

Study on Advanced Simulation of Non-Normal Customer Credit Data and Development of New Approaches for In-Depth Credit Risk Analysis

Alicia Gardette¹ and Olivier Gassner²

¹ Atheïa, 17 rue Choiseul, Paris, France - alicia.gardette@atheia.fr

² Atheïa, 17 rue Choiseul, Paris, France - olivier.gassner@atheia.fr

Abstract. Credit risk analysis is fundamental to financial stability, yet the scarcity of granular data and the inadequacy of traditional Gaussian-based simulation methods present significant barriers to model development. This paper presents a framework for the advanced simulation of non-normal credit risk data to train and evaluate credit risk analysis methods. This paper details a novel methodology for generating synthetic credit datasets that faithfully replicate socio-demographic distributions, accounting rules, and industry-standard data models. By utilizing the Foldnes-Olsson method for non-normal variance-covariance matrices, a time-reset Markov chain for Probability of Default (PD) simulation, and a two-step Beta-Recovery Speed approach for Loss Given Default (LGD), we demonstrate a robust alternative to conventional simulation. Our results show that the simulated data accurately captures macroeconomic dependencies and historical economic cycles, providing a high-fidelity tool for applied research.

Keywords: Synthetic Data, Credit Risk, Non-Normal Distribution.

1 Introduction

Synthetic data offer a way to bypass data scarcity, providing a technical solution when real-world information is restricted by privacy issues or compromised by bias [1]. They can thus be valuable in multiple areas, such as healthcare [2], cybersecurity [3], supply chain [4], or finance [5]. In particular, credit risk analysis is a cornerstone of finance, where model accuracy is vital for loan portfolio management. However, governance and security remain critical hurdles, as financial institutions only share such data with national or international supervisory authorities (e.g., ECB, PRA, ACPR). Several studies addressed this limitation by developing solutions to compensate for the lack of accessible, high-quality datasets, for example through adversarial learning [6], robust support vector machines [7], or by considering additional data such as social media data [8] or Artificial Intelligence news [9].

Until recently, the literature on generating synthetic data for the banking and financial sector was scarce. From 2018 to 2023, 72 papers were published on this topic and,

among them, only 9 dealt with credit risk. Moreover, this literature appears difficult to collect as many articles do not explicitly mention finance as a research focus [10]. In the last couple of years, the literature starts to highlight the importance and the need of generating synthetic data in credit risk. The first stake is to mitigate the imbalance of default events, by recreating statistical default characteristics. A second issue is to enhance credit scoring with multi-modal sources of information, such as text loan applications, online transaction patterns, customer communications... Synthetic data can also help to generate stress test scenarios for regulatory compliance, enabling banks to simulate portfolio effects more faithfully and more aggressively to allocate accordingly their capital [11].

Our research explores advanced methods for simulating non-normal credit data, focusing on robust models that integrate the diversity and heterogeneity of real-world portfolios. Specifically, the goals and corresponding contributions are:

- Goal 1. Provide sophisticated tools (both a method for the generation of synthetic data, and the resulting dataset) for in-depth risk analysis while respecting credit data granularity across temporal, sectoral, and geographical dimensions. The complexity of credit data—often characterized by non-normal distributions—challenges traditional simulation methods. Standard approaches based on normality assumptions fail to capture the asymmetries and generally prevent modeling such nuances.
- Goal 2. Discuss the generalizability of the method to other fields.

In order to achieve these goals, we apply the following methodology. We focus on credit risk analysis through various data types: contractual (e.g., loan amount, duration), behavioral (e.g., payment delays), and socio-demographic (e.g., age, income). These simulations aim to replicate i) accounting rules applied by credit institutions, ii) socio-demographic and sectoral composition of portfolios, and iii) standard industry data models. The project addresses major scientific bottlenecks, such as the consistent reproduction of interdependencies between socio-demographic variables (e.g., age and family status) and the incorporation of macroeconomic factors (e.g., CPI or disposable income) into portfolio dynamics.

The rest of this paper is structured as follows. Section 2 presents the related work. Section 3 introduces the proposed framework and Section 4 presents the simulation assessment and controls. Finally, Section 5 concludes this paper.

