Université Paris-Panthéon-Assas

Generating Synthetic Health Insurance Data Using Copulas

Nom de l'auteur:
Pidburachynskyi Arsen

Tuteur Universitaire:
Skalli Ali

Date de Soutenance:
11 september 2024

Generating Synthetic Health Insurance Data Using Copulas

Raison Sociale et Adresse de l'Organisme d'Accueil:
AXA FRANCE IARD
RCS NANTERRE 722 057 460
313 TERRASSES DE L'ARCHE
92727 NANTERRE CEDEX
FRANCE

Nom de la Maitresse d'Apprentissage:
Ho Mei Hung

Nom de l'Auteur:
Pidburachynskyi Arsen

Résumé

Mots-clés: Données tabulaires, Données synthétiques, Copules gaussiennes, Assurance santé, Modélisation de dépendances

Dans le domaine de l'assurance santé internationale, la manipulation de grandes quantités de données est essentielle pour la modélisation des risques et la tarification. AXA Life and Health International Solutions gère des données provenant des nombreuses sources internationales, ce qui introduit des défis en matière de cohérence, de qualité et de conformité réglementaire. De plus, ces données sensibles nécessitent une strictes de confidentialité.

Nous allons nous concentrer sur la génération de données synthétiques, une approche qui permet de créer des jeux de données réalistes sans exposer la confidentialité des informations. En utilisant des méthodes basées sur les copules, notamment copule gaussien, nous avons pu modéliser les dépendances complexes entre les variables tout en reproduisant les caractéristiques des distributions observées dans les données réelles.

Les données synthétiques générées ont été évaluées à l'aide de comparaisons des métriques univariées et multivariées par rapport aux données réelles, avec une baseline basée sur des modèles de mélanges gaussiens (GMM). Les résultats indiquent que l'utilisation de copules, notamment les copules gaussiennes, permet de générer des données synthétiques qui reproduisent raisonnablement bien les structures de dépendance et certaines propriétés des données originales. Cependant, des écarts ont été observés, soulignant la nécessité d'explorer davantage ces méthodes pour améliorer leur précision et leur capacité à capturer toutes les nuances des données réelles.

Contents

1 Introduction

This master's thesis is based on the work conducted during my apprenticeship at AXA France, a leading global insurance company. AXA France deals with vast amounts of tabular data, which are crucial for various aspects of the insurance business, including pricing, risk assessment, and customer behavior prediction. However, handling such data presents significant challenges, especially concerning data privacy, missing values, and the presence of anomalies. Insurance data is highly sensitive, containing personal information that must be protected under strict privacy regulations, making it difficult to work with the data directly. Missing data is common in insurance datasets due to factors such as non-disclosure, errors during data collection, or inconsistencies across different data sources. On top of that, insurance data often contains anomalies or outliers, such as unusually high claims.

Generating synthetic tabular data is a potential solution to these challenges. By creating synthetic datasets that mimic the statistical properties of real data, AXA can overcome privacy issues, handle missing values more effectively, and model anomalies without exposing sensitive information. This approach offers several benefits, including improved data privacy by enabling model development without compromising sensitive information, enhanced model development by filling gaps caused by missing values and allowing for the modeling of rare events or anomalies, and scalability by creating large datasets to train complex models. The primary objective of this thesis is to explore and implement methods for generating synthetic tabular data within an insurance context.

To achieve this, the thesis will be structured as follows: It will start with a overview of the context in which AXA operates, the challenges faced with real-world data, and the specific objectives of this study. This will include a discussion of the importance of synthetic data generation in addressing these problems. The next section will look into the mathematical and statistical foundations of synthetic data generation, particularly focusing on high-dimensional data and large samples, covering theoretical aspects of data generation techniques that can handle the complexity and scale of health insurance data. Then we compare the results of synthetic datasets generated using gaussian copula with those generated using Gaussian Mixture Models (GMM). The analysis will highlight the advantages and limitations of the copula approach, exploring its effectiveness and potential drawbacks. Additionally, the discussion will include other possible approaches to synthetic data generation

2 Context

In the insurance industry, data is essential for making decisions related to pricing, risk assessment, and customer management. As the amount and complexity of data grow, insurance companies increasingly use analytical techniques to gain insights, support decision-making, and remain competitive. This thesis is based on work conducted at AXA Life and Health International Solutions, a division of AXA France responsible for managing the company's international health insurance business. At AXA Life and Health International Solutions, we work with large amounts of tabular data, including customer demographics, claim histories, policy details, and other relevant information, which are crucial for accurate underwriting and pricing of insurance products.

Due to the international nature of AXA Life and Health International Solutions, the data comes from various international sources, each with different regulations, standards, and levels of data quality. This diversity poses problems in maintaining data consistency and compliance. Therefore, it is important to use reliable data management and analysis methods to handle and integrate data from multiple countries. By applying data generation and analysis techniques, the company could improve its ability to provide effective insurance solutions while adhering to multiple regulatory requirements and maintaining acceptable quality of data.

2.1 Objectives

The primary objective of this thesis is to explore and implement methods for generating synthetic tabular data that can reflect the statistical properties and dependencies observed in real-world insurance census datasets. By doing so, the results of this work could address following challenges :

- **Data Privacy**: Insurance data contains sensitive personal information that must be protected under strict privacy regulations. Synthetic data generation offers a solution by creating realistic without actually containing any confidential information, datasets that can be used for analysis and modeling without compromising privacy.
- **Handling Missing Data and Anomalies**: Insurance datasets often have missing values or anomalies, such as extreme claims or unusual customer behavior, which can distort analysis and modeling. Synthetic data can help fill in gaps and provide robust models for outlier scenarios.
- **Scalability and Flexibility**: Synthetic data generation allows for the creation of large-scale datasets needed to train complex models. This scalability is important for developing predictive models that can handle a wide range of scenarios and conditions.

2.2 Potential Applications

The development of synthetic data generation techniques has several potential applications in the insurance industry, particularly in areas such as:

- **Pricing and Underwriting**: Accurate pricing models are essential for the profitability of insurance products. Synthetic data can help refine these models by providing additional data points that capture a bigger range of risk factors and customer behaviors.
- **Risk Management**: By using synthetic data to simulate various risk scenarios, insurers can better understand potential exposures and develop strategies to mitigate them. This is especially valuable in assessing the impact of rare but catastrophic events, such as high claims in health insurance industry.
- **Regulatory Compliance and Reporting**: Insurance companies are often required by regulators to prove that their pricing models are robust and fair. Using synthetic data offers a transparent and reproducible method for testing these models, helping ensure they comply with regulatory standards.

3 Synthetic data generation

3.1 Notations and Mathematical background

In this work, we will use the following notations:

- $n \in \mathbb{N}$: the number of data points or observations. This represents the total size of the dataset we are working with. $d \in \mathbb{N}$ is the dimension of the data points, or the number of features. This indicates the number of variables or attributes each data point has. Each data point will be denotes as a tuple in the form $(x_1, \ldots, x_d) \in \mathbb{R}^d$: The individual features or attributes of a data point. Each x_i represents a specific feature in the d-dimensional space. When considering the index j of each d-tuple in the dataset, we will denote $y_j = (x_1^j, \ldots, x_d^j) \in \mathbb{R}^d$ and the corresponding dataset $\mathscr{D} = \{y_j\}_{j=1}^n$.
- $\theta \in \Theta$ corresponds to the set of parameters for the parametric distribution model, belonging to the set Θ of plausible parameters. These parameters define the shape and other characteristics of the distribution.
- $p_\theta(x_1, \ldots, x_d) : \mathbb{R}^d \to \mathbb{R}_+$: The estimated probability density function (PDF), which depends on the parameters θ. This function represents the likelihood of observing a data point $(x_1, \ldots, x_d)$ given the parameters θ. The empirical cumulative density function (CDF) associated to each of the variables x_i, $i \in \{1, \ldots, d\}$, when using n

samples will be denoted as F_i^n, and their estimations F_{θ_i}. We will also denote the data log-likelihood as $\mathscr{L}(\theta; \mathscr{D})$ which corresponds to the logarithm of the likelihood of the observations in the dataset when using parameter θ.

The primary goal of this work is to generate synthetic data that accurately reflects the characteristics of the original data. To achieve this, we focus on fitting a parametric distribution to the data. The parametric distribution is represented as:

$$p_\theta(x_1, \ldots, x_d),$$

where θ represents the set of parameters that define the distribution. The choice of a parametric distribution allows for efficient sampling, making it easier to generate synthetic data that maintains the statistical properties of the original dataset.

