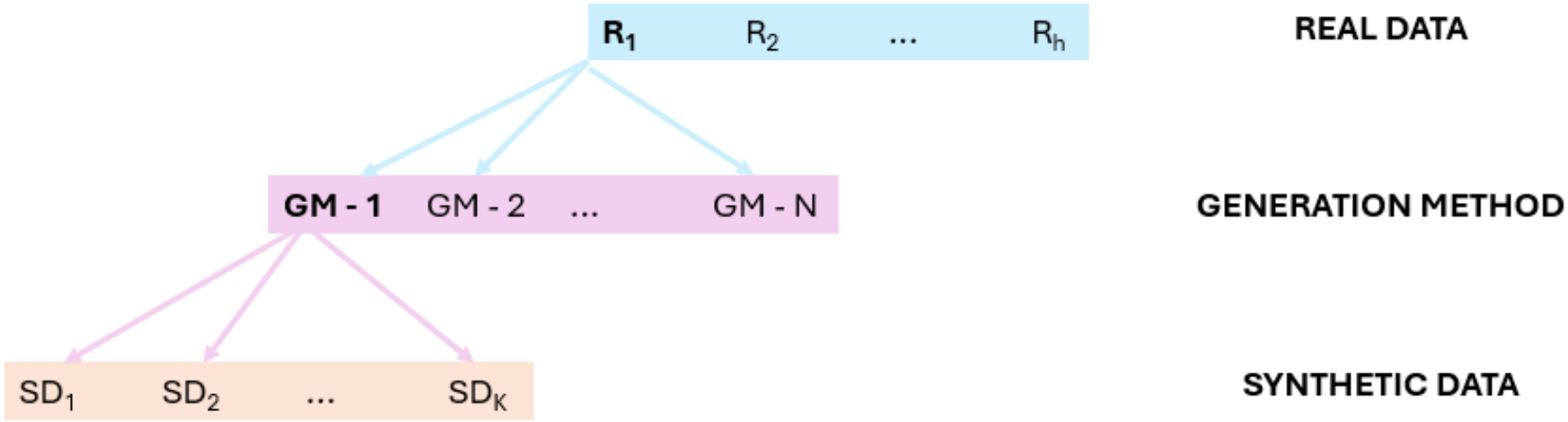


SIMULATION STUDY

Generation algorithm
Randomness level

Methodology

2. Evaluate reliability of metrics

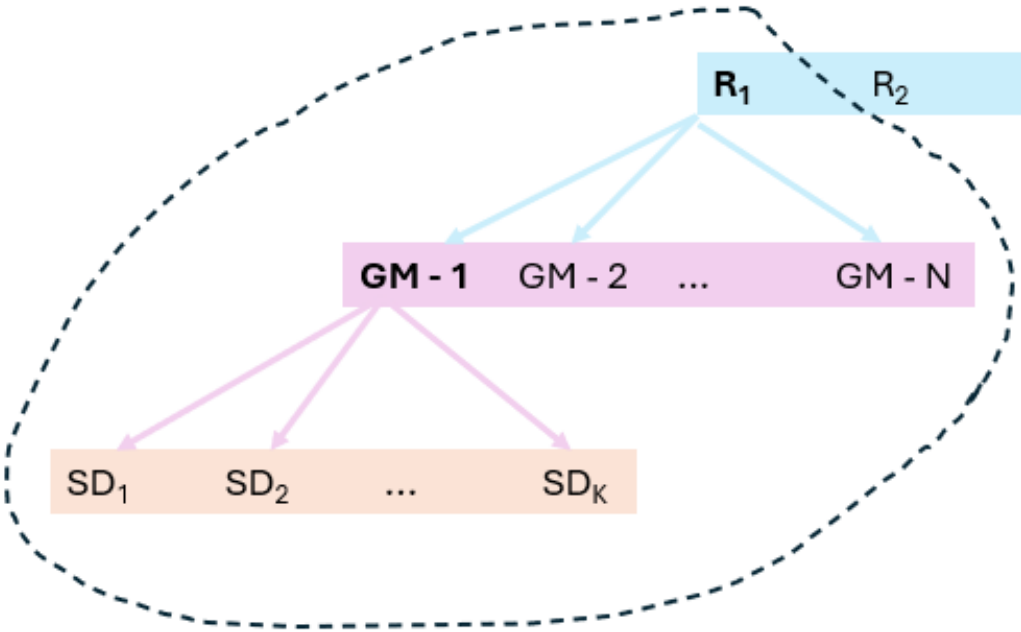


SIMULATION STUDY

Generation algorithm
Randomness level

Methodology

2. Evaluate reliability of metrics



REAL DATA

GENERATION METHOD

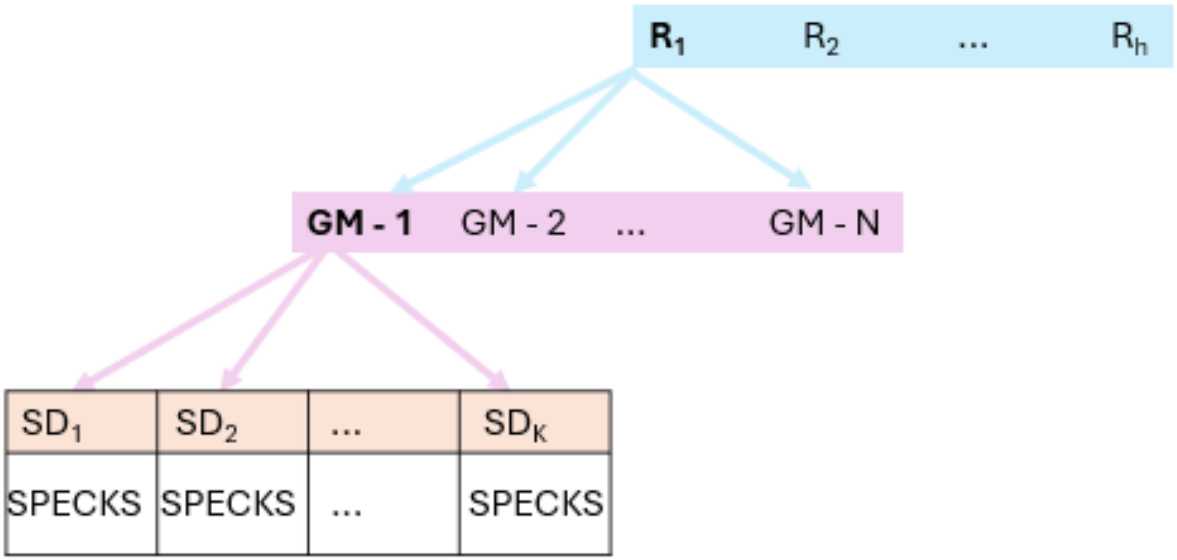
SYNTHETIC DATA

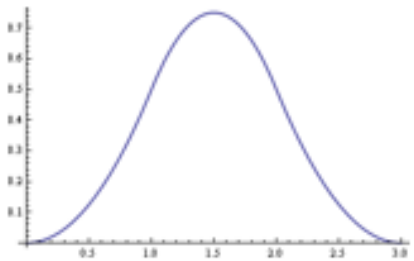
SIMULATION STUDY

Generation algorithm
Randomness level

Methodology

2. Evaluate reliability of metrics





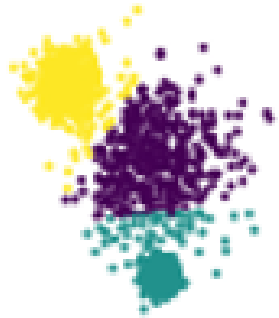
SIMULATION STUDY

Generation algorithm
Randomness level

Methodology

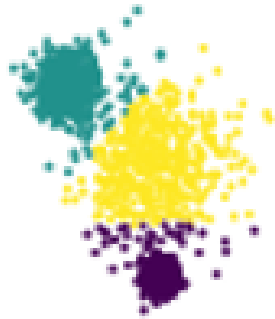
3. The most suitable metric for specific statistical analyses

REAL DATA



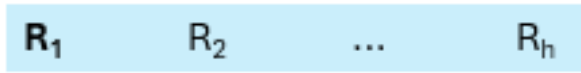
Do we get the same result?

