Imperial College London

Department of Earth Science and Engineering

MSc in Environmental Data Science and Machine Learning

Independent Research Project

Project Plan

Copula-VAE Hybrid Synthesis for Small-Sample Geoscience Data

By

Xiangxin Zhao

Email: xz2821@ic.ac.uk

GitHub username: esemsc-xz2821

Repository: https://github.com/ese-ada-lovelace-2024/irp-xz2821

Supervisors:

Dr. Paulo Lopes

Dr. Pablo Brito Parada

June 2025

AI Acknowledgement Statement

Tool name: Chat GPT-o3

Provider: Open AI

URL: https://chatgpt.com/share/684b5d05-3ef4-8003-8f3b-38b74b866e8a

Brief: Used to recommend papers.

Statement: I used OpenAI's ChatGPT-o3 (https://chat.openai.com/) to recommend papers. This generative AI tool supported my learning process, but the submitted work is my own and it reflects my own understanding and effort as I declared by signing the Academic Integrity Declaration.

Table of Content

Abstract

As machine and deep learning develop, data has become a core digital asset and competitive advantage. However, in domains such as finance and rock-blasting, costly, privacy-constrained data acquisition leads to small-sample datasets that impede model deployment. We therefore develop a two-stage pipeline: statistical copulas learn global distributions, a Variational Autoencoder (VAE) -based module refines local structure, and multidimensional criteria evaluate the resulting synthetic data.

Our approach first applies several copula families to model marginal distributions and extreme-value dependence, generating an initial synthetic data. A VAE or VAE with a Bayesian Gaussian Mixture (VAE-BGM) then refines these samples, adding detailed structure while checking physical constraints. The synthetic data will be evaluated by: (i) statistical fidelity (Kolmogorov–Smirnov Distance, Total Variation Distance, coverage), (ii) downstream utility using Train-Synthetic-Test-Real (TSTR) and Train-Real-Test-Synthetic (TRTS) error gaps on Extreme Gradient Boosting (XGBoost), benchmarked against published results, and (iii) robustness against resampling and model choice.

We hypothesize that the pipeline can keep per-variable Kolmogorov–Smirnov (KS) distances below 0.15 (vs 0.30+ for a raw-copula baseline) and cut the TRTS root-mean-squared-error (RMSE) gap on XGBoost by 5–8 % when a 20–40 % synthetic blend is used. Metric variance across ten resamples is anticipated to stay within ±2 percentage points.

1.Problem Description

1.1 Background

Data scarcity is a persistent obstacle to machine and deep learning application in geoscience. Field logging is costly and slow, so available industrial datasets are small. One blast-vibration study had only 110 cases; after clustering, the authors still needed SMOGN oversampling to balance classes before XGBoost training [1]. Another blast-fragmentation project used 91 records and noted that a larger neural network "could not be trained reliably" without more data [2]. Similar size limits appear in open-pit-slope instability work based on artificial neural networks [3].

Small samples hinder model performance in three ways:

1. Tail dependencies cannot be learned, leading to poor risk estimation for extreme events.
2. Higher-order feature interactions remain under-represented, limiting the expressiveness of deep models.
3. Models easily overfit, yielding optimistic in-sample metrics but weak generalisation.