In parametric modeling, we assume that the data can be described by a certain distribution family, characterized by a finite set of parameters θ. The objective is to estimate these parameters in such a way that the distribution p_θ closely matches the observed data. More formally, we are looking for θ that maximizes the data log-likelihood:

$$\mathscr{L}(\theta; \mathscr{D}) \triangleq \sum_{(x_1,\ldots,x_d)\in\mathscr{D}} \log p_\theta(x_1, \ldots, x_d).$$

Some common parametric distributions used in modeling include:

- **Normal (Gaussian) Distribution**: Suitable for modeling data with a symmetric distribution around a mean value. It is defined by its mean μ and variance σ^2.
- **Beta Distribution**: Used for modeling data that is bounded within a certain range, typically [0, 1]. It is characterized by two shape parameters, α and β.
- **Gamma Distribution**: Appropriate for modeling skewed distributions, often used in scenarios where data is strictly positive. It is defined by a shape parameter α and a rate parameter β.
- **Poisson Distribution**: Commonly used for modeling count data, representing the number of events occurring within a fixed interval of time or space.
- **Multivariate Gaussian Distribution**: Used for modeling multi-dimensional data, where each dimension follows a Gaussian distribution and there is a correlation structure between the dimensions.

The problem of synthetic data generation can be stated as follows. Given a real dataset $\mathscr{D}$ consisting of n data points in a d-dimensional space, our objective is to estimate a parametric

distribution p_θ that captures both the marginal distributions of each feature and the dependence structure between them. Once this distribution is estimated, we can generate new data points $(x_1, \ldots, x_d)$ by sampling from p_θ. The generated synthetic data should resemble the original data in terms of statistical properties, including the distribution of individual features and the relationships between features.

By achieving this, we aim to create synthetic datasets that can be used for various applications, such as data augmentation, privacy-preserving data sharing, and testing analytical models, while maintaining the integrity of the underlying statistical relationships observed in the real data.

In Section 3.2 we present a method for performing this data generation with Gaussian Mixture Models. In the real datasets considered in this thesis, the marginal distributions F_i can exhibit a wide range of shapes. For example, some may have heavy tails, and may deviate significantly from a Gaussian distribution. While modeling these individual marginals might be feasible, capturing the dependencies between them presents a more complex challenge, especially in high dimension. This motivates the use of copulas in Section 3.3, which will be discussed in the section as a tool to model those multivariate dependencies.

3.2 Gaussian Mixture Models (GMM)

Gaussian Mixture Models (GMM) [15] are a type of probabilistic model that represent a population as a mixture of multiple Gaussian distributions, each defined by its mean and variance. Formally, a GMM is defined as:

Definition (Gaussian Mixture Model)**.** *A Gaussian Mixture Model for a set of n data points $X = \{x_1, x_2, \ldots, x_n\}$ is a weighted sum of K Gaussian components, expressed as:*

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

where π_k are the mixing weights (summing to 1), $\mathcal{N}(x \mid \mu_k, \Sigma_k)$ is the k-th Gaussian component with mean μ_k and covariance Σ_k.

GMMs can be useful for modeling data that has multiple subpopulations, each represented by a Gaussian distribution. The flexibility of GMMs allows them to model complex data distributions using this Gaussian prior.

To determine the optimal number of components K in the GMM, we use model selection criteria such as the Akaike Information Criterion (AIC) and the Negative Log-Likelihood (NLL), as described in Appendix D. These criteria help balance model complexity and fit, ensuring that the GMM accurately captures the data distribution without overfiting.

Nevertheless, the GMMs are based on the prior that marginal distribution of each variable is Gaussian, which may not be verified in real-world data, which may be heavy-tailed for instance. This motivates the use of Copula, as described in 3.3.

3.3 Approach with copulas

We present hereafter an introduction to Copulas inspired by [11], as a way to model the dependence between the variables in the dataset. This will include a presentation of the algorithm used for fitting and sampling from those Copulas, thereby generating the wanted data.

Definition. *A d-dimensional copula,* $C : [0,1]^d :\to [0,1]$ *is a cumulative distribution function (CDF) with uniform marginals.*

The practical usefulness of copulas is based on the following important property:

Theorem 3.1. *(Sklar's Theorem [13], 1959) Consider a d-dimensional CDF,* F*, with marginals* $F_1, \ldots, F_d$*. Then there exists a copula,* C*, such that*

$$F\left(x_1, \ldots, x_d\right) = C\left(F_1\left(x_1\right), \ldots, F_d\left(x_d\right)\right), \tag{1}$$

for all $x_i \in [-\infty, \infty]$ *and* $i = 1, \ldots, d$*.*

This means that while marginal distributions can often be modeled separately and relatively easily, copulas provide a way to model the dependence structure between these variables.

When the Marginals Are Continuous: If the marginal distributions $F_1, \ldots, F_d$ are continuous, it can be shown that:

$$F_i(F_i^{-1}(y)) = y.$$

By evaluating at $x_i = F_i^{-1}(u_i)$ and using this transformation, we obtain the characterization:

$$C(u_1, \ldots, u_d) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)).$$

To generate synthetic data for d variables $Z_1, \ldots, Z_d$, we need to follow two main steps for fitting the joint distribution using Copulas:

1. **Identify Suitable Marginal Distributions:** We need to find appropriate marginal distributions $F_1, \ldots, F_d$ for each variable. A straightforward method is to approximate these marginals using empirical distribution functions based on the observed data.

2. **Model the Joint Distribution:** Capturing the relationships between the variables involves modeling the joint distribution function $F(z_1, \ldots, z_d)$. According to Sklar's

Theorem, this joint distribution can be expressed as:

$$F(z_1, \ldots, z_d) = C(F_1(z_1), \ldots, F_d(z_d)),$$

where C is the copula function representing the dependence between the variables.

Copulas can be seen as distribution functions in their own right. Specifically, when all variables are continuous, the copula C describes the joint distribution of the transformed variables $U_1 = F_1(Z_1), \ldots, U_d = F_d(Z_d)$. This property significantly simplifies the tasks of estimation and simulation within the model.

Steps for Estimation and Simulation:

To estimate the copula function C, the following steps are taken:

1. **Estimate the Marginal Distributions:** Estimate the marginal distributions $\hat{F}_1, \ldots, \hat{F}_d$ for each variable using the data.
2. **Create Pseudo-Observations:** Use the estimated marginals to create pseudo-observations $\hat{U}_1 = \hat{F}_1(Z_1), \ldots, \hat{U}_d = \hat{F}_d(Z_d)$.
3. **Estimate the Copula:** From these pseudo-observations, estimate the copula function C.

Once the copula $\hat{C}$ and the marginal distributions $\hat{F}_1, \ldots, \hat{F}_d$ are estimated, synthetic data can be generated by:

1. **Simulating Random Variables:** Generate random variables $U_1, \ldots, U_d$ from the estimated copula $\hat{C}$.
2. **Transforming Back to the Original Scale:** Convert these to the original data space using $Z_1 = \hat{F}_1^{-1}(U_1), \ldots, Z_d = \hat{F}_d^{-1}(U_d)$.

These steps allow for the creation of synthetic data that maintains both the marginal distributions and the dependence structure of the original data.

3.3.1 Gaussian Copula

The Gaussian copula is a widely used and straightforward type of copula model. It is popular because of its simplicity, ease of interpretation, and straightforward implementation. The Gaussian copula is based on the multivariate normal distribution and uses the probability integral transformation to model dependencies between random variables, given a specific covariance matrix, denoted as Σ. The formal definition of the Gaussian copula is given as follows.

Definition (Gaussian Copula)**.** *The Gaussian copula* $C^{\Sigma}_{Gauss}(u)$ *for a given covariance matrix* Σ *is expressed for* $u_1, \ldots, u_d \in (0,1)^d$ *as:*

$$C^{\Sigma}_{Gauss}(u) = \Phi_{\Sigma}\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right) \tag{2}$$

where $\Phi : s \in \mathbb{R} \mapsto \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s} \exp(-\frac{t^2}{2})\mathrm{d}t$ *represents the cumulative distribution function (CDF) of the standard normal distribution* $\mathcal{N}(0,1)$ *and* Φ_{Σ} *is the joint CDF of a multivariate distribution* $\mathcal{N}(\mathbf{0}_d, \Sigma)$*. If we define the transformation vector* V *as*

$$V = \left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right),$$

the density function of the Gaussian copula, $c^{\Sigma}_{Gauss}(u)$*, is given by:*

$$c^{\Sigma}_{Gauss}(u) = \frac{1}{\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2} V \cdot \left(\Sigma^{-1} - I\right) \cdot V^T\right). \tag{3}$$

The Gaussian copula is commonly chosen due to the normal distribution's well-understood properties, making it a suitable option for modeling dependencies. It has been applied in various fields where modeling the relationships between variables is essential, for instance in finance in [8] or in the insurance context in [3]..

3.4 Related Work in Tabular Data Generation

The generation of synthetic data is increasingly important for research and development, especially to maintain privacy and protect sensitive information. Various methods have been proposed for generating synthetic data, each offering different pros and cons.

Copulas for Synthetic Data Generation Copulas are a popular method for generating synthetic data because they effectively model dependencies between variables. By separating the modeling of marginal distributions from the dependency structure, copulas can capture complex relationships in data. This makes them useful for high-dimensional datasets where understanding dependencies is crucial. For example, Copulas could be used to generate synthetic training data for machine learning emulators in weather and climate [9].

