FROM REAL TO SYNTHETIC DATA ENSURING QUALITY

Nora Amama Ben Hassun[1], Jordi Cortés Martínez[1], and Daniel Fernández Martínez[1]

[1]Universitat Politècnica de Catalunya - BarcelonaTech (UPC)

What are Synthetic Data (SD)?

Synthetic data (SD) is artificially generated to **replicate the statistical properties** of real data. Its primary goal is to **keep the quality** of the real data while **mitigating privacy risks** by preventing the disclosure of sensitive information.

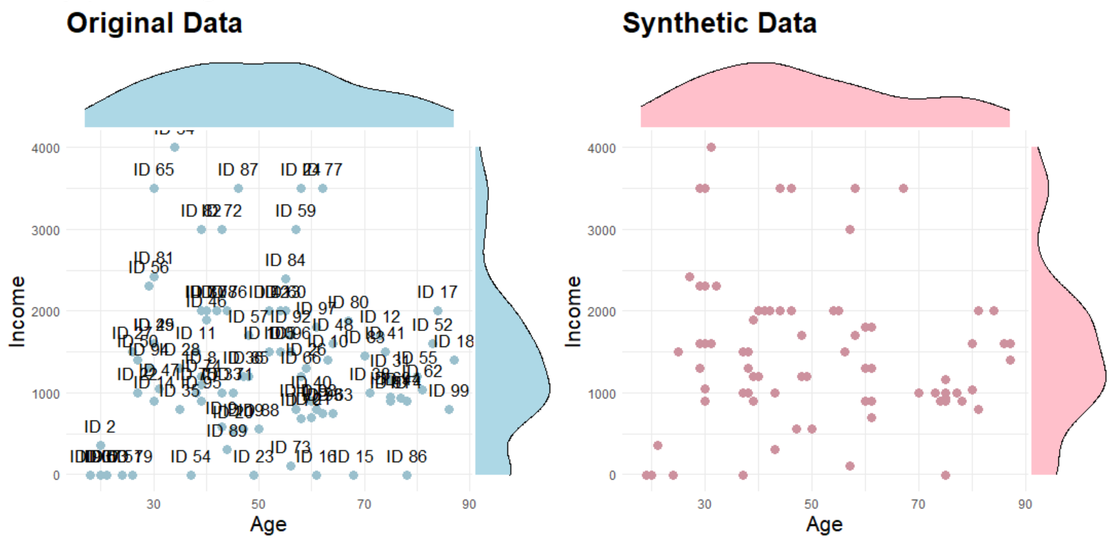

Fig. 1: Comparison of original and synthetic data distributions. The left panel shows the `Income` vs. `Age` distribution for the original dataset, while the right panel displays the synthetic dataset replicating its statistical properties. The synthetic data preserves distributions, ensures privacy, and maintains representativeness, with density overlays highlighting distributional similarity.

What is High-Quality SD?

High-quality synthetic data replicates the **distribution** of the original data, is suitable for **specific analyses**, and ensures no identification of **individuals** or **sensitive variables**.

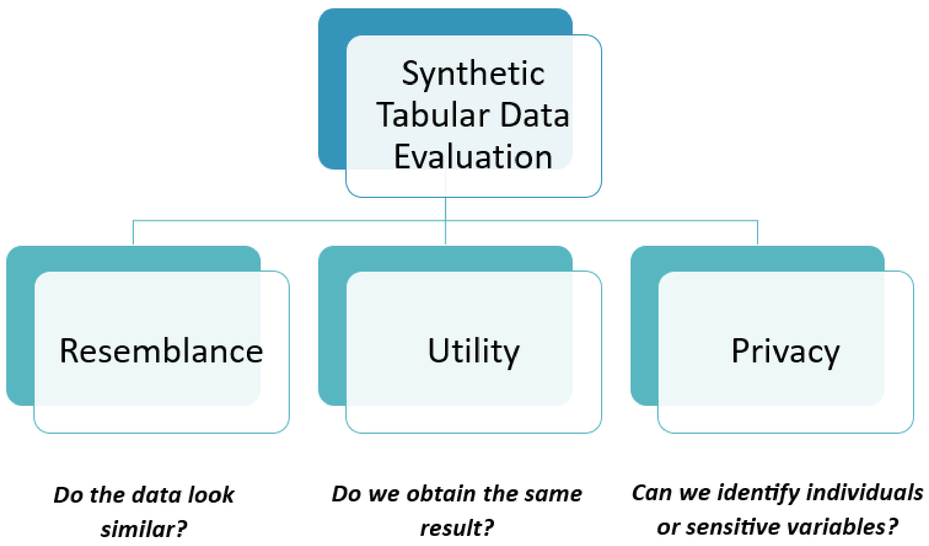

Fig. 4: Key criteria for evaluating synthetic tabular data: Resemblance assesses how similar the synthetic data is to the original; Utility evaluates whether the synthetic data produces comparable results in analysis; and Privacy ensures that sensitive information or individuals cannot be identified.