2 Literature Review

To tackle data scarcity, several methods are proposed in the literature. Historically, a first set of methods is based on a statistical approach: simulations based on Monte-Carlo to acknowledge rare and extreme events [12], the use of copulas to model the interdependency between variables [13], sequential methods to successively integrate variables [14], non-parametric methods to overcome the normality hypothesis which is not proven in credit risk data [15]. A second set of methods to generate synthetic data relies on an artificial intelligence approach: Generative Adversarial Networks (GANs)

and Variational Autoencoders (VAEs) [10], Large Language Models (LLM) [16], diffusion methods [11].

However, the methods explored in the literature suffer from multiple challenges and limitations. The first one deals with the quality of the generated data as the methods struggle to generate realistic data and data that preserve the expert-logic relationship expected [17]. Another challenge faced by the generation of synthetic data in credit risk is to reflect non-linear dynamism, especially when the methods have difficulties to integrate macro-economic factors [15, 18]. A third challenge of synthetic data concerns the representativeness of the generated data, as the methods manage to reproduce the main pattern of the observed features but fail to replicate extreme points, especially when the observed distributions present complexity or irregularity [19]. One last limitation of using AI algorithms to generate synthetic data in credit risk is called “double opacity” and is linked to the regulatory obligation to have explainable and interpretable credit risk models. Indeed, using a black box method to generate data on which a machine learning model, also seen as a black box by the regulators, makes it difficult to explain such credit risk models and they might, as well, suffer from over-fitting [20].

The review of literature shows a growing concern for the generation of synthetic data in credit risk. However, even if several methods are under scrutiny, they are facing multiple challenges that threaten their direct application in the strictly regulated domain of credit risk.

3 Proposed Framework

3.1 General Concepts

The objective is to replicate the behavior of economic agents (or counterparties) within the credit repayment process. The risk associated with this behavior is decomposed into two risk parameters: Probability of Default (PD) and Loss Given Default (LGD). Rather than directly simulating default occurrences, **we model an underlying random variable whose realizations represent the complete or partial repayment cash flows**, including payment delays. Main steps of the simulation process are briefly introduced hereafter and detailed in the subsequent sections.

The simulation consists of mapping a-priori defined risk parameters (e.g., PD, LGD) into socio-demographic, sectoral, and contractual risk characteristics, while strictly preserving the correlations across these three dimensions. These correlations ensure that the portfolio structure is faithfully reflected, both at the time of credit granting and throughout the lifecycle of the contract. Expected correlations are user-defined and may be guided by national or international statistics institutions. By applying scores conditional to specific counterparty features, the model enables the simulation of repayment cash flows and associated behavioral information. The relationship between risk drivers and risk parameters, in terms of intensity, direction and non-linearity, can be fine-tuned by the user.

Then, the application of accounting rules to these repayment flows (or payment delays), both prior to and during potential default, leads to the construction of a credit

amortization schedule and conditions the occurrence of default as well as the resulting economic loss. A similar mechanism is applied to simulate the utilization of revolving facilities, thereby conditioning the exposure at risk and the credit maturity. This method enables the simultaneous simulation of explanatory variables (risk characteristics) and a target random variable—such as a default rate (resp. loss, exposure at risk)—whose expected value corresponds to the a-priori defined PD (resp. LGD, EAD).

Finally, theoretical PD is converted into conditional probabilities of arrears. Random drawings into Bernoulli law allow for simulating arrears. Default is triggered when arrears occur during a 90-day period, replicating one of the major features of the Basel Definition of default. Realized LGD is drawn into a Beta law, whose parameters can be adjusted for reflecting expected distribution in terms of convexity and asymmetry. Economic loss is derived from the realized LGD and broken down into regular cash-flows ventilated through the time spent in default. Cash-flows dynamics are set up according to a predefined recovery profile modelled via the well-known method of recovery speed.

3.2 Non-Normal Obligor Panel Simulation

We drew inspiration from the research of Foldnes and Olsson for the simulation of a panel of obligors presenting complex dependencies [21]. The method proposed by the authors enables the generation of a matrix of non-standard distributions—characterized by significant skewness or heavy tails—derived from a linear combination of independent non-normal variables.

