Master

Actuarial Science

Master's Final Work

Dissertation

Modelling Dependencies in Airport Passenger Claim Data Using Copulas

Roberto Carcache Flores

February 2022

Master

Actuarial Science

Master's Final Work

Dissertation

Modelling Dependencies in Airport Passenger Claim Data Using Copulas

Roberto Carcache Flores

Supervision:

Alexandra Bugalho de Moura

Manuel Cidraes Castro Guerra

February 2022

GLOSSARY

CBS Checked Baggage Claim Severity. i, 8, 26, 35

CBX Checked Baggage Claim Counts. i, 8, 26

CML Canonical Maximum Likelihood. i, 3

CPS Security Checkpoint Claim Severity. i, 8, 26, 34, 35

CPX Security Checkpoint Claim Counts. i, 8, 26

ECDF Empirical Cumulative Distribution Function. i, 31

GLM Generalized Linear Models. i, 2, 4

GOF Goodness of Fit. i, 32

JEL Journal of Economic Literature. i, ii

LTDC Lower Tail Dependence Coefficient. i, 21, 22

MLE Maximum Likelihood Estimation. i, ii, 3, 12, 14, 28

NCT Noncentral Student t distribution. i, 18, 19

PDS Property Damage Claim Severity. i, 8, 32, 47

PDX Property Damage Claim Counts. i, 8

PLS Property Loss Claim Severity. i, 8, 32, 47

PLX Property Loss Claim Counts. i, 8

PMF Probability Mass Function. i, 14

TSA Transportation Security Administration. i, v, vi, 1, 3–10, 21, 22, 27, 36, 40, 41

TVaR Tail Value at Risk. i, ii, 1, 29, 32, 35, 47, 48

UTDC Upper Tail Dependence Coefficient. i, v, 21–26

VaR Value at Risk. i, ii, 1, 3, 29, 30, 32, 33, 35, 41, 47, 48

ABSTRACT

Every year, thousands of passengers flying through U.S. airports file claims to the Transportation Security Administration (TSA). The objective of this dissertation is to use copulas to model dependencies in counts and severities for TSA claims, during the years 2007-2015. Initially, monthly claim counts and amounts are aggregated from daily records, according to their type and site. These monthly series are detrended and fit into different probability distributions using Maximum Likelihood Estimation (MLE), to obtain the corresponding parameters.

Once the marginal distributions are obtained, it is possible to fit them into different bivariate copulas. These bivariate copulas are used to determine different tail dependence measures and to highlight non-linear dependencies between the variables. The final procedure involves fitting multivariate copulas and performing simulations. These simulations contrast risk measures like monthly Value at Risk (VaR) and Tail Value at Risk (TVaR) estimates for the different copulas used, along with the independence case and historical values.

The results show modelling claims with copulas can yield higher risk measures than the historical values, for random variables with heavy-tailed distributions. The choice of the copula used is also important in this sense, as different copulas generate different simulated risk measures. All of the data processing and modelling is performed using different open source Python libraries.

KEYWORDS: Copulas; Risk Modelling; Airport Claims; Dependencies

JOURNAL OF ECONOMIC LITERATURE (JEL) CODES: G22, C14, C15, C46

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I want to thank my parents, friends, and family for all their support during these years. I also want to thank my supervisors, Professor Alexandra and Professor Manuel for their guidance and patience with me throughout this process.

1 INTRODUCTION

Every day, the TSA screens more than 2 million passengers in the U.S.. These screenings involve security checkpoints for passengers, luggage scans, among others. Passengers may file a claim if they are injured, and if their property is lost or damaged during the screening process. These claims fall under different types and occur not just in different airports, but also in different sites within each airport. Thus, each claim type can be viewed through different random variables, with their own probability distributions that model claim severity and frequency.

The objective of this dissertation is to model the TSA claim counts and severities using copulas. This procedure involves two main steps. The first step is to determine the best parametric distribution for each individual claim frequency and severity random variables. Afterwards, these marginal distributions are incorporated into different copulas, that model the non-linear claim dependencies. This methodology is different than the traditional risk model approach, which assumes independence between claim counts and severities.

Through the use of copulas, it is possible to identify that checked baggage claim counts and severities have upper tail dependence, for example. Furthermore, the copulas are used to simulate more sensitive estimates of risk measures like VaR and TVaR. These copula-based estimates are then contrasted with the historical values, along with the independence case. Thus, this dissertation also presents a practical component in terms of loss reserving using copulas.

The results show modelling claims with copulas can yield higher risk measures than the historical values, in the case of random variables modeled with heavy-tailed distributions. The simulated risk measures also show sensitivity to the type of copula used. All of the data transformation and modelling is performed with open source Python libraries. The general outline of this work is as follows.

Chapter 2 consists of a theoretical overview of copulas, including applications for modelling claim dependencies. Chapter 3 is dedicated to reviewing the TSA database, and the data transformation that allows the modelling of claim dependencies. Chapter 4 presents the univariate data analysis, first for the claim types, and then for the claim sites. Chapter 5 focuses on the copulas used to model the TSA claims, presenting the main results obtained in terms of tail dependence and the simulations performed. Lastly, Chapter 6 presents the main conclusions.

2 LITERATURE REVIEW

There are various approaches in actuarial science for modelling claim counts, severities, and dependencies between different types of risk. In the case of claim counts and claim amounts for a given risk, traditional risk theory assumes independence among these variables (see Valdez (2014)). This approach can involve modelling claim frequency and severity through separate Generalized Linear Models (GLM), as covered in Frees et al. (2016). The issue with the independence assumption is that risks often have common elements, or dependencies, which can jointly affect their distributions. Dorey & Joubert (2005) provides a theoretical overview of how dependencies can be modelled.

Thus, copulas may be used to model the dependence structure between different types of risks, as well as between claim counts and claim severities. From the most basic definition, copulas can be regarded as functions that join multivariate distribution functions to their one-dimensional marginals (Nelsen (2006)). Specifically, a n-dimensional copula C is a multivariate distribution function on the n-dimensional hyper-cube $[0,1]^n$ with uniformly distributed marginals, per Czado (2018). This can be formalized in the following way:

$$C(u_1, u_2, ..., u_n) = Pr(U_1 \leq u_1, U_2 \leq u_2, ..., U_n \leq u_n) \tag{1}$$

where $U_1, U_2, ..., U_n$ are n uniform random variables on (0,1) and their joint distribution function is represented by the copula $C(u_1, u_2, ..., u_n)$.

The flexibility of copulas is that they also allow for applications in data that is not uniformly distributed. According to Sklar (1973), if G is an n-dimensional joint distribution function with 1-dimensional margins $F_1, F_2, ..., F_n$, then there exists a copula function C from the unit cube to the unit interval such that:

$$G(x_1, x_2, ..., x_n) = C(F_1(x_1), F_2(x_2), ..., F_n(x_n)) \tag{2}$$

for all real n-tuples of the random variables $x_1, x_2, ..., x_n$

This is the definition that will be used in this dissertation. The copula contains information about the structure of the dependency in a standardized form across the unit square, while the marginal distributions help to characterize individual risks (Hochrainer-Stigler et al. (2018)). Thus, this approach to modelling claim dependencies essentially consists of two parts.

The first part involves determining the appropriate marginal distribution for each risk

that will be modelled. This includes a parametric estimation of different pre-defined statistical distributions, using MLE for the estimation of the corresponding parameters. A clear example of this procedure can be found in Omari et al. (2018). The authors selected different statistical distributions and fit them to automobile claims. A similar procedure will be shown in Chapter 4.

The second part involves selecting and fitting a copula for these marginal distributions. Manner (2007) provides a concise overview of different copulas belonging to the elliptical and Archimedean families. Since the copulas being considered are parametric families of functions, the process of fitting also involves parameter estimation and goodness-of-fit testing. For two dimensional copulas, Vandenberghe et al. (2010) demonstrates how to estimate parameters through Canonical Maximum Likelihood (CML). Chapter 5 will be dedicated to this analysis.

In the case of copulas with higher dimensions, Oh (2014) includes a general theoretical overview of copula parameter estimation using composite likelihood estimation. An application of composite likelihood to estimate the parameters of a multivariate Gaussian copula can be found in Shi et al. (2016). Meanwhile, Hofert et al. (2012) present the likelihood estimation for the parameters of high dimensional Archimedean copulas. In this dissertation, the parameter estimation for higher dimensional copulas was done by MLE, using the *Copulae* Python library, developed by Bok (2019).