Generative Adversarial Networks (GANs) for Synthetic Data GANs were originally developed for generating realistic images, but their use has expanded to include synthetic tabular data generation, notably with Conditional GAN [16] [2]. GANs consist of two neural networks: a generator, which creates synthetic data, and a discriminator, which evaluates how real the data looks. This setup helps the generator improve over time, producing increasingly realistic data. While GANs are powerful, applying them to tabular data could be challenging because of the mixed data types and complex dependencies typically found in

such data. Additionally, training GANs can be difficult and requires a lot of computational resources, which can be a barrier for some applications.

Autoencoders for Data Generation Autoencoders are neural networks that learn to compress data into a lower-dimensional form and then reconstruct it. Variational Autoencoders (VAEs) [7] extend this idea by adding a probabilistic element, which allows for the generation of new data points that are similar to the original data. While autoencoders are useful for capturing the underlying structure of data, they may not handle the complex dependencies between variables as effectively as copulas or GANs as claimed in [5].

4 Experimental results

4.1 Dataset Considered

This section provides a descriptive analysis of the dataset, which consists of $n = 28,633$ rows. The dataset is an extract from the census of MSH portfolio and includes information about individuals who have utilized generalist consultation benefits. It covers various aspects such as demographics, policy details, and usage statistics. You can find a mapping of variables and features in Appendix A

The key variables in the dataset include:

- **policy_year**: The year in which the policy was active.
- **gender**: Gender of the beneficiary (e.g., male, female).
- **expat_country**: The country where the expatriate resides.
- **frequency_group**: Categorization of the frequency of consultations.
- **severity_group**: Categorization of the severity of claims.
- **beneficiary_relationship**: Relationship of the beneficiary to the policyholder (e.g., self, spouse, child).
- **age_range**: Age group of the beneficiary.
- **nb_conjoints**: Number of spouses covered under the policy.
- **age_range_18**: Age group of the beneficiary (alternative categorization).
- **exposure**: The amount of time covered by the policy.
- **quantity**: Number of claims or consultations.

- **cost**: The total cost associated with the claims or consultations.

To better understand the structure and characteristics of the data, we use various visualizations. These graphs illustrate key aspects of the data, including age distribution, cost and quantity trends, exposure over time, and distribution across different groups.

Age Range Distribution Figure 1 shows the proportion of different age ranges in the dataset.

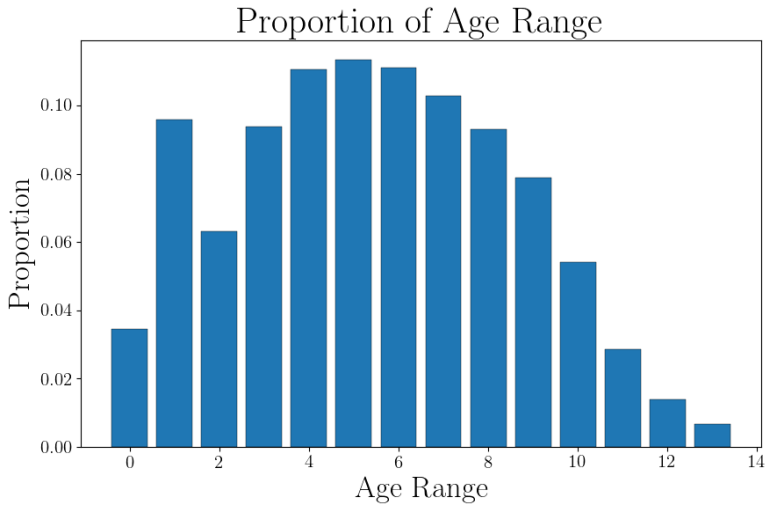

Figure 1: Proportion of Age Range

The age range distribution reflects the number of individuals in each respective age group. As seen in Figure 1, the majority of the population in the dataset is between 20 and 60 years old, with fewer observations outside of these age ranges.

Total Cost and Quantity per Policy Year Figure 2 depicts the total cost and total number of consultations per policy year.

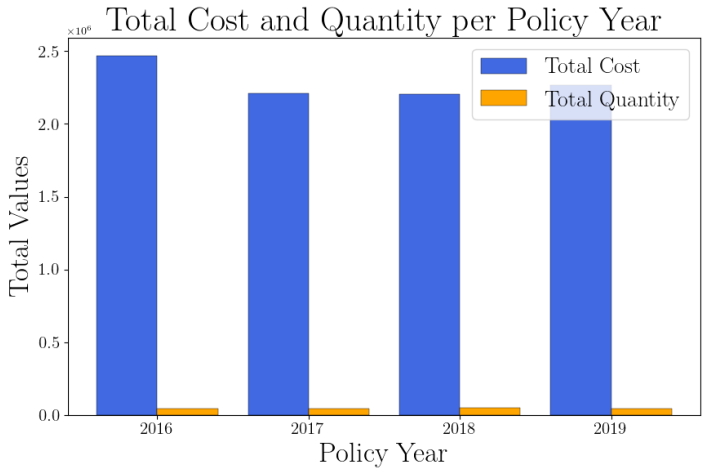

Figure 2: Total Cost and Quantity per Policy Year

The total cost represents the aggregate amount spent on generalist claims over the years 2016, 2017, 2018, and 2019. The total quantity indicates the total number of claims made during the same period. As shown in Figure 2, the number of claims is relatively low compared to the total cost, which can be attributed to the significant exposure in countries with higher medical costs.

Evolution of Exposure Over Policy Year Figure 3 illustrates the changes in exposure over different policy years.

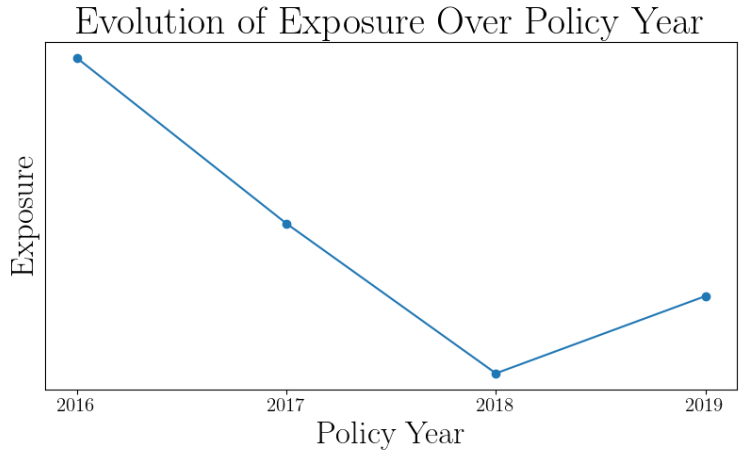

Figure 3: **Evolution of Exposure Over Policy Year**

Exposure represents the duration covered by the policy. For example, an individual with a policy covering only nine months in 2017 would have an exposure value of 0.75, indicating

three-quarters of a year. As shown in Figure 3, there was a significant decrease in exposure from 2016 to 2018, which is not directly reflected in the total cost (Figure 2) due to high medical inflation and the nature of the countries covered by the portfolio.

Frequency Group Distribution Figure 4 presents the distribution of different frequency groups within the dataset.

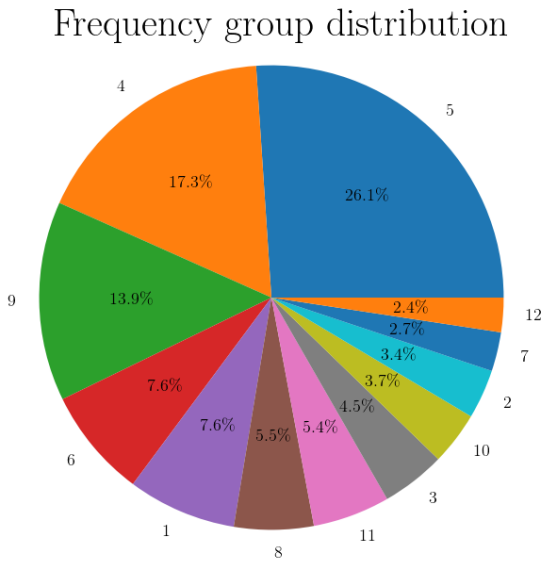

Figure 4: **Frequency Group Distribution**

The frequency group indicates the likelihood of an individual using the benefits, taking into account factors such as age, gender, and country of origin. This classification is based on the historical experience of the MSH portfolio. As shown in Figure 4, there are fewer observations for individuals in the highest and lowest frequency groups.

Marital Status Distribution Figure 5 shows the distribution of marital status within the dataset.