How are they Generated?

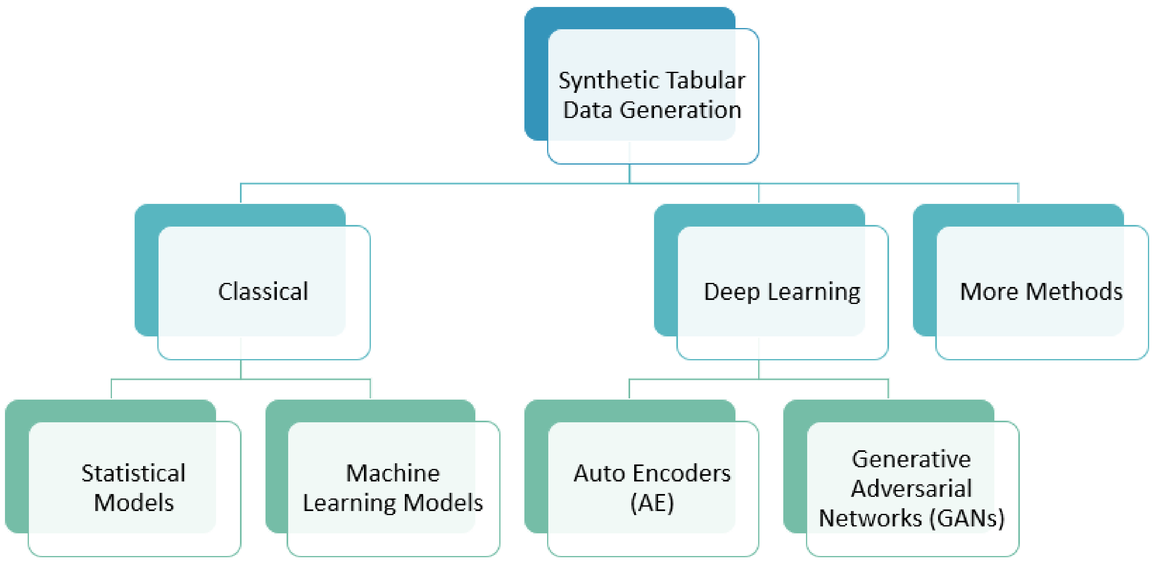

Fig. 2: Hierarchy of methodologies for synthetic tabular data generation. Classical methods, such as Statistical Models and Machine Learning Models, focus on traditional approaches, while Deep Learning techniques, including Auto Encoders (AE) and Generative Adversarial Networks (GANs). Additional methods fall under 'More Methods,' showcasing diverse strategies to meet specific requirements

Existing Validation Metrics for SD

Various validation metrics are implemented in the `synthpop` package, including resemblance metrics such as Propensity Score Mean-Squared Error (**pMSE**), Kolmogorov-Smirnov Statistic (**SPECKS**), and Voas-Williamson Utility Measure (**VW**), as well as utility metrics like Confidence Interval Overlap (**CIO**). These metrics evaluate the similarity and utility of synthetic data.