This method relies on the numerical resolution of a system of equations involving skewness, kurtosis, the variance-covariance matrix, and the generating variables matrix. This resolution utilizes the Pearson distribution system, a family of probability distributions defined by Pearson’s differential equations, which are used to generate distributions with specific moments—namely mean, variance, skewness, and kurtosis. This family comprises seven distribution types, including asymmetric Beta and Gamma distributions. The skewness and kurtosis for each variable, as well as the variance-covariance matrix, are specified as inputs by the user.

3.3 Idiosyncratic Factors Scoring

A wide variety of approaches may be used for designing a scoring model intending to reflect idiosyncratic factors. Our proposal relies on two main components. First, a scorecard defines a linear relationship mapping the model’s exogenous variables and their respective parameters to a latent variable—the score, as described by equation (1). Second, an activation function performs a bijection—linear for LGD and non-linear for PD—that maps each score value to the corresponding final risk parameter (cf. (2)).

$$Score = f(X|\beta) = \beta_0 + \sum_i(\beta_i X_i) \quad (1)$$

$$Parameter = g(Score|\gamma) \quad (2)$$

This parameterization requiring the dichotomization of each explanatory variable allows for designing non-linear association patterns without involving complex transformation functions. Linear activation functions, such as those utilized for the LGD parameter, are straightforward to configure. The process involves assigning location and scale parameters to map the score range onto the desired LGD range, defined by LGD_{Min} and LGD_{Max} , as exposed in (3).

$$LGD = g(s) = (s - Score_{Min}) \cdot \left(\frac{LGD_{Max} - LGD_{Min}}{Score_{Max} - Score_{Min}} \right) + LGD_{Min} \quad (3)$$

Non-linear activation functions are more complex to parameterize. It remains a function that maps a theoretical risk level to a score, while maintaining an equivalence relationship between a score point difference and an odds ratio across the entire domain of the function. This type of function can be expressed as a “Points to Double the Odds” function (PDO), generalized as it follows:

$$PD = g(s|\{\mu, \alpha, \beta\}) = \beta^{(s-\mu)/\alpha} \cdot (1 + \beta^{(s-\mu)/\alpha})^{-1} \quad (4)$$

Where:

- μ defines the median score such as $g(s|\{\mu, \alpha, \beta\}) = \frac{1}{2}$
- α defines the number of points to multiply the odds
- β defines the odds multiplier

3.4 Arrears Probability Optimization

The simulation of the transition toward default is modelled with a homogeneous semi-Markov chain, referred to as “clock-reset” [22], aiming at decomposing the 90-day past due default trigger into three probabilities—in a monthly snapshot approach—linked to arrears, under specific constraints. Those probabilities (i.e., $P0$ for the probability of an arrear in t , without an arrear in $t-1$; $P1$ for the probability of an arrear in t given an arrear in $t-1$; $P2$ for the probability of an arrear in t given an arrear in $t-1$ and in $t-2$) are to be determined such as their product approximates an instantaneous arrear probability $PD(t)$ that is derived from an annual probability of default $PD_{12}(t)$, according to (5).

$$PD(t) = \prod_{a=0}^2 (P_a(t)) = 1 - (1 - PD_{12}(t))^{(1/12)} \quad (5)$$

This formula assumes that the probability of default is uniformly distributed across each month of the year. The optimization process objective function computes the absolute difference between $PD(t)$ and $\prod_{a=0}^2 (P_a(t))$, adding a penalty if the constraint $P_0 < P_1 < P_2$ is not met, to ensure that the probability of an arrear increases with consecutive arrears in the previous periods. Optimization relies on the BFGS (Broyden, Fletcher, Goldfarb and Shanno) method, which is an iterative method for solving non-linear optimization problems [23].

