GDS: a SWIFT-specialised TRGAN for synthetic financial transaction data generation

11^e Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle, Dijon

Romain A. Alfred, & Thomas Lemonnier
SKAIZen Lab, SKAIZen Group, Paris, France

June 30th, 2025

Summary

1 Introduction

Context

- Why generate SWIFT messages?
- SWIFT message: format of a banking transaction transiting through the SWIFT network.
- In Europe: financial institutions are subject to a strict customer data regulations (GDPR: General Data Protection Regulation).
- Consequences: lack of testing and training data.
  - Non-coverage of all test cases (for development project),
  - Difficulty to obtain generalisable AI model.

Objectives

- To generate large volume of data from small samples,
- To respect the statistical structure of real data.
- Example: to reproduce a typical week of SWIFT data flows using only a few messages.
- Applications:
  - Load test,
  - Robustness load,
  - Fraud detection,
  - Anti-money laundering (AML) and combating the financing of terrorism (CFT).

Scientific challenges

- Confidentiality and fidelity,
- Heterogeneity of data types,
  - A SWIFT message combines categorical and continuous data, as well as unstructured fields.
- Non-Gaussian distributions,
  - Continuous variables are often multimodals (transaction amounts, temporality) whereas neural networks are most often optimised for Gaussian inputs.
- Complex relationships,
  - Strong dependence of transaction amounts to time and counterparties,
  - Strong dependence of missing data to SWIFT message typologies.
- High cardinalities and class imbalance,
  - Difficulty to accurately generate variables with low modalities (fraud detection, AML-CFT): mode collapse.

Summary

2 State of the art

VAE

2 State of the art

- Variational Auto-Encoders (VAEs): generative models based on learning of latent data representations (Kingma & Welling, 2019).
  - Step 1 – encoder: transformation of data into latent space using a neural network,
  - Step 2 – decoder: generation of synthetic data from latent spaces by a neural network.

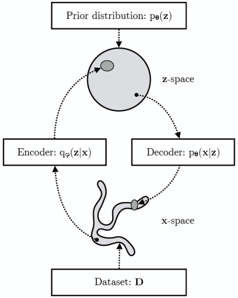

Figure: Modelling of the data processing of a Variational Auto-Encoder. Reproduced from Kingma & Welling (2019: 333).

Diffusion models

2 State of the art

- Diffusion models: implicit generative models (Ho & Abbeel, 2020).
  - Step 1 – chain addition of Gaussian noise,
  - For each addition: noise prediction by a neural network,
  - Step 2 – denoising using the precedent noise assessments.

Figure: Diffusion model.

Graph models

- Graph recurrent neural networks (GRNNs) (You et al., 2018).
  - Each node (a counterparty) has a hidden state that captures its transaction history,
  - Each node is associated with a RNN which is updated according to the transactions performed.
- Variational Graph Auto-Encoders (VGAEs): adaptation of VAE to graphs (Kipf & Welling, 2016).
  - Step 1 – encoder: production of a latent space for each node by a GNN,
  - Step 2 – decoder: generation of a synthetic graph from the latent spaces by a GNN.
- Graph Generative Adversarial Networks (GraphGANs): adaptation of GAN to graphs (Wang et al., 2018).
  - Step 1 – generator: generation of neighbouring nodes knowing an input node,
  - Step 2 – discriminator: prediction of the probability that a node is the neighbour of the input node.

- Generative Adversarial Networks (GANs) (Goodfellow et al., 2014).
  - Step 1 – generator: generation of synthetic data by a neural network trying to mislead the discriminator,
  - Step 2 – discriminator: prediction of the probability that a data is real.

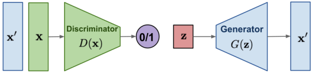

Figure: GAN.

CGAN

- Conditional Tabular Generative Adversarial Networks (CGANs): to address the problem of the generation of under-represented modalities (Xu & Veeramachaneni, 2018; Xu et al., 2019).
  - Conditional sampling,
  - Sampling by diversified mini-batches.
- High proportion of pre-parameterisation.
- CTAB-GAN: improving of the CGAN by Zhao et al. (2021).
  - Use of Gaussian mixture models during the data pre-processing.