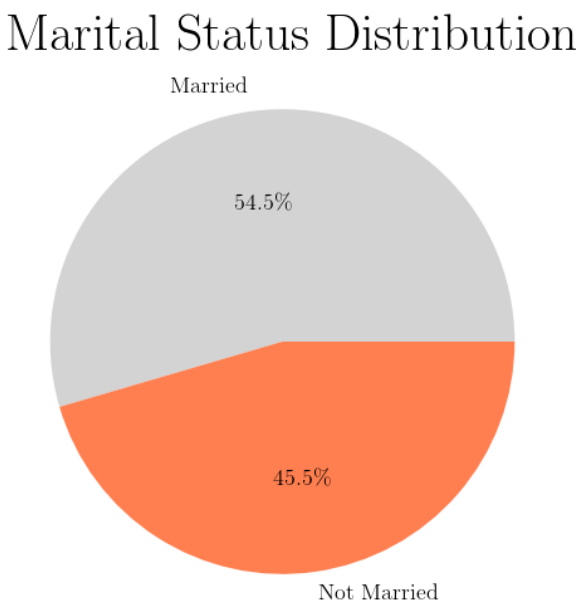

Figure 5: **Marital Status Distribution**

Marital status indicates whether the policyholder is married. As observed in Figure 5, the dataset is relatively balanced, with a similar number of observations for both married and unmarried individuals.

Severity Group Distribution Figure 6 illustrates the distribution of severity groups.

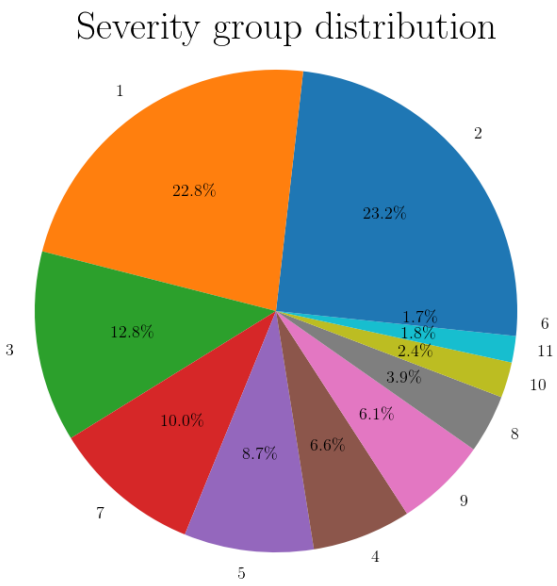

Figure 6: **Severity Group Distribution**

The severity group represents the cost of medical treatment in different countries, based on historical data from the portfolio. A value of 1 indicates the least expensive countries for treatment, while a value of 11 represents the most expensive (e.g., USA, Hong Kong). As shown in Figure 6, the dataset includes a range of severity groups, reflecting the varied cost of treatment across different regions.

Distribution of cost, quantity and exposure Figure 7 illustrates the distribution of cost,quantity and exposure.

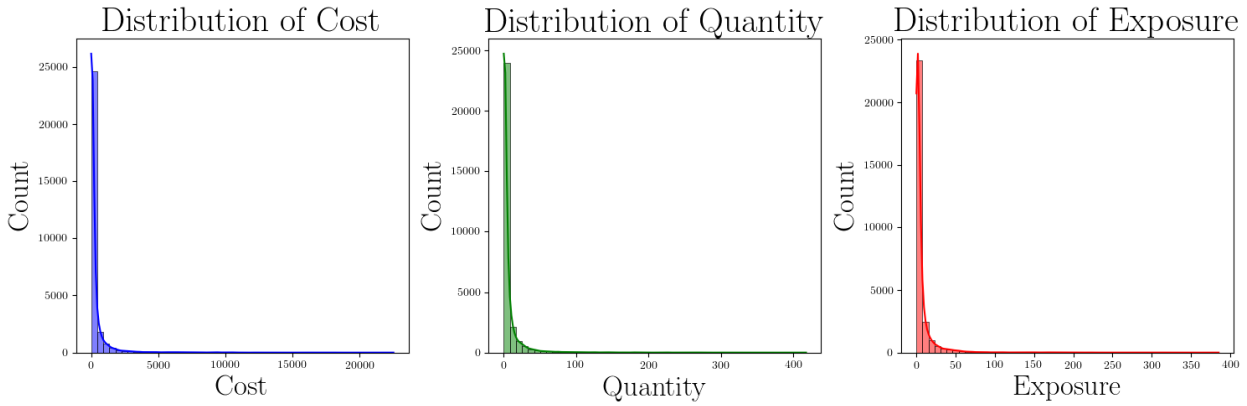

Figure 7: **Distribution of cost, quantity and exposure.**

As seen in Figure 7, the distributions of cost, quantity, and exposure are all heavily skewed to the right, indicating that these variables are heavy-tailed. Most of the observations are concentrated near the lower values, with a sharp drop-off as the values increase. This pattern suggests that while the majority of claims and consultations have relatively low costs, quantities, and exposure durations, there are occasional instances with significantly higher values, leading to the heavy-tailed nature of these distributions.

4.2 Experimental protocol

4.2.1 Data generation

In our setup, F_θ is fitted using various families of distributions to capture the behavior of continuous variables effectively. These distributions include:

- **Beta Distributions** Beta(α, β): Suitable for modeling variables within [0, 1].
- **Gamma Distributions** $\Gamma(\alpha, \beta)$: Used for skewed distributions.
- **Gaussian Kernel Density Estimation (KDE)**: Non-parametric density estimation.

- **Gaussian Distributions** $\mathcal{N}(\mu, \sigma^2)$: For normally distributed data.
- **Truncated Gaussian Distributions**: Gaussian distributions limited to a range.
- **Student's t Distributions**: For data with heavier tails.
- **Uniform Distributions**: Modeling equally likely outcomes within a range.
- **Log-Laplace Distributions** $LL(\mu, b)$: For log-normal behavior.
- **Exponential Distributions**: Modeling time until an event.
- **Levy Distributions**: For highly skewed data with heavy tails.

The first eight distributions are supported by the Copulas Library, while custom tests were implemented for the exponential and Levy distributions to broaden the scope of fitting options. Parameters for each distribution are estimated using Maximum Likelihood Estimation (MLE), which maximizes the likelihood of the observed data under the model.

In order to choose the best-fitting marginal distribution for each variable, we use the Kolmogorov-Smirnov (KS) test. The algorithm works as described in Algorithm 1.

Data: Dataset $\{x_j^i\}_{i=1}^n$ for each variable $i \in \{1, \ldots, d\}$
Result: Best-fitting distribution for each variable
for $i \in \{1, \ldots, d\}$ **do**
  for *distribution* $p_\theta \in \mathcal{F}$ **do**
    Fit parameters θ_i using data $\{x_i^j\}_{j=1}^n$;
    Compute the predicted CDF F_{θ_i};
    Compute the empirical CDF F_i^n;
    Perform KS test between F_{θ_i} and F_i^n;
    Store KS test statistic;
  end
  Select the distribution with the minimum KS test statistic;
end

Algorithm 1: Algorithm for Selecting Best-Fitting Univariate Distributions

Once the best-fitting marginal distributions are selected for each variable, the next step is to model the dependencies between these variables. For this purpose, we use the Multivariate Gaussian copula. The Gaussian copula is chosen due to its simplicity and the straightforward interpretation of its parameters, which are based on the correlation structure of the data.

The Gaussian copula relies on a correlation matrix, Σ , to capture the dependency structure between the variables. This correlation matrix has an analytical from, which means that

it can be directly calculated from the data by calculating the pairwise correlations between the variables. This matrix encapsulates the linear relationships among all pairs of variables.

By using the Multivariate Gaussian copula, we can model the dependencies between variables independently from the marginal distributions. This approach allows us to keep the relationships observed between different variables in the real data while still taking into account different marginal behaviors of each variable. This is useful for generating synthetic data that needs to reflect both the individual characteristics of each variable and their interactions.

Using the Multivariate Gaussian copula, we construct a joint distribution that captures these dependencies effectively, enabling us to generate synthetic datasets that preserve the basic statistical properties of the original data.

4.2.2 Evaluation Metrics

In this section, we present the metrics used to evaluate the quality of the generated synthetic data. First, to assess the dependency structure of the generated dataset, we employ Pearson's correlation, Kendall's Tau, and Spearman's rank correlation. These measures will help us understand how well the synthetic data preserves the relationships between variables as observed in the real data. Second, to evaluate the individual fitted marginals, we use the Kolmogorov-Smirnov (KS) test to compare the empirical distributions of the generated data against those of the original data. Additionally, we will compare key descriptive statistics such as the mean, median, and the first and third quartiles to understand how well the synthetic data matches the overall characteristics of the real dataset. Finally, we compare the results of the copula-generated data with data generated using a Gaussian Mixture Model (GMM) to provide a baseline for the evaluation.

Kolmogorov-Smirnov Test (KS-test) The KS-test [4] aims to test whether two (one-dimensional) cumulative distribution functions (CDFs) are equal. For each feature i, we compare the empirical CDF F_i^n with n samples from $\mathscr{D}$ and the estimated CDF F_θ. In 1D, the empirical CDF is given by:

$$F_i^n(x) = \frac{\mid x_i^j \mid j \in \{1, \ldots, n\},\ x_i^j \leq x\} \mid}{n} = \frac{1}{n} \sum_{j=1}^{n} 1_{(-\infty, x]} \left(x_i^j\right),$$

With n samples, the test statistic D_i^n is computed as:

$$D_i^n = \sup_x \left| F_i^n(x) - F_\theta(x) \right|,$$

Under the null hypothesis that the sample comes from the distribution F_θ, we use the critical values of the Kolmogorov distribution to decide if we reject the null hypothesis. If:

$$\sqrt{n}D_i^n > K_\alpha,$$

where K_α is the critical value at the significance level α, we reject the null hypothesis.