Once the parameters have been estimated, either for the bivariate or the high dimensional case, copulas provide a flexible framework to analyze the dependence structure of different variables. In the simplest case, copulas allow for the exploration of asymptotic tail dependence, as the association of variables may be stronger at the tails and go beyond correlation (Shemyakin et al. (2019)). A summary of asymptotic tail dependence, especially of non-parametric estimation, is presented by Ferreira (2013). Furthermore, Nelsen (1997) includes a comprehensive deduction of asymptotic tail dependence coefficients for Archimedean copulas. Such measures are presented with more detail in Chapter 5.

Other applications of modelling dependencies with copulas include the use of simulations. Aussenegg & Cech (2012) used copulas to simulate returns for different financial assets, obtaining non-parametric estimates of a portfolio's VaR. A similar application can be found in Cheng et al. (2007), where the authors used copulas to simulate risk measures, like VaR, for the Chinese stock market. Brechmann et al. (2013) provides a more general deduction of conditional simulations using elliptical and Archimedean copulas. Chapter 5 is dedicated to presenting the results from similar simulations using the TSA data.

It is important to highlight some alternative approaches within the copula framework, such as the use of vine copulas. This approach utilizes a pair-copula construction, to

decompose multivariate density into products of conditional densities, known as vines (Czado (2018)). While vine copulas are beyond the scope of this dissertation, Haff et al. (2010) provide a thorough introduction into pair-copula constructions. An application of vine copulas can also be found in Hernández et al. (2016), who estimated asymptotic tail dependencies between different macro-economic sectors.

Another alternative approach is the use of copula regressions, which apply the principles of GLM regressions into the copula framework. Ding (2015) presents a comprehensive overview of copula regressions, while Masarotto & Varin (2017) focus on the computational estimation of Gaussian copula regressions. An application of copula regressions for insurance claim counts can be found in Safari-Katesari & Zaroudi (2020).

The different methodological approaches presented above demonstrate the flexibility of the copula framework. This dissertation will focus on parametric copula estimation, using conditional simulations to calculate risk measures for the monthly severity of different claim types and sites. The next chapter describes the TSA database, detailing the methods used to process it.

3 METHODS AND DATA

This chapter describes the TSA database and the methods used to process it. In the first section, a general introduction to the database is provided. The second section details the sample that was used for this dissertation, along with statistics for key variables. The third section deals with the aggregation of daily claim records into monthly series. In the final section, an overview of the detrending process of the aggregated series is given.

3.1 Introduction to the database

The main database compiles different records of TSA claims, published by the Department Department of Homeland Security (2019). This database is comprised of 204,262 claims made during the years 2002-2015. Each claim includes the following 11 variables:

1. **Claim number**: a unique identifier for each claim.
2. **Date received**: the date the claim was received by the TSA.
3. **Incident date**: the date the claim incident took place.
4. **Airport code**: a code for the airport where the incident took place.
5. **Airport name**: the name of the airport where the incident took place.
6. **Airline name**: the name of the passenger's airline.
7. **Claim type**: a categorical label used by the TSA to identify the type of claim.
8. **Claim site**: a categorical label used by the TSA to identify the site within the airport where the claim took place.
9. **Item**: a written description of the passenger's item that was damaged or lost.
10. **Close amount**: the monetary amount the TSA agrees to pay for the claim, should the claim be fully or partially paid.
11. **Disposition**: this is the status of the claim, specifying if it will be paid in full, settled, or denied.

The close amount is the target variable for this dissertation, in order to model claim severity. This variable is closely related to the claim disposition, as claims that were rejected had zeros or were left blank. Figure 1 presents a breakdown of all claims by their

disposition. We can see most claims are denied, approximately 47%. The claims that were approved in full and settled represented around 40% of all claims. The remaining claims, shown in purple, were missing values. Therefore, a subset of the claims approved in full and settled will be taken, as these represent the claims paid by the TSA.

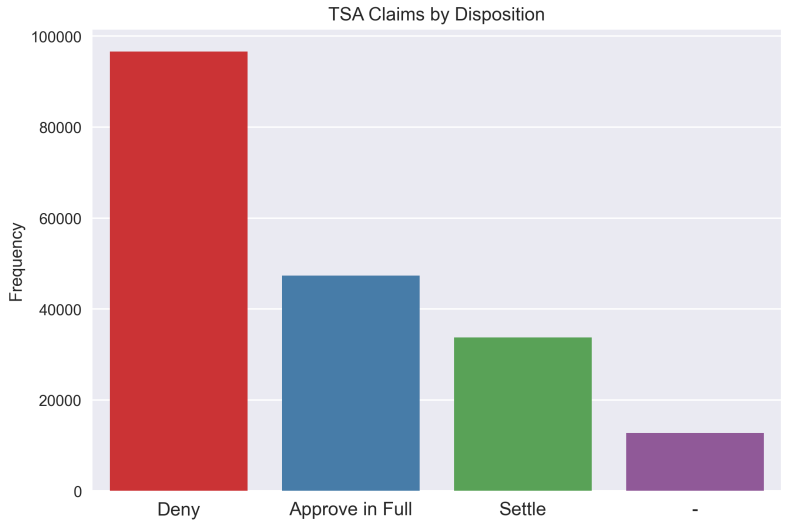


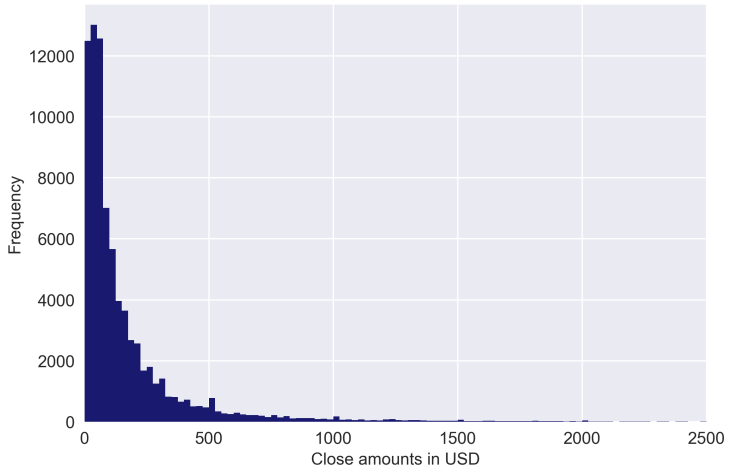
FIGURE 1: TSA claims by disposition

The removal of claims that were denied is justified, since the objective of this dissertation is not to use this data for rate-making purposes, but to study the dependencies of the claims that were paid. Additionally, from a technical standpoint, this is also feasible as this is a very big database. The following subsection details the characteristics of this sample and provides statistics of key variables.

3.2 Details on the main data subset used

The subset generated for the claims paid by the TSA includes a total of 81,107 observations. The first key variable that will be analyzed is the close amount paid by the TSA. This series had to be cleaned, using Python, as it included unwanted characters like semicolons and dollar signs. A histogram is presented in Figure 2 of all the close amounts, along with a table with some general statistics.

It is important to remark that Figure 2 is truncated for aesthetic purposes, as there are very low frequencies for close amounts above $2000. Furthermore, it is also possible to appreciate the high dispersion for this data, as the coefficient of variation is approximately 3.80. The range for the data is also very high. Notice how the mode of the claims is $50, but the maximum amount the TSA paid to a passenger was $125000.



Statistic	Value in USD
Minimum	0
Maximum	125,000
Mean	199
Median	80
Standard Deviation	755
Mode	50

FIGURE 2: Histogram and descriptive statistics of daily close amounts paid

These close amounts can be broken down into claims by type, or by site. Figure 3 illustrates the distribution of claims by type. Most claims fall into the passenger property loss and property damage claim types. In fact, these two claim types represent approximately 94% of all claims for the data subset used.