- Directional Acyclic Tabular Generative Adversarial Networks (DAT-GANs): includes an ”expert” dimension (Lederrey et al., 2021).
  - Addition of a directional acyclic graph containing information about the causal links between input variables,
  - Generation of ”root” variables, then generation of variables linked to the ”root” variables by a causal relationship, and so on.

TRGAN

2 State of the art

- Transactional Generative Adversarial Networks (TRGANs): specialising in synthetic transactions generation (Zakharov et al., 2024).
  - Step 1 – creation of a conditional vector composed of the encoded real data, along with the date variables which are mathematically transformed.
    - Real data: addition of a transaction frequency variable,
    - Date variable: transformation by cosine and sine functions to introduce a notion of cyclicity.
  - Step 2 – generator,
  - Step 3 – evaluation of the generator by a discriminator,
  - Step 4 – synthetic data generation by a supervisor on the basis of real and generator's synthetic data,
  - Step 5 – evaluation of the supervisor's results by the discriminator.

Synthesis

- The non-graph generative models are mainly oriented towards computer vision.
- The generative graph models are mainly:
  - Molecular analysis oriented,
  - Focused on relational structures (edges) rather than on variables.
- Among the transaction-oriented models: no emphasis on the relational aspect.
- Selection of the TRGAN for:
  - Its introduction of the notion of temporal cyclicity of transactions,
  - Its dual learning principle (generator then supervisor).

Summary

3 Modelisation

GDS

3 Modelisation

- Construction of a TRGAN (Zakharov et al., 2024) customised for SWIFT messages.

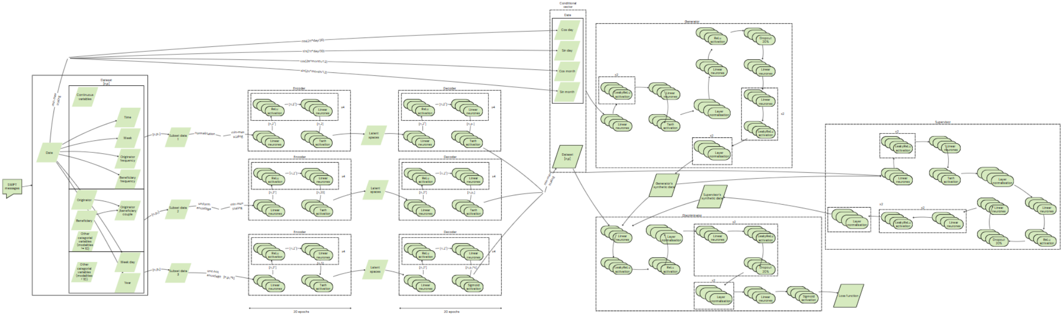

Figure: GDS model.

- Construction of a TRGAN (Zakharov et al., 2024) customised for SWIFT messages.

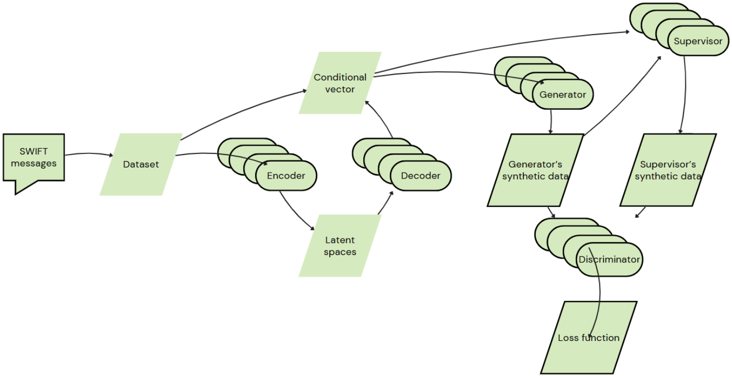

Figure: GDS model.