Kendall's Tau Kendall's Tau (τ) [6] is a statistic that measures the ordinal association between two variables. It is based on the relative ordering of data rather than the actual values, making it robust to non-linear relationships and less sensitive to outliers.

Given a dataset with n pairs (x_i, y_i), Kendall's Tau is calculated by considering all possible pairs of observations and counting the number of concordant and discordant pairs:

- A pair of observations (x_i, y_i) and (x_j, y_j) is concordant if both $x_i > x_j$ and $y_i > y_j$, or if both $x_i < x_j$ and $y_i < y_j$. - A pair is discordant if $x_i > x_j$ and $y_i < y_j$, or if $x_i < x_j$ and $y_i > y_j$.

The formula for Kendall's Tau is:

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)}$$

where: - C is the number of concordant pairs, - D is the number of discordant pairs, - n is the total number of pairs.

Kendall's Tau ranges from -1 (perfect disagreement) to 1 (perfect agreement), with 0 indicating no association.

Spearman's Rank Correlation Coefficient Spearman's rank correlation coefficient (ρ) [14] measures the strength and direction of the monotonic relationship between two variables. It is a non-parametric measure, meaning it does not assume a specific distribution for the data.

To calculate Spearman's ρ, each value in the dataset is converted to its rank. If x_i and y_i are the ranks of the i-th pair, Spearman's rank correlation is calculated as the Pearson correlation coefficient of the ranks:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where: - $d_i = \text{rank}(x_i) - \text{rank}(y_i)$ is the difference between the ranks of each observation, - n is the number of observations.

This coefficient ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation.

Pearson's Correlation Pearson's correlation coefficient (r) measures the linear relationship between two continuous variables. It assesses how well the relationship between the variables can be described using a straight line.

Given a dataset of n pairs (x_i, y_i), Pearson's correlation coefficient is defined for $u_1, \ldots, u_d \in [0,1]^d$ as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where: - x_i and y_i are the individual sample points, - $\bar{x}$ and $\bar{y}$ are the means of x and y respectively.

Pearson's r ranges from -1 to 1, where: - $r = 1$ indicates a perfect positive linear relationship, - $r = -1$ indicates a perfect negative linear relationship, - $r = 0$ indicates no linear relationship.

Pearson's correlation assumes that the data is normally distributed and that the relationship between the variables is linear.

Descriptive Statistics Comparison In addition to the dependency metrics, we will also compare the descriptive statistics of the real and synthetic data. This comparison the quartiles of the distributions. By comparing these statistics, we can assess how well the synthetic data captures the central tendency and dispersion of the real data. This helps in understanding whether the synthetic data accurately represents the overall characteristics of the original dataset.

4.3 Results

4.3.1 Visualization

This section presents a series of visualizations comparing the real data with the synthetic data. The visualizations help us understand how well the synthetic data captures the characteristics of the real data in terms of distribution and relationships among key variables.

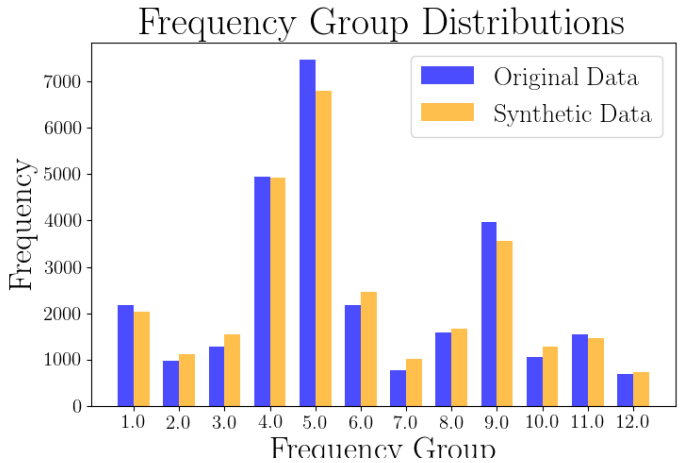

Figure 8: **Comparison of Frequency Group: Real Data vs. Synthetic Data**

Comparison of Frequency Group Figure 8 shows the distribution of individuals across different frequency groups for both real and synthetic datasets. The distributions are fairly similar, indicating that the synthetic data successfully replicates the overall structure of frequency groups observed in the real data. However, slight discrepancies in certain groups suggest that some nuances of the real data may not be fully captured.

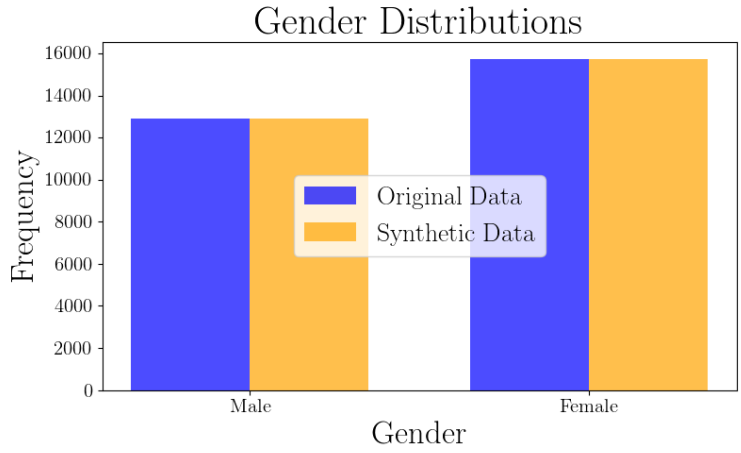

Figure 9: Comparison of Gender Counts: Real Data vs. Synthetic Data

Comparison of Gender Counts In Figure 9, the number of male and female individuals in both the real and synthetic datasets is compared. The bar heights indicate that the frequencies in the synthetic data closely matches that of the real data, showing the model's ability to maintain balance between two populations

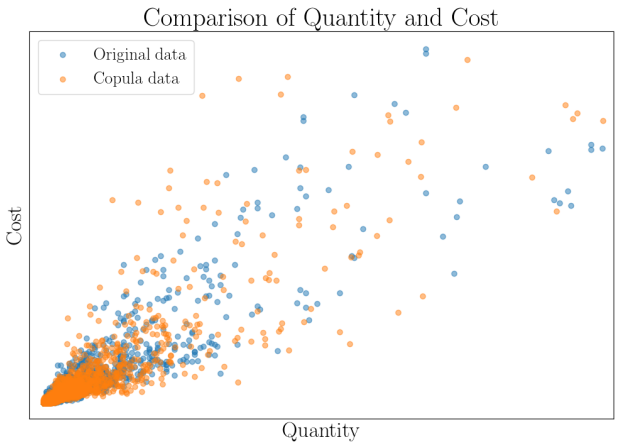

Figure 10: **Comparison of Quantity and Cost: Original vs. Synthetic Data**

Relationship Between Quantity and Cost Figure 10 provides a scatter plot comparing the relationship between the number of claims (quantity) and the associated costs in the real and synthetic datasets. The overall pattern indicates a positive relationship between quantity and cost, which is preserved in the synthetic data. This consistency suggests that the synthetic data captures the core economic behavior of claims.

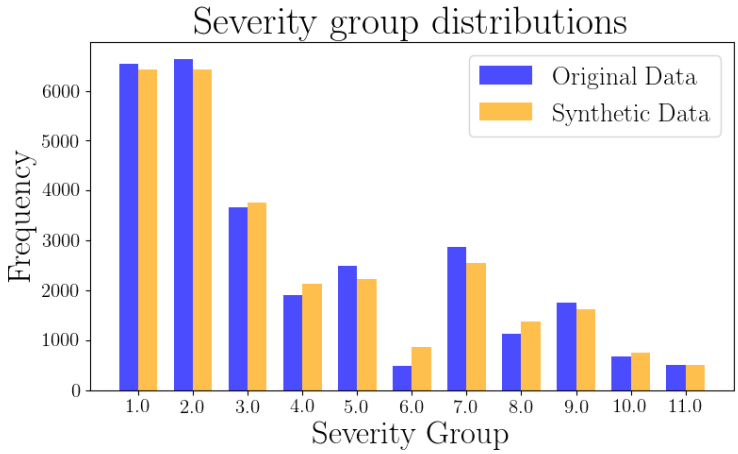

Figure 11: **Comparison of Severity Group: Real Data vs. Synthetic Data**

Comparison of Severity Group In Figure 11, the distribution of severity groups between the real and synthetic datasets is shown. While the overall trend is similar, with most data points falling into certain severity categories, there are minor variations. This could imply that while the synthetic data models severity adequately, it may not fully capture the extremes.