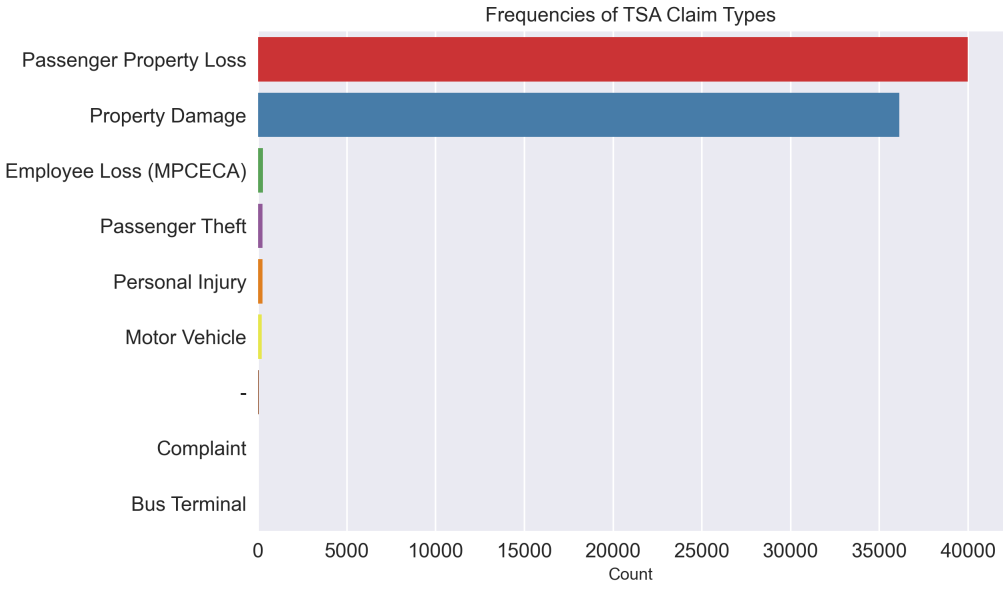


FIGURE 3: TSA claims by type for the data subset used

A similar result occurs when plotting the TSA claims by site in Figure 4, as most claims take place in the checked baggage and security checkpoint claim sites. These sites represent approximately 98% of all claims in the subset.

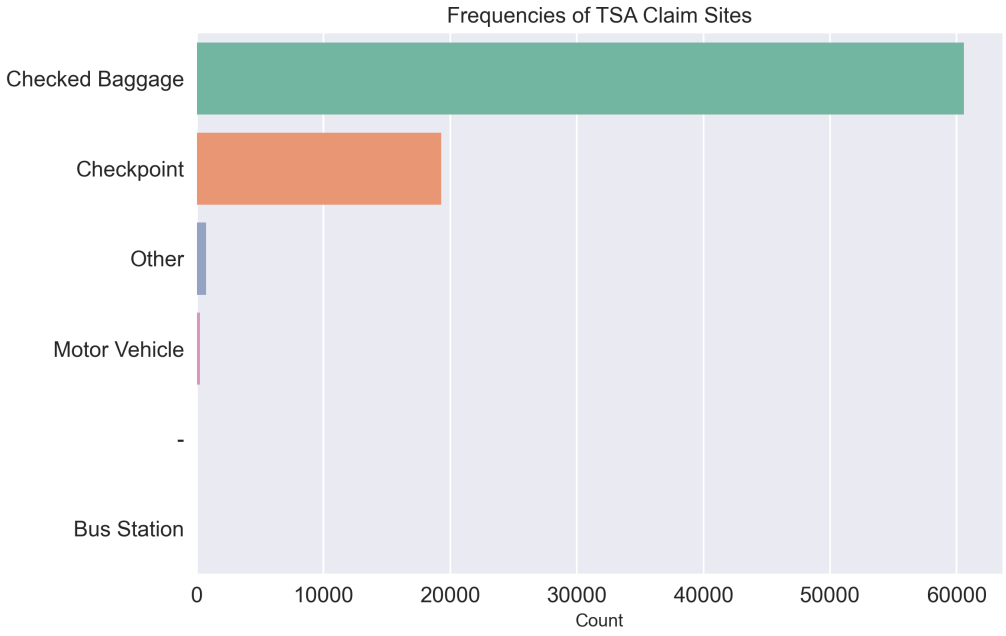


FIGURE 4: TSA claims by site for the data subset used

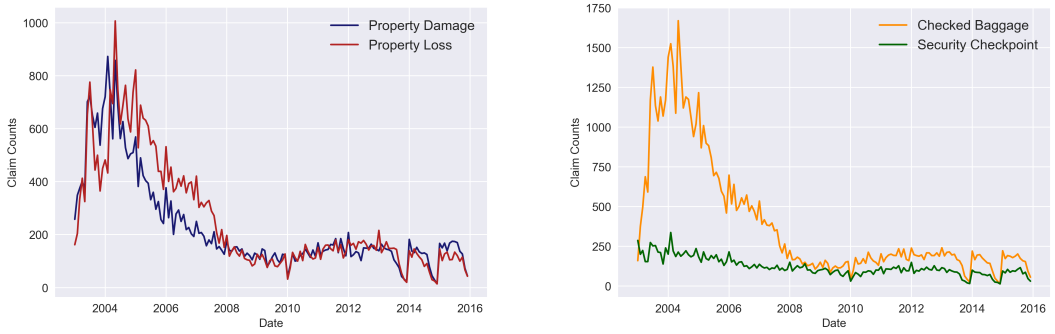
The following section is dedicated to explaining the aggregation procedure for obtaining claim counts and severities, using the close amounts of the two main claim types and sites, shown in Figures 3 and 4, respectively.

3.3 Monthly aggregation of claim counts and severities

The aggregation process consists initially in formatting the received dates on Python, and then aggregating close amounts from daily to monthly records. This is done twice for each target variable. First, the aggregation is done using sums, in order to obtain the claim severities for the defined period. Then, the aggregation is performed using frequencies, to obtain the number of claims for each month. The data allows for monthly records, starting from January, 2003 until December, 2015; a total of 156 months.

This aggregation generates the following variables: Property Damage Claim Counts (PDX), Property Damage Claim Severity (PDS), Property Loss Claim Counts (PLX), Property Loss Claim Severity (PLS), Checked Baggage Claim Counts (CBX), Checked Baggage Claim Severity (CBS), Security Checkpoint Claim Counts (CPX), and Security Checkpoint Claim Severity (CPS). The plots in Figure 5 show the counts for the claims by type, followed by the claims by site. We can observe that property loss and property

damage claim counts are closely related. This is not the case for the claim counts by site, as the checked baggage claims are clearly higher until around 2008.



(a) Counts by type (b) Counts by site

FIGURE 5: Monthly TSA claim counts from 2003-2015

Table I includes some descriptive statistics for the aggregated claim counts using the abbreviation presented earlier. It confirms the clear differences in scale of the checked baggage and security checkpoint claims. The table also shows the similarities between the two claims types, as illustrated in Figure 5(a). In the case of the aggregated claim counts, none of the series have a coefficient of variation above 1.

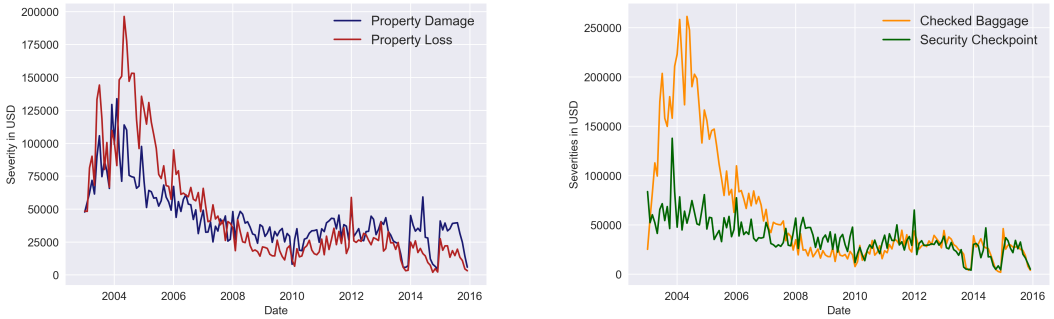
Statistic	PDX	PLX	CBX	CPX
Minimum	15.0	14.0	14.0	14.0
Maximum	873.0	1007	1670.0	337.0
Mean	229.7	255.4	388.1	120.9
Standard Deviation	182.1	208.6	377.3	56.0
Coefficient of Variation	0.79	0.82	0.97	0.46
Skewness	1.70	1.29	1.55	0.98
Kurtosis	2.17	0.82	1.39	1.38

TABLE I: Descriptive statistics for aggregated TSA claim counts

One potential issue with these aggregated series, is the clear trend component that is observed for all of the claim counts. This may be due to changes in how the claims were registered, as there were inconsistencies in the database from 2009 onward, in comparison to the previous years. This trend component is addressed in the final section of this chapter.

Figure 6 shows the aggregate severities obtained for the claims by type, followed by

the claims by site. All of the series are given in U.S. dollars. There is less of a difference when it comes to the severities of the two claim sites. The graph also shows a clear negative trend component for all series.



(a) Severities by type (b) Severities by site

FIGURE 6: Monthly TSA claim severities in USD from 2003-2015

Table II highlights some descriptive statistics for all of the aggregate claim severities and illustrates the real cost incurred by the TSA during this period. For example, in the case of the property damage claims, the TSA paid an average $43,393.38 per month to settle property damage claims. This sum is $520,720.56 per year.