Real data analysis

3 Modelisation

- Existence of multiple SWIFT message formats. For the ISO 15022 format:
  - MT1**: customer payments and cheques,
  - MT2**: financial institution transfers,
  - MT3**: foreign exchange, money markets and derivatives,
  - MT4**: collateral claim,
  - MT5**: securities markets,
  - MT6**: commodities, syndication and reference data,
  - MT7**: documentary credits and guarantees,
  - MT8**: travellers cheques,
  - MT9**: cash management and customer status.

Data pre-processing
3 Modelisation

- According to the methodology of Zakharov et al. (2024) used to pre-process data, we have a division of real data into three categories for three different pre-processing neural networks:
  — Continuous,
  — Categorial with high modalities,
  — Categorial with low modalities.

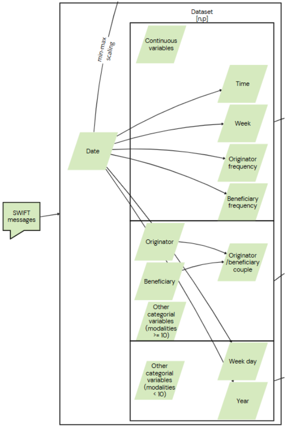

Figure: Real data.

Data pre-processing

- Computing of additional variables:
  - Initiator/beneficiary couple: introduction of a relational dimension,
  - Initiator frequency and beneficiary frequency: introduction of a customer behaviour dimension,
  - Schedule, week day, week, year: addition of precision in the temporal dimension of the generation.

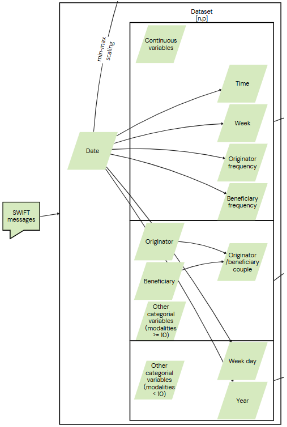

Figure: Real data.

Data pre-processing

3 Modelisation

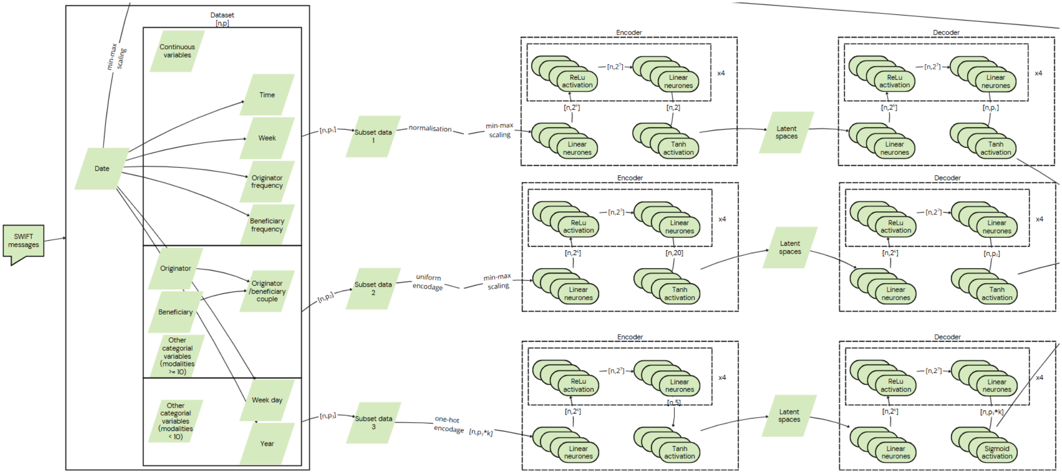

Figure: Pre-processing of real data.