3.5 Beta Distribution Configuration for Realized LGD Simulation

Distribution Parameters. We retained the Beta law for the theoretical distribution of LGD. Its two parameters α and β allow to simulate asymmetrical convex distributions, and its theoretical moments are easily computable. The convex shape of LGD distribution observed in real portfolios is obtained if both parameters are lower than 1. Finally, the expected value is aligned with the theoretical LGD.

$$\alpha < 1; \beta < 1; E(X) = \frac{\alpha}{\alpha + \beta} = LGD_{th} \quad (6)$$

α and β are linked to the expected value μ and the variance σ^2 of the Beta law as per (7), whose estimates are obtained by the method of the moments as exposed in (8).

$$\begin{cases} \mu = \frac{\alpha}{\alpha + \beta} \\ \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \end{cases} \quad (7)$$

$$\begin{cases} \hat{\alpha} = \left(\frac{\hat{\mu}(1 - \hat{\mu})}{\hat{\sigma}^2} - 1 \right) \hat{\mu} \\ \hat{\beta} = \left(\frac{\hat{\mu}(1 - \hat{\mu})}{\hat{\sigma}^2} - 1 \right) (1 - \hat{\mu}) \end{cases} \quad (8)$$

Where $\hat{\mu}$ and $\hat{\sigma}^2$ are respectively the estimators of the mean and variance of the sample. Common practice to rely on a scale parameter ϕ , which is easier to use than the variance and which defines the relationship reported in (9).

$$\begin{cases} \alpha = \mu \cdot \phi \\ \beta = (1 - \mu) \cdot \phi \end{cases} \Leftrightarrow \begin{cases} \mu = \frac{\alpha}{\alpha + \beta} \\ \phi = \alpha + \beta \end{cases} \quad (9)$$

Influence of the Scale Parameter. Beta law configuration objective is to determine the value of ϕ such as to obtain the desired shape of the curve (e.g. concave, convex, bimodal). Figure 1 illustrates how the density of the center of the distribution increases with ϕ and shows that for low values of μ , the smoothing effect increases, whereas bimodality decreases with the value of ϕ .

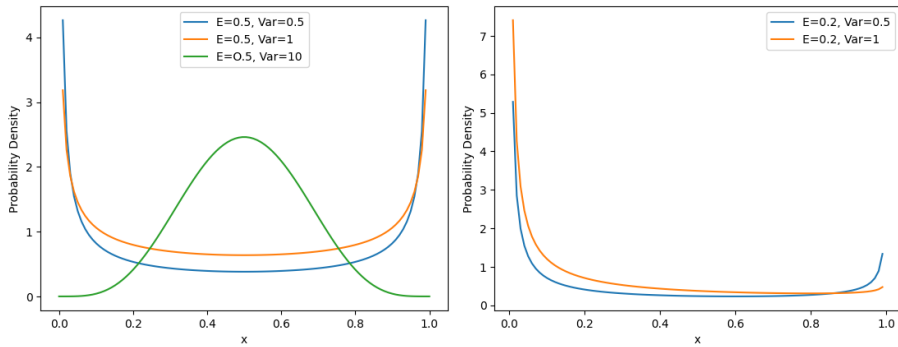


Fig. 1. Effect of ϕ on distribution convexity.

Configuration of the Scale Parameter. Different curve profiles can be combined by expressing ϕ as a function of μ . For instance, if the objective is to obtain a curve profile that combines (i) a significant smoothing effect on low values of μ (i.e. high values of ϕ) and (ii) low density at the center of the distribution (i.e. low values of ϕ) for median values of μ , then ϕ should be defined as a decreasing function of μ (e.g., inverse exponential, inverse square root, opposite natural logarithm).

An advanced approach consists in combining two different curve profiles (e.g., inverse square root, inverse exponential) to merge, within a single joint distribution, a significant smoothing effect on low losses and a low density at the center of the distribution. For instance, for a given μ , the following decomposition better captures those two characteristics than a single function:

- Beta distribution defined with parameters $\{\mu_1 = 0,5\mu; \phi_1 = \frac{1}{\sqrt{\mu_1}}\}$ emphasizing the smoothing effect on the lower part of the distribution.
- Beta distribution defined with parameters $\{\mu_2 = 1,5\mu; \phi_2 = \frac{1}{e^{\mu_2}}\}$ limiting the density of the distribution at its center.

By assigning an equal weight to those two distributions, the mean of the joint distribution is equal to the theoretical LGD. This configuration is illustrated by figure 2. Profile A is defined by $\mu = 0.25$, whereas profiles A1 and A2 are defined by $\mu_1 = 0.125$ and $\mu_2 = 0.375$, respectively. The blue area represents the empirical distribution of the resulting combination, that is closer to expected shape—in terms of smoothness and central density—than profile A.