Statistic	PDS	PLS	CBS	CPS
Minimum	4465.96	2033.88	1964.96	4121.36
Maximum	133835.07	196328.29	261424.43	137850.99
Mean	43393.38	45801.33	59298.65	37450.70
Standard Deviation	23012.96	41028.61	59463.31	18528.23
Coefficient of Variation	0.53	0.90	1.00	0.49
Skewness	1.40	1.50	1.70	1.35
Kurtosis	2.80	1.64	2.05	4.96

TABLE II: Descriptive statistics for aggregated TSA claim severities

It is important to note that the claims by type will be separated from those by site when modelling. This is done to prevent double counting, especially when it comes to calculating risk measures. The next section summarizes the detrending process used for all the series.

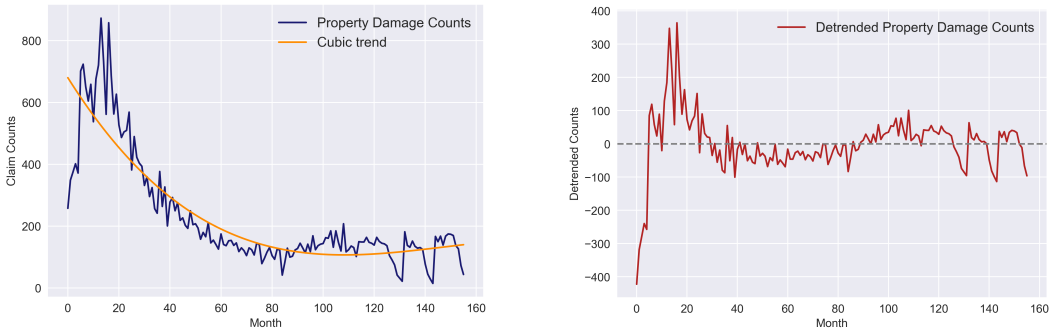
3.4 Overview of claim detrending process

As discussed in the previous section, all of the claim counts and severities exhibited a negative trend component. This trend was corrected using a polynomial detrending method. The method consists in setting up a linear regression, and uses exponentiated time indexes (t) to estimate the values of an indexed variable, $\tilde{y}$. For this detrending process, a third order exponentiated time index is used in the regression:

$$\tilde{y} = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \epsilon \tag{3}$$

where each $\alpha_{j=0,1,2,3}$ represents a coefficient and ϵ is an error term.

The detrended series are obtained by subtracting these estimates from the observed values, i.e. $y - \tilde{y}$. In practice, this was done on Python using the *scikit-learn* library, developed by Pedregosa et al. (2011). A visual example is shown in Figure 7 of how the property damage counts are detrended. The yellow line, shown in Figure 7(a), is the trend obtained through equation (3). The detrended counts, shown in Figure 7(b), represent the difference between the observed values and the trend.



(a) Detrending process

(b) Detrended property damage counts

FIGURE 7: Illustration of detrending process for property damage counts

This explains why there are negative values in the detrended series, a drawback of this method. It is less intuitive to estimate risk measures for detrended severities, for example, as these do not represent monetary values, just trend deviations. Therefore, results for both the raw and detrended claims will be presented from now on. Additionally, for a statistical summary of all the detrended series, see Appendix A.

4 UNIVARIATE DATA ANALYSIS

This chapter is dedicated to univariate data analysis, particularly the process used to determine the marginal distributions for the claim counts and severities. First some preliminaries to the main data analysis are presented.

4.1 Preliminaries

As highlighted in the literature review, this dissertation looks to fit predefined marginal distributions into different copulas. To determine the best marginal for each variable, the process of fitting distributions was facilitated through the use of *SciPy*, a Python library developed by Virtanen et al. (2020). The fit process with MLE is detailed in Omari et al. (2018) as follows.

Suppose $X_1, X_2, .., X_n$ is a random sample of independent and identically distributed observations drawn from an unknown population. Let $X = x$ denote a realization of a random variable or vector $\mathbf{X}$, with probability mass or density function $f(x; \theta)$; where θ is a vector or a scalar of unknown parameters which will be estimated. The likelihood function $L(\theta)$, is the probability mass or density function of the observed data $\mathbf{x}$, expressed as a function of the unknown parameter(s) θ.

Given that $X_1, X_2, ..., X_n$ have a joint density function $f(X_1, X_2, ..., X_n|\theta)$ for every observed sample of independent observations $\{x_{i=1,2,...,n}\}$, the likelihood function is defined by:

$$L(\theta) = L(\theta|x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{N} f(x_i|\theta) \tag{4}$$

The maximum likelihood estimate $\tilde{\theta}$ of parameter(s) θ is obtained through maximizing the likelihood function $L(\theta)$:

$$\tilde{\theta} = arg\max_{\theta} L(\theta) \tag{5}$$

Since the logarithm of the likelihood function is a monotonically non-decreasing function of $\mathbf{X}$, maximizing $L(\theta)$ is equivalent to maximizing the log of the likelihood function $l_x(\theta)$, given by:

$$l_x(\theta) = \log L(\theta) = \log \prod_{i=1}^{N} f(x_i|\theta) = \sum_{i=1}^{N} \log f(x_i|\theta) \tag{6}$$

Once the parameters are estimated for different distributions, these distributions are compared using the χ^2 test for the goodness-of-fit (see page 475 of Klugman et al. (2008)). Under the null hypothesis of the test, the data follows the specified distribution that is being tested. To compute the χ^2 test, the data is grouped into k bins or classes. The test statistic is defined as:

$$\chi^2 = \sum_{j=1}^{k} \frac{(E_j - O_j)^2}{E_j} \overset{a}{\sim} \chi^2_{(k-1-v)} \tag{7}$$

where $j = 1, 2, ..., k$ represent predetermined bins, O_j is the observed frequency in the data for bin j, and E_j is the expected frequency for bin j, given by:

$$E_j = N(F(X_U) - F(X_L)) \tag{8}$$

where F is the cumulative distribution function for the distribution tested, X_U is the upper bound for bin j, X_L is the lower bound for bin j, and N is the sample size.

Under the null hypothesis, the test statistic is asymptotically distributed as a χ^2 distribution. Thus, the critical value for the test comes from upper tail of the χ^2 distribution with $k - 1 - v$ degrees of freedom, where k is the total number of bins, and v is the total number of parameters for the distribution being tested. This test is sensitive to the choice of the bins selected. For this dissertation, 20 bins were selected using a quantile based approach; see Appendix 2 for more details on this procedure.

4.2 Demonstration of marginal selection process for property damage claims

The objective of the marginal selection process is to find the distributions that best model different series. To illustrate the selection process, an example will be done with the raw property damage claims, depicted in Figure 8.

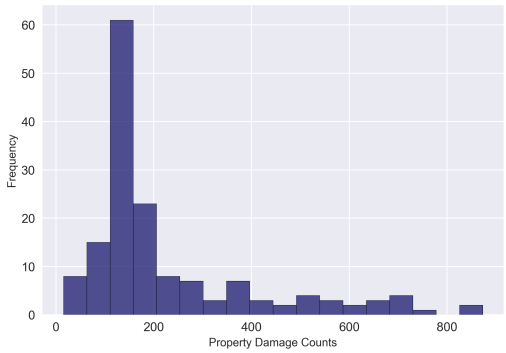


FIGURE 8: Histogram of property damage counts

A Python loop is used to test 8 different distributions, including the Poisson, Negative Binomial, Geometric, and Binomial. This loop estimates parameters, using a *SciPy* MLE optimizer proposed by Haberland (2020) specific to discrete distributions. Then, the χ^2 statistic from equation (7) is calculated for a distribution with these estimated parameters, along with its p-value[1].

Table III summarizes the results for the property damage counts, including notation for the Probability Mass Function (PMF) of four distributions tested, and the estimated parameters. The four discrete distributions are poor fits, as all they have p-values of zero. For the purposes of this work, the distribution with the lowest χ^2 statistic will be used, although future research could include exploring mixture distributions to find better fits. This was not expanded upon as the purpose of this dissertation is to model dependencies and simulate risk measures which focus on the claim severities.

Distribution	SciPy PMF notation	Parameter(s)	χ^2 statistic	p-value
Poisson	$f(k) = \exp(-\mu)\frac{\mu^k}{k!}$	$\mu = 526.31$	2015.36	0.00
Geometric	$f(k) = (1-p)^{k-1}p$	$p = 0.0003$	1429.73	0.00
Binomial	$f(k) = \binom{n}{k}p^k(1-p)^{n-k}$	$n = 323220$; $p = 0.0007$	1344.43	0.00
Negative Binomial	$f(k) = \binom{k+n-1}{n-1}p^n(1-p)^k$	$n = 2.14$; $p = 0.01$	110.52	0.00

TABLE III: Parametric estimation of distributions for raw property damage counts

Figure 9 shows a histogram for the property damage counts in light blue. The bins or intervals from this histogram, are then used to obtain the probabilities for the distributions from Table III. The Negative Binomial, shown in blue, seems to be a relatively better fit to the observed, than the other distributions. For example, the Poisson distribution probabilities center around an interval of 400-600 claims, which is a poor fit to the observed. The Binomial distribution, meanwhile, has a similar behavior as its probabilities center around an interval of 150-250 claims, showing it is also a very poor fit to the observed.