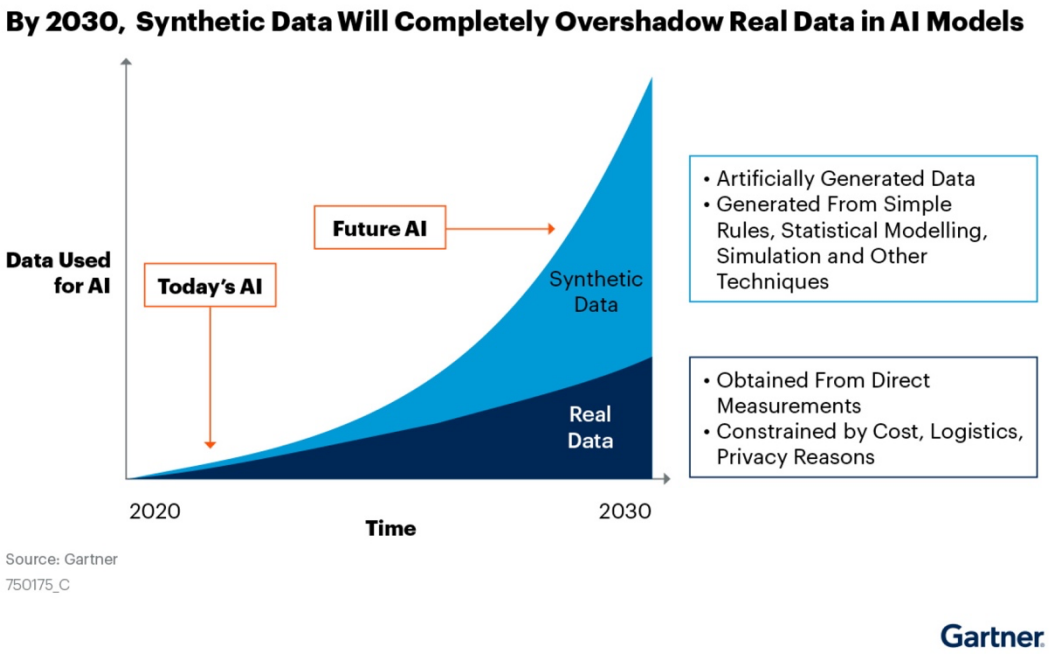
To cope with small samples, engineers use statistical resampling tools such as Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN), but these methods only re-weight existing points and cannot create truly new patterns [4]. Others down-scale the model itself—fewer network layers or a Principal Component Analysis (PCA) pre-filter—to fit the data [5]. These fixes improve stability yet leave the core problem unchanged: the dataset is still too small to capture tail events and high-order dependence. Generating realistic synthetic data therefore becomes a necessary next step.

1.2 Significance

Most geoscience datasets for blasting data, fragmentation and slope stability hold fewer than 200 rows. Small samples make current models miss extreme values and cross-variable relationships. For a 162-blast dataset, the best published model still reported Anderson-Darling statistic (AD) = 0.456 and a p-value of 0.02, confirming that tail behaviour is not captured well. The AD test weights extreme quantiles more heavily than the Kolmogorov–Smirnov distance, making it a regular gauge of tail fit [6]. Our Copula-VAE pipeline tackles this by reproducing the global distribution with copulas and adding fine detail with a VAE model. Our target is to push the Anderson–Darling statistic below 0.22 and, with the extra rows, allow deeper networks–Deep Neural Network, Transformers–to learn higher-order dependence and handle both normal and extreme states in a single mine [7].

Field data are also costly: one blast geophone or slope laser scanner is roughly $7 000 plus yearly upkeep, so operators deploy only a few—coverage is sparse, and cycles are long [8]. Reliable synthetic tables can potentially replace many extra sensors, reducing data-acquisition spending by up to 50 % and avoiding site downtime.

In short, a sound synthetic-data pipeline raises sample size without extra field work, unlocks more complex models, captures extreme events that matter for safety, and saves both money and time for industry users.



Source: Gartner, graphic ID 750175_C, via NVIDIA blog (accessed 10 Jun 2025).

1.3 Review of Existing Work

Different Methods of Mean Fragment Size Prediction in Rock Blasting

We conclude several works on mean-fragment-size (MFS) prediction which are all built on very small datasets. Across these studies, metrics focus on MSE/R^2. None evaluates Anderson–Darling, KS, or tail dependence, and none imposes physical limits on explosive charge or burden. Thus, a method that enlarges these small tables, preserves tail behaviour and respects blasting constraints is still missing—the gap our Copula-VAE pipeline aims to fill.