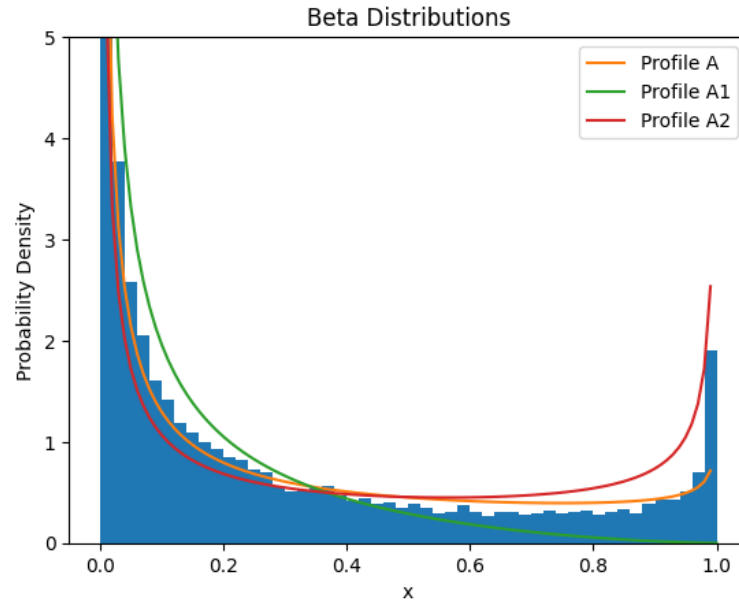


Fig. 2. Probability density function and empirical distribution depending on the combination of two underlying distributions.

3.6 Recovery profile definition

Recovery profile defines the dynamics of the cash-flows, i.e. the way the amount of capital recovered by the financial institution is collected across the snapshots spent in default. One generally distinguishes two recovery profiles: one for the cured facilities (i.e., returning to performing status without economic loss) and one for the non-cured facilities (i.e., closing with economic loss). As an infinite number of cash-flow distributions can lead to the same realized LGD, we need to provide a model for both cured and non-cured facilities.

Cured Facilities. Economic loss is assumed to be exclusively explained by discounting effect. We propose to spread cash-flows uniformly throughout the default period, which corresponds to a linearly increasing recovery curve, in line with an up-to-date loan repayment schedule leading to return into performing.

Discounting effect is a function of the time spent in default. This duration is determined such that the sum of the net cash-flows (current installment, arrear, and artificial cash-flow) is equal to the EAD and such that the sum of discounted cash-flows is equal to the difference between EAD and the economic loss. We retained a grid search strategy for defining such duration.

Non-Cured Facilities. Total recovery amount is set as the difference between EAD and economic loss. Real portfolios empirically evidence the decreasing recovery speed of cash-flows, such as the cumulative recovery profile assumes a concave shape. We propose to use the well-known recovery speed approach for modelling the cash-flow sequence. The cumulative recovery rate is then defined by equation (10).

$$CRR_n = (1 - (1 - v)^n) \cdot (1 - LGD_{Realized}) \quad (10)$$

In the first period, a proportion of this amount (i.e., the recovery speed v) is recovered. In the second period, the remaining balance to recover is computed as the difference between the initial amount and the amount recovered in the first period. Then, the same proportion of this remaining balance is recovered, and so on. A minimum recovery amount per period is defined to ensure convergence.

The recovery speed can either be deterministic or stochastic. In this latter case, expressed by equation (11), the recovery speed v_i is defined by a random variable whose expected value is equal to a theoretical recovery speed defined a priori. A scale parameter (e.g. the variance) can influence the variability of the profiles.

$$CRR_n = (1 - \sum_{i=1}^n (1 - v_i)) \cdot (1 - LGD_{Realized}) \quad (11)$$

4 Simulation Assessment and Controls

Each step of the simulation process is subject to in-depth controls ensuring the alignment of simulated contracts with expected behavior in terms of risk level and dynamics. This section exposes the main controls that are performed at granular or portfolio level.