Data pre-processing

3 Modelisation

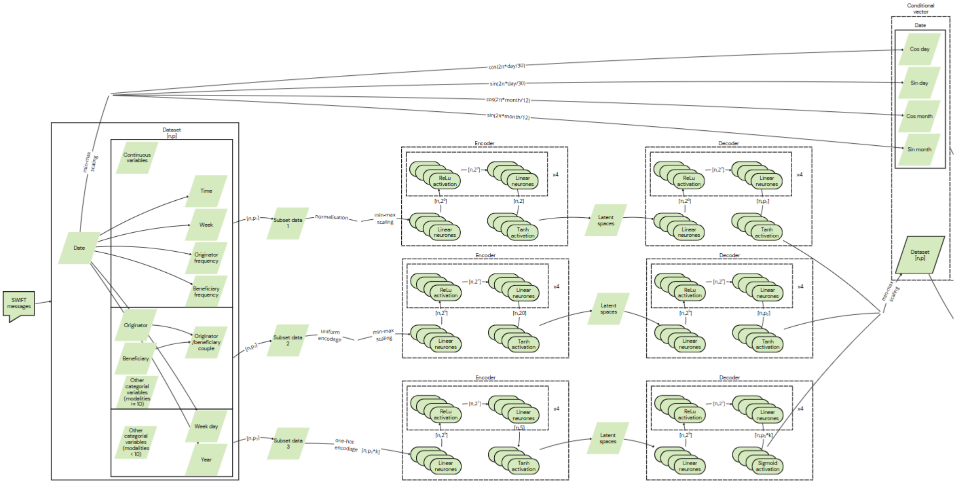

Figure: Construction of the conditional vector.

- TRGAN model of Zakharov et al. (2024) consists of a sequential passage through a generator, a discriminator, a supervisor, then a second time through the discriminator, in an antagonistic training approach.

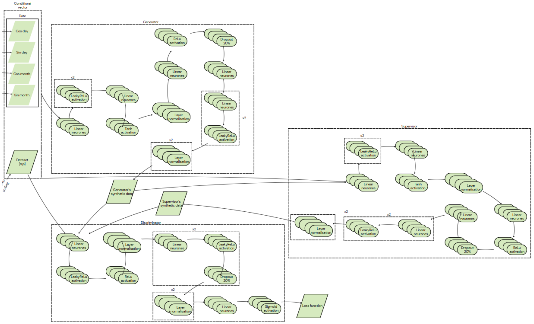

Figure: TRGAN (Zakharov et al., 2024).

Summary

4 Results

Metrics for continuous variables

- Let X be real data, $\hat{X}$ synthetic data, X_j a variable of the real data, $n = dim(X_j)$, $(n, p) = dim(X)$ and F_{X_j} the distribution function of the variable X_j.
- Mean absolute percentage error:

$$MAPE_j = \frac{100}{n} \sum_{i=1}^{n} \frac{X_{ij} - \hat{X}_{ij}}{X_{ij}}$$

- Kolmogorov-Smirnov test (Hodges, 1958):

$$KS_j = \frac{1}{n} \underset{x}{sup} |F(\hat{X}_{ij} \leq x) - F(\hat{X}_{ij} \leq x)|$$

Metrics for categorical variables

- χ^2 test statistic:

$$\chi_j^2 = \frac{1}{100} \sum_{x \in modalities(X_j)} (F(\hat{X}_{ij} = x) - F(X_{ij} = x))^2$$

- Kullback-Leibler divergence (1951):

$$KL_j = \sum_{x \in modalities(X_j)} F(\hat{X}_{ij} = x) ln \frac{F(\hat{X}_{ij} = x)}{F(X_{ij} = x)}$$

Global metric

- Similarity score:

$$s = \frac{1}{2p}\sum_{j=1}^{p}\left(\left(\left(1 - \frac{MAPE_j}{100}\right)^+ + KS_j\right)\mathbf{1}_{\{\mathbf{R}\}} - \left(\left(1 - \frac{\chi_j^2}{100}\right)^+ + KL_j\right)\mathbf{1}_{\{\mathbf{String}\}}\right)$$

Real data

- Number of observations: 10 000 SWIFT messages,
- Number of variables: 11,
- 59% of MT103 and 41% of MT202,
- Continuous variable: transaction amount,
- Categorial variables=: message type, initiator, beneficiary, date, schedule, sender, receiver, currency, fraud, transaction reference,
- Objective: to generate 2 millions of messages.