Author	Core Method & Result	Strength & Limit
Krop et al. 2024 [1]	- 110 blasts (97 train, 13 hold-out) and combined hierarchical clustering, SMOGN oversampling and XGBoost - MSE 0.0002 and $R^2 = 0.98$	- Combines clustering, hyper-opt and data-augmentation, top accuracy - Relies on SMOGN only, no tail or physics check
Kulatilake et al. 2010 [2]	- Trained a one-hidden-layer neural network on two groups of 35 and 56 blasts - $R = 0.841$, $R^2 = 0.708$, std-err = 0.092	- ANN beats empirical rule with tiny network - Dataset tiny; no augmentation; seven inputs only
Sui et al. 2025 [9]	- Built a stacking ensemble (RF + XGBoost) on 97 blasts - $R^2 = 0.943$, RMSE = 0.044	- Stacking lifts single-model performance; external validation on five blasts - Still <100 rows; ensemble ignores tail extremes
Sharma et al. 2017 [5]	- Applied PCA or stepwise screening followed by linear regression on 76 blasts data with another 19 for testing - $R^2 = 0.734$ and 5 % RMSE	- Variable screening keeps model simple - Linear form, lowest accuracy, no interaction terms

Engineering Tools: Copula, VAE, TabVAE-GMM and other models

Copula

Sklar's theorem [10] states that any multivariate cumulative-distribution function

$$F(\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{d}) = C(\mathrm{F}_1(\mathrm{x}_1),\ldots,\mathrm{F}_\mathrm{d}(\mathrm{x}_\mathrm{d}))$$

can be decomposed into a copula C that handles dependence, marginals keep geology-specific units. We can therefore preserve each variable's range while adding realistic joint tails—something simple oversampling cannot do.

Variational Autoencoder (VAE)

Kingma and Welling [11] introduced VAEs, which learn generative models by maximising evidence lower bound (ELBO) over latent variables. VAEs train stably on datasets with fewer than 200 rows—a typical small-sample scenario—and yield a smooth latent space for further refinement.

VAE + Bayesian Gaussian Mixture (VAE-GMM)

Apellaniz et al. [12] fit a Bayesian Gaussian Mixture on a trained VAE's latent codes. This extension captures mixed discrete-continuous shapes and reduces Kolmogorov–Smirnov error by > 20 % relative to CTGAN or TVAE, while leaving the original VAE prior and training loop unchanged.

TSTR – Train Synthetic Test Real

TSTR is a utility check for synthetic data [13]. We generate a synthetic training set, train a normal predictive model only on that synthetic set, and measure its performance on untouched real data. If the score—accuracy, RMSE, AUROC, etc.—is close to what we get when training on real data, the synthetic set has preserved the information the task needs.

Generative Adversarial Networks (GAN) & Conditional GAN (CGAN)

Recent GAN-based models for tabular synthesis include TGAN and CTGAN. TGAN demonstrates that a vanilla GAN can jointly model mixed continuous-and-categorical columns and beats Bayesian-network generators on several public datasets [14]. CTGAN refines this idea with a conditional sampler that balances rare categories, achieving better results on a

benchmark of 15 datasets [15]. These works highlight GANs' strength at preserving inter-column correlations, but they neither enforce physical constraints nor target tail fidelity—gaps our Copula + VAE hybrid explicitly addresses.

1.4 Objective

O1 — Statistical fidelity. Generate synthetic tables whose per-variable Anderson–Darling (AD) statistic is < 0.22 (baseline ≈ 0.45), and whose Kolmogorov–Smirnov (KS) distance is < 0.15.