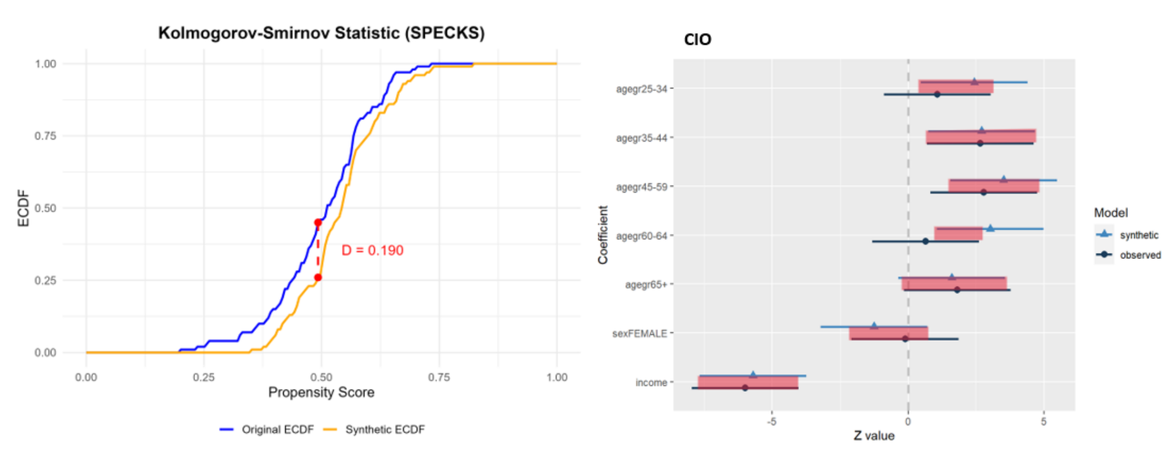

Fig. 5: Validation metrics applied to assess resemblance and utility in synthetic data. Left: Kolmogorov-Smirnov statistic (SPECKS). Right: Z-values comparing synthetic and observed coefficients.

Applications in the Energy Sector

Utilities:

- **Reduce the risk** of exposing sensitive data.
- **Accelerate model development** and validation.
- Ensure compliance with **energy data regulations**.

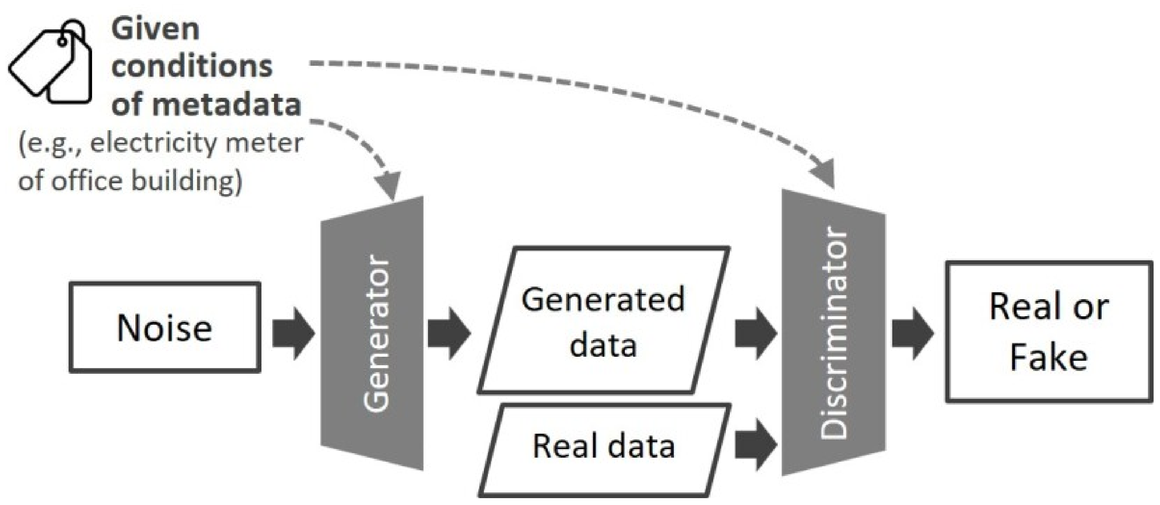

Fig. 3: Synthetic data generated for energy consumption analysis using GANs, ensuring privacy and maintaining statistical accuracy.

Future Steps

- Collect **existing metrics** and incorporate **measures from other disciplines** into synthetic data evaluation.
- Assess which **metrics** are more **reliable** for **data generation methods** and determine the **best metric** for specific **statistical analyses** (utilities mesures).
- Evaluate **missing data handling**: Identify which **imputation methods** perform best depending on **several factors**.

References

Apellániz P. A., Jiménez A., Arroyo Galende B., Parras J., Zazo S. (2024, 5). Synthetic Tabular Data Validation: A Divergence-Based Approach. *IEEE Access 12*, 103895–103907. URL: `https://ieeexplore.ieee.org/document/10613395/`, doi:10.1109/ACCESS.2024.3434582.

Murtaza H., Ahmed M., Khan N. F., Murtaza G., Zafar S., Bano A. (2023, 5). Synthetic data generation: State of the art in health care domain. *Computer Science Review 48*. doi:10.1016/j.cosrev.2023.100546.

Osorio-Marulanda P. A., Epelde G., Hernandez M., Isasa I., Reyes N. M., Iraola A. B. (2024). Privacy mechanisms and evaluation metrics for Synthetic Data Generation: A systematic review. *IEEE Access*. doi:10.1109/ACCESS.2024.3417608.

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`.

Financiado por la Unión Europea NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

España | digital 2026