Super learner for tabular synthetic data generation

Journées des biostatistiques - Montpellier 2025

Marie-Félicia Béclin, Joint work with Julie Josse & Jean-Baptiste Woillard

4 novembre 2025

Introduction

Introduction : Sensitive Medical Data

Privacy and Medical Data

- **In health dataset medical information are sensitives**, data breaches.
- **General regulations** (e.g., adoption of GDPR by the EU in 2018, HIPAA in USA), data cannot be kept after a certain delay and can be shared under strict conditions.

Introduction : Current Solutions

Pseudonymisation and Anonymisation

- **Pseudonymisation** (replacing directly identifiable data, e.g. names with codes). Pseudonymisation remains vulnerable to linkage attacks.
- **Anonymisation** : irreversible removal of data identifiability. Fully anonymized data are the goal, can be shared according to GPDR but can lose utility - See Supplementary Material.

Federated learning with Differential Privacy (DP)

Promising approach but have some drawbacks

- **Federated learning** requires specific architectural configurations,
- **Differential Privacy**[1] considered the gold standard for privacy protection, necessitates a DP-compliant version of each ML or stat. algorithm used.

Synthetic data

- The goal is to enable data sharing with a better utility than anonymization, especially for educational or open science purposes.

1. Dwork, "Differential privacy." In *Proceedings of the International Colloquium on Automata, Languages, and Programming*, pages 1–12, 2006. Springer.

Privacy versus Utility

Privacy has a cost on the utility of the analysis, ideally it should not destroy it.

Utility

Utility is a measure of how well synthetic data retains the statistical properties and practical usefulness of real data :

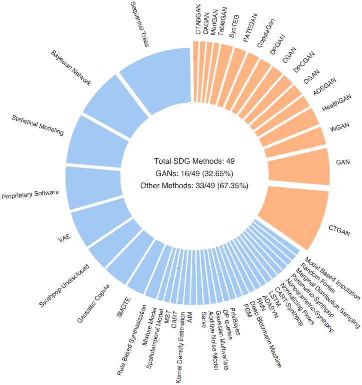
- Statistical utility : Synthetic data should preserve all the statistical properties of the original data (marginal and joint distribution).
- Task-based utility : regression's coefficients, supervised learning, etc.

Privacy

- Privacy can be defined as "the ability of an individual to withhold information about themselves".
- The goal is to preserve the right of individuals to control how their personal data is shared and to ensure that it is kept secure and only accessed by authorized parties.

Existing methods

The Rise of Synthetic Data Generation with GANs



Synthetic data generation methods [2]

49 different synthetic data generation methods : GANs (32.65 % of the total) and other techniques reports in that figure in a review based on 92 studies. They founded 48 utility metrics and 9 methods to evaluate privacy.

2. Kaabachi, Bayrem, et al. "A scoping review of privacy and utility metrics in medical synthetic data." NPJ digital medicine 8.1 (2025)

Concurrent Methods

In our analysis, we further categorize these into :

- **Synthpop** : One of the first packages available (in R) since 2016 [3].
- **Avatar** : A SMOTE-like method based on k-nearest neighbors [4]. Octopize startup, not free, re-implemented using custom Python code.
- **Conditional Tabular Generative Adversarial Network (CTGAN)** : A more recent approach, available in several Python libraries, with support for various data structures (longitudinal, survival, etc.).
  - CTGAN from Synthcity [5].
  - Synthetic Data Vault [6].
  - TabGAN [7]

---

3. Nowok et al., "synthpop : Bespoke creation of synthetic data in R." *Journal of Statistical Software*, 74 :1–26, 2016.
4. Guillaudeux et al., "Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis." *NPJ Digital Medicine*, 6(1) :37, 2023. Nature Publishing Group UK, London.
5. Qian et al., "Synthcity : a benchmark framework for diverse use cases of tabular synthetic data." *Advances in Neural Information Processing Systems*, 36 :3173–3188, 2023.
6. Patki et al., "The Synthetic Data Vault." In *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, October 2016. doi : 10.1109/DSAA.2016.49.
7. Ashrapov, I. (2020). Tabular GANs for uneven distribution. arXiv preprint arXiv :2010.00638. ; https ://github.com/Diyago/Tabular-data-generation/tree/master

Synthpop Method [9]

The **Synthpop** approach models the joint distribution of variables and then samples from it sequentially.

Algorithm : Synthpop Generation of a Sample

1: Draw $x_{\text{syn}}^{(1)}$ from the empirical distribution of $\{x_{\text{true},1}^{(1)}, \ldots, x_{\text{true},n}^{(1)}\}$
2: **for** $k = 2$ to p **do**
3: Draw $x_{\text{syn}}^{(k)} \sim \widehat{P}\left(X^{(k)} \mid x_{\text{syn}}^{(1)}, \ldots, x_{\text{syn}}^{(k-1)}\right)$
4: **end for**

CART is the default model. Although it is not inherently distributional, it approximates draws from a conditional distribution by sampling from the leaves [8].

Pros and Cons :

- **+ Flexible** : Allows different models per variable and supports known data structure.
- **– Sensitive to variable order** : The generation process depends on variable sequencing, which may affect synthetic data quality.

8. Future work may explore distributional tree models such as DRF : Cevid, Domagoj, et al. "Distributional random forests : Heterogeneity adjustment and multivariate distributional regression." *Journal of Machine Learning Research* 23.333 (2022) : 1-79.
9. Nowok et al., "synthpop : Bespoke creation of synthetic data in R." *Journal of Statistical Software*, 74 :1–26, 2016.

Avatar [11]

1. **Normalization** and projection using **Principal Component Analysis (PCA)** to reduce the dimensionality of the dataset.
2. Apply the k-**nearest neighbors algorithm** to the PCA-transformed dataset to select the k neighbors $x_i^1, \ldots, x_i^k$ of each original sample x_i.
3. Generate synthetic data in the latent space using : $\tilde{x}_i = \frac{\sum_{j=1}^{k} P_i^j x_i^j}{\sum_{j=1}^{k} P_i^j}$, where P_i^j is the weight attributed to neighbor x_i^j of the original sample x_i.
4. The weight is computed as : $P_i^j = \frac{1}{d_i} \times R_i^j \times C_i^j$, where :
   - $d_i = d(x_i, x_i^j)$ is the distance between x_i and its neighbor x_i^j,
   - $R_i^j \sim \mathcal{E}xp(1)$ is a random weight following an exponential distribution,
   - $C_i^j = \left(\frac{1}{2}\right)^{\sigma(j)}$ is a contribution term, where σ is a random permutation in $\mathcal{S}_k$, the space of bijections over $\{1, \ldots, k\}$.

Avatar can handle **categorical features** and **missing data** using Multiple Correspondence Analysis (MCA), Factorial Analysis of Mixed Data (FAMD) and matrix completion methods (iterative SVD) [10].

10. Josse, J. and Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. Journal de la société française de statistique, 153(2), 79–99. - Avatar creators are former student of F. Husson.

11. Guillaudeux et al., "Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis." *NPJ Digital Medicine*, 6(1) :37, 2023. Nature Publishing Group UK, London.

CTGAN [12]

- **Generator** : Receives a random noise vector and conditional vector. Outputs synthetic data rows.
- **Discriminator** : Takes real and synthetic data as input. Outputs probability of data being real.
- **Conditional Vector** : Ensures that generated data matches the distribution of real data. Helps in handling imbalanced discrete columns.

Training process :

- Sample a batch of real data and corresponding conditional vectors.
- Generate synthetic data using the generator.
- Train the discriminator to distinguish between real and synthetic data.
- Train the generator to fool the discriminator.
- Repeat until convergence.

12. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K. : Modeling tabular data using conditional gan. In : Advances in Neural Information Processing Systems (2019)

Energy distance [13]

We suppose that your original data X_i are i.i.d copies from $X \sim P$, we generate synthetic data from a distribution H.

$$\begin{aligned} d(H,P) &= 2\mathbb{E}_{Y\sim P, X\sim H}(||Y-X||_2) - \mathbb{E}_{Y\sim P, Y'\sim P}(||Y-Y'||_2) \\ &\quad -\mathbb{E}_{X\sim P, X'\sim P}(||X-X'||_2). \end{aligned}$$

Wasserstein distance

$$W_2(P,H) = \left(\inf_{\gamma\in\Pi(P,H)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^2 \, d\gamma(x,y) \right)^{1/2}.$$

13. Székely, "E-statistics : The energy of statistical samples." *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05) :1–18, 2003.

Tasked Based / Inferential utility

- Distance from the parameter from the value of the original dataset
- Coverage : Proportion of times that the confidence intervals of the synthetic data include the true values (possible in simulations)
- Overlap of confidence interval of estimators (Drechsler and Reiter , 2009 [14] ; Nowok, 2015 [15])

14. Drechsler, J., and Reiter, J. P. (2009). Disclosure risk and data utility for partially synthetic data : An empirical study using the German IAB Establishment Survey. Journal of Official Statistics, 25(4), 589.
15. Nowok, B. (2015). Utility of synthetic microdata generated using tree-based methods. UNECE Statistical Data Confidentiality Work Session, 1-11.
16. Decruyenaere et al "The real deal behind the artificial appeal : Inferential utility of tabular synthetic data." *arXiv preprint* arXiv :2312.07837, 2023.

Privacy : Metrics based on distances between sample from the original data $\mathcal{X}$ and the synthetic data $\mathcal{S}$

Distance to Closest Record (DCR)

Measures the Euclidean distance between each synthetic record and its closest real record.

$$DCR(s) = \min_{x \in \mathcal{X}} d(s, x)$$

The metric can be 5th percentile of the vector (Zhao et al. (2021)[17], or the mediane Guillaudeux et al. (2023)[18] and Kotelnikov et al. (2022)[19]).

Nearest Neighbor Distance Ratio (NNDR)

Ratio between the distance to the closest and the second closest real record for each synthetic record $s \in \mathcal{S}$.

$$\mathrm{NNDR}(s) = \frac{d\left(s, x_{\text{nearest}}\right)}{d\left(s, x_{\text{second-nearest}}\right)}$$

17. Zhao, Z., Kunar, A., Birke, R., Chen, L. Y. (2021, November). Ctab-gan : Effective table data synthesizing. In Asian Conference on Machine Learning (pp. 97-112). PMLR.
18. Guillaudeux et al., "Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis." *NPJ Digital Medicine*, 6(1) :37, 2023. Nature Publishing Group UK, London.
19. Kotelnikov, A., Baranchuk, D., Rubachev, I., Babenko, A. (2023, July). Tabddpm : Modelling tabular data with diffusion models. In International Conference on Machine Learning (pp. 17564-17579). PMLR.

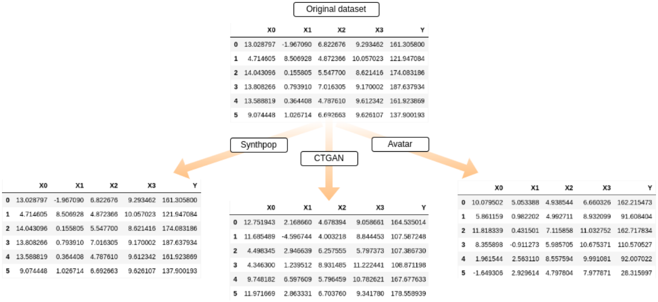
Superlearner ideas

Purpose of a Superlearner

Motivation :

- Different synthetic data generation methods perform better on different types of data structures.
- Combining them can potentially improve both utility and privacy.

Goal : Leverage multiple synthetic datasets to build a more robust and privacy-preserving synthetic dataset.



	X0	X1	X2	X3	Y
0	13.028797	-1.967090	6.822676	9.293462	161.305800
1	4.714605	8.506928	4.872366	10.057023	121.947084
2	14.043096	0.155805	5.547700	8.621416	174.083186
3	13.808266	0.793910	7.016305	9.170002	187.637934
4	13.588819	0.364408	4.787610	9.612342	161.923869
5	9.074448	1.026714	6.692663	9.626107	137.900193

	X0	X1	X2	X3	Y
0	13.028797	-1.967090	6.822676	9.293462	161.305800
1	4.714605	8.506928	4.872366	10.057023	121.947084
2	14.043096	0.155805	5.547700	8.621416	174.083186
3	13.808266	0.793910	7.016305	9.170002	187.637934
4	13.588819	0.364408	4.787610	9.612342	161.923869
5	9.074448	1.026714	6.692663	9.626107	137.900193

	X0	X1	X2	X3	Y
0	12.751943	2.168660	4.678394	9.058661	164.535014
1	11.685489	-4.596744	4.003218	8.844453	107.587248
2	4.498345	2.946639	6.257555	5.797373	107.386730
3	4.346300	1.239512	8.931485	11.222441	108.871198
4	9.748182	6.597609	5.796459	10.782621	167.677633
5	11.971669	2.863331	6.703760	9.341780	178.558939

	X0	X1	X2	X3	Y
0	10.079502	5.053388	4.938544	6.660326	162.215473
1	5.861159	0.982202	4.992711	8.932099	91.608404
2	11.818339	0.431501	7.115858	11.032752	162.717834
3	8.355898	-0.911273	5.985705	10.675371	110.570527
4	1.961544	2.563110	8.557594	9.991081	92.007022
5	-1.649306	2.929614	4.797804	7.977871	28.315997

Figure 1 – Superlearner idea

Superlearner ideas

Aggregation Combination Strategies :

- **STATIS-based Approaches :**
  - Dual STATIS
- **Barycenter Methods :**
  - **MMD Barycenter** : based on maximum mean discrepancy in kernel space

Superlearner ideas

- **Original Dataset** :
  - $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, with n samples from $\mathbb{R}^p$.
  - The data table is denoted as $\boldsymbol{X}$.

- **Synthetic Datasets** :
  - K methods generate K datasets : $\mathcal{S}_i = \{s_1^{(i)}, \ldots, s_n^{(i)}\}$
  - Each dataset has p variables (columns).
  - Each synthetic dataset has n samples same as the original.
  - Samples $s_j^{(i)}$ are not necessarily derived from x_j.
  - Synthetic datasets : $\boldsymbol{S}^{(1)}, \ldots, \boldsymbol{S}^{(K)}$.
- Weights $\delta = (\delta_1, \ldots, \delta_K)$; Δ is a $K \times K$ diagonal matrix of weights for each table.

Dual Statis [20]

- Find a trade-off table that "summarizes" the synthetic data tables $\boldsymbol{S^{(1)}}, \ldots, \boldsymbol{S^{(K)}}$ represent the data tables.
- For each table l, compute the scalar product matrix :

$$V^{(l)} = (S^{(l)})^\top S^{(l)} \in \mathbb{R}^{p \times p}, \quad V_{ij}^{(l)} = \langle C_i^{(l)}, C_j^{(l)} \rangle$$

- $V^{(l)}$ **is a symetric correlation matrix between column** $p \times p$**,** $V_{i,j}^{(l)} = V_{j,i}^{(l)} = corr(\textbf{variable i}, \textbf{variable j})$ (because we previously centered and scaled datatables)

20. Lavit, C., Escoufier, Y., Sabatier, R., Traissac, P. (1994). The act (statis method). Computational Statistics Data Analysis, 18(1), 97-119.

Dual Statis : How to aggregate correlation matrix V ?

- Different way to define a "trade-off table"
  - Weighted mean : $\bar{V} = \frac{1}{K}\sum_{i=1}^{K} \delta_i V_i$
- Maximization of the weighted sum of Hilbert-Schmidt similarities, $\bar{V} = \sum_{l=1}^{K} \gamma_l V^{(k)}$ with $\|\gamma\|^2 = 1$. The optimal γ solves the maximization problem :

$$\gamma^* = \arg\max_{\|\gamma\|^2=1} \sum_{k=1}^{K} \delta_k \left\langle \sum_{l=1}^{K} \gamma_l V^{(l)}, V^{(k)} \right\rangle_{\mathrm{HS}}^2$$

- Rayleigh quotient maximization problem
- $\Omega_{k,l} = \mathrm{trace}(V^{(k)} V^{(l)})$.
- τ_1 : first normalized eigenvector of $\Omega\Delta$ (in decreasing order of the eigenvalues), with all coordinates positive [21].
- Weighted mean with eigenvector τ : $\bar{V} = \frac{1}{K}\sum_{i=1}^{K} [\tau_1]_i V_i$

21. This eigenvalue always exists according to Frobenius theorem

How to Reconstruct the Table $\bar{S}$

Singular Value Decomposition

- For each data table $S^{(l)}$, we assume the decomposition :

$$S^{(l)} = Q^{(l)}\Sigma^{(l)}P^{(l)\top}$$

where $\Sigma^{(l)}$ is diagonal.

- Then, define matrix :

$$V^{(l)} = S^{(l)\top}S^{(l)} = P^{(l)}\Lambda^{(l)}P^{(l)\top}$$

$$W^{(l)} = S^{(l)}S^{(l)\top} = Q^{(l)}\Gamma^{(l)}Q^{(l)\top}$$ [22]

where $\Lambda^{(l)} = \Sigma^{(l)\top}\Sigma^{(l)}$ $\Gamma^{(l)} = \Sigma^{(l)}\Sigma^{(l)\top}$ and is diagonal.

- Our goal is to reconstruct the matrix $\bar{S}$ such that :

$$\begin{aligned} \bar{V} &= \bar{S}^\top\bar{S} \\ \bar{S} &= \bar{Q}\bar{\Sigma}\bar{P}^\top \end{aligned}$$

We can then find $\bar{P}$ and $\bar{\Sigma}$ from its eigendecomposition.

22. The original STATIS method interprets the W matrix as being tailored for multiple data tables comprising the same set of individuals, each characterized by different variables.

Constructing $\bar{S}$ and Basis Alignment

- We need to find a matrix $\bar{Q}$, so for that we use the SVD from the best synthetic datatable $S^{(l^*)}$:

$$W^{(l^*)} = S^{(l^*)} S^{(l^*)\top} = Q \Gamma Q^\top$$ [23]

- Define :

$$\bar{\Sigma} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\lambda_p} \\ 0 & 0 & \dots & 0 \end{bmatrix} \in \mathbb{R}^{n \times p}$$

- Final trade-off matrix :

$$\bar{S} = Q \bar{\Sigma} \bar{P}^\top$$

23. An alternative is to use the trade-off matrix $\bar{W}$, capturing correlations between individuals as in the original STATIS framework. However, this assumes that individual i in a synthetic dataset represents the same real individual i, which generally does not hold except in the Avatar scenario. This approach remains feasible if synthetic datasets are appropriately paired

A basis mismatch remains, as illustrated in Figure 2. To address this, we need to change the basis and identify the optimal orthogonal transformation that aligns the synthetic data table $\bar{\boldsymbol{S}}$ as closely as possible with $\bar{\boldsymbol{X}}$.

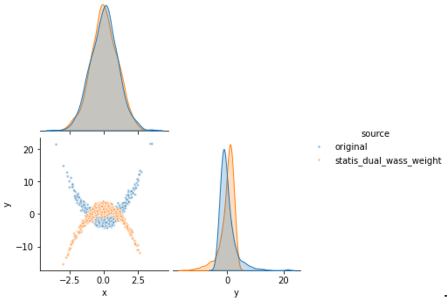


Figure 2 – Example of STATIS

Orthogonal Procrustes Problem

- Given $\boldsymbol{X}, \bar{S} \in \mathbb{R}^{n \times p}$, find $T \in \mathcal{O}(p)$ that minimizes :

$$\min_{T \in \mathcal{O}(p)} \| T\bar{S} - \boldsymbol{X} \|_F$$

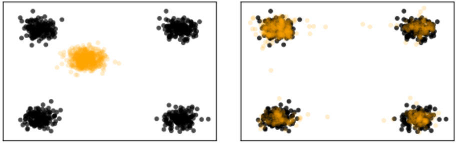
- **Solution (Schönemann, 1966) :**
  1. Compute $A = \boldsymbol{X}\bar{\boldsymbol{S}}^\top$.
  2. Perform SVD : $A = U\Sigma V^\top$.
  3. The optimal T is $T = UV^\top$.

This T is the best orthogonal matrix (in Frobenius norm) aligning $\boldsymbol{X}$ with $\bar{\boldsymbol{S}}$.

Definition : The *Fréchet mean* (or barycenter) of a set of probability distributions $\{\mu_1, \ldots, \mu_K\}$ with weights $\beta = (\beta_1, \ldots, \beta_K)$ is the distribution $\bar{\mu}$ minimizing the weighted sum of squared distances :

$$\bar{\mu} \in \arg\min_{\mu} \sum_{i=1}^{K} \beta_i \, d^2(\mu, \mu_i)$$

where d is a distance between probability distributions (e.g., Wasserstein or MMD).



(a) Wasserstein (b) MMD

Figure 3 – Barycenter (orange) of four Gaussians (black) with respect to (a) W ; (b) MMD. Best-left Gaussian has three times the weight of the others : $\beta = [3/6, 1/6, 1/6, 1/6]$.. [24]

24. Cohen et al., "Estimating barycenters of measures in high dimensions." *arXiv preprint arXiv :2007.07105*, 2020.

Proposition (MMD Barycenter) [25] **:**
Let $\mu_1, \dots, \mu_P \in \mathcal{M}_1^+(\mathcal{X})$ be probability measures and let $\beta \in \Delta_P$ be a weight vector (i.e., $\beta_p \geq 0, \sum_p \beta_p = 1$). If the discrepancy $D = \mathrm{MMD}^2$, then the barycenter μ^* is given by :

$$\mu^* = \sum_{p=1}^{P} \beta_p \mu_p \in \mathcal{M}_1^+(\mathcal{X})$$

The MMD barycenter is simply the weighted *mixture* of the input measures.

Sampling Procedure :

1. Sample index $z \sim \mathrm{Categorical}(\beta)$
2. Draw a sample $x \sim \mu_z$

Implication : Sampling from the MMD barycenter is computationally trivial — no optimization is needed.

25. Cohen et al., "Estimating barycenters of measures in high dimensions." *arXiv preprint arXiv :2007.07105*, 2020.

Simulations

- Linear dataset with low correlation (called :"lowcorr").
  - The original dataset is simulated as $X_i \overset{\text{iid}}{\sim} \mathcal{N}(m, \Sigma)$.
  - $Y = X\beta + \epsilon$
- Linear dataset with high correlation (called :"highcorr")
  - Same process but with different correlation value.
- Non-linear dataset
  - $X_0 \sim \mathcal{N}(0, 1) \quad$ (500 samples)
  - $X_1 \sim \mathcal{N}(0.9, 0.51^2) \quad$ (500 samples)
  - $X_2 = (X_1 - 0.9)^2 + \epsilon_2, \quad \epsilon_2 \sim \mathcal{N}(0, 0.25^2)$
  - $X_3 \sim \chi^2(6) + \exp(X_0)$
  - $Y = 2.5 \cdot \exp\left(-1.3X_0 - 2X_1 - 1.2X_2 - 0.03X_3\right) + \epsilon_y, \quad \epsilon_y \sim \mathcal{N}(0, 0.02^2)$
- Breast Cancer dataset [a]

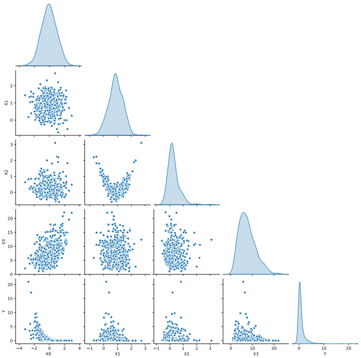


Figure 4 – Pairplot NLS dataset

a. Wolberg, W., Mangasarian, O., Street, N., and Street, W. (1993). Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository.

Evaluating Superlearners Using Best Methods - Energy distance

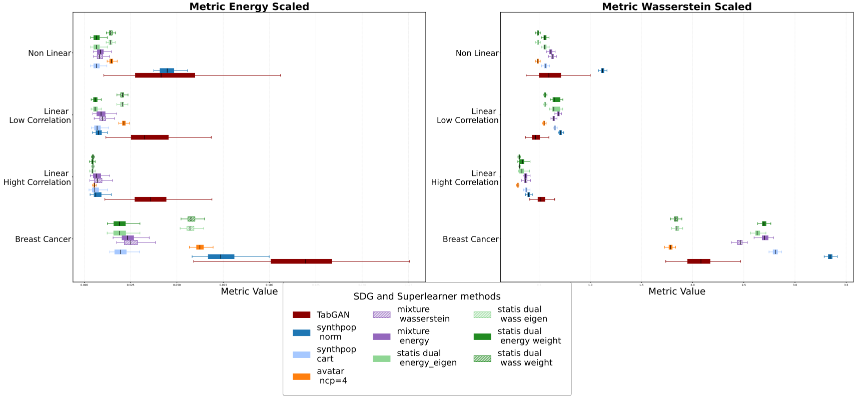


Figure 5 – Wasserstein and Energy distance with superlearners with best methods (**Synthpop CART**, **Synthpop norm**, **Avatar**) on different datasets.

Evaluating Privacy Metrics in Superlearners Using Best Methods on Linear Data

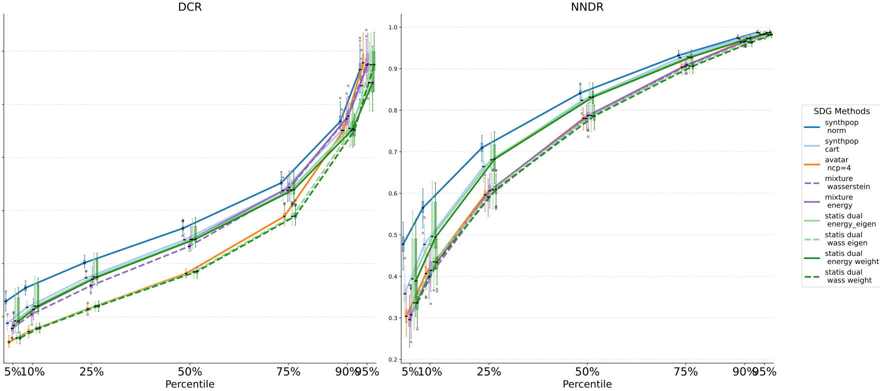


Figure 6 – Privacy results for linear dataset

Trade-off Utility-Privacy

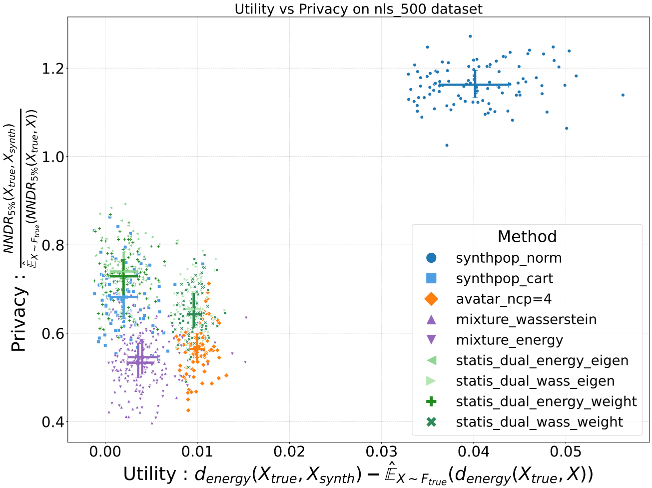


Figure 7 – Privacy vs. utility trade-off on the Non linear dataset ; Dataset size : 500.

Task-based Utility in Nonlinear Regression on nls dataset

In the *Non Linear* dataset, the outcome variable in the last column is simulated as : $Y = a \cdot \exp(-\beta X) + \epsilon_y, \quad \epsilon_y \sim \mathcal{N}(0, \sigma^2)$.

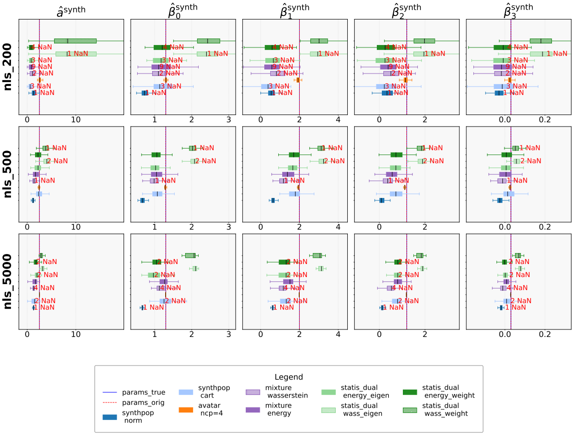


Figure 8 – Estimated biases of parameters $\hat{a}$ and $\hat{\beta}_j$ in the regression [26]

26. Fox, J., Weisberg, S. (2019). Nonlinear regression, nonlinear least squares, and nonlinear mixed models in R. population, 150, 200. $Y = a \cdot \exp(-\beta X) + \epsilon_y$ across different methods and sample sizes on the nls dataset.

Conclusion

- Different utility metrics can yield contrasting rankings of methods.
- **Statis** that aggregate well-performing models can offer acceptable utility across multiple evaluation metrics.
- **Avatar** remains the top performer across most metrics, especially in task-based metrics.
- For some datasets, we saw that **Statis** can be the best trade-off between privacy and utility, but not for task-based metrics.
- Many methods which exhibit strong utility ; as shown by low Energy or Wasserstein distances can fail on specific tasks, such as non-linear regression, where biases persist even with large sample size.
- Superlearners show encouraging results. Next work will consist of developping these methods for more complex data (multimodal, handling missing values and longitudionnal data).

Supplementary Material

Anonymization and **Pseudonymization** are solutions to protect sensitive data, enhance its confidentiality during sharing, and limit risks related to data processing and breaches.

Pseudonymization

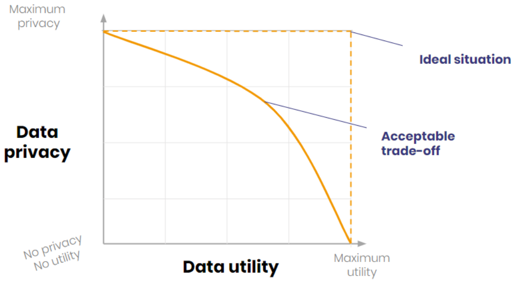
- Replacing directly identifiable data (e.g. first name) with indirectly identifiable data (e.g. code).
- Reversible operation.
- GDPR applies.

Anonymization

- Set of techniques to make impossible to identify the person by any means.
- Complete removal of data identifiability.
- Irreversible operation.
- GDPR does not apply.

Drawbacks of fully anonymized data

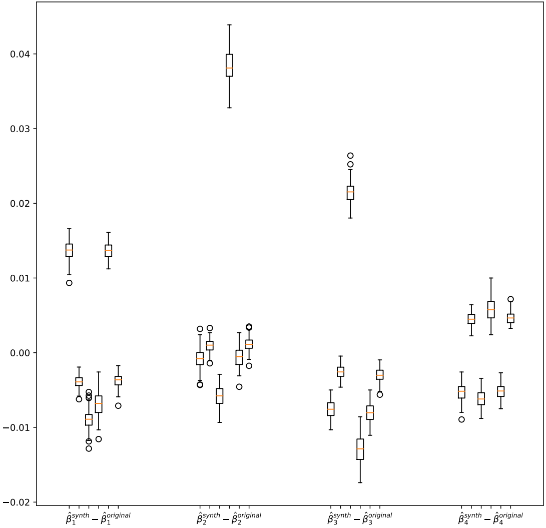
- **Loss of Data Utility :** Fully anonymized data may lose significant value and utility because essential information might be removed to ensure anonymity.



- **Irreversible Process :** Once data is fully anonymized, it cannot be reverted to its original form.
- **Complexity and Cost :** The process of fully anonymizing data can be complex and costly
- **Regulatory Challenges :** Ensuring compliance with diverse and evolving privacy regulations can be challenging when dealing with fully anonymized data.

Synthpop : Sensibility to order

Results for linear regression for unlineary correlated predictors
with data synthesized by synthpop from real data set with different shuffled columns.



More details on pMSE

Propensity MSE (pMSE) [27] **: Can we distinguish the original and the synthetic data ?** :

- $\widehat{p}_i$ is the estimated probability by a binary classifier for the sample i to be a synthetic data.
- $c = \frac{n_{syn}}{n_{orig}+n_{syn}}$ the probability to draw a synthetic data into the concatenated original and synthetic dataset
- $pMSE = \sum_{i=1}^{n_{orig}+n_{syn}} (\widehat{p}_i - c)^2$
- Sensible to the classifier used (CART, RF, logistic regression) [28].
- Theoritical result [29] , under the null hypothesis : Z_{synth} is generated by the true distribution $f(z|\theta)$; $pMSE = aF$ with $F \sim \chi^2(k-1)$.
- Derived pMSE metric :

27. **woo2009global**.
28. Snoke et al., "General and specific utility measures for synthetic data." *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 181(3) :663–688, 2018. Oxford University Press.
29. Snoke et al., "General and specific utility measures for synthetic data." *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 181(3) :663–688, 2018. Oxford University Press.

Supplementary simulations results : Evaluating Superlearners Using Best Methods - Breast Cancer Dataset

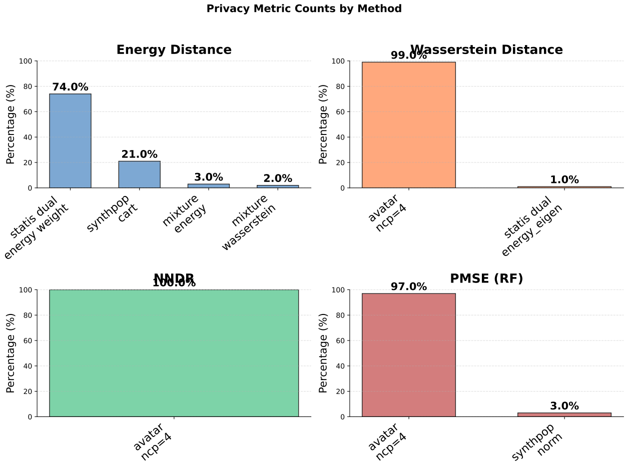
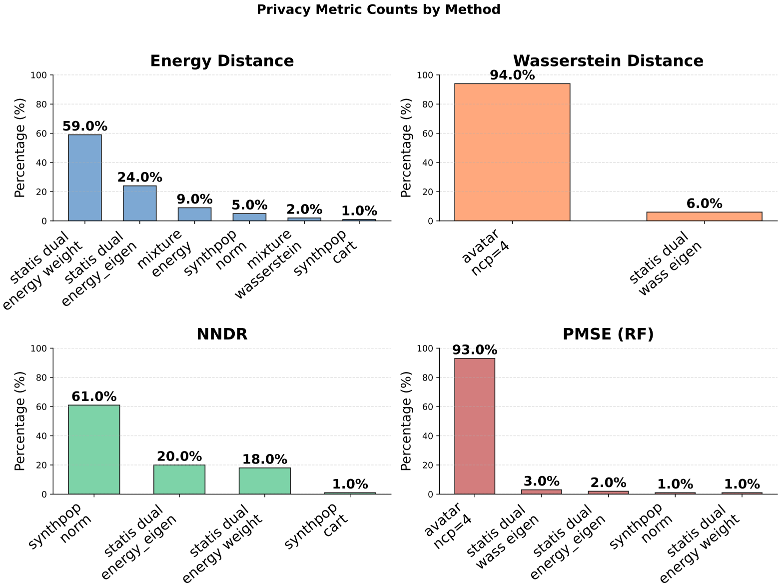


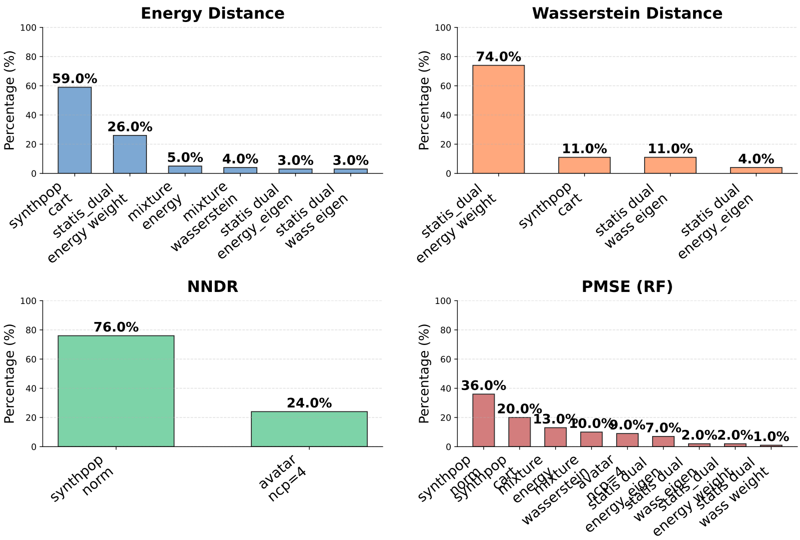
Figure 10 – Privacy and utility result on breast cancer dataset

Supplementary simulations results : Evaluating Superlearners Using Best Methods - Low correlation linear dataset



Supplementary simulations results : Evaluating Superlearners Using Best Methods - NLS dataset

Privacy Metric Counts by Method



Supplementary simulations results : Breast Cancer Dataset

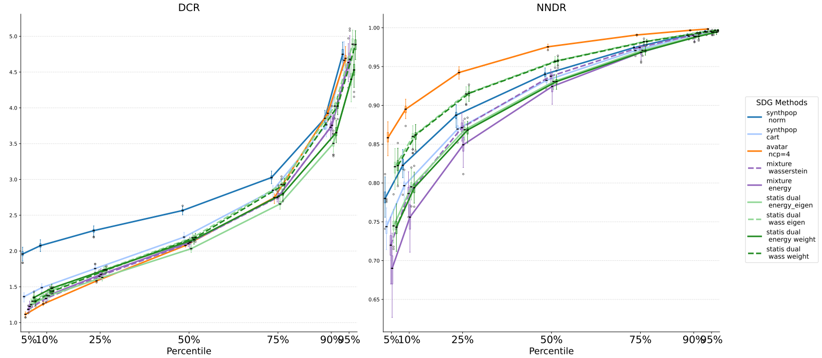


Figure 13 – Privacy results on breast cancer dataset

Supplementary simulations results : Evaluating Superlearners Using Best Methods - NLS dataset