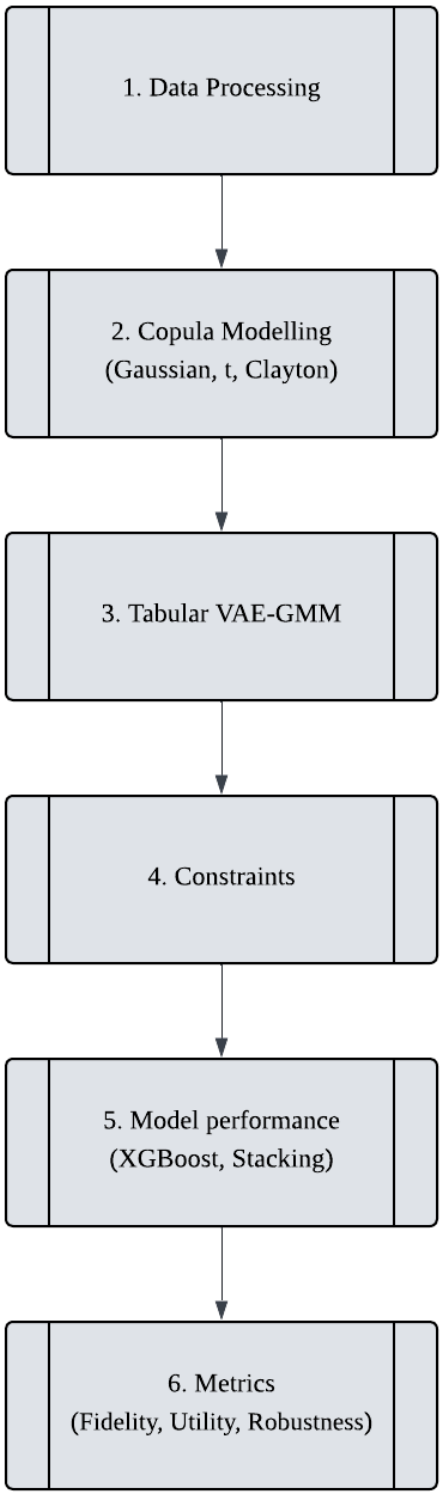
O2 — Downstream utility. Keep the TSTR RMSE gap on the XGBoost prediction task within 5-8 % when a 20-40 % synthetic blend is used.

O3 — Robustness. Ensure the variance of KS and RMSE across ten random resamples stays below ±2 percentage points.

O4 — Benefit-interval mapping. Systematically vary the real: synthetic training ratio from 9: 1 to 1: 9 and plot predictive accuracy. Identify any benefit interval where synthetic augmentation improves performance. Accept that the optimum may occur at 100 % real data if quality targets in O1–O4 are not met.

O5 — Deliverables. Provide open-source code, a reproducible Jupyter report, and one case-study dataset release.

2.Methodology



1. **Data Processing**: We merge blasting data from Dr. Lopes into one table with 586 rows. Because each file lists a different set of columns, we keep only the 11 features shared by all sites to make the data consistent. We remove rows lacking PPV or charge weight and discard obvious errors such as negative burden values. All length fields are converted to metres and charge weights to kilograms. The cleaned table is then split 70 % / 30 % into 410 train rows and 176 test rows, stratified by explosive type.

2. **Copula Modelling:** We fit three copula families (Gaussian, t, and Clayton) on the clean training set by using maximum-likelihood. Then, select the model with the lowest AIC. We sample about 5 000 joint rows from the model and map each uniform margin back to the original units. This generates a "rough" synthetic data set that already captures one - and two-dimensional dependence but lacks higher-order structure.

3. **Tabular VAE-GMM**: We run two refinement models in sequence and keep the one that meets our fidelity target (KS < 0.15).

Option A – plain TabVAE

The rough copula sample data goes into a VAE with two fully-connected layers (64 and 32 units, ReLU), a 16-dim latent space, and $\beta = 0.5$ in the ELBO. After 300 epochs the decoder outputs new rows directly.

Option B – VAE-GMM (preferred)

If Option A does not achieve the KS target, we keep the same VAE but, once training is done, fit a 10-component Bayesian Gaussian Mixture to its latent codes. We then draw fresh latent points from the mixture and decode them into the final table. This extra step gives the model a richer, non-Gaussian prior and usually sharpens rare-value regions.

We compare the two options on KS, AD and the TSTR gap; whichever scores better is carried forward to the constraint-checking stage.

4. **Constraints**: Every synthetic data is passed through a rule-based filter: vibration and charge weights are clipped at physically plausible maxima; burden and spacing obey burden $\leq$ spacing; integer features (e.g., number of cartridges) are rounded. Rows that violate any hard constraint are discarded. If necessary, resample until the target synthetic size is reached. If more than 50 % of rows are discarded, the VAE model is reviewed and retrained.