Univariate and multivariate results

- Similarity score: 99.585%,
- $KS_{Amount} = 0.0398.$

Différence des Corrélations (Réelle - Synthétique)

	Originator	Sender	Receiver	Beneficiary	Date	Time	Type	Currency	Value	Flag
Originator	0	0.04324495	−0.02381661	−0.0999511	−0.1609677	−0.04376505	−0.004944448	−0.01072261	−0.3547278	0.02025631
Sender	0.04324495	0	−0.02194392	−0.1123518	−0.1563858	−0.04376374	−0.003287369	−0.002292697	−0.3219987	0.01484719
Receiver	−0.02381661	−0.02194392	0	0.05869428	−0.001838543	0.006130999	−0.05197954	−0.1675487	−0.1099415	0.01369106
Beneficiary	−0.0999511	−0.1123518	0.05869428	0	−0.01625657	0.01675086	0.002379114	−0.06351123	−0.08223024	0.03041623
Date	−0.1609677	−0.1563858	−0.001838543	−0.01625657	0	−0.004722803	−0.03272004	0.2252684	−0.4525308	0.01635531
Time	−0.04376505	−0.04376374	0.006130999	0.01675086	−0.004722803	0	0.002477824	0.005582936	−0.01042	301.2995µ
Type	−0.004944448	−0.003287369	−0.05197954	0.002379114	−0.03272004	0.002477824	0	0.01524166	−0.1188554	−0.04599615
Currency	−0.01072261	−0.002292697	−0.1675487	−0.06351123	0.2252684	0.005582936	0.01524166	0	0.1693431	−876.4004µ
Value	−0.3547278	−0.3219987	−0.1099415	−0.08223024	−0.4525308	−0.01042	−0.1188554	0.1693431	0	0.04678552
Flag	0.02025631	0.01484719	0.01369106	0.03041623	0.01635531	301.2995µ	−0.04599615	−876.4004µ	0.04678552	0

Variables constantes (exclues de la corrélation) : Transaction Ref

Figure: Difference between correlation matrices for real and synthetic data.

Results graph view

4 Results

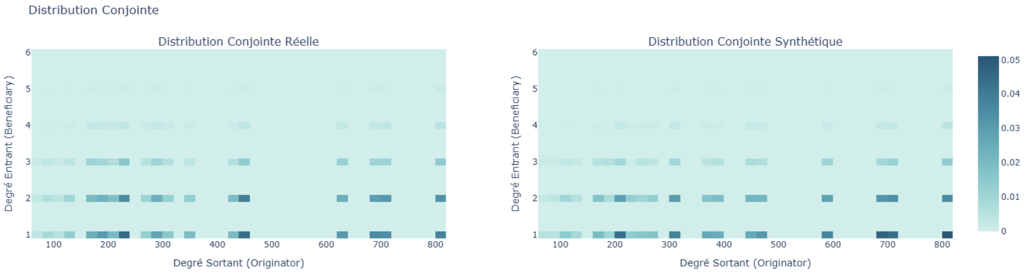

Figure: Joint distribution of real and synthetic data.

Summary

5 Conclusion

Conclusion

- Relatively well-preserved graph structure,
- Efficiency of our TRGAN model adjusted to SWIFT data,
- Required adjustments:
  - Correlation between the amount and date variables (over-correlation, and therefore also between the currency and date variables),
  - Temporal distribution of data (good preservation of the intra-daily distributions but improvement required at the daily and supra-daily levels).
- Development of a DAT-TRGAN: combination of a DAT-GAN and a TRGAN to include causal relationships between variables,
  - Particularly interesting regarding the sequencing of SWIFT messages of different categories.
- To divide the generation according to the main typologies of SWIFT messages,
- Development of a metric combining assessment of the relational structure, and univariate and multivariate joint distributions.

Summary

6 References

References

- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, *3*(11).
- Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan & H. Lin (Eds.), *NIPS'20: proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 6840-6851), Red Hook: Curran Associates Inc.
- Hodges, J. L. Jr. (1958). The significance probability of the Smirnov Two-Sample Test. *Arkiv fiur Matematik*, *3*(43): 469-486.
- Kingma, D. P., & Welling, M. (2019). An introduction to Variational Autoencoders. *Foundations and Trends in Machine Learning*, *12*(4): 307-392.
- Kipf, T., & Welling, M. (2016). *Variational Graph Auto-Encoders* [Conference paper]. Bayesian Deep Learning: NIPS 2016 Workshop, Barcelona.