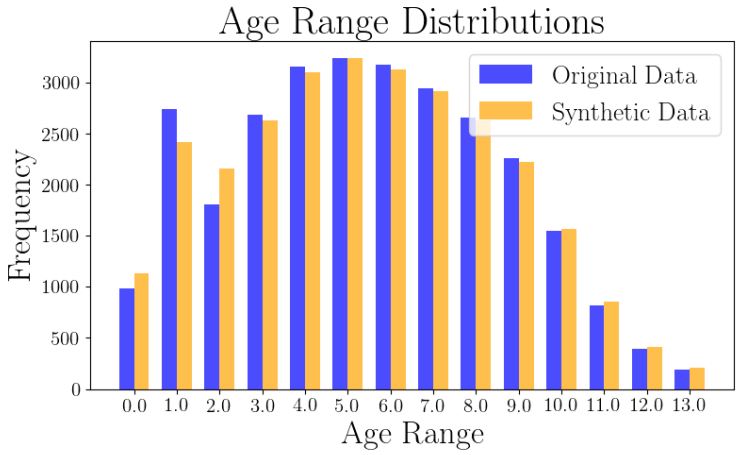

Figure 12: **Comparison of Age Range: Real Data vs. Synthetic Data**

Comparison of Age Range Figure 12 compares the age distribution in the real and synthetic datasets. The age distribution is replicated well in the synthetic data, maintaining a similar shape and spread. This consistency ensures that age-related analysis remains valid in synthetic data-driven studies.

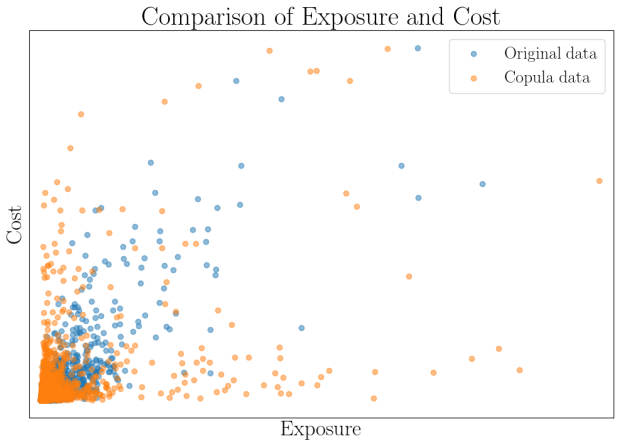

Figure 13: **Comparison of Exposure and Cost: Original vs. Synthetic Data**

Relationship Between Exposure and Cost The scatter plot in Figure 13 shows the relationship between exposure and cost. Both datasets exhibit a clustered pattern with a few outliers, which is consistent across real and synthetic data. The synthetic data captures the general trend, suggesting that the model handles exposure-related cost variations adequately.

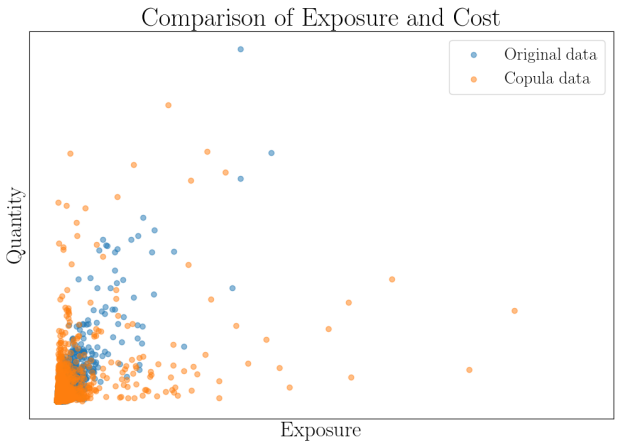

Figure 14: Comparison of Exposure and Quantity: Original vs. Synthetic Data

Relationship Between Exposure and Quantity Figure 14 presents a comparison between exposure and quantity. The scatter plot indicates that as exposure increases, the quantity of claims also increases. This trend is evident in both the real and synthetic datasets, highlighting the model's ability to preserve key relationships within the data.

The visualizations demonstrate that the synthetic data generated captures many of the essential characteristics of the real data. The distributions of key variables and the relationships between them are mostly maintained, though there are minor discrepancies in some areas, such as specific frequency and severity groups. These differences could be attributed to the variability and complexity of the real-world data that may not be fully replicated by our synthetic data generation process. Overall, the synthetic data provides a reasonably accurate representation of the original dataset, making it potentially useful for various analytical purposes while ensuring data privacy.

4.3.2 Quantitative Results

In this section, we present the quantitative results of our evaluation metrics. We compared the performance of synthetic data generated using the Gaussian copula with that generated using Gaussian Mixture Models (GMM). The comparison was made against the original dataset to assess the quality of synthetic data.

Differences in Marginal Distributions **Quantile-quantile plot results.** Figure 15 displays the Quantile-quantile plots comparing the generated data with the real data quantiles. These plots provide a visual assessment of how well the synthetic data, generated using both Copula and GMM methods, aligns with the original data distribution across the 13

features.

Overall, most of the features show points lying close to the diagonal line, indicating that both methods replicate the distribution of the original data fairly well. For features such as *policy_year*, *gender*, and *beneficiary_relationship*, the alignment is almost perfect, demonstrating that both methods capture these categorical distributions accurately.

However, for features like *frequency_group* and *severity_group*, some deviations from the diagonal are observed, particularly in the higher quantiles. This suggests that while the general trend of these features is captured, the extremes are less accurately modeled. The Copula method tends to align slightly better in these cases, hinting at its potential advantage in modeling spread for more extreme values.

Features such as *quantity* and *cost* exhibit greater deviation from the diagonal, especially towards the higher end, highlighting the challenge of accurately replicating skewed distributions with high-value outliers. This indicates areas where synthetic data generation can be improved, particularly in accurately capturing tails and rare events.

We see that, both synthetic generation methods (Copula and GMM) do a reasonable job at approximating the distribution of the original data across most features, with some slight advantages of Copula in handling extreme values. However, deviations, especially in higher quantiles for features like *frequency_group*, *severity_group*, *quantity*, and *cost*, highlight areas where synthetic data generation could be improved, particularly in modeling the tails and rare high-value cases.

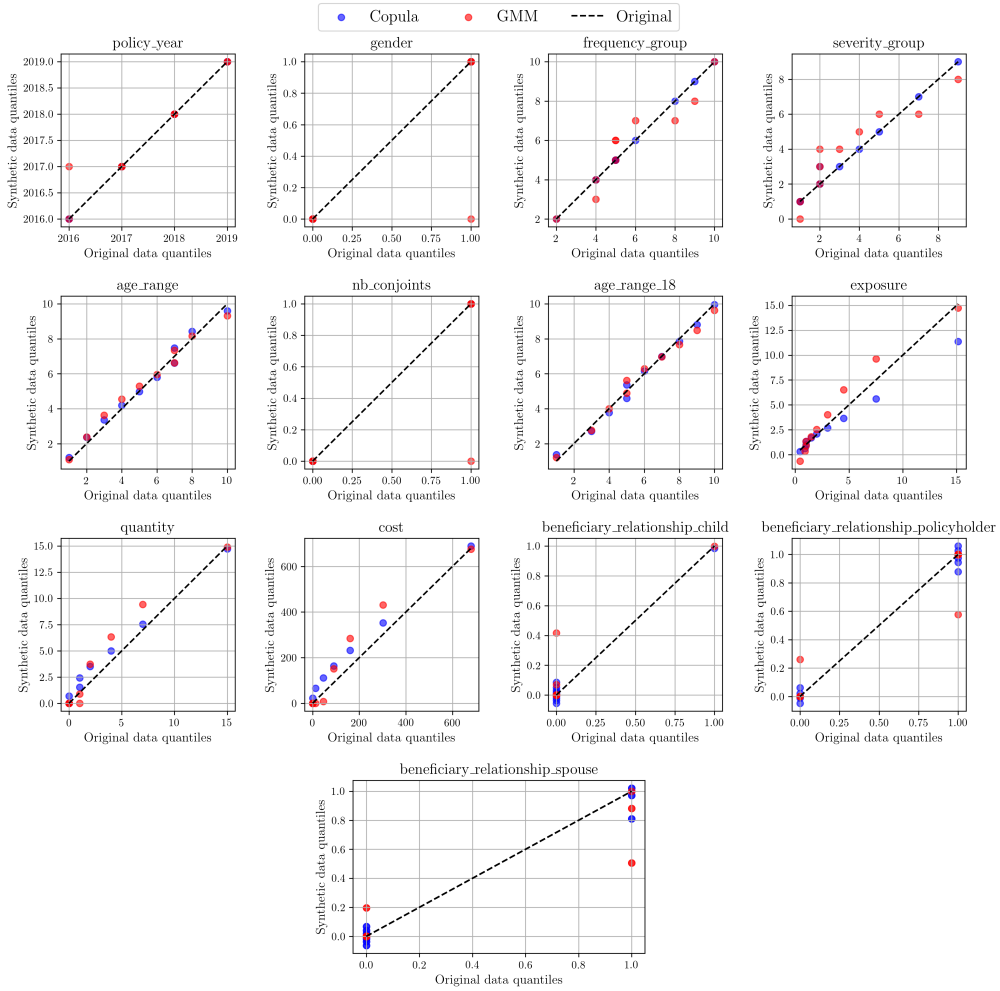

Figure 15: **Quantile/quantile plots.** Quantiles of the original data distribution *v.s.* quantiles of the generated data plots for the 13 different features of the dataset. Red dots are quantiles from the GMM baseline, and blue points from the Copula.