SYNTHETIC DATA



Methodology

3. The most suitable metric for specific statistical analyses



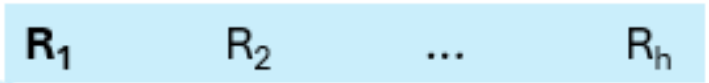
REAL DATA

SIMULATION STUDY

Proportion of data
Sample size
Variable type
Outliers
Missings

Methodology

3. The most suitable metric for specific statistical analyses



REAL DATA

SYNTHETIC DATA

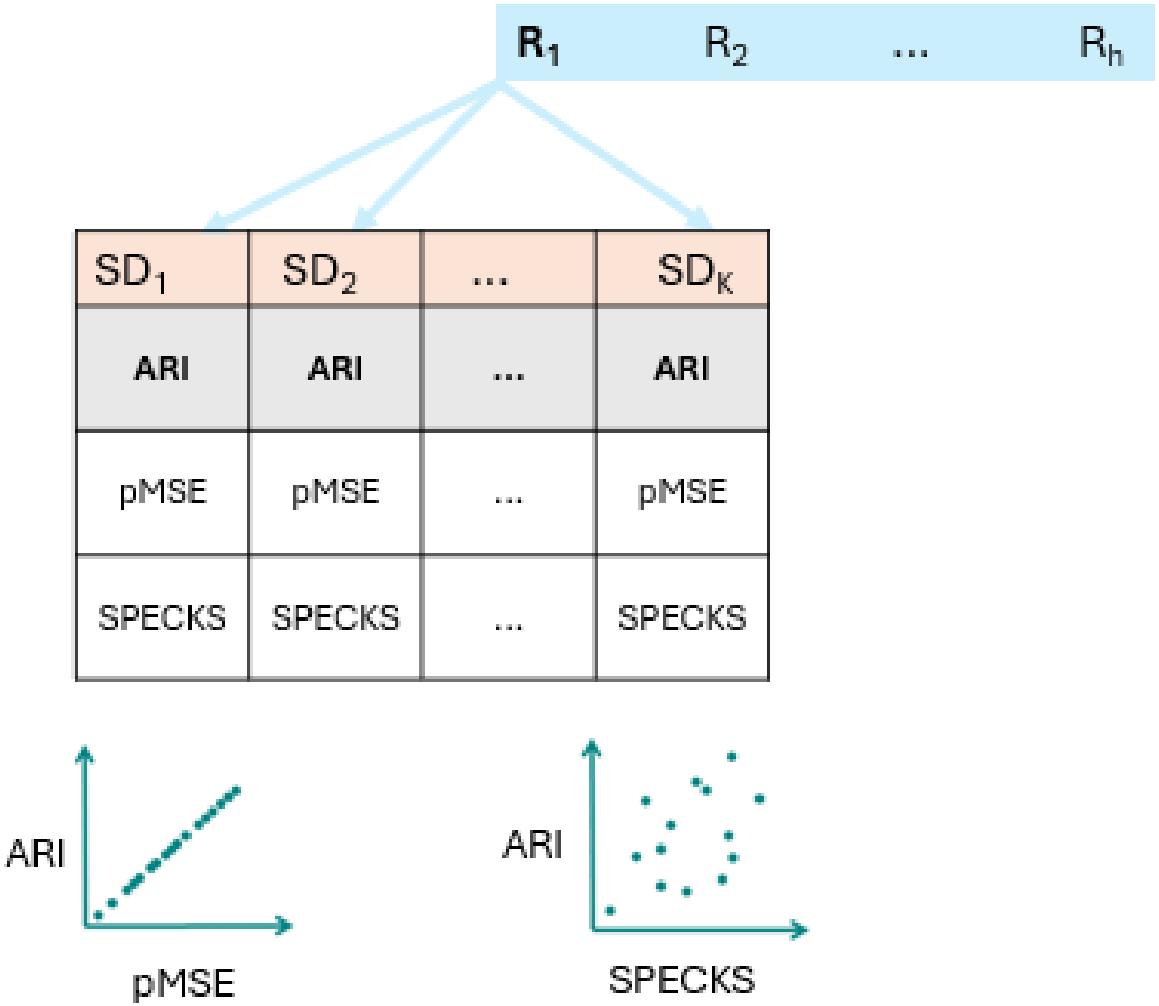
SD_1	SD_2	...	SD_K
ARI	ARI	...	ARI
pMSE	pMSE	...	pMSE
SPECKS	SPECKS	...	SPECKS

SIMULATION STUDY

Proportion of data
Sample size
Variable type
Outliers
Missings

Methodology

3. The most suitable metric for specific statistical analyses



SIMULATION STUDY

Proportion of data
Sample size
Variable type
Outliers
Missings

References

Nowok B., Raab G. M., Dibben C. (2016, 10). Synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software 74*. doi:10.18637/jss.v074.i11.

Raab G. M. (2022, 6). Utility and Disclosure Risk for Differentially Private Synthetic Categorical Data.

Raab G. M., Nowok B., Dibben C. (2021, 9). 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, 6). 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.

Questions?

Thank you for your attention!

I am happy to answer your questions.

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**.

¡Muchas gracias!