Generating Financial Losses From Extreme Events via GANs

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Group Member: Zhengqi Zhu, Ziyun Zhou

European Investment Bank — The EU bank

Abstract

We used Generative Adversarial Networks (GANs) to model financial impacts of extreme events, improving on traditional methods. By processing EM-DAT and WEO data, and testing different GAN architectures, we generated realistic synthetic financial loss data, enhancing risk assessment and management.

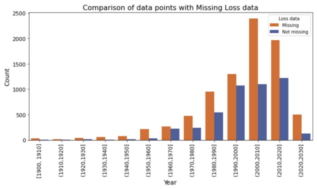
Project Introduction

Data:

Sources: Financial loss data from EM-DAT (1900-2022) and GDP data from WEO and the World Bank (1960-2021).
Content: EM-DAT includes disaster type, location, and impact; GDP data provides economic context for normalizing losses.

Column Name	Description
ISO	Country Code to identify each country
Year	The year in which the disaster occurred. This data spans from 1900 to 2022.
Start Month	The month when the disaster event started. This column is numeric, representing months from 1 (January) to 12 (December).
Start Day	The day when the disaster event started. This column is numeric, representing days from 1 to 31.
Losses	Financial losses incurred due to the disaster, expressed in US dollars. This column may contain values that reflect the economic impact of the event.

Preprocessing: Extensive data cleaning, interpolation for missing values, and normalization were conducted to ensure accuracy and consistency.
Purpose: Prepared data for effective GAN model training, enabling reliable synthetic data generation and meaningful cross-country comparisons.

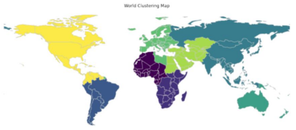


Embedding

Purpose: Reduce dimensionality of high-dimensional financial loss data to improve clustering and GAN performance.
Techniques: Applied methods like PCA, Robust PCA, Kernel PCA, and Autoencoders.
PCA/Robust PCA: Captured linear relationships and reduced noise, with Robust PCA effectively handling outliers.
Kernel PCA: Captured complex, nonlinear relationships, enhancing clustering quality for datasets with intricate patterns.
Autoencoders: Used neural networks to learn efficient data representations, particularly effective for nonlinear data and feature extraction.
Outcome: Improved clustering and GAN training by emphasizing key data features while reducing noise and dimensionality.

Clustering

Techniques: Employed K-Means++, K-Medoids, DBSCAN, and sliding window clustering.
Data Integration: Used geographical, economic, and time-series data for more accurate clustering.
Dynamic Time Warping (DTW): Managed time-series data to better capture temporal patterns.
Sliding Window Clustering: Updated clusters over time, focusing on recent data to improve relevance.
Hierarchical Clustering: Stabilized initial centroid selection in K-Means, leading to more consistent clusters.



Outcome: Enabled GAN models to better capture regional and temporal variations, improving synthetic data accuracy.

Imputation

Purpose: Addressed missing or incomplete financial loss data to ensure robust datasets for GAN training.
Brownian Bridge: Smoothly interpolated missing time-series data but struggled with extreme values.

The Brownian Bridge process is defined as follows: given a Brownian motion process $B(t)$ with $B(0) = a$ and $B(T) = b$, the Brownian Bridge process $B_{a,b}(t)$ can be expressed as:

$$B_{a,b}(t) = B(t) - \frac{t}{T}B(T) + \frac{T-t}{T}B(0)$$

Here, T represents the total time interval, and t is a specific point within that interval. This ensures that the path starts at a and ends at b, maintaining the statistical properties of a Brownian motion path.

FlowGAN: Effectively handled heavy-tailed distributions by generating synthetic data for lower and higher distribution segments.
Parametric Methods: Focused on accurately capturing the tail properties of the data, enhancing the realism of generated synthetic data.

$$\text{Loss} = w_1 \times \text{ClassicLoss} + w_2 \times |\text{IQR}(x_{\text{fake}}) - \text{IQR}(x_{\text{real}})| + w_3 \times |\text{tail}(x_{\text{fake}}) - \text{tail}(x_{\text{real}})| + w_4 \times |\min(x_{\text{fake}}) - \min(x_{\text{real}})|$$

(IQR stands for interquartile range)

Outcome: Produced robust and reliable datasets, leading to more accurate and effective GAN model training.

Conclusion

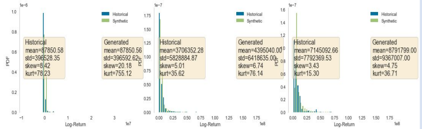
Result of Loss Simulation by Sliding Window Clustering and FlowGAN Imputation

Metric	TTGAN (150 epochs)	TAGAN (150 epochs)
$W_1^{(1)}$	9.3043e+04	7.1292e+04
$W_1^{(5)}$	3.1255e+05	2.2909e+05
$W_1^{(20)}$	9.2908e+05	5.4873e+05
$W_1^{(50)}$	1.6011e+06	1.0590e+06
$W_1^{(100)}$	2.6599e+06	1.8404e+06
$W_1^{(200)}$	4.7359e+06	3.5527e+06
$Skewness$	7.6932e+00	3.9525e+00
$Kurtosis$	3.8710e+02	1.5191e+02
VaR	1.4848e+05	7.7859e+04
ES	2.1368e+05	1.4501e+05
ACF	1.1220e+00	6.3157e-01
ACF^{abs}	1.1469e+00	4.7522e-01
ACF^{sq}	1.3260e+00	8.9719e-01
Lev	1.2023e+00	7.7017e-01

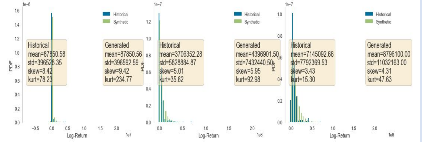
Effective Data Generation: The study successfully generated realistic synthetic financial loss data from extreme events using advanced GAN architectures.
Improved Risk Modeling: The use of GANs, combined with sophisticated data preprocessing, significantly enhanced the accuracy of financial risk assessments.
Tailored GAN Models: Different GAN models, such as FlowGAN, TAGAN, and TTGAN, were optimized to handle the unique challenges posed by extreme event data.

GANs

Purpose: Generate realistic synthetic financial loss data for extreme events, improving risk assessment.
DCGAN (Deep Convolutional GAN): Enhanced extremeness of generated data, aligning with real-world financial loss distributions.
QuantGAN: Targeted financial time series data but struggled with capturing heavy tails accurately.
TAGAN (Temporal Attention GAN): Excelled in capturing temporal dependencies for individual loss data but faced challenges with cumulative losses.



FlowGAN aids imputation by splitting data based on quantiles, followed by targeted methods. Combining FlowGAN with TAGAN generates improved outputs, but Wasserstein distance improvement remains limited.
TTGAN (Temporal Transformer GAN): Used transformer architecture to better model long-term dependencies in time series data, improving sequence accuracy.



TTGAN shows improved results after 200 epochs of training, but further epochs lead to overfitting. he Wasserstein distance stabilizes at around 20,000, yet narrow range issues persist.

EIB Advisors: Giuseppe Bonavolontà, Oleg Reichmann
Research Advisors: Dr. Ali Hirsa | **Teaching Assistant:** Miao Wang