5. **Model Performance**: By checking the down-stream model performance, we can judge the quality of synthetic data.

   Baseline: train XGBoost or a stacking ensemble on the real train data set only and compare the result of previous papers.

   Mixed: train the model with same structure on mixed data set with different mixture ratio (1:9, 2:8, 3:7, …, 9:1). Performance is always measured on the untouched real test set. This design shows whether synthetic data helps predictive accuracy.

6. **Metrics**: We focus on three aspects – fidelity, utility and robustness.

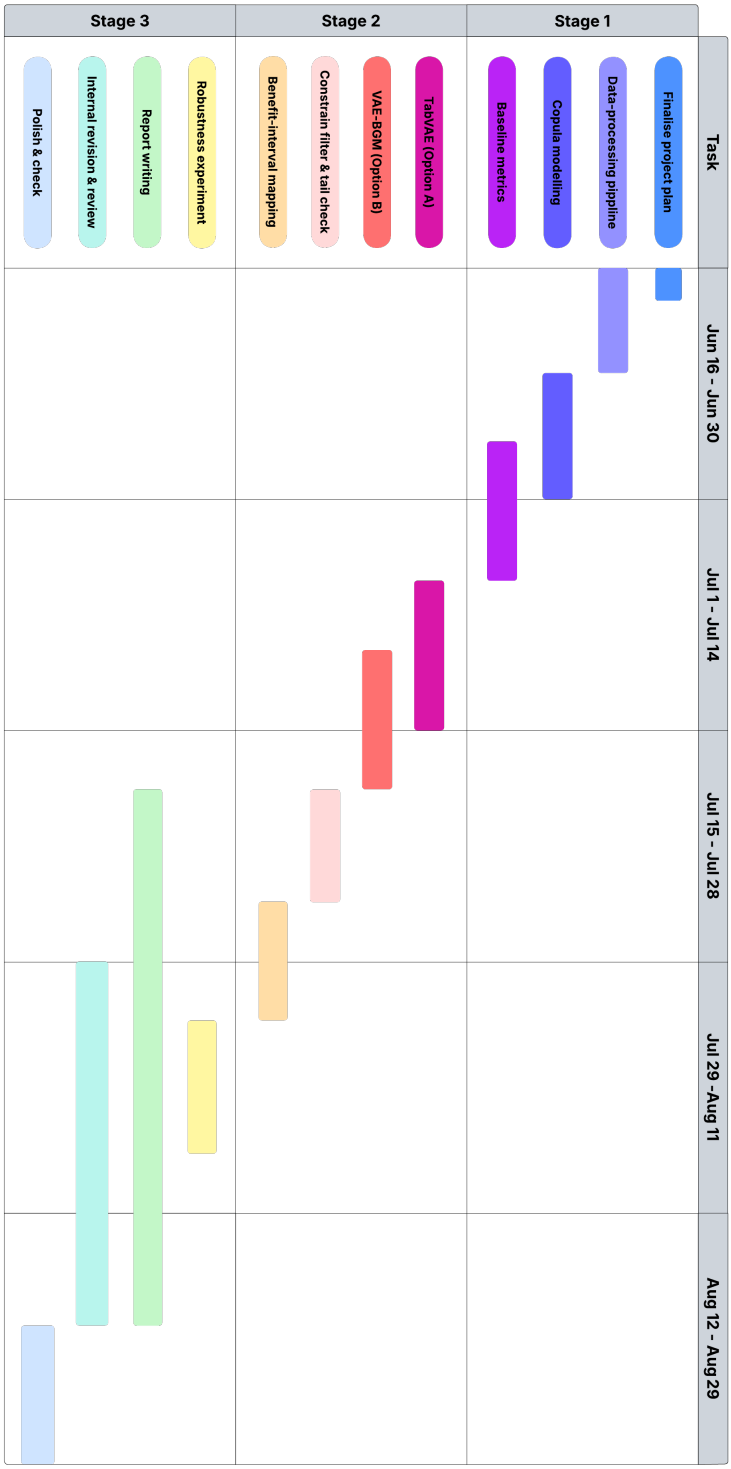
   Fidelity: per-column KS and Anderson–Darling, target AD < 0.22.