[1]The complete code used for this process and for the dissertation is available in this repository: `https://github.com/rcarcacheflores/ISEG-Dissertation`

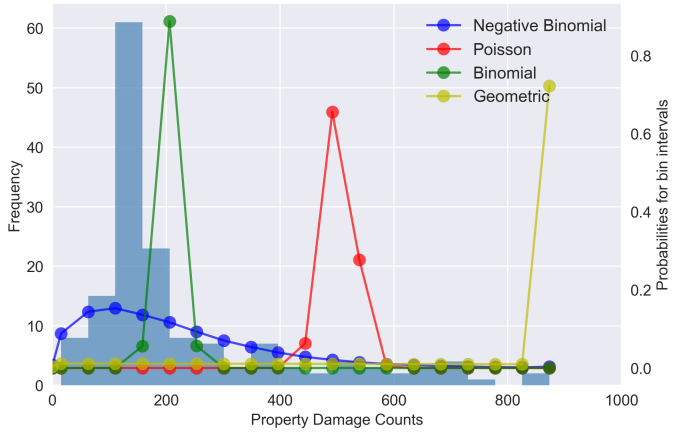


FIGURE 9: Distribution comparison of property damage counts

A similar procedure is used for the severity series, making use of the built-in *SciPy* optimizer to test 27 continuous distributions. The results are presented in Table IV.[2] The table shows the four continuous distributions with the lowest χ^2 statistic, but only the Log-Laplace and Johnson SU distributions are statistically significant at a 1% significance level:

Distribution	Parameter(s)	χ^2 statistic	p-value
Log-Laplace	$c = 2.48$; $S = 37891.93$	23.73	0.095
Johnson SU	$a = -2.88$; $b = 2.44$ $S = 27045.30$	26.68	0.031
Generalized Logistic	$c = 8.71$ $S = 16070.20$	34.66	0.004
Lognormal	$s = 0.59$; $S = 37495.36$	36.82	0.003

TABLE IV: Parametric estimation of distributions for raw property damage severities

Figure 10 shows the densities for severity values ranging from 0 to 200000, for each distribution presented in Table IV, and compares them with the original. The Log-Laplace distribution, shown in red, mirrors the peak density of the property damage severity.

[2]See Appendix 3 for an exhaustive presentation of the distributions used with SciPy notation

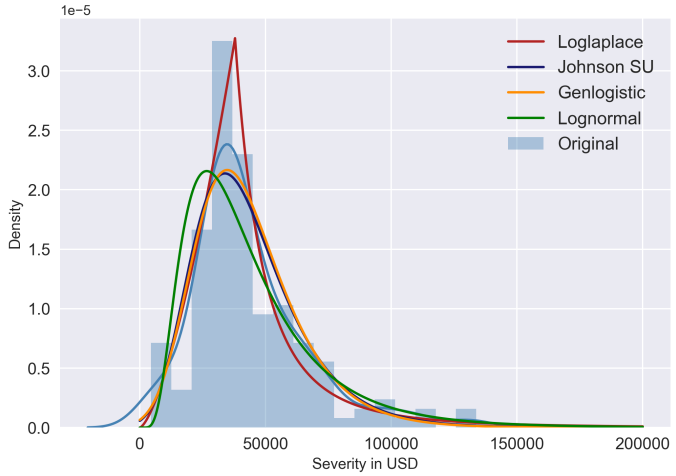


FIGURE 10: Distribution comparison of property damage severities

This example shows the general procedure used to perform the univariate data analysis. The next section summarizes these findings for all the claims by type.

4.3 Results for the univariate analysis of claims by type

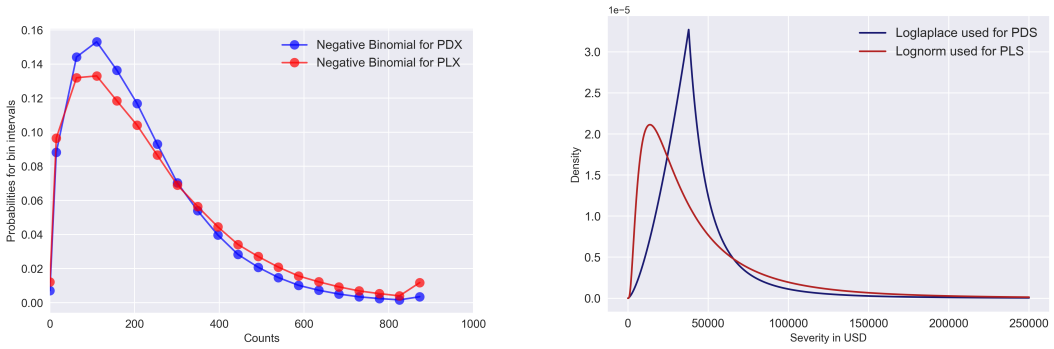
After performing the selection process, the best fit marginal distributions for the raw claim types are presented in Table V. This table shows only the distributions found for the severities were good fits at a 1% significance level. Both claim count series are best modelled with Negative Binomial distributions, at least in comparison to the discrete distributions tested.

Variable	Best fit	Parameter(s)	χ^2 statistic	p-value
PDX	Negative Binomial	$n = 2.14$; $p = 0.01$	110.52	0.00
PLX	Negative Binomial	$n = 1.76$; $p = 0.007$	97.06	0.00
PDS	Log-Laplace	$c = 2.48$; $S = 37891.93$	23.73	0.11
PLS	Lognormal	$s = 0.91$; $S = 31402.12$	29.39	0.02

TABLE V: Parametric estimation of distributions for raw claims by type

Figure 11(b) shows the two distributions fit for the claim severities, while the discrete distributions used for the counts are shown in Figure 11(a). The Log-Laplace distribution,

used to model property damage severity, has a heavier tail than the Lognormal distribution used for property loss severity.



(a) Distributions for raw counts by type

(b) Distributions for raw severity by type

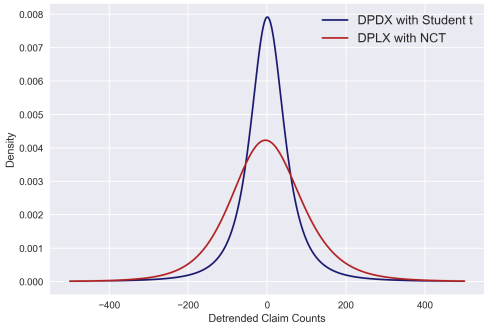
FIGURE 11: Best fit distributions for raw claims by type

These results will now be contrasted by fitting distributions for the detrended claims by type. The only difference in the Python loop programmed for the detrended claims, is the inclusion of a location parameter for all the continuous distributions tested. This is done to allow for distributions with supports containing negative values. As before, the selection criteria is the χ^2 test with $(20 - 1 - v)$ degrees of freedom.

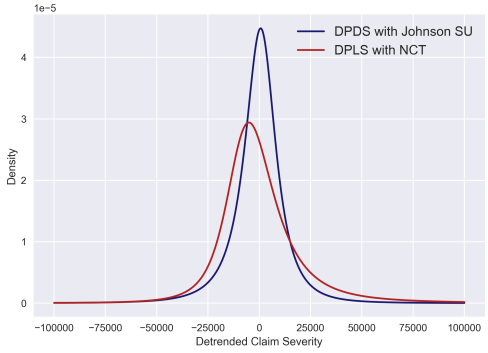
Table VI summarizes the best fits for the detrended claims by type. The shapes of these estimated distributions are visualized in Figure 12. Overall, detrending the claim types yields distributions with good fits, according to the χ^2 test. The distributions used to model detrended property loss counts and severity have heavier tails, compared to those distributions used for property damage.

Variable	Best fit	Parameter(s)	χ^2 statistic	p-value
DPDX	Student t	$df = 2.33; L = 0.60;$ $S = 45.41$	17.33	0.36
DPLX	NCT	$df = 4.96; nc = 0.16;$ $L = -16.63; S = 89.65$	19.64	0.19
DPDS	Johnson SU	$a = 0.11; b = 1.05;$ $L = 1211.59; S = 9305.03$	13.45	0.56
DPLS	NCT	$df = 2.36; nc = 0.66;$ $L = 11186.73; S = 11886.15$	17.11	0.31

TABLE VI: Parametric estimation of distributions for detrended claims by type



(a) Distributions for detrended counts by type



(b) Distributions for detrended severity by type

FIGURE 12: Best fit distributions for detrended claims by type

As for the best fit distributions, the Student t and the Noncentral Student t distribution (NCT), were studied by Choi & Yoon (2020) for their use in modelling equity linked securities, which also have positive and negative values.