Kolmogorov-Smirnov Test Results. Figure 16 displays the p-values from the Kolmogorov-Smirnov test, comparing the CDFs of the real and synthetic data for both methods across different features.

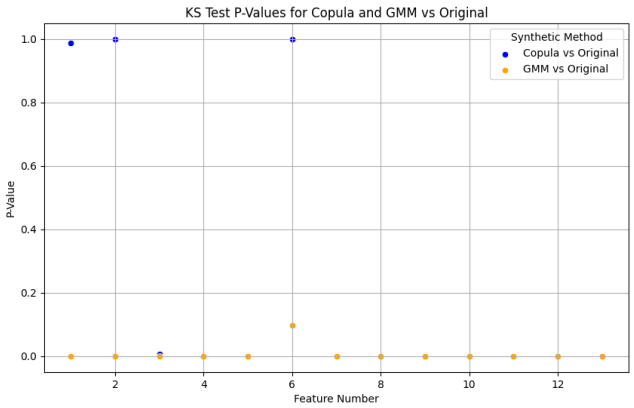

Figure 16: KS Test P-Values for Copula and GMM vs. Original Data

The KS Test P-values give us insight into how closely the synthetic data resembles the original data's distribution. Here's what we can gather from the results:

Features like *policy_year* (1) and *nb_conjoints* (6) show high p-values for the Copula method, close to 1. This indicates that the synthetic data generated with Copula matches the real data quite well for these features. High p-values mean we cannot reject the idea that the synthetic and original samples come from the same distribution, suggesting a good fit.

Most features, particularly when using the GMM method, have p-values near zero. This indicates significant differences between the synthetic and original data distributions, implying that the GMM struggles to accurately reproduce the original data, especially for features like *frequency_group* (3), *severity_group* (4), *quantity* (9), and *cost* (10). Low p-values suggest that these synthetic datasets are not capturing the complexities of the original data, especially at the extremes. The Copula method has a mixed performance for some features. While it replicates features like *policy_year* and *nb_conjoints* well, it still shows low p-values for features such as *frequency_group* and *severity_group*. This indicates that, although Copula may be better than GMM in some cases, it still has limitations, particularly in capturing more complex or nuanced distributions.

The results indicate that Copula generally outperforms GMM in capturing the distributions of the original data, especially for certain features. However, both methods face difficulties with features that have complex distributions or contain extreme values. Improving these methods to better handle skewed data and outliers could enhance the quality of synthetic data generation.

Correlation Differences Figure 17 shows the differences in Kendall's Tau, Spearman's Rank, and Pearson's correlation coefficients between the original and synthetic datasets. The

histograms represent the distribution of differences for each method.

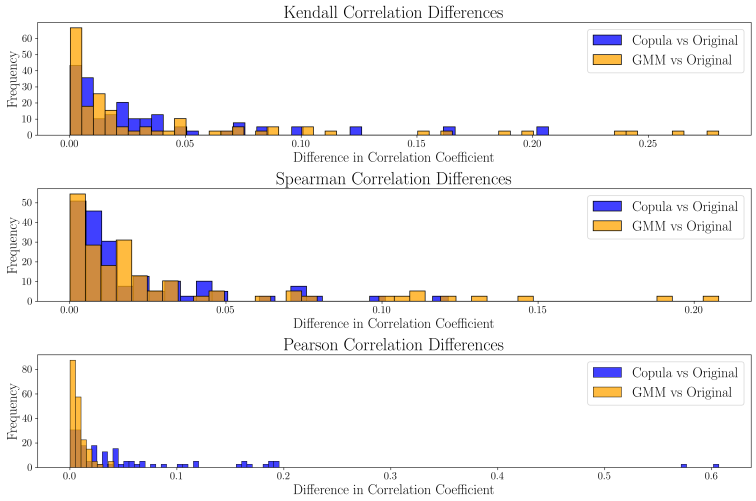

Figure 17: Differences in Correlation Coefficients: Copula vs. GMM vs. Original Data

The plots in Figure 17 compare the differences in correlation coefficients (Kendall's Tau, Spearman's Rank, and Pearson's) between the original data and the synthetic datasets generated using the Copula and GMM methods. The differences are measured in terms of how much the correlations deviate from the original data, with separate bars for each method.

Kendall Correlation Differences: The first plot shows the differences in Kendall's Tau correlation. Most of the differences are clustered around zero, indicating that both Copula and GMM methods are reasonably close to capturing the ordinal relationships present in the original data. However, the mean error for Copula (0.0316) is slightly lower than for GMM (0.0395), suggesting that Copula does a marginally better job at preserving the rank-based relationships.

Spearman Correlation Differences: The second plot represents the differences in Spearman's rank correlation. Similar to Kendall's Tau, the differences are mostly small, with Copula showing a mean error of 0.0188 compared to 0.0282 for GMM. This indicates that Copula is again slightly more effective at maintaining the monotonic relationships seen in the original data.

Pearson Correlation Differences: The third plot, showing Pearson correlation differences, has a noticeably different pattern. Here, the Copula method shows a broader spread of differences, with some reaching up to 0.6. This reflects in the higher mean error for Copula (0.0559) compared to GMM (0.0047). The broader spread in Copula is due to Pearson's sensitivity to extreme values, and since the synthetic data generated using Copula often includes

more extreme values (a result of fitting heavy-tailed marginals), the Pearson correlation is more affected. This suggests that while Copula might be better for capturing the distribution and dependency structures in general, it can introduce more variability when it comes to linear relationships.

Overall, the Copula method tends to preserve the dependency structures of the original data better than GMM, as seen in the Kendall and Spearman correlation differences. However, when it comes to capturing linear relationships (as measured by Pearson), Copula introduces more variability due to the generation of extreme values, highlighting the trade-off between accurately modeling marginal distributions and preserving linear correlations.

Analysis and Interpretation The results indicate that the Gaussian copula method provides a reasonable approximation of both marginal distributions and the dependence structure of the original data, often outperforming GMM. The minimal differences in mean and median suggest that the central tendency is well-preserved. The larger discrepancies observed in the quartiles indicate that extreme values and tails are less accurately replicated, a common challenge in synthetic data generation.

Overall, these quantitative results demonstrate that while both copula and GMM methods are effective, the Gaussian copula has a slight edge in preserving the real data's characteristics. This finding aligns with our expectation that the copula's ability to model dependencies independently of marginals can result in more accurate synthetic data, especially when the focus is on maintaining the integrity of relationships within the dataset.

5 Limitations and Future Work

5.1 Limitations

While the synthetic data generation methods and evaluation techniques used in this study have shown promise, several limitations need to be acknowledged:

Algorithmic Complexity One of the significant limitations of the methods employed is the algorithmic complexity, especially when dealing with high-dimensional data. The process of fitting copulas and generating samples can become computationally expensive as the number of variables increases. This is particularly important in the inversion of the cumulative distribution function (CDF) required for sampling. When dealing with large datasets with many features, the computational complexity can increase exponentially, making real-time or large-scale applications difficult.

Scalability with Number of Features As the dimensionality of the dataset increases, the complexity of capturing and modeling the dependencies between variables also increases. High-dimensional copulas require a large amount of data to estimate the dependence structure in an accurate way. In practice, this may limit the effectiveness of the copula approach when applied to datasets with a large number of features, as it can lead to overfitting or the inability to accurately capture relationships between variables.

Assumptions of Independence and Distribution The approaches employed in this study rely on assumptions about independence and distribution that may not accurately reflect the complexities of real-world data. For instance, assuming that marginal distributions can be independently modeled and then combined using copulas might fail to capture the intricate dependencies found in complex insurance datasets. Moreover, the use of Gaussian copulas, which assume certain types of dependencies like linear or monotonic relationships, limits our ability to accurately model the full range of interactions between variables. As a result, we are primarily able to capture linear relationships, potentially overlooking more complex, non-linear interactions that could be significant in understanding the data.

5.2 Future Work

To address these limitations and further improve the applicability of synthetic data generation methods, we have several options for future research:

Exploration of Conditional Generative Adversarial Networks (GANs) One promising direction for future work is the exploration of Conditional Generative Adversarial Networks (GANs) [10] for synthetic data generation. GANs have the potential to capture complex, non-linear dependencies between features without the need for explicit parametric forms. Conditional GANs can be conditioned on specific attributes or outcomes, allowing for more accurate synthetic data generation that aligns closely with real-world data distributions. However, unlike copulas, GAN models are subject to the black box effect, meaning we cannot directly observe the relationships between variables.

Integration of Neural Network-Based Approaches Neural networks, particularly deep learning models, offer another avenue for enhancing synthetic data generation. Techniques such as Variational Autoencoders (VAEs) [7] or other neural network-based models can be employed to learn latent representations of data, from which new samples can be generated. These approaches can be particularly beneficial in capturing intricate dependencies and higher-order interactions between variables, which traditional methods may not effectively model.

Optimization of Computational Efficiency Future research should also focus on optimizing the computational efficiency of synthetic data generation processes. This could involve improving algorithms for copula fitting and sampling or exploring parallel computing and cloud-based solutions to work with large-scale data. Reducing the computational burden will make these methods more applicable to real-world scenarios, where quick and scalable solutions are often required.

Empirical Validation and Benchmarking Finally, further empirical validation and benchmarking against real-world data and alternative synthetic data generation methods are crucial. This will involve comparing the performance of copula-based methods, GANs, neural networks, and hybrid approaches across different datasets and contexts. Establishing benchmarks and performance metrics will help identify the most effective methods for various types of insurance data and applications. As well as measuring the performance of synthetic data in real models.

Appendices

A Mapping of variables and features

Code	Age Range
0	Between 0 and 4 years old
1	Between 4 and 20 years old
2	Between 20 and 25 years old
3	Between 25 and 30 years old
4	Between 30 and 35 years old
5	Between 35 and 40 years old
6	Between 40 and 45 years old
7	Between 45 and 50 years old
8	Between 50 and 55 years old
9	Between 55 and 60 years old
10	Between 60 and 65 years old
11	Between 65 and 70 years old
12	Between 70 and 75 years old
13	Between 75 and 120 years old

Table 1: Age Range Codes and Corresponding Age Intervals

Feature	Code
policy_year	1
gender	2
frequency_group	3
severity_group	4
age_range	5
nb_conjoints	6
age_range_18	7
exposure	8
quantity	9
cost	10
beneficiary_relationship_child	11
beneficiary_relationship_policyholder	12
beneficiary_relationship_spouse	13

Table 2: Feature Mappings in the Dataset

B Dataset prepossessing

B.1 Categorical Data

Categorical variables present another challenge, as they cannot be directly modeled using copulas. To incorporate categorical data into our analysis, we employed a transformation technique known as One-Hot Encoding [5]. This method converts each category into a separate binary variable, effectively capturing the interaction and dependency structures between different categories. By using One-Hot Encoding, we can include categorical variables in the copula model while preserving the nuances of their relationships.

B.2 Correction of negative values

To ensure the coherence of the data, we first addressed the presence of negative values of the cost and quantity, in our generated dataset. Negative values are not meaningful in this context, as they do not reflect realistic scenarios in insurance claims and policy usage. These negative values were not present in the original dataset but were introduced during the synthetic data generation process using Gaussian copulas. This is a known issue with Gaussian copulas, which can produce values outside the range of the original data due to their reliance on the normal distribution. The normal distribution is symmetric around its mean, meaning it allows for values on both sides of the mean, including negative values When generating synthetic data, this inherent symmetry can result in negative values. This represents one of the limitations of our approach.

C Gaussian Copulas

A Gaussian copula is constructed using the multivariate normal distribution. It enables us to model the dependency structure separtly from the marginal distributions of each variable. That enables us to build models with arbitrary marginal distributions and Gaussian-like dependence

The Gaussain copula $C^{\Sigma}_{\text{Gauss}}$ for a given correlation matric Σ is defined as :

$$C^{\Sigma}_{\text{Gauss}}(u_1, \ldots, u_d) = \Phi_{\Sigma}\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right)$$

Where :

- Φ_{Σ} is the cumulative distribution function (CDF) of a centered multivariate normal distribution with correlation matrix Σ.
- Φ^{-1} is the inverse CDF (quantile function) of the standard normal distribution.

- $u_1, \ldots, u_d$ are the uniform marginals on the interval $[0, 1]$.

The correlation matrix Σ plays a crucial role in defining the Gaussian copula. It encapsulates the linear dependencies between the variables. The elements σ_{ij} of the matrix represent the correlation coefficients between variables i and j. In the Gaussian copula, Σ is the only parameter that needs to be estimated from the data, which makes the model more computationally efficient

D GMMs

The Gaussian Mixture Model (GMM), was implemented based on the open-sourced Scikit-Learn Implementation [12]. Following the documentation, the number of modes was selected based on the value of the Akaike information criterion (AIC) 18. It is defined as :

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

where k is the number of parametrs in the model and $\hat{L}$ is maximized value of likelyhood function for the model. The AIC balances model fit and complexity, with lower values indicating a better model. In our case, the optimal numbers of modes was 9, as this minimized the AIC. In addition to AIC, the Negative Log-Likelihood(NLL) was used to mesure wellness of fit the GLM to the data (Figure 18). NLL measures the model's ability to explain the data with lower values indicating a better fit. The NLL is defined as :

$$\text{NLL} = -\sum_{i=1}^{n} \log p(x_i|\theta)$$

where $p(x_i|\theta)$ represents the probability of observing data point x_i under the model parameters θ. The NLL metric confirms the choice of 9 modes, showing that this configuration provides a good representation of the data while avoiding overfitting [1] (See Figure 19).

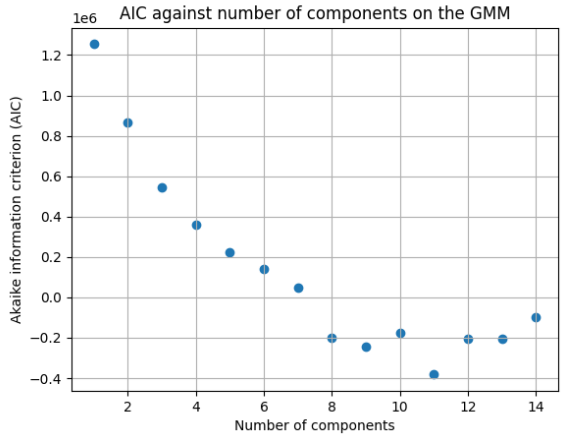

Figure 18: AIC value after fiting on the dataset for different number of modes. In this thesis, we therefore considered 11 modes in the GMM implementation.

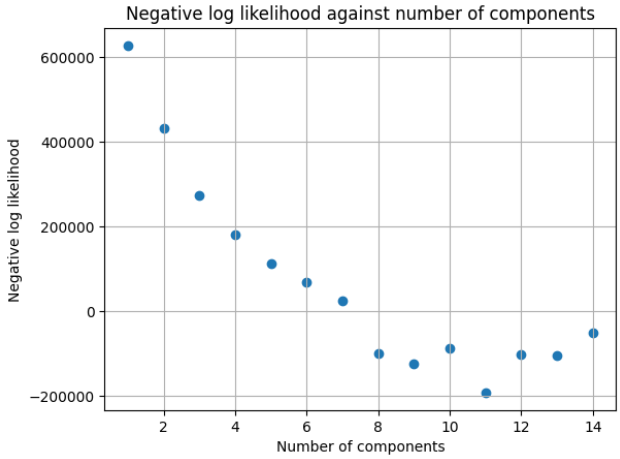

Figure 19: NLL value after fiting on the dataset for different number of modes. In this thesis, we therefore considered 11 modes in the GMM implementation.

References

[1] Hirotugu Akaike. Akaike's information criterion. *International encyclopedia of statistical science*, pages 25–25, 2011.

[2] Insaf Ashrapov. Tabular gans for uneven distribution, 2020.

[3] Claudia Czado, Rainer Kastenmeier, Eike Christian Brechmann, and Aleksey Min. A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, 2012(4):278–305, 2012.

[4] Wayne W Daniel. Kolmogorov–smirnov one-sample test. *Applied nonparametric statistics*, 2, 1990.

[5] Regis Houssou, Mihai-Cezar Augustin, Efstratios Rappos, Vivien Bonvin, and Stephan Robert-Nicoud. Generation and simulation of synthetic datasets with copulas, 2022.

[6] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[7] Diederik P Kingma. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[8] Yannick Malevergne and Didier Sornette. Testing the gaussian copula hypothesis forfinancial assets dependences. *Quantitative finance*, 3(4):231, 2003.

[9] David Meyer, Thomas Nagler, and Robin Hogan. Copula-based synthetic data generation for machine learning emulators in weather and climate: application to a simple radiation model, 12 2020.

[10] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[11] Roger B Nelsen. *An introduction to copulas*. Springer, 2006.

[12] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[13] M Sklar. Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, volume 8, pages 229–231, 1959.

[14] Charles Spearman. The proof and measurement of association between two things. 1961.

[15] Zhiqiang Wang, Mathieu Ritou, Catherine M. da Cunha, and Benoît Furet. Contextual classification for smart machining based on unsupervised machine learning by Gaussian mixture model. *International Journal of Computer Integrated Manufacturing*, 33(10-11):1042–1054, July 2020.

[16] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*, 2019.