References

- Kullback, S., & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22: 79-86.
- Lederrey, G., Hillel, T., & Bierlaire, M. (2021). DATGAN: integrating expert knowledge into deep learning for population synthesis. In *Proceedings of the* 21*st* *Swiss Transport Research Conference* (pp. 1-21). Ascona: Swiss Transport Research Conference.
- Wang, H., Wang, J., Wang, J., Zhao, M., Zhang, W., Zhang, F., Xie, X., & Guo, M. (2018). GraphGAN: graph representation learning with Generative Adversarial Nets. In S. A. McIlraith & K. Q. Weinberger (Eds.), *AAAI'18/IAAI'18/EAAI'18: proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (pp. 2508-2515). New Orleans: AAAI Press.

References

- Xu, L., & Veeramachaneni, K. (2018). *Synthesizing tabular data using Generative Adversarial Networks* [Work paper].
- Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). *Modeling tabular data using Conditional GAN* [Papier de conférence]. 33^{rd} Conference on Neural Information Processing Systems 2019, Vancouver.
- You, J., Ying, R., Ren, X., Hamilton, W. L., & Leskovec, J. (2018). GraphRNN: generating realistic graphs with Deep Auto-Regressive Models. In J. Dy & A. Krause (Eds.), *Proceedings of the* 35^{th} *International Conference on Machine Learning (ICML 2018), volume 6* (pp. 4320-4329). Red Hook: Curran Associates Inc.

References

- Zakharov, K., Stavinova, E., & Lysenko, A. (2024). TRGAN: a Time-Dependent Generative Adversarial Network for synthetic transactional data generation. In *ICSeB'23: proceedings of the 2023* 7^{th} *International Conference on Software and E-Business* (pp. 1-8). New York: Association for Computing Machinery.
- Zhao, Z., Kunar, A., Birke, R., & Chen, L. Y. (2021). CTAB-GAN: effective table data synthesizing. *Proceedings of the* 13^{th} *Asian Conference on Machine Learning, PMLR*, 157: 97-112.

Summary

7 Annexes

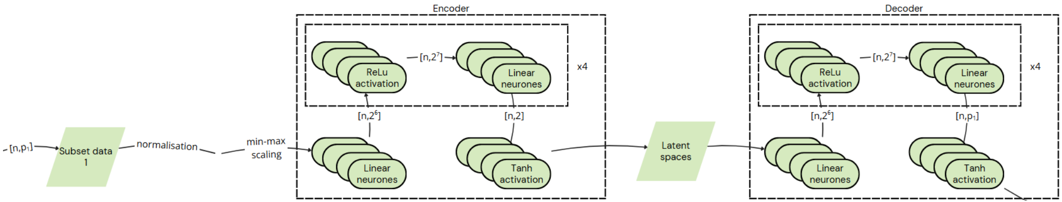

Figure: Pre-processing neural network of continuous variables (Zakharov et al., 2024).

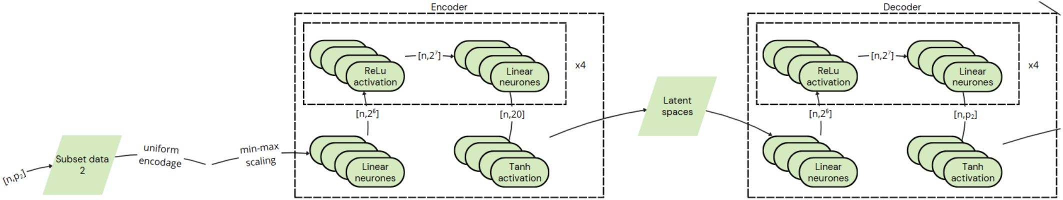

Figure: Pre-processing neural network of categorical variables with high modalities (Zakharov et al., 2024).