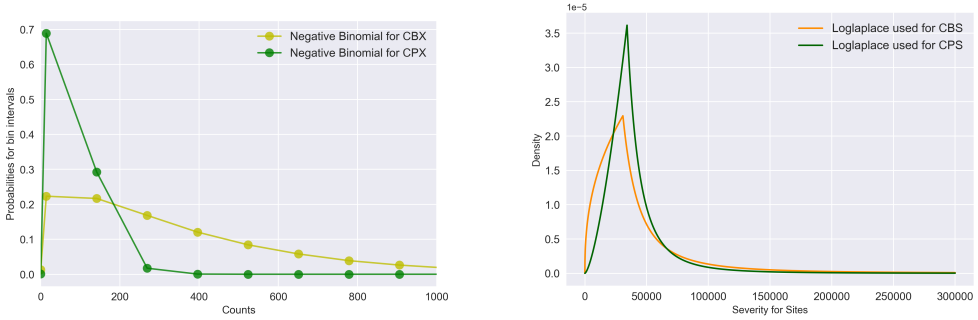
4.4 Results for the univariate analysis of claims by site

The same procedure from Section 4.3 is applied to the claims by site. Table VII summarizes the best fits found for the raw claim sites. The fits are poor according to the χ^2 test, except for the Log-Laplace distribution used to model the Security Checkpoint Severity.

Variable	Best fit	Parameter(s)	χ^2 statistic	p-value
CBX	Negative Binomial	$n = 1.40$; $p = 0.004$	147.49	0.00
CPX	Negative Binomial	$n = 4.68$; $p = 0.037$	47.83	0.00
CBS	Log-Laplace	$c = 1.41$; $S = 30811.30$	47.56	0.00
CPS	Log-Laplace	$c = 2.47$; $S = 34138.79$	21.37	0.16

TABLE VII: Parametric estimation of distributions for raw claims by site

Figure 13 shows the Log-Laplace distribution, used to model the checked baggage severity, has a heavier tail than the Log-Laplace distribution used for the security checkpoint severity.



(a) Distributions fit for raw counts (b) Distributions fit for raw severity

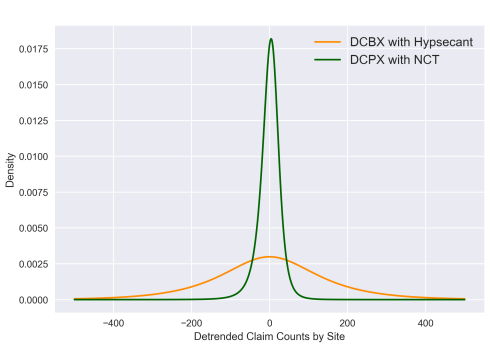
FIGURE 13: Best fit distributions for raw claims by site

The last step of this chapter is to verify these results for the detrended claims by site, using the methods highlighted in Section 4.3. Table VIII summarizes the best distributions fit for the detrended claims by site. All of the distributions are good fits according to the χ^2 test at a 1% significance level.

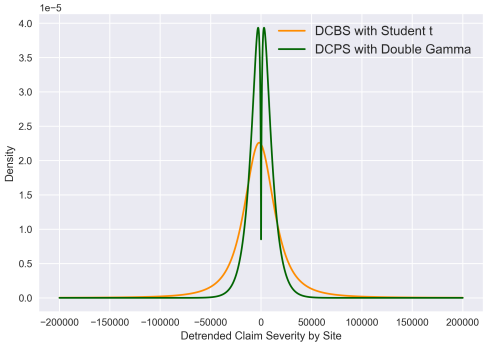
Variable	Best fit	Parameter(s)	χ^2 statistic	p-value
DCBX	Hyperbolic Secant	$L = 1.52;$ $S = 106.49$	32.21	0.01
DCPX	NCT	$df = 3.89; nc = -0.51;$ $L = 12.84; S = 20.32$	11.44	0.72
DCBS	Student t	$df = 2.22; L = -1857.89;$ $S = 15819.19$	16.82	0.42
DCPS	Double Gamma	$a = 2.21; L = -1857.88;$ $S = 15819.19$	16.82	0.39

TABLE VIII: Parametric estimation of distributions for detrended claims by site

Meanwhile, Figure 14 illustrates the considerable differences between these distributions. Notice how the checked baggage counts and severity distributions have heavy tails, in comparison to those distributions of the security checkpoints. This makes sense as there are also clear differences in the observed values of these series (recall Figure 6(b)).



(a) Distributions for detrended counts by site



(b) Distributions for detrended severity by site

FIGURE 14: Best fit distributions for detrended claims by site

Overall, these results suggest detrending the series facilitates finding good distribution fits, for both the claim types and sites. This insight on the marginal distributions of all these series is carried on to the next chapter. The marginals obtained are used in bivariate and higher dimensional copulas, to model claim dependencies by type and by site.

5 MODELS FOR CLAIM DEPENDENCE

This chapter presents the copulas used to model the claim dependencies for the TSA database. The first section is dedicated to preliminary definitions, mainly related to the formulation of asymptotic tail dependence coefficients. The second and third sections summarize the results obtained using bivariate copulas for the claims by type and by site, respectively. The main simulation results for the multivariate copulas ($n > 2$ dimensions) for the claims by type are included in section 4, while those for the claims by site are shown in the final section.

5.1 Preliminary definitions

As highlighted in Chapter 2, one of the most immediate uses of copulas is the estimation of asymptotic tail dependence coefficients. Upper tail dependence relates to dependence in the upper-right-quadrant of a bivariate distribution, while lower tail dependence takes place in the lower-left-quadrant. Frahm et al. (2005) provides the following general definition for both measures.

Let (X, Y) be a random pair with joint cumulative distribution function F, along with marginals $G(X)$ and $H(Y)$. The UTDC, provided this limit exists, is given by:

$$\lambda_U = \lim_{t \to 1^-} P\{G(X) > t | H(Y) > t\} \tag{9}$$

Meanwhile, Lower Tail Dependence Coefficient (LTDC) follows a similar deduction:

$$\lambda_L = \lim_{t \to 0^+} P\{G(X) \leq t | H(Y) \leq t\} \tag{10}$$

These coefficients estimate the probability that one margin exceeds a high or low threshold under the condition that the other margin exceeds a high or low threshold. Two variables are said to be upper tail dependent if $\lambda_U > 0$ and upper tail independent if $\lambda_U = 0$. The same logic applies to lower tail dependence. As we are modelling claims, more emphasis will be placed on estimating upper tail dependence coefficients.

These definitions can also be expressed in copula notation. If the copula C defines the joint distribution of these two variables such that $F(x, y) = C(G(X), H(y))$; then the upper tail dependence is given by:

$$\lambda_U = \lim_{t \to 1^-} \frac{1 - 2t + C(t, t)}{1 - t} \tag{11}$$

These coefficients can be estimated empirically through the use of the empirical copula [3]. This process requires the selection of a threshold or quantile ($\frac{k}{N}$), which can be optimized as shown by Frahm et al. (2005), or be selected *a priori*. Computationally, the estimation of both the optimal and empirical tail coefficients is available in the *Pycop* library, developed by Nicolas (2021).

Furthermore, parametric estimates of the tail dependence coefficients can also be made, depending on the copula used to model the joint distribution. For example, in the case of the Archimedean copulas used in this chapter, the Clayton copula only has a LTDC, while the Gumbel copula only has an UTDC. The next section presents the results for the bivariate copulas used to model the TSA claims by type.

5.2 Bivariate copulas for claims by type

The first test performed is estimating the empirical UTDC for different pairs of raw claim types. For example, the first pair includes property damage and property loss counts. This pair will be denoted as $\tilde{C}(F_e(PDX), F_e(PLX))$, where $\tilde{C}$ is the empirical copula, and F_e are the empirical distributions of each random variable. For these claim types, a threshold of $\frac{k}{N} = 0.90$ will be used to find the upper tail dependence coefficients.

This threshold is selected as there are relatively few observations for each variable, and choosing a higher threshold may yield unreliable UTDC, per Frahm et al. (2005). For example, Figure 15 plots the UTDC for $\tilde{C}(F_e(PDX), F_e(PLX))$ using different thresholds ($\frac{k}{N}$). Notice how from around the 92% quantile, the UTDC becomes unstable, which may be due to a low number of observations in the data for these quantiles.