   Utility: We implement TSTR / TRTS strategy and check RMSE gaps. Hence, we search for the "benefit interval," the mix ratio giving the lowest gap.

   Robustness: standard deviation of KS and RMSE over 10 random seeds; goal $\pm$ 2 percentage points or less.

   Together these metrics verify that the synthetic data look right, work for real-world tasks, and behave consistently across runs.

3.Project Timeline



References

[1] I. Krop, T. Sasaoka, H. Shimada, and A. Hamanaka, “Optimizing mean fragment size prediction in rock blasting: A synergistic approach combining clustering, hyperparameter tuning, and data augmentation,” Eng, vol. 5, no. 3, pp. 1905–1936, 2024, doi:10.3390/eng5030102.

[2] P. H. S. W. Kulatilake, Q. Wu, T. Hudaverdi, and C. Kuzu, “Mean particle size prediction in rock blast fragmentation using neural networks,” *Engineering Geology*, vol. 114, no. 3–4, pp. 298–311, 2010, doi: 10.1016/j.enggeo.2010.05.008.

[3] M. Z. Naghadehi, R. Jimenez, R. KhaloKakaie, and S.-M. E. Jalali, “A new open-pit mine slope instability index defined using the improved rock engineering systems approach,” *International Journal of Rock Mechanics and Mining Sciences*, vol. 61, pp. 1–14, 2013, doi: 10.1016/j.ijrmms.2013.01.012.

[4] P. Branco, L. Torgo, and R. P. Ribeiro, “SMOGN: a pre-processing approach for imbalanced regression,” in *Proc. 1st Int. Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA), Skopje*, Macedonia, Sep. 2017, vol. 74 of *Proc. Machine Learning Research*, pp. 36–50

[5] S. K. Sharma and P. Rai, “Establishment of blasting design parameters influencing mean fragment size using state-of-art statistical tools and techniques,” *Measurement*, vol. 96, pp. 34–51, 2017, doi: 10.1016/j.measurement.2016.10.047.

[6] S. Hosseini, J. Khatti, B. O. Taiwo, Y. Fissha, K. S. Grover, H. Ikeda, M. Pushkarna, M. Berhanu, and M. Ali, “Assessment of the ground vibration during blasting in mining projects using different computational approaches,” *Scientific Reports*, vol. 13, Art. no. 18582, Oct. 2023, doi: 10.1038/s41598-023-46064-5.

[7] D. M. Hawkins, “The distribution of the Anderson–Darling statistic,” *Communications in Statistics—Simulation and Computation*, vol. 53, no. 12, pp. 6022–6026, Aug. 2023, doi: 10.1080/03610918.2023.2245174.

[8] BlastersTool, “Seismographs | Blasting Supplies,” *BlastersTool.com*. Accessed: Jun. 10, 2025. [Online]. Available: https://blasterstool.com/blasting-supplies/seismographs.html

[9] Y. Sui, Z. Zhou, R. Zhao, Z. Yang, and Y. Zou, “Open-Pit Bench Blasting Fragmentation Prediction Based on Stacking Integrated Strategy,” *Appl. Sci.*, vol. 15, no. 3, Art. no. 1254, Feb. 2025, doi: 10.3390/app15031254.

[10] R. B. Nelsen, *An Introduction to Copulas*, 1st ed. New York, NY, USA: Springer, 1999.

[11] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114*, Dec. 2022.

[12] P. A. Apellaniz, J. Parras, and S. Zazo, "An improved tabular data generator with VAE-GMM integration," in *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*, 2024, doi: 10.23919/EUSIPCO.2024.9

[13] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional GANs," *arXiv:1706.02633*, Jun. 8, 2017.

[14] L. Xu and K. Veeramachaneni, "Synthesizing Tabular Data using Generative Adversarial Networks," *arXiv* preprint arXiv:1811.11264, Nov. 2018.

[15] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular Data using Conditional GAN," *arXiv* preprint arXiv:1907.00503, Jul. 2019.