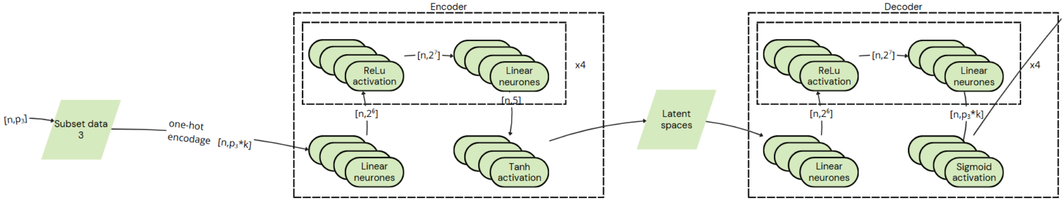

Figure: Pre-processing neural network of categorical variables with low modalities (Zakharov et al., 2024).

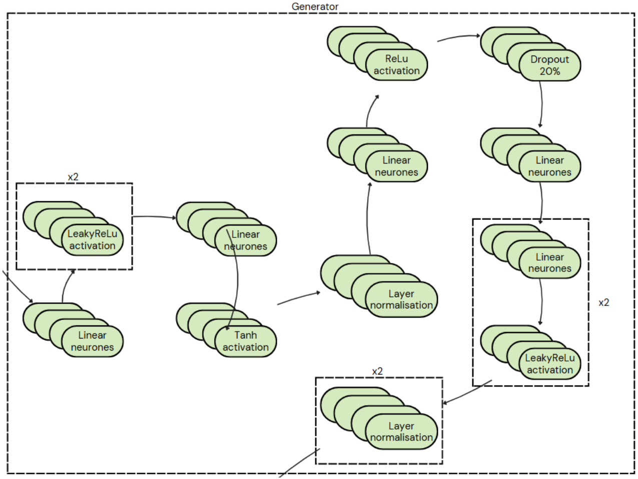

Figure: Generator (Zakharov et al., 2024).

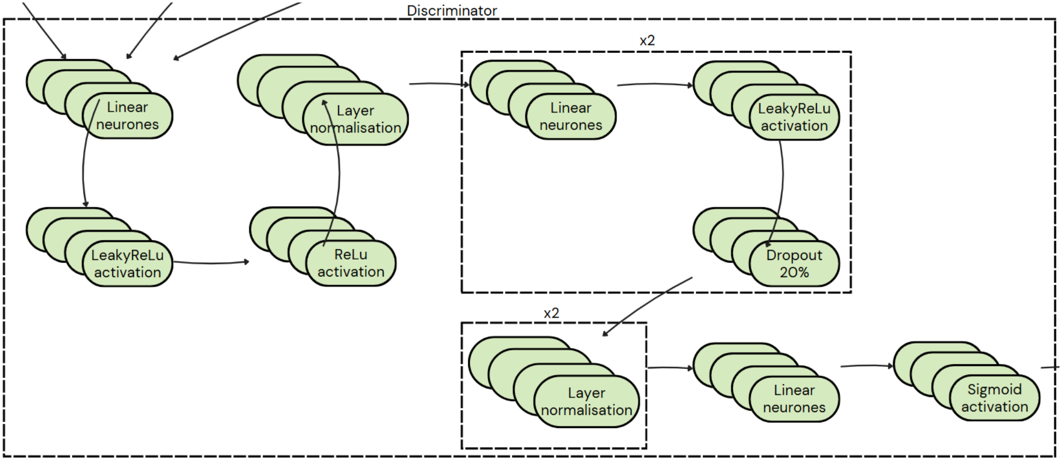

Figure: Discriminator (Zakharov et al., 2024).