4.1 Probabilities of Default

The theoretical model performs a bijection between the idiosyncratic factors of the counterparties and a latent variable (i.e. the score). Simulation quality assessment consists in analyzing the accuracy of the migrations from performing to default status, through confidence intervals. Table 1 proves the precision of the simulation as, out of 18 tests, only one empirical default rate falls outside the confidence interval, representing a proportion consistent with the significance level of the interval.

Table 1. Default Rate and Probability of Default Confidence Intervals.

Score	PD12	Denominator	Numerator	%default	Lower bound	Upper bound	Inside
0.24	3.0%	4949	131	2.6%	2.5%	3.5%	TRUE
0.32	4.3%	3446	167	4.8%	3.6%	5.0%	TRUE
0.39	5.8%	3358	187	5.6%	5.0%	6.6%	TRUE
0.40	6.0%	68	6	8.8%	0.4%	11.6%	TRUE
0.44	7.2%	1564	102	6.5%	5.9%	8.5%	TRUE
0.47	8.1%	1290	102	7.9%	6.6%	9.6%	TRUE
0.52	10.0%	133	14	10.5%	4.9%	15.1%	TRUE
0.55	11.3%	340	28	8.2%	7.9%	14.7%	TRUE
0.59	13.3%	1290	179	13.9%	11.4%	15.2%	TRUE
0.60	13.8%	92	18	19.6%	6.8%	20.8%	TRUE
0.67	18.1%	48	12	25.0%	7.2%	29.0%	TRUE
0.69	19.5%	455	67	14.7%	15.9%	23.1%	FALSE
0.75	24.2%	506	118	23.3%	20.5%	27.9%	TRUE
0.77	25.9%	106	22	20.8%	17.6%	34.2%	TRUE
0.84	32.5%	351	110	31.3%	27.6%	37.4%	TRUE
0.85	33.5%	60	20	33.3%	21.6%	45.4%	TRUE
0.92	41.0%	33	14	42.4%	24.2%	57.8%	TRUE
1.00	50.0%	355	167	47.0%	44.8%	55.2%	TRUE

4.2 Macroeconomic Dynamics

One of the main contributions of this project lies in the incorporation of macroeconomic dynamics within the credit portfolio. Sensitivity of obligors is not homogeneous inside a portfolio, as riskiest profiles are the most exposed to macroeconomic conditions degradation. Time series of default rates (see Fig. 3) reflect the influence of macroeconomic trends (e.g., sovereign debt crisis in 2011, Covid-19 pandemic) with a lag consistent with the 12-month observation period required by the Basel definition of default. We clearly observe the highest sensitivity of riskiest profiles, as supported by empirical evidence.

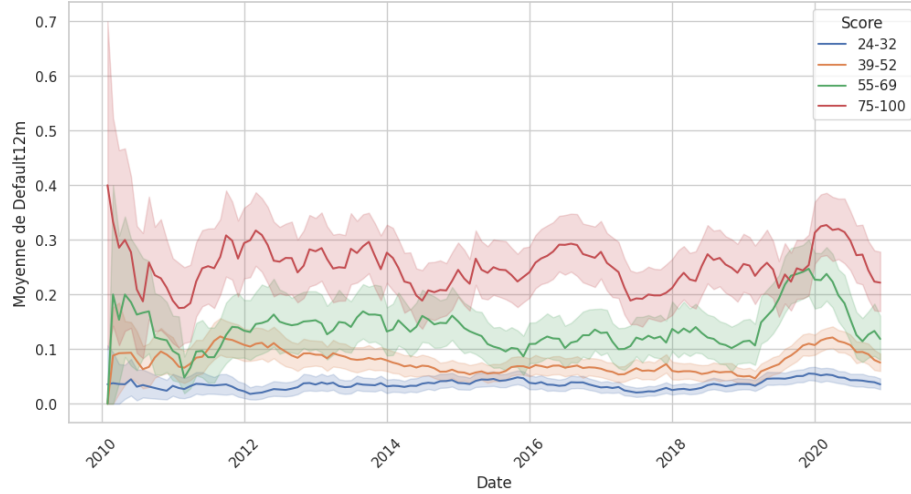


Fig. 3. Time series of 12-month Basel default rates by risk profiles – Portfolio sensitive to macroeconomic factors.