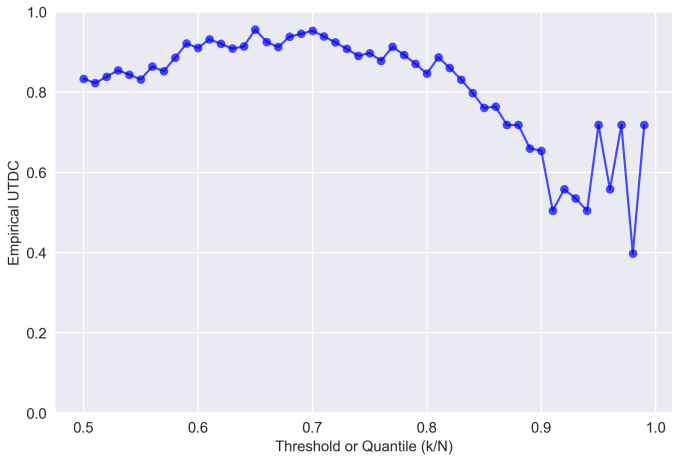


FIGURE 15: Empirical UTDC for $\tilde{C}(F_e(PDX), F_e(PLX))$

[3] See Appendix 4 for a complete formulation of the copulas used and their tail dependence coefficients

Table IX summarizes the UTDC obtained with $\frac{k}{N} = 0.90$ for these claim type pairs. This table indicates the property damage counts and severities have a high UTDC of 0.91. In the case of the property loss counts and severities, the two variables have strong upper tail dependence with an UTDC of 0.97.

Empirical Copulas	Empirical UTDC
$\tilde{C}(F_e(PDX), F_e(PLX))$	0.6538
$\tilde{C}(F_e(PDS), F_e(PLS))$	0.7179
$\tilde{C}(F_e(PDX), F_e(PDS))$	0.9102
$\tilde{C}(F_e(PLX), F_e(PLS))$	0.9743

TABLE IX: Empirical upper tail coefficients for each copula pair of raw claim types

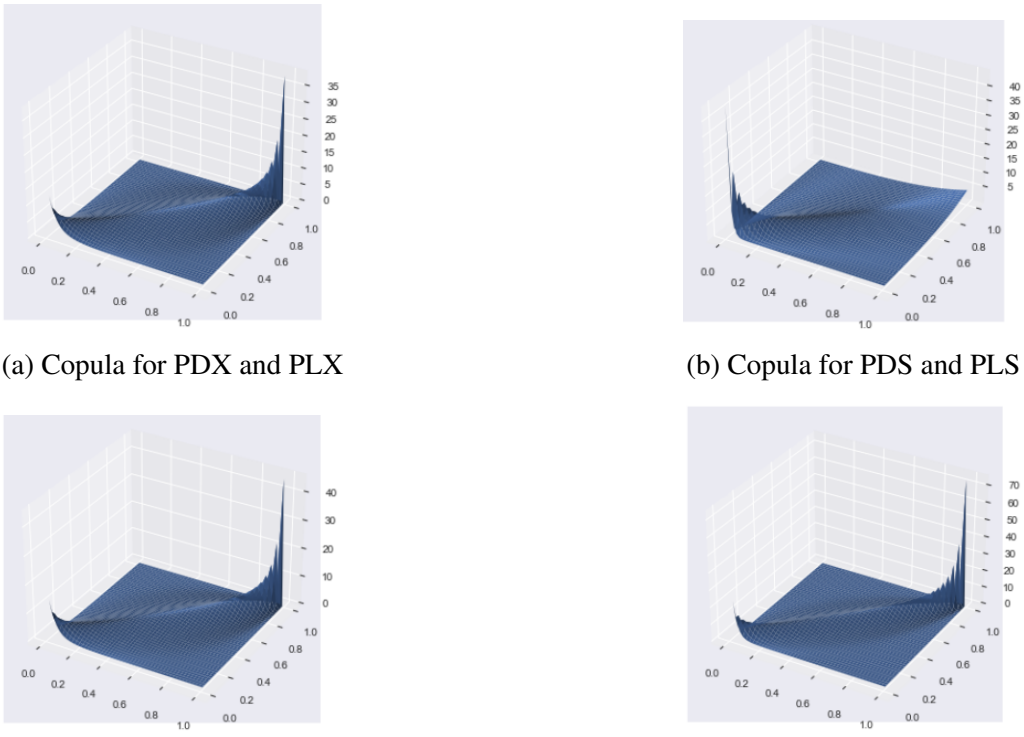
The next step is to compare these results by finding the best parametric copulas C_p, to model the marginal distributions found in Table V of Chapter 4, denoted F_p, for each random variable. The selection process of the best fitted copula is done using the *copulas* library, developed by Alvarez et al. (2018). This library compares the fit of copulas from the Archimedean family, while also testing against the independence case.

The results obtained for these variable pairs are summarized in Table X. All of the pairs are best modelled with Gumbel's copula, which has upper tail dependence, confirming the results from Table IX. The only diverging result is the copula for property damage and property loss severities, best modelled by Clayton's copula, which only has lower tail dependence. In future research, copulas with both upper and lower tail dependence could be tested to verify if this may explain this result.

Parametric Copulas	Best Fit	Parameter (θ)	LTDC	UTDC
$C_p(F_p(PDX), F_p(PLX))$	Gumbel	3.5358	0	0.7834
$C_p(F_p(PDS), F_p(PLS))$	Clayton	3.4264	0.8168	0
$C_p(F_p(PDX), F_p(PDS))$	Gumbel	4.0587	0	0.8137
$C_p(F_p(PLX), F_p(PLS))$	Gumbel	6.3517	0	0.8847

TABLE X: Summary of bivariate copulas modelled with marginals for raw claim types

Figure 16 visualizes the densities in three dimensions for the copulas from Table X. This plot helps to visualize the dependence structure for these Archimedean copulas.



(a) Copula for PDX and PLX

(b) Copula for PDS and PLS

(c) Copula for PDX and PDS

(d) Copula for PLX and PLS

FIGURE 16: Densities for bivariate copulas modelling raw claim types

These results will now be contrasted for the detrended claims by type. Table XI shows the UTDC at the 90% quantile for the detrended claims. The results for the detrended claims also show strong tail dependence for the property loss counts and severities. In the case of detrended property damage counts and severities, the UTDC is lower than the coefficient found in Table IX.

Empirical Copulas	Empirical UTDC
$\tilde{C}(F_e(DPDX), F_e(DPLX))$	0.5897
$\tilde{C}(F_e(DPDS), F_e(DPLS))$	0.4615
$\tilde{C}(F_e(DPDX), F_e(DPDS))$	0.6538
$\tilde{C}(F_e(DPLX), F_e(DPLS))$	0.8461

TABLE XI: Empirical upper tail coefficients for each pair of detrended claim types

Table XII summarizes the best fit copulas used to model these detrended claim types, after incorporating the marginals from Table VI. These results differ from those found in Table XI, as the detrended property loss counts and severities are best modelled with Clayton's copula, which has lower tail dependence.

Parametric Copulas	Best Fit	Parameter (θ)	LTDC	UTDC
$C_p(F_p(DPDX), F_p(DPLX))$	Clayton	2.0748	0.7159	0
$C_p(F_p(DPDS), F_p(DPLS))$	Gumbel	1.6089	0	0.4615
$C_p(F_p(DPDX), F_p(DPDS))$	Gumbel	2.4345	0	0.6706
$C_p(F_p(DPLX), F_p(DPLS))$	Clayton	6.5805	0.9000	0

TABLE XII: Summary of bivariate copulas modelled for detrended claim types

These findings are explored with more detail in Figure 17. This plot visualizes the dependence structure of the first Clayton's copula, used to model the two detrended claim counts by type ($C_p(F_p(DPDX), F_p(DPLX))$). The observed values from the parametric marginals of these detrended counts, are shown in red. This plot illustrates the lower tail dependence of the marginals is effectively modelled with Clayton's copula.