Training
7 Annexes

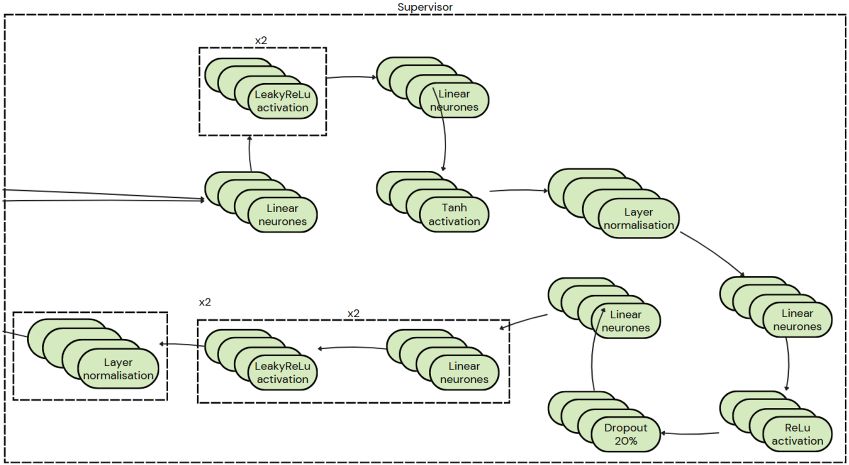

Figure: Supervisor (Zakharov et al., 2024).

Results
7 Annexes

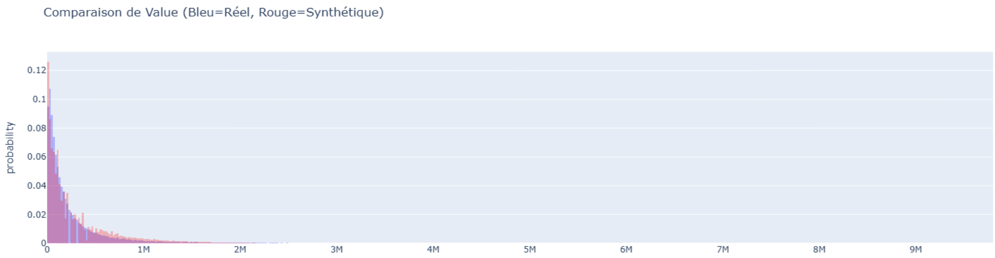

Figure: Real and synthetic distributions of the amount variable.

Results
7 Annexes

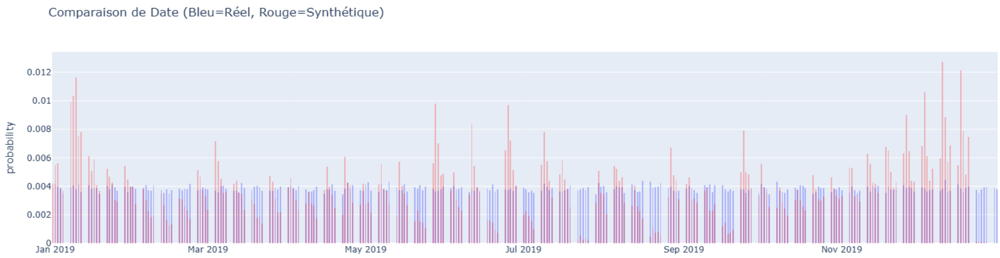

Figure: Real and synthetic distributions of the date variable.

Results
7 Annexes

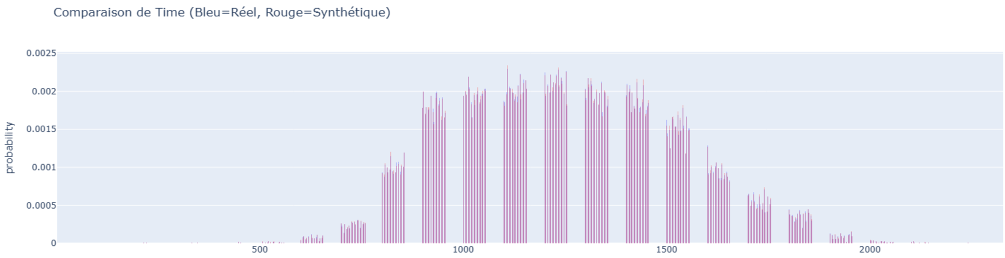

Figure: Real and synthetic distributions of the schedule variable.

How to cite

Alfred, Romain A., & Lemonnier, Thomas (2025). GDS: a SWIFT-specialised TRGAN for synthetic financial transaction data generation [Conference presentation]. 11^{e} Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle, Dijon.