Industrial Engineering and Operations Research
COLUMBIA | ENGINEERING
THE CENTER FOR ARTIFICIAL INTELLIGENCE IN BUSINESS ANALYTICS & FINANCIAL TECHNOLOGY

# Generating Financial Losses From Extreme Events via GANs

Researchers: Wenhan Yang, Yueru He, Noah Dawang
EIB Advisors: Giuseppe Bonavolontà and Oleg Reichmann
Research Advisors: Dr. Ali Hirsa and Miao Wang

European Investment Bank

## Abstract

Analysis of financial losses due to extreme events is an important component of risk management. Due to scarce data, non-stationarity and unique distributions of losses data, prediction is difficult.

As a continuation of the exploration of generating high quality synthetic data using GAN models, we tried three types of GANs and various clustering techniques to increase data size for GAN training.

## Project Introduction

**Previous Teams' Work:** There are 2 teams before our research. The first team built a pipeline for distribution shifting data with DCGAN and extreme loss generation with conditional DCGAN; The second team implemented the first team's pipeline and validated the distribution shifting procedure and verified that repeated applications lead to the generalized Pareto distribution shape.

**Data:** Financial Losses of Natural Disasters (EM-DAT) between 1900 and 2022

**New GAN models:** Temporal Transformer GAN (TTGAN), Temporal Attention GAN (TAGAN) treats data input as time series, and use series of convolutional and attention layer to capture local patterns; FlowGAN works well on imputing losses data when combined with parametric imputation methods.

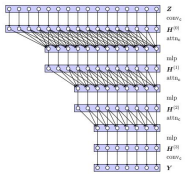
**Clustering Efforts**: Because all three GANs are data-hungry, we would like to increase the sample size by first clustering countries and regions with similar characteristics together. We tried several clustering methods, including K-previous Means, K-Medoids, Multidimensional Hierarchical Clustering, Sliding Windows Clustering, and DBSCAN.

## GAN Performances

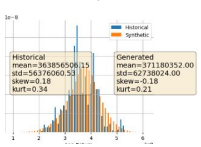
### TAGAN

The TAGAN incorporates a generator architecture composed of convolutional layers augmented by an attention layer strategically inserted to enhance local feature capture.

Generator Architecture



500 Epoch Test Results With FlowGAN Imputation

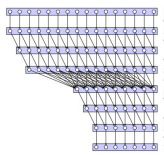



TAGAN's generator employs convolutional layers and an attention layer for effective feature capture. FlowGAN aids imputation by splitting data based on quantiles, followed by targeted methods. Combining FlowGAN with TAGAN generates improved outputs, but Wasserstein distance improvement remains limited.

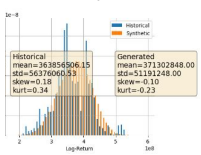
### TTGAN

TTGAN utilizes of transformers within both the generator and discriminator. The transformer architecture includes convolutional layers, two-layer MLPs, and attention layers, making TTGAN a suitable choice for analyzing and synthesizing time series data.

Generator Architecture



200 Epoch Test Results With FlowGAN Imputation




TTGAN shows improved results after 200 epochs of training, but further epochs lead to overfitting. The Wasserstein distance stabilizes at around 20,000, yet narrow range issues persist. Loss plots indicate TTGAN's inability to capture fat tails in the data.

## Country Clustering Efforts

### Distance Metrics

We employed Euclidean and Dynamic Time Warping (DTW) distance metrics. Euclidean distance measures multi-dimensional point distances, while DTW handles sequential data, accommodating variations in series lengths. Besides, K-Medoids, a clustering method, customizes distance metrics to identify points closest to cluster centers, making clusters more robust.

### K-previous Means

Hierarchical clustering was employed to mitigate K-means randomness, while engineered latitude, longitude, and loss features were tested. K-previous means outperformed region-based clustering from last team. Attempts with DTW distance and FlowGAN-imputed data showed persistent challenges, including several clusters only have one or small number of countries in each. Efforts to balance cluster dominance were inconclusive, revealing intricate data complexities.

### K-Medoids

K-Medoids clustering only requires a distance metrics so a more general metric spaces can be incorporated rather than just euclidean space. We use "great circle distance" to approximate distance on earth's sphere.

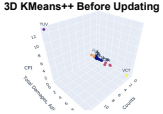
Great Circle Distance with Other Distance Metrics

Great Circle Distance Only

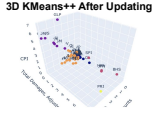
### Multidimensional Sliding Windows

We employed KMeans++ for better initialization for distant data point selection. "Sliding window clustering" improved cluster updates, focusing on recent extremes. Multidimensional feature engineering included interactions for both numerical and categorical variables, while reclustering prevented single-cluster dominance.

3D KMeans++ Before Updating



3D KMeans++ After Updating



### DBSCAN

We tested DBSCAN in parallel with K-previous means to analyze clustering patterns. DBSCAN can identify and exclude outliers. We applied DBSCAN using FlowGAN-imputed loss data, with Euclidean distance yielding 3 clusters, and DTW distance yielding 2 clusters, suggesting the persist dominating country issue in clustering isn't outlier-induced.

## Severity Modelling

### FlowGAN

FlowGAN is a GAN which restricts itself to invertible activation functions, allowing an easy formula for maximum likelihood estimation on top of the usual GAN loss. We used a FlowGAN to train the lower quantile of disasters

### Shallow GAN

A shallow GAN is the simplest form of the GAN, consisting of two feedforward neural networks as the generator and discriminator, with no fancy components. We used a shallow GAN with a special "tail loss" to train the upper quantile of disasters

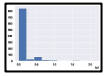
### Combining the two GANs

We used a stratified sampler to generate synthetic data using both the FlowGAN and the shallow GAN

## Conclusion

### Results

The Stratified Sampler can generate both realistic and extreme natural disaster losses on a cluster and a country level.



### Challenges

More works needed on high dimensional clustering

## Future Work

- Further tuning of TAGAN and TTGAN parameters
- Focusing on multidimensional clustering instead of single-dimension, as the multidimensional clustering suggesting a relief of dominating country issue
- Incorporating subcategorization of regions with the usage of external datasets and features