4.3 Econometric Model Parameters

To assess the heterogeneity of the simulated risk profiles, a logistic regression is fitted on all risk characteristics that are simulated, either included or not in the definition of the theoretical score. Table 2 reports the model's parameters' estimates. This control proves that only the variables included in the theoretical score (i.e., housing status, age, income, employment seniority, loan amount) have significant parameters, confirming the correct discriminatory power of the risk profiles. Moreover, the ordinal ranking of risk levels is maintained.

Table 2. Parameters estimates and significance test p-values.

Variable	Coef	Std Err	z	P> z	[0.025	0.975]
Intercept	-3,8174	0,134	-28,417	0.000	-4,081	-3,554
Marital status - Single	0,0355	0,058	0,614	0,539	-0,078	0,149
Housing status - Rental	0,7202	0,059	12,268	0.000	0,605	0,835
SPC - Farmers	0,3295	0,243	1,356	0,175	-0,147	0,806
SPC – Managers & Professionals	0,1723	0,124	1,392	0,164	-0,07	0,415
SPC – Service Workers	0,1760	0,121	1,455	0,146	-0,061	0,413
SPC – Manual Workers	0,1076	0,128	0,844	0,398	-0,142	0,358
SPC – Associate Professionals	-0,0093	0,124	-0,075	0,940	-0,251	0,233
Age bin <= 30	1,0132	0,132	7,699	0.000	0,755	1,271
Age bin 30 - 40	0,6907	0,117	5,922	0.000	0,462	0,919
Age bin 40 - 50	0,6315	0,107	5,868	0.000	0,461	0,802
Age bin 50 - 60	0,5402	0,094	5,772	0.000	0,357	0,724

Loan amount - bin < 8 000	-0,1575	0,062	-2,536	0,011	-0,279	-0,035
Income - < 850	1,5032	0,093	16,146	0.000	1,321	1,685
Income - 850 - 1 500	0,6735	0,086	7,869	0.000	0,506	0,841
Income - 1 500 - 2 000	0,0604	0,086	0,701	0,484	-0,109	0,229
Emp. Seniority - < 12	-0,0529	0,089	-0,596	0,551	-0,227	0,121
Emp. Seniority - 12 - 24	-0,1882	0,081	-2,328	0,020	-0,347	-0,030

5 Conclusion

Recent literature highlights the lack of high-quality granular data in the field of credit risk modelling. Our research underscores that such data, with significant historical depth, are indispensable for conducting robust applied research and objectively evaluating modeling performance. A major contribution of this study is the shift from direct risk parameter simulation to modeling the underlying repayment cash flows, which dictates default occurrences and economic loss. To the best of our knowledge, this inductive method has never been used in such framework. In addition, managing the temporal dimension enables the creation of balanced panels which are essential for a precise analysis of credit risk and allows for high sensitivity to macroeconomic conditions. By incorporating the Foldnes-Olsson method to capture skewness and heavy tails, the proposed approach strictly adheres to recent academic recommendations regarding the non-normality of financial distributions. As a next step, the high-fidelity database we have established may serve as a premier foundation for applying advanced synthetic data methods, such as Large Language Models (LLM) and generative AI. Furthermore, this advanced framework is highly adaptable across all financial sectors, climate risk assessment, and various industrial fields such as supply chain.

Acknowledgments. This study was funded by Atheia.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Capasso, M.: Synthetic data as meaningful data. On Responsibility in data ecosystems. *Big Data & Society*, 12(4), 20539517251386053 (2025)
2. Rahman, M. I., Hossain, M. R., Sayem, S. M., & Emdad, F. B.: Exploring the role of synthetic data in the future of ai in healthcare: A scoping review of frameworks, challenges, and implications. *Intelligence-Based Medicine*, 100342 (2025)
3. Artioli, P., Dentamaro, V., Galantucci, S., Magri, A., Pellegrini, G., & Semeraro, G.: SIEVE: Generating a cybersecurity log dataset collection for SIEM event classification. *Computer Networks*, 266, 111330 (2025)
4. Long, Y., Kroeger, S., Zaeh, M. F., & Brintrup, A.: Leveraging synthetic data to tackle machine learning challenges in supply chains: challenges, methods, applications, and research opportunities. *International Journal of Production Research*, 1-22 (2025)