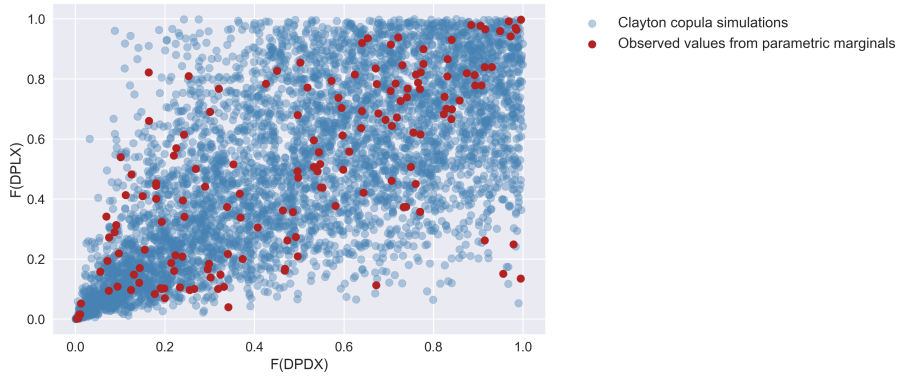


FIGURE 17: Simulated and observed values for $C_p(F_p(DPDX), F_p(DPLX))$

Although the results changed for the detrended claim types, in comparison to the raw claims, the most important finding is that all the variable pairs show some form of tail dependence. This confirms the importance studying non-linear dependence structures between different types of risks such as these ones.

5.3 Bivariate copulas for claims by site

The first step is to calculate the empirical UTDC. This was estimated for a threshold of $\frac{k}{N} = 0.89$, as there were fewer observations for higher quantiles in the case of the claim sites. Table XIII summarizes these results. This table shows the checked baggage counts and severities have the strongest tail dependence with a probability of 0.95. The rest of the variable pairs have similar coefficients with probabilities around 0.60-0.66.

Empirical Copulas	Empirical UTDC
$\tilde{C}(F_e(CBX), F_e(CPX))$	0.6596
$\tilde{C}(F_e(CBS), F_e(CPS))$	0.6013
$\tilde{C}(F_e(CBX), F_e(CBS))$	0.9510
$\tilde{C}(F_e(CPX), F_e(CPS))$	0.6596

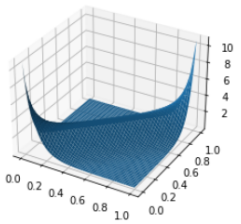
TABLE XIII: Empirical tail coefficients for each copula pair of raw claims by site

The next step is to find the best fitting parametric copulas for these pairs, summarized in Table XIV. This table confirms the checked baggage counts and severity are best modelled with Gumbel's copula, which has an UTDC, showcased in Table XIII.

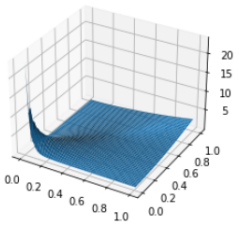
Parametric Copulas	Best Fit	Parameter (θ)	LTDC	UTDC
$C_p(F_p(CBX), F_p(CPX))$	Frank	10.7825	0	0
$C_p(F_p(CBS), F_p(CPS))$	Clayton	1.8235	0.6837	0
$C_p(F_p(CBX), F_p(CBS))$	Gumbel	7.6771	0	0.9055
$C_p(F_p(CPX), F_p(CPS))$	Clayton	3.9156	0.8377	0

TABLE XIV: Summary of bivariate copulas modelled for raw claim sites

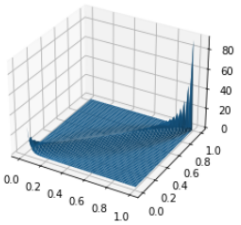
Figure 18 visualizes these dependencies by plotting the densities of the copulas shown in Table XIII, and 18(c) illustrates the UTDC for Gumbel's copula.



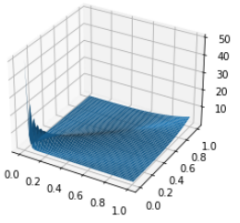
(a) Copula for CBX and CPX



(b) Copula for CBS and CPS



(c) Copula for CBX and CBS



(d) Copula for CPX and CPS

FIGURE 18: Densities for bivariate copulas modelling raw claim sites

The last step of this section is to contrast these results with the detrended claims by site, recalling the parametric distributions for these variables are presented in Table VIII. In this case, only the parametric tail dependence coefficients will be shown, along with the copulas with the best fits.

Table XV summarizes the results obtained for the detrended claim sites. This table shows that the variable pairs for the detrended claim sites all have some form of tail dependence. This table also confirms the existence of upper tail dependence between the checked baggage counts and severity, a result shown in Table XIV for the raw claim sites.

Parametric Copulas	Best Fit	Parameter (θ)	LTDC	UTDC
$C_p(F_p(DCBX), F_p(DCPX))$	Clayton	1.1649	0.5515	0
$C_p(F_p(DCBS), F_p(DCPS))$	Clayton	0.4155	0.1886	0
$C_p(F_p(DCBX), F_p(DCBS))$	Gumbel	5.6024	0	0.8683
$C_p(F_p(DCPX), F_p(DCPS))$	Clayton	1.6393	0.6552	0

TABLE XV: Summary of bivariate copulas modelled for detrended claim sites

Figure 19 illustrates in more detail the dependence structure for the Gumbel's copula used to model the detrended checked baggage claim counts and severity. The observed values obtained from the parametric marginal distributions of each series are shown in red. In this case, the fit does not seem as adequate as the one shown in Figure 17.

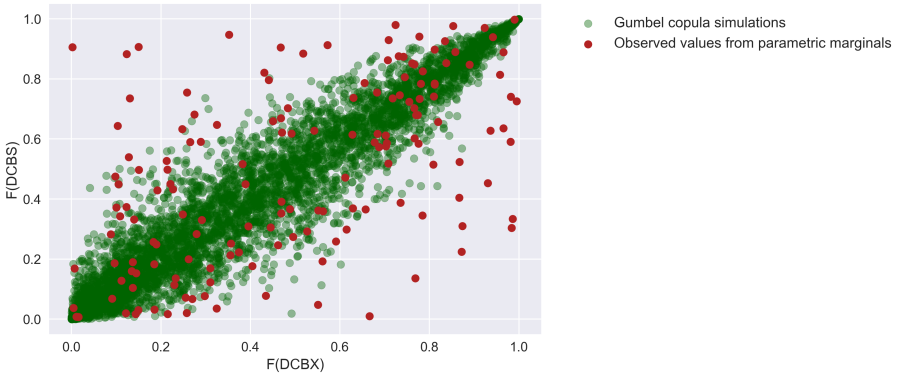


FIGURE 19: Simulated and observed values for $C_p(F_p(DCBX), F_p(DCBS))$

In the next two sections, the final analysis is performed using multivariate copulas, in order to simulate risk measures for the TSA claims.

5.4 Simulating risk measures for claim types with copulas

This section is dedicated to using multivariate copulas to simulate risk measures for the claim types. This analysis is performed for the raw claims, as it is more intuitive to work with these severities, in comparison to the detrended claims. The first step is fitting the copulas to the raw claims and finding their parameters with MLE using the *copulae* library, developed by Bok (2019).

Each copula has four dimensions, corresponding to the marginals of the raw claim counts and severities. Using the notation from the previous section, this can be written as: $C_p(F_p(PDX), F_p(PLX), F_p(PDS), F_p(PLS))$. Thus, the copula models the dependence structure of these four marginals.

According to Hasebe (2013), if the marginal distributions are fixed, and the numbers of estimated copula parameters are the same, choosing the copula with the smallest information criterion is equivalent to choosing the copula with the largest log-likelihood value. This is a useful way to compare the fit of the Archimedean copulas, as they only have one parameter and the same marginals.

Table XVI summarizes the estimated parameters and log-likelihood values for the copulas fit. This table shows Gumbel's copula is a better fit than Clayton's copula for the raw claim types. Additionally, the two elliptical copulas have higher likelihoods than the Archimedean copulas, and although they are not directly comparable just based on this measure, they can be considered better fits. All of these copulas will be used to contrast how different dependence structures can affect two risk measures.

Type of copula	Parameter(s)	Log-likelihood
Gaussian	$\Sigma = \begin{pmatrix} 1 & 0.92 & 0.91 & 0.91 \\ 0.92 & 1 & 0.84 & 0.95 \\ 0.91 & 0.84 & 1 & 0.88 \\ 0.91 & 0.95 & 0.88 & 1 \end{pmatrix}$	484.23
Student t	$\Sigma = \begin{pmatrix} 1 & 0.90 & 0.93 & 0.90 \\ 0.90 & 1 & 0.83 & 0.96 \\ 0.93 & 0.83 & 1 & 0.85 \\ 0.90 & 0.96 & 0.85 & 1 \end{pmatrix}$; $df = 7.75$	468.64
Clayton	$\theta = 2.77$	330.74
Gumbel	$\theta = 3.07$	405.79

TABLE XVI: Summary of multivariate copulas for marginals of raw claim types

The first risk measure estimated is the Value at Risk (VaR). The VaR_p of random variable X, at the $100p\%$ level, is the $100p$ quantