June 11, 2025.
Vianey DARSEL
Etienne CÔME
Latifa OUKHELLOU

Population Synthesis with Deep Generative Models - is it worth it? Exploring new models and metrics.

13th Symposium of the European Association for Research in Transportation.

Université Gustave Eiffel

LABORATOIRE GRETTIA
GÉNIE DES RÉSEAUX
DE TRANSPORT TERRESTRES
ET INFORMATIQUE AVANCÉE

Introduction

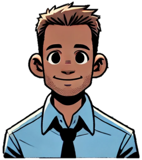
Use of a synthetic population

A Synthetic Population is necessary for any agent-based models

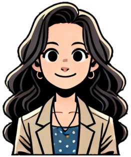
Different use cases with agent-based models:

- Transport Simulation (W. Axhausen et al., 2016)
- Epidemic Simulation (Kerr et al., 2021)
- Social Interaction Model (Macal et al., 2014)
- Poverty modelisation (Gisby et al., 2023)
- ...

How to pass from individuals to data ?

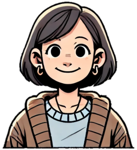


Mike, 40 yo, no degree, 2 cars



Lisa, 29 yo, Master Deg., 0 car

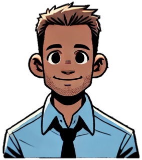
Jonathan, 12 yo, no degree, 0 car



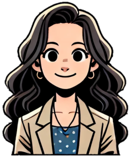
Zoe, 16 yo, no degree, 1 car

Université Gustave Eiffel

How to pass from individuals to data ?

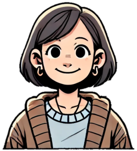


Mike, 40 yo, no degree, 2 cars



Lisa, 29 yo, Master Deg., 0 car

Jonathan, 12 yo, no degree, 0 car



Zoe, 16 yo, no degree, 1 car

→

Sex	Age	Education level	Number of cars	...
M	40	No Degree	2	...
F	29	Master Degree	0	...
M	13	No Degree	0	...
F	16	No Degree	1	...
...	...	...	...	...

Different sources of data

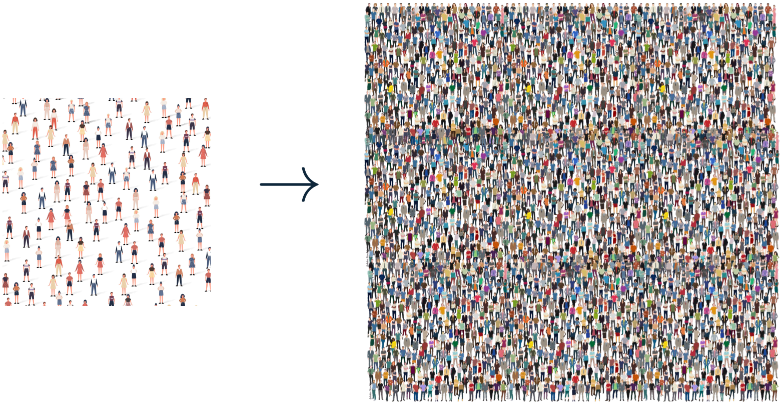
In transport, we have 2 main sources of data for population synthesis, whose size depends on the country:

- Household Travel Survey (HTS) $\sim 0.03\%$ of the total population
- Census Data $\sim 1\%$ of the total population[1]

Population synthesis: using algorithm to generate a full synthetic population from limited datasets

[1] In France, we can get in open-access a total reconstruction of the population

Population Synthesis in one scheme



A generated synthetic population must respect several criteria :

Criterion	**Goal**
Distribution	Comparing the distribution of the generated population with the true population.
Realism	Verifying that each generated sample is realistic.
Originality	Capacity to generate unseen samples.

Which metrics to evaluate a synthetic population?

A generated synthetic population must respect several criteria :

Criterion	Goal
Distribution	Comparing the distribution of the generated population with the true population.
Realism	Verifying that each generated sample is realistic.
Originality	Capacity to generate unseen samples.

Which metrics to evaluate a synthetic population?

In the last 15 years, many algorithms have been applied to generate a synthetic population, from reproduction models to Deep Generative Models.

Which algorithm is recommended for Population Synthesis?

Université Gustave Eiffel

Menu

Population Synthesis Evaluation

Data Notations

$\mathbf{X}_{train}$ $\rightarrow$ Data for model training

$\mathbf{X}_{test}$ $\rightarrow$ Data for model evaluation

$\mathbf{X}_{gen}$ $\rightarrow$ Data generated by the model

Université Gustave Eiffel

Distribution evaluation: $SRMSE_3$

We propose using the mean of the Standardized Rooted Mean Squared Error (SRMSE) on the distributions of all possible combinations of three variables.

$$
\begin{aligned}
SRMSE_{ijk}(\mathbf{X}_{gen}, \mathbf{X}_{test}) &= \sqrt{\sum_{x^{ijk}} \left(f_{gen}(x^{ijk}) - f_{test}(x^{ijk})\right)^2 \times |\Omega_i| \times |\Omega_j| \times |\Omega_k|} \\
SRMSE_3(\mathbf{X}_{gen}, \mathbf{X}_{test}) &= \frac{1}{\binom{n}{3}} \sum_{(i,j,k) \in \binom{\{1,\dots,n\}}{3}} SRMSE_{ijk}(\mathbf{X}_{gen}, \mathbf{X}_{test})
\end{aligned} \tag{1}
$$

Arguments:

Distribution evaluation: $SRMSE_3$

We propose using the mean of the Standardized Rooted Mean Squared Error (SRMSE) on the distributions of all possible combinations of three variables.

$$SRMSE_{ijk}(\mathbf{X}_{gen}, \mathbf{X}_{test}) = \sqrt{\sum_{x^{ijk}} \left(f_{gen}(x^{ijk}) - f_{test}(x^{ijk})\right)^2 \times |\Omega_i| \times |\Omega_j| \times |\Omega_k|}$$
$$SRMSE_3(\mathbf{X}_{gen}, \mathbf{X}_{test}) = \frac{1}{\binom{n}{3}} \sum_{(i,j,k)\in\binom{\{1,\dots,n\}}{3}} SRMSE_{ijk}(\mathbf{X}_{gen}, \mathbf{X}_{test}) \quad (2)$$

Arguments:

- Most widely used metric (SRMSE)

Distribution evaluation: $SRMSE_3$

We propose using the mean of the Standardized Rooted Mean Squared Error (SRMSE) on the distributions of all possible combinations of three variables.

$$SRMSE_{ijk}(\mathbf{X}_{gen}, \mathbf{X}_{test}) = \sqrt{\sum_{x^{ijk}} \left(f_{gen}(x^{ijk}) - f_{test}(x^{ijk})\right)^2 \times |\Omega_i| \times |\Omega_j| \times |\Omega_k|}$$
$$SRMSE_3(\mathbf{X}_{gen}, \mathbf{X}_{test}) = \frac{1}{\binom{n}{3}} \sum_{(i,j,k) \in \binom{\{1,\dots,n\}}{3}} SRMSE_{ijk}(\mathbf{X}_{gen}, \mathbf{X}_{test}) \quad (3)$$

Arguments:

- Most widely used metric (SRMSE)
- Considering the trivariate distributions allows grabbing marginals, and bivariate distributions without exploding the computation time (trivariate)

Université Gustave Eiffel

Distribution evaluation: $SRMSE_3$

We propose using the mean of the Standardized Rooted Mean Squared Error (SRMSE) on the distributions of all possible combinations of three variables.

$$SRMSE_{ijk}(\mathbf{X}_{gen}, \mathbf{X}_{test}) = \sqrt{\sum_{x^{ijk}} \left(f_{gen}(x^{ijk}) - f_{test}(x^{ijk})\right)^2 \times |\Omega_i| \times |\Omega_j| \times |\Omega_k|}$$
$$SRMSE_3(\mathbf{X}_{gen}, \mathbf{X}_{test}) = \frac{1}{\binom{n}{3}} \sum_{(i,j,k)\in\binom{\{1,\dots,n\}}{3}} SRMSE_{ijk}(\mathbf{X}_{gen}, \mathbf{X}_{test}) \quad (4)$$

Arguments:

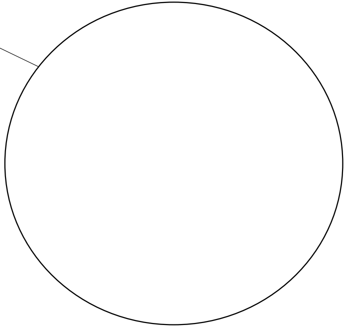
- Most widely used metric (SRMSE)
- Considering the trivariate distributions allows grabbing marginals, and bivariate distributions without exploding the computation time (trivariate)
- Balanced metric on all combinations (mean)

Realism evaluation: Structural zero definition

Structural Zeros: generated samples that should not have been generated (Borysov et al., 2019).

Realism evaluation: Structural zero definition

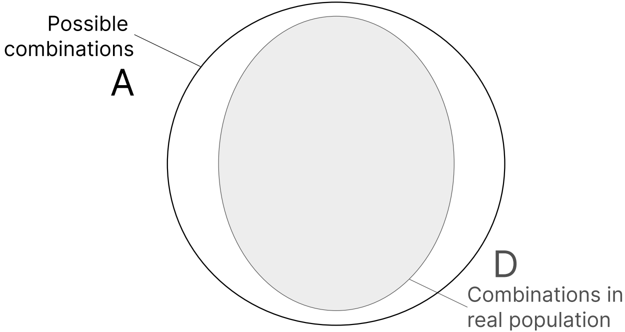
Structural Zeros: generated samples that should not have been generated (Borysov et al., 2019).



All possible combinations of modalities theoretically

Realism evaluation: Structural zero definition

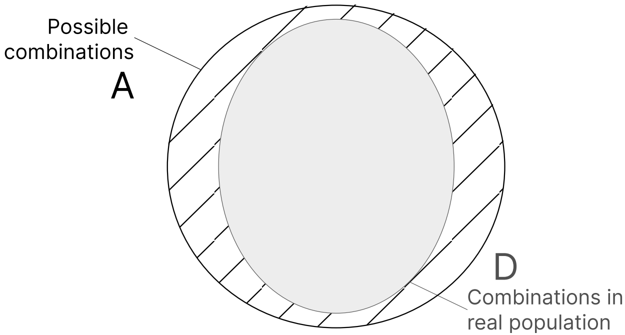
Structural Zeros: generated samples that should not have been generated (Borysov et al., 2019).



All combinations that exist in the real population

Realism evaluation: Structural zero definition

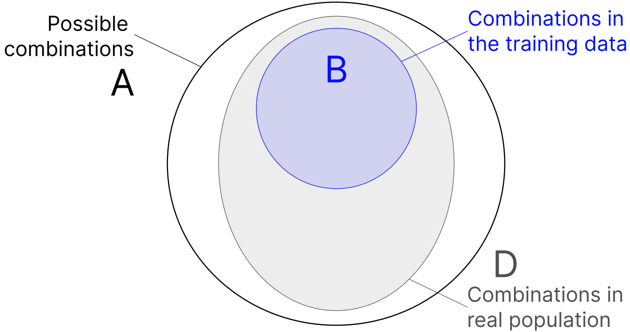
Structural Zeros: generated samples that should not have been generated (Borysov et al., 2019).



A structural zero is a sample that is a combination of attributes that do not exist in the real population.

Realism evaluation: Detection method in the literature

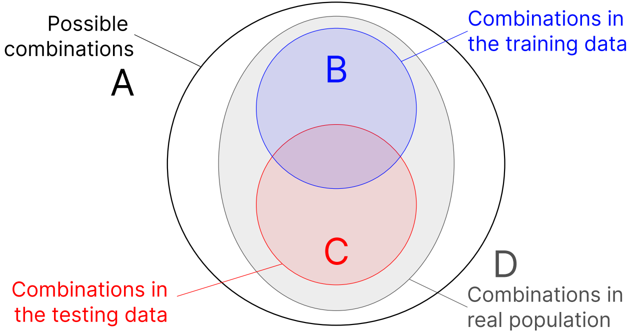
Structural Zeros: generated samples that should not have been generated (Borysov et al., 2019).



D is not accessible in practice.

Realism evaluation: Detection method in the literature

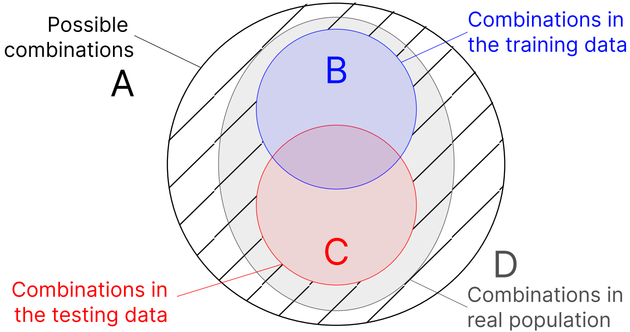
Structural Zeros: generated samples that should not have been generated (Borysov et al., 2019).



D is not accessible in practice.

Realism evaluation: Detection method in the literature

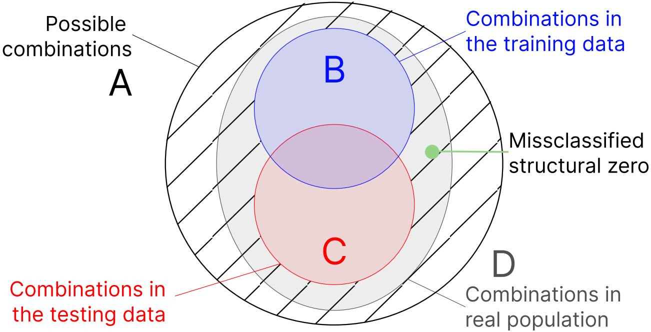
Structural Zeros: generated samples that should not have been generated (Borysov et al., 2019).



An approximation is done in the literature: all samples that do belong to neither the training data, nor the testing data is a structural zero.

Realism evaluation: Detection method in the literature

Structural Zeros: generated samples that should not have been generated (Borysov et al., 2019).



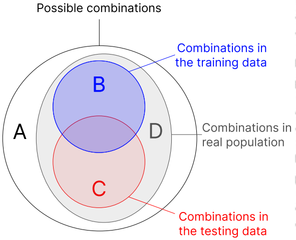
This detection can lead to false positive structural zeros. This phenomenon grows with the number of attributes.

Realism evaluation: *SSCIOT*

New detection method: at least one couple of its attributes is absent from both training and testing sets

Share of Samples with a Couple of Instances that is Out of Testing data:

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



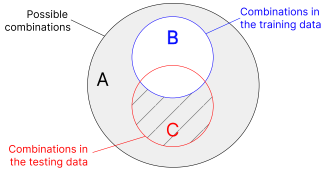
Argument:

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

C_{ij} is the restriction of C to the variables i and j

Originality evaluation: *SSOTT*

Sampling Zeros: Generated samples that are in real population, but not in training data (Garrido et al., 2020).



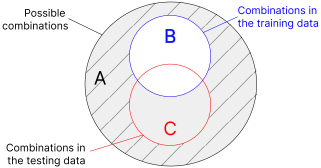
Detection method:
generated samples that belong to testing data, but not to training data

Originality evaluation: *SSOTT*

Sampling Zeros: Generated samples that are in real population, but not in training data (Garrido et al., 2020).

Transform into a minimizing metric:
Share of Samples Out of Training and Testing:

$$SSOTT = 1 - \frac{\sum_{x \in \mathbf{X}_{gen}} \mathbb{1}_{x \in C \setminus B}}{\sum_{x \in \mathbf{X}_{gen}} \mathbb{1}_{x \in \bar{B}}} = \frac{\sum_{x \in \mathbf{X}_{gen}} \mathbb{1}_{x \in A \setminus (B \cup C)}}{\sum_{x \in \mathbf{X}_{gen}} \mathbb{1}_{x \in \bar{B}}} \quad (6)$$



Our target area.

Argument:

- Minimizing metric with 0 as minimal score

Models in Population Synthesis

Models in Population Synthesis

3 types of models in Population Synthesis:

- Reproduction Models (Beckman et al., 1996; Voas and Williamson, 2000; Guo and Bhat, 2007)

Models in Population Synthesis

3 types of models in Population Synthesis:

- Reproduction Models (Beckman et al., 1996; Voas and Williamson, 2000; Guo and Bhat, 2007)
- Probabilistic Models (Farooq et al., 2013; Sun and Erath, 2015; Hu et al., 2018)

Models in Population Synthesis

3 types of models in Population Synthesis:

- Reproduction Models (Beckman et al., 1996; Voas and Williamson, 2000; Guo and Bhat, 2007)
- Probabilistic Models (Farooq et al., 2013; Sun and Erath, 2015; Hu et al., 2018)
- Deep Generative Models (Borysov et al., 2019; Garrido et al., 2020; Kim and Bansal, 2023)

Deep Generative Models in Population Synthesis

State-of-the-art Deep Generative Model for image synthesis	**First Implementation with Tabular data in Population Synthesis**
Variational Auto Encoder (Kingma and Welling, 2013)	Borysov et al. (2019)

Table: Chronological state-of-the-art Deep Generative Model and its implementation in Population Synthesis

Deep Generative Models in Population Synthesis

State-of-the-art Deep Generative Model for image synthesis	**First Implementation with Tabular data in Population Synthesis**
Variational Auto Encoder (Kingma and Welling, 2013)	Borysov et al. (2019)
Generative Adversarial Network (Goodfellow et al., 2014)	Garrido et al. (2020)

Table: Chronological state-of-the-art Deep Generative Model and its implementation in Population Synthesis

Université Gustave Eiffel

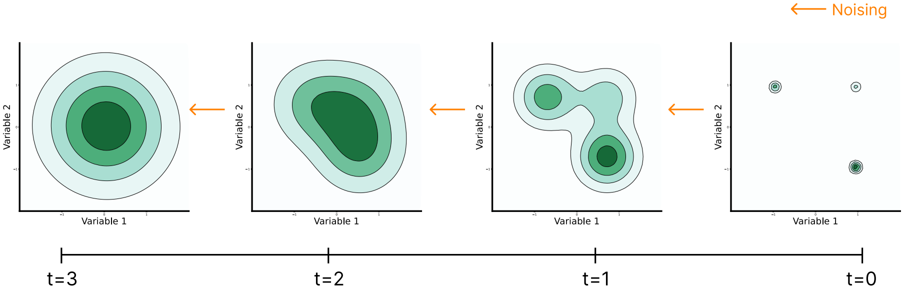
Deep Generative Models in Population Synthesis

State-of-the-art Deep Generative Model for image synthesis	**First Implementation with Tabular data in Population Synthesis**
Variational Auto Encoder (Kingma and Welling, 2013)	Borysov et al. (2019)
Generative Adversarial Network (Goodfellow et al., 2014)	Garrido et al. (2020)
Diffusion Model (Song and Ermon, 2019)	?

Table: Chronological state-of-the-art Deep Generative Model and its implementation in Population Synthesis

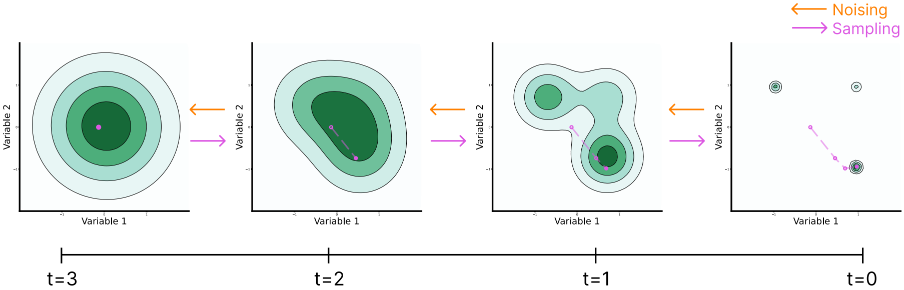
Introduction to Diffusion models

Philosophy of Diffusion Model



Noising data is easy

Philosophy of Diffusion Model



Diffusion aims to learn how to denoise a signal

Sampling = Denoising
Training = Noising

Mathematical derivations of the models (Song et al., 2022)

Noising: $d\mathbf{X}^t = \mathbf{f}(\mathbf{X}^t, t)dt + g(t)d\mathbf{w}$

Denoising: $d\mathbf{X}^t = [\mathbf{f}(\mathbf{X}^t, t) - g(t)^2\nabla_{\mathbf{x}} \log p_t(\mathbf{X}^t)]dt + g(t)d\bar{\mathbf{w}}$

$d\mathbf{X}^t = [\mathbf{f}(\mathbf{X}^t, t) - g(t)^2\nabla_{\mathbf{x}} \log p_t(\mathbf{X}^t)]dt + g(t)d\bar{\mathbf{w}}$

where:

- $\mathbf{f}$ is the drift function.
- g is the diffusion coefficient.
- $\mathbf{w}$ and $\bar{\mathbf{w}}$ are standard Wiener processes.

Université Gustave Eiffel

But only designed for continuous data

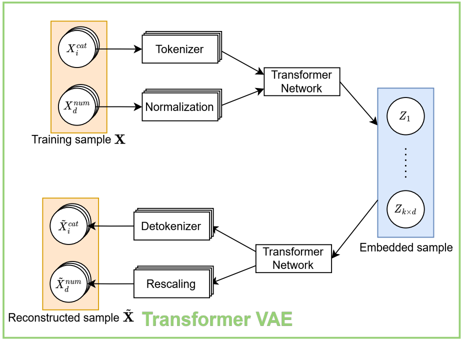
Some terms cannot be adapted for categorical variables:

$$d\mathbf{X}^t = \mathbf{f}(\mathbf{X}^t, t)dt + g(t)d\mathbf{w} \tag{7}$$

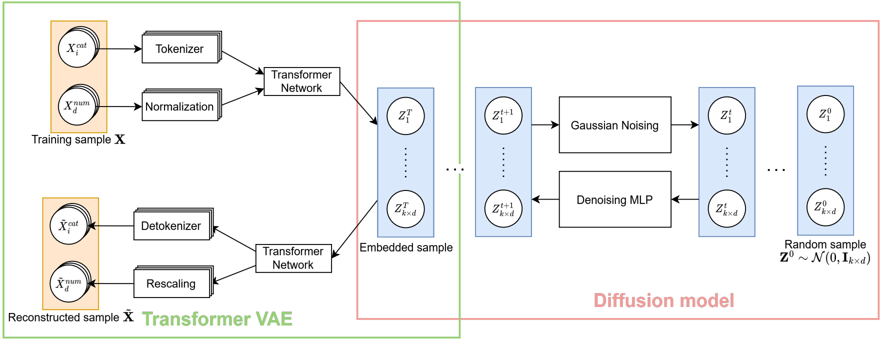
$$d\mathbf{X}^t = [\mathbf{f}(\mathbf{X}^t, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{X}^t)]dt + g(t)d\bar{\mathbf{w}} \tag{8}$$

Unlike VAE and GAN, diffusion cannot be apply directly for tabular data synthesis

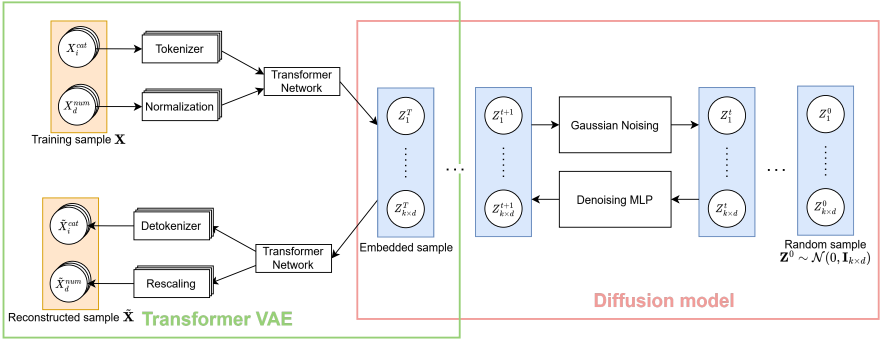
Adapted Diffusion model for Tabular Data Synthesis: TabSyn by Zhang et al. (2023)



Adapted Diffusion model for Tabular Data Synthesis: TabSyn by Zhang et al. (2023)



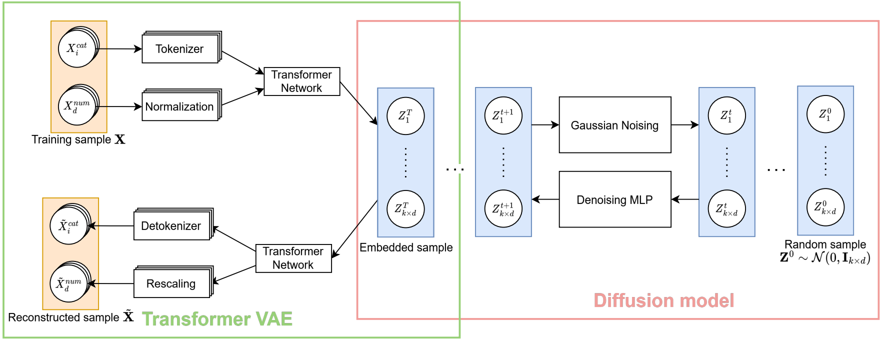
Adapted Diffusion model for Tabular Data Synthesis: TabSyn by Zhang et al. (2023)



- The model we used in our experiments

Université Gustave Eiffel

Adapted Diffusion model for Tabular Data Synthesis: TabSyn by Zhang et al. (2023)



- The model we used in our experiments
- It was the model with the best performance for tabular data generation at the publication of the article

Université Gustave Eiffel

Benchmark

Benchmark data: French census data in 2015 in Île-de-France

Attributes description: 12 attributes (3 numerical and 9 categorical)

Two scenarios for training data:

- Census data scenario: 1% of the population
- Household Travel Survey scenario: 0.03% of the population

Testing data: 23% of the total population

Attribute	Data type
Age	integer
Sex	binary
Last diploma	category
Number of persons in the household	integer
Type of household	category
Type of professional activity	category
Family link	category
Married	boolean
Department	category
Number of cars	integer
Socio-professional category	category
Type of accommodation	category

Université Gustave Eiffel

Experiments

1) Comparing Deep Generative Models with various data encoding

2) Comparing the best model from the first experiments with Probabilistic models

Benchmark of Deep Generative Models: protocol

Comparison of five Deep Generative Models:

- Variational Auto Encoder (with and without Transformer VAE embedding)
- Generative Adversarial Network (with and without Transformer VAE embedding)
- Diffusion Model

With three different encodings for numerical variables:

- All variables are categorical
- All variables are categorical, except for age, which is continuous.
- All numerical variables are continuous.

Benchmark of Deep Generative Models: protocol

5 DGMs

VAE (Raw data)
VAE (Embedding data)
GAN (Raw data)
GAN (Embedding data)
Diffusion

3 data encodings

All numerical as Categorical
Only Age as Continuous
All numerical as Continuous

2 data scenarios

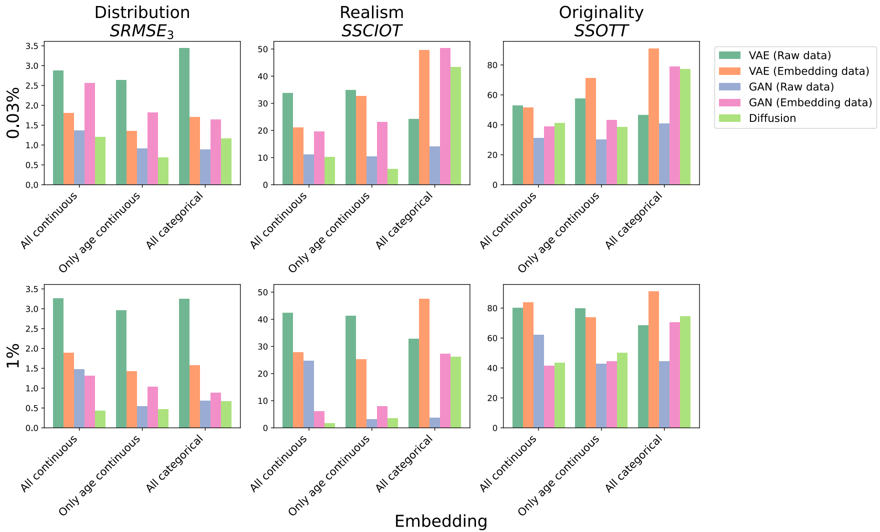
Census data (1%)
HTS (0.03%)

3 metrics

$SRMSE_3$
SSCIOT
SSOTT

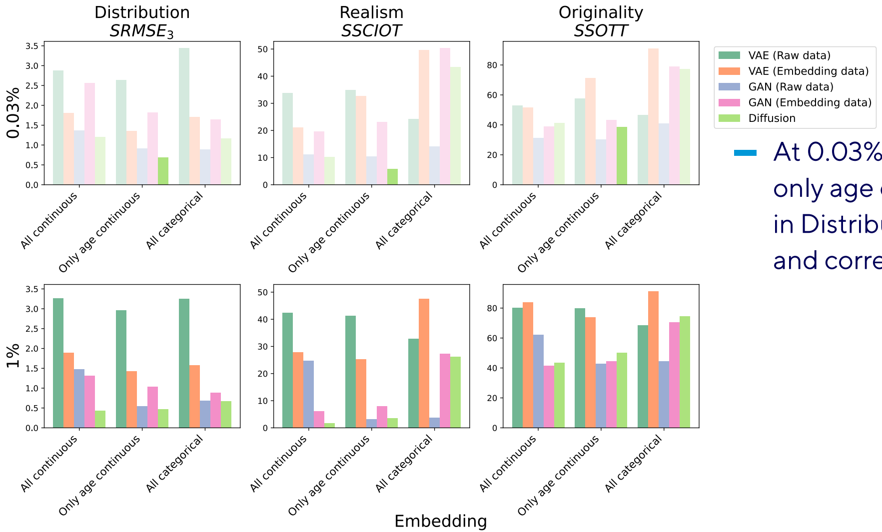
Université Gustave Eiffel

Results



Université Gustave Eiffel

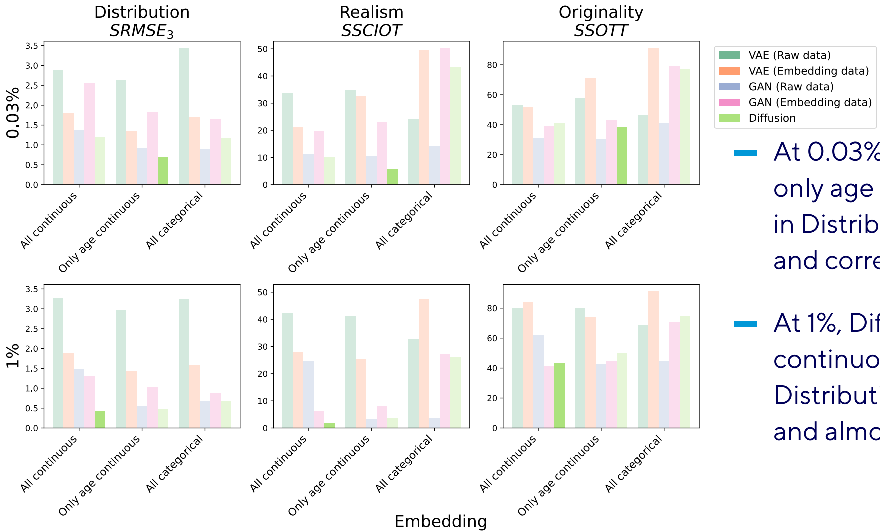
Results



- At 0.03%, Diffusion with only age continuous is best in Distribution and Realism, and correct in Originality

Université Gustave Eiffel

Results



- At 0.03%, Diffusion with only age continuous is best in Distribution and Realism, and correct in Originality
- At 1%, Diffusion with all continuous is best in Distribution and Realism, and almost in Originality

Comparison between Deep Generative Model and Probabilistic Models: protocol

5 Models

Monte Carlo Markov Chain (freq.)
Monte Carlo Markov Chain (Bayesian)
Bayesian Network (tree)
Bayesian Network (hill)
Diffusion

2 data scenarios

Census data (1%)
HTS (0.03%)

3 metrics

$SRMSE_3$
$SSCIOT$
$SSOTT$

Université Gustave Eiffel

Results

Training size of 0.03% of the total population

Model	Distribution *SRMSE*	Realism *SSCIOT*	Originality *SSOTT*
MCMC (freq.)	28.9	**0%**	NC
MCMC (Bayesian)	2.81	96.45%	99.99%
BN (hill)	0.773	4.62%	50.61%
BN (tree)	0.79	0.69%	48.67%
Diffusion	**0.693**	5.74%	**38.56%**

Training size of 1% of the total population

Model	Distribution *SRMSE*	Realism *SSCIOT*	Originality *SSOTT*
MCMC (freq.)	6.34	**0%**	NC
MCMC (Bayesian)	2.84	98.07%	100.0%
BN (hill)	0.432	0.37%	46.29%
BN (tree)	0.676	3.4%	60.54%
Diffusion	**0.422**	1.82%	**43.46%**

Table: Comparison of the best DGM with probabilistic models on three criteria.

Conclusion

Conclusion

- We present three metrics evaluating the distribution, the realism, and the originality of a synthetic population.

Conclusion

- We present three metrics evaluating the distribution, the realism, and the originality of a synthetic population.
- We introduce Diffusion models for Population Synthesis.

Conclusion

- We present three metrics evaluating the distribution, the realism, and the originality of a synthetic population.
- We introduce Diffusion models for Population Synthesis.
- Our benchmark indicates that Diffusion stands out as the top deep generative model for population synthesis. Its performance is comparable to that of the leading probabilistic models.[5]

[5] In our experimental framework

Perspectives

Perspectives

- Diffusion is better, but at which cost ?

Model	Training + Sampling Time
BN (tree)	7 seconds
BN (hill)	9 seconds
Diffusion	78 minutes

Table: Time for training and sampling for the different models for a training set of 1%. For diffusion, 76 minutes are spent for the training.

Université Gustave Eiffel

Perspectives

- Diffusion is better, but at which cost ?

Model	Training + Sampling Time
BN (tree)	7 seconds
BN (hill)	9 seconds
Diffusion	78 minutes

Table: Time for training and sampling for the different models for a training set of 1%. For diffusion, 76 minutes are spent for the training.

- One important criterion that is omitted: Privacy

Université Gustave Eiffel

Perspectives

- Diffusion is better, but at which cost ?

Model	Training + Sampling Time
BN (tree)	7 seconds
BN (hill)	9 seconds
Diffusion	78 minutes

Table: Time for training and sampling for the different models for a training set of 1%. For diffusion, 76 minutes are spent for the training.

- One important criterion that is omitted: Privacy
- See the impact of Deep Generative Models (and Diffusion), on more complex tasks, such as Population Synthesis at the Household generation or in a time perspective

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.

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.

Garrido, S., Borysov, S. S., Pereira, F. C., and Rich, J. (2020). Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C: Emerging Technologies*, 120:102787.

Gisby, J., Kiknadze, A., Roitner-Fransecky, I., and Mitterling, T. (2023). Fighting poverty with synthetic data.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks.

Guo, J. Y. and Bhat, C. R. (2007). Population Synthesis for Microsimulating Travel Behavior. *Transportation Research Record*, 2014(1):92–101. Publisher: SAGE Publications Inc.

Hu, J., Reiter, J. P., and Wang, Q. (2018). Dirichlet Process Mixture Models for Modeling and Generating Synthetic Versions of Nested Categorical Data. *Bayesian Analysis*, 13(1):183–200. Publisher: International Society for Bayesian Analysis.

Kerr, C. C., Stuart, R. M., Mistry, D., Abeysuriya, R. G., Rosenfeld, K., Hart, G. R., Núñez, R. C., Cohen, J. A., Selvaraj, P., Hagedorn, B., George, L., Jastrzębski, M., Izzo, A. S., Fowler, G., Palmer, A., Delport, D., Scott, N., Kelly, S. L., Bennette, C. S., Wagner, B. G., Chang, S. T., Oron, A. P., Wenger, E. A., Panovska-Griffiths, J., Famulare, M., and Klein, D. J. (2021). Covasim: An agent-based model of COVID-19 dynamics and interventions. *PLOS Computational Biology*, 17(7):e1009149. Publisher: Public Library of Science.

Kim, E.-J. and Bansal, P. (2023). A deep generative model for feasible and diverse population synthesis. *Transportation Research Part C: Emerging Technologies*, 148:104053.

Kingma, D. P. and Welling, M. (2013). Auto-Encoding Variational Bayes.

Macal, C. M., North, M. J., Collier, N., Dukic, V. M., Wegener, D. T., David, M. Z., Daum, R. S., Schumm, P., Evans, J. A., Wilder, J. R., Miller, L. G., Eells, S. J., and Lauderdale, D. S. (2014). Modeling the transmission of community-associated methicillin-resistant Staphylococcus aureus: a dynamic agent-based simulation. *Journal of Translational Medicine*, 12(1):124.

Song, J., Meng, C., and Ermon, S. (2022). Denoising Diffusion Implicit Models. arXiv:2010.02502 [cs].

Song, Y. and Ermon, S. (2019). Generative Modeling by Estimating Gradients of the Data Distribution. *Advances in Neural Information Processing Systems*, 32.

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

Voas, D. and Williamson, P. (2000). An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography*, 6(5):349–366.

W. Axhausen, K., Horni, A., and Nagel, K., editors (2016). *The Multi-Agent Transport Simulation MATSim*. Ubiquity Press. Accepted: 2016-12-31 23:55:55.

Zhang, H., Zhang, J., Shen, Z., Srinivasan, B., Qin, X., Faloutsos, C., Rangwala, H., and Karypis, G. (2023). Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space.

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

Université Gustave Eiffel

LABORATOIRE GRETTIA
GÉNIE DES RÉSEAUX
DE TRANSPORT TERRESTRES
ET INFORMATIQUE AVANCÉE

Derivation SRMSE

Let consider

- $\mathbf{Z} = (Z_1, Z_2, ..., Z_m)$ a multi-categorical variable, where Z_k has h_k modalities $(1, ...h_k)$.
- True frequencies: $\forall u \in \{1, ..., h_k\}, f_{k,u} = \frac{1}{h_k}$
- Estimated frequencies: $\forall k \in [1, ..., m], \forall u \in \{1, ..., h_k\}, \hat{f}_{k,u} = (1 + (-1)^u e_k) f_{k,u}$.

$$SRMSE(\mathbf{Z}, \hat{\mathbf{Z}}) = \frac{1}{m}\sqrt{\sum_{k=1}^{m}\sum_{u=1}^{h_k}\left(f_{k,u} - \hat{f}_{k,u}\right)^2} \times \sqrt{\sum_{k=1}^{m} h_k} = \frac{1}{m}\sqrt{\sum_{k=1}^{m}\sum_{u=1}^{h_k}\frac{e_k^2}{h_k^2}} \times \sqrt{\sum_{k=1}^{m} h_k} = \frac{1}{m}\sqrt{\sum_{k=1}^{m}\frac{e_k^2}{n_k}} \times \sqrt{\sum_{k=1}^{m} n_k}. \quad (9)$$

$$SRMSE_3(\mathbf{Z}, \hat{\mathbf{Z}}) = \frac{1}{m}\sum_{k=1}^{m}\sqrt{\sum_{u=1}^{h_k}\left(f_{k,u} - \hat{f}_{k,u}\right)^2 \times h_k} = \frac{1}{m}\sum_{k=1}^{m}\sqrt{\sum_{u=1}^{h_k}\left(\frac{e_k^2}{h_k^2}\right) \times h_k} = \frac{1}{m}\sqrt{\sum_{k=1}^{m} e_k^2} \quad (10)$$

Université Gustave Eiffel

Comparing SRMSE metrics

Comparison $SRMSE_3$ with $SRMSE$

0.03% of the total population

	Marginal		Bivariate		Trivariate	
	SRMSE	$SRMSE_3$	*SRMSE*	$SRMSE_3$	*SRMSE*	$SRMSE_3$
BN (hill)	**0.0575**	**0.0449**	**0.454**	**0.268**	**1.49**	**0.774**
BN (tree)	0.0615	0.0458	0.499	0.27	1.62	0.777
MCMC (freq.)	3.17	2.68	12.7	9.63	40.9	29.4
MCMC (Bayesian)	1.37	0.749	3.45	1.54	7.15	2.76
VAE (Raw data)	0.653	0.51	1.93	1.32	4.67	2.82
VAE (Embedding data)	0.475	0.348	1.46	0.873	3.61	1.8
GAN (Raw data)	0.504	0.251	1.4	0.613	3.17	1.26
GAN (Embedding data)	0.638	0.366	1.83	0.888	4.24	1.81
Diffusion	0.218	0.155	0.654	0.383	1.58	0.818

1% of the total population

	Marginal		Bivariate		Trivariate	
	SRMSE	$SRMSE_3$	*SRMSE*	$SRMSE_3$	*SRMSE*	$SRMSE_3$
BN (hill)	0.119	0.0412	**0.392**	**0.178**	**1.06**	**0.496**
BN (tree)	**0.118**	**0.0383**	0.571	0.27	1.74	0.778
MCMC (freq.)	2.13	1.27	6.65	3.33	16.6	7.14
MCMC (Bayesian)	1.3	0.741	3.36	1.54	7.09	2.79
VAE (Raw data)	1.33	0.767	4.11	1.9	10.5	4.27
VAE (Embedding data)	0.386	0.261	1.17	0.636	2.89	1.33
GAN (Raw data)	0.451	0.232	1.22	0.519	2.7	1.02
GAN (Embedding data)	0.179	0.118	0.518	0.281	1.24	0.595
Diffusion	0.166	0.107	0.463	0.252	1.08	0.522

Table: Impact on the measurements of using $SRMSE_3$ rather than $SRMSE$

0.03% of the total population

	Marginal		Bivariate		Trivariate	
	SRMSE	$SRMSE_3$	*SRMSE*	$SRMSE_3$	*SRMSE*	$SRMSE_3$
BN (hill)	1	1	1	1	1	1
BN (tree)	2	2	2	2	3	2
MCMC (freq.)	9	9	9	9	9	9
MCMC (Bayesian)	8	8	8	8	8	7
VAE (Raw data)	7	7	7	7	7	8
VAE (Embedding data)	4	5	5	5	5	5
GAN (Raw data)	5	4	4	4	4	4
GAN (Embedding data)	6	6	6	6	6	6
Diffusion	3	3	3	3	2	3

1% of the total population

	Marginal		Bivariate		Trivariate	
	SRMSE	$SRMSE_3$	*SRMSE*	$SRMSE_3$	*SRMSE*	$SRMSE_3$
BN (hill)	2	2	1	1	1	1
BN (tree)	1	1	4	3	4	4
MCMC (freq.)	9	9	9	9	9	9
MCMC (Bayesian)	7	7	7	7	7	7
VAE (Raw data)	8	8	8	8	8	8
VAE (Embedding data)	5	6	5	6	6	6
GAN (Raw data)	6	5	6	5	5	5
GAN (Embedding data)	4	4	3	4	3	3
Diffusion	3	3	2	2	2	2

Table: Impact on the ranking of using $SRMSE_3$ rather than *SRMSE*

Complementary Results

Results with Training size of 0.03% of the total population (DGMs)

DGM	Continuous data representation	Distribution *SRMSE*	Originality *SSOTT*	Realism *SSCIOT*
Diffusion	All continuous	1.21	42.91%	10.16%
	Only age continuous	**0.693**	38.56%	**5.74%**
	All categorical	1.16	77.3%	43.22%
GAN (Embedding data)	All continuous	2.56	38.98%	19.55%
	Only age continuous	1.82	43.2%	23.17%
	All categorical	1.64	79.05%	50.31%
GAN (Raw data)	All continuous	1.37	31.11%	11.08%
	Only age continuous	0.915	**30.22%**	10.31%
	All categorical	0.881	40.8.6%	14.11%
VAE (Embedding data)	All continuous	1.81	51.69%	20.99%
	Only age continuous	1.35	71.18%	32.64%
	All categorical	1.7	90.9%	49.68%
VAE (Raw data)	All continuous	2.88	52.86%	33.72%
	Only age continuous	2.64	57.63%	34.91%
	All categorical	3.44	46.57%	24.12%

Table: Comparison of the different DGMs and encodings on three criteria. For each metric, the optimal value is the smallest one and is highlighted in bold.

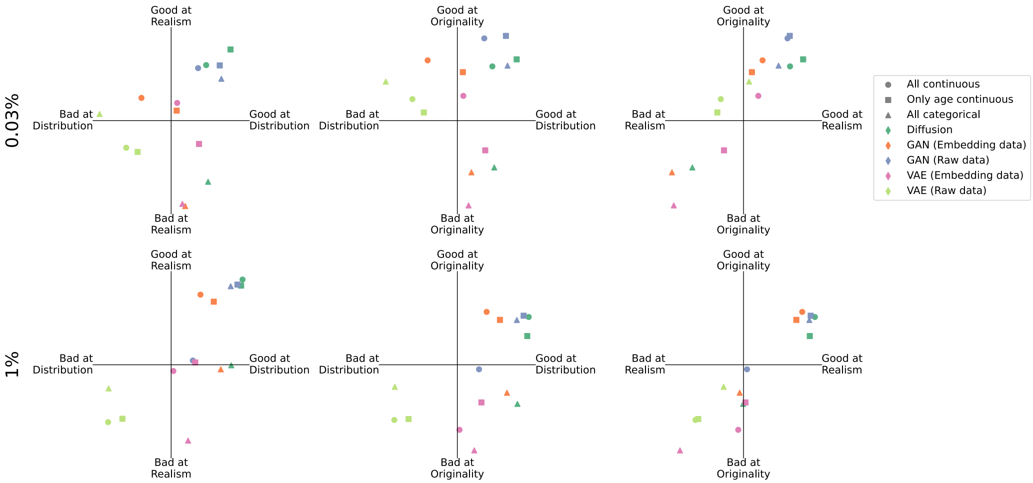
Université Gustave Eiffel

Results with Training size of 1% of the total population (DGMs)

DGM	Continuous data representation	Distribution *SRMSE*	Originality *SSOTT*	Realism *SSCIOT*
Diffusion	All continuous	**0.422**	43.46%	**1.82%**
	Only age continuous	0.471	49.97%	3.47%
	All categorical	0.678	74.76%	26.1%
GAN (Embedding data)	All continuous	1.32	**41.64%**	6.08%
	Only age continuous	1.04	44.5%	8.08%
	All categorical	0.893	70.57%	27.32%
GAN (Raw data)	All continuous	1.46	61.9%	24.76%
	Only age continuous	0.543	42.64%	3.09%
	All categorical	0.685	44.54%	3.68%
VAE (Embedding data)	All continuous	1.89	83.84%	27.8%
	Only age continuous	1.43	74.04%	25.35%
	All categorical	1.58	91.2%	47.58%
VAE (Raw data)	All continuous	3.26	80.31%	42.34%
	Only age continuous	2.96	79.98%	41.37%
	All categorical	3.25	68.45%	32.76%

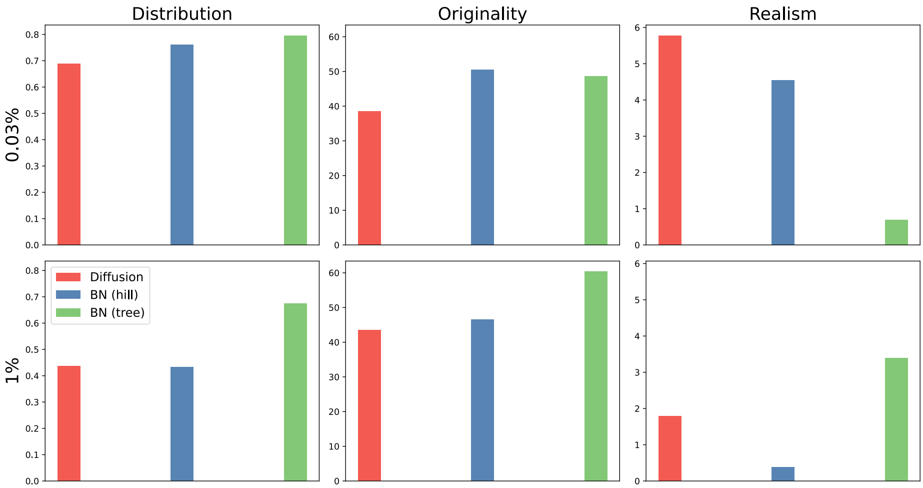
Table: Comparison of the different DGMs and encodings on three criteria. For each metric, the optimal value is the smallest one and is highlighted in bold.

Comparison with two criteria at the same time for DGMs



Université Gustave Eiffel

Bar chart comparing Bayesian Network with Diffusion



Université Gustave Eiffel

Architectures

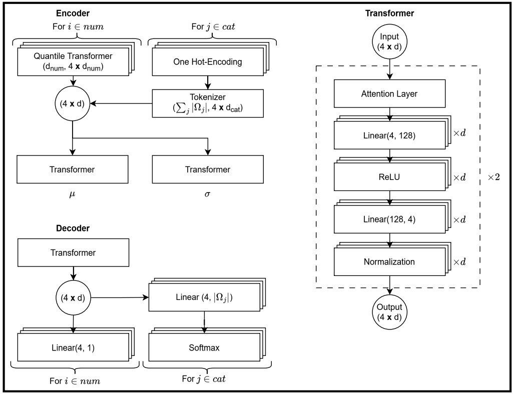


Figure: Transformer VAE architecture

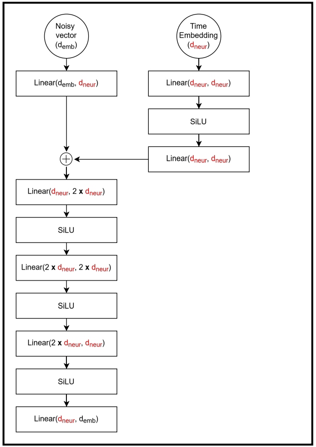


Figure: Diffusion architecture

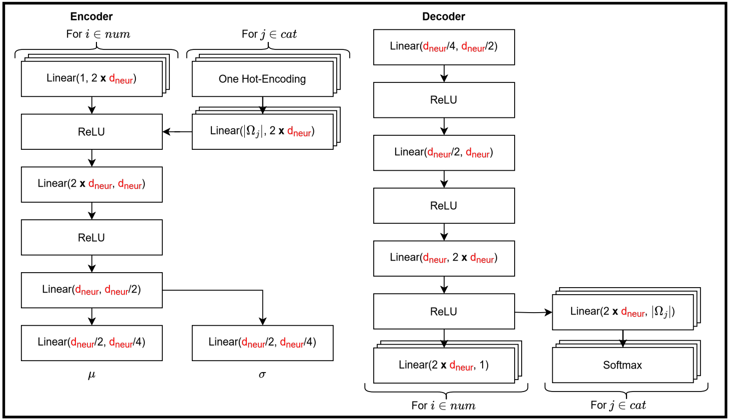


Figure: VAE (raw data) architecture

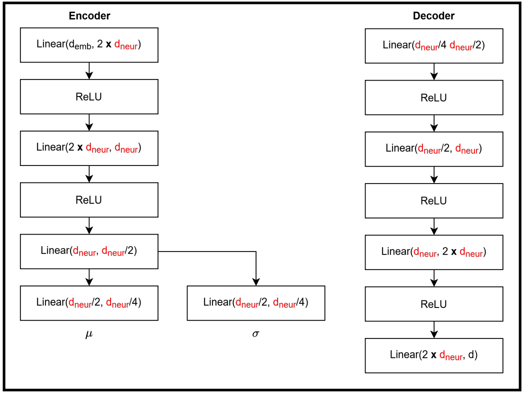


Figure: VAE (embedding data) architecture

Université Gustave Eiffel

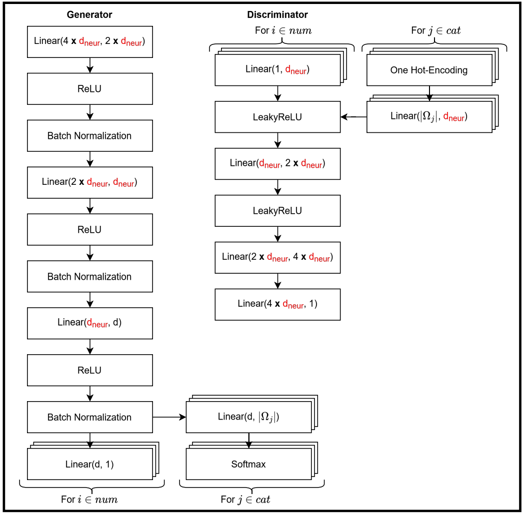


Figure: GAN (raw data) architecture

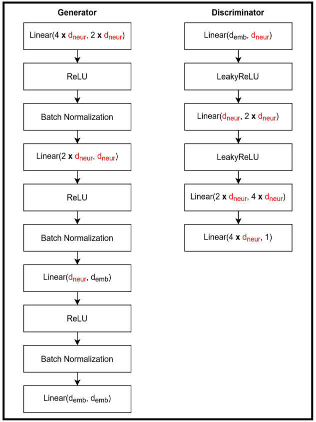


Figure: GAN (embedding data) architecture