5. Assefa, S. A., Dervovic, D., Mahfouz, M., Tillman, R. E., Reddy, P., & Veloso, M.: Generating synthetic data in finance: opportunities, challenges and pitfalls. In Proceedings of the first ACM international conference on AI in finance (pp. 1-8) (2020)
6. Lin, J., & Yang, L.: Mitigating algorithmic bias in credit scoring support systems through adversarial learning. *Decision Support Systems*, 114653 (2026)
7. Yan, D., & Zhang, M.: The nature of statistical learning Prediction of Financial Distress for Chinese Listed Manufacturing Companies given Uncertain Financial Accounting Data Environments. *Information Systems Frontiers*, 1-27 (2025)
8. Zhou, Z., Chen, Z., Zhang, Y., Lu, T., & Lu, X.: Social media meets FinTech platforms: How do online emotions support credit risk decision-making?. *Decision Support Systems*, 195, 114471 (2025)
9. Zhang, J., Wang, Z., Zhao, W., & Yan, Z.: The Impact of AI News on Default Risk of Chinese Listed Companies: Forecasting Based on Deep Learning. *Information Systems Frontiers*, 1-16 (2025)
10. Meldrum J., Suleiman B., Rabhi F., & Muhammad J. A.: New Money: A Systematic Review of Synthetic Data Generation for Finance. *arXiv:2510.26076* (2025)
11. Kakkar S.: Advancing credit risk modeling through generative artificial intelligence: methods, applications and, challenges. *International Journal of Financial Data Science (IJFDS)* Volume 3, Issue 2, July-December 2025, pp. 30-48 (2025)
https://doi.org/10.34218/IJFDS_03_02_003.
12. Van den Bergh, O. C.: Assessing concentration and diversification in portfolio credit risk models (2020)
13. Gallo, V.: Large portfolios credit risk analysis with LT-Archimedean copulas and application to a case of securitised ABS (Doctoral dissertation, Politecnico di Torino) (2023)
14. Nicolaie, M. A., Füssenich, K., Ameling, C., & Boshuizen, H. C.: Constructing synthetic populations in the age of big data. *Population Health Metrics*, 21(1), 19. (2023)
15. Kowsar, M. M., Mohiuddin, M., & Islam, S.: Mathematics for finance: A review of quantitative methods in loan portfolio optimization. *International Journal of Scientific Interdisciplinary Research*, 4(3), 01-29 (2023)
16. Golec M., & AlabdulJalil M.: Interpretable LLMs for Credit Risk: A Systematic Review and Taxonomy. *arXiv:2506.04290v2* (2025)
17. Garg S., & Torra V.: Knowledge Informed GANs for Privacy-Preserving Synthetic Tabular Data. United Nations Economic Commission for Europe, conference in October 2025.
18. Anagnostou, I.: Risk management in trading activities through the lens of complex systems theory. Universiteit van Amsterdam (2020)
19. Adiputra I. N. M., Lin P.C., & Wanchai P.: The Effectiveness of Generative Adversarial Network-Based Oversampling Methods for Imbalanced Multi-Class Credit Score Classification. *Electronics* 2025, 14, 697 (2025). <https://doi.org/10.3390/electronics14040697>
20. Spears T., Bondo Hansen K., Xu R., & Millo Y.: Governing synthetic data in the financial sector. *Finance and Society* (2025), 1–17. <https://doi.org/10.1017/fas.2025.10017>.
21. Foldnes, N., & Olsson, U. H.: A Simple Simulation Technique for Nonnormal Data with Prespecified Skewness, Kurtosis, and Covariance Matrix. *Multivariate Behavioral Research*, 51(2–3), 207–219 (2016). <https://doi.org/10.1080/00273171.2015.1133274>
22. Asanjarani A, Lique B, Nazarathy Y.: Estimation of semi-Markov multi-state models: a comparison of the sojourn times and transition intensities approaches. *Int J Biostat.* (2021) Jan 6;18(1):243-262. doi: 10.1515/ijb-2020-0083. PMID: 35641138.
23. Nocedal J., and Wright S. J.: Numerical Optimization. Springer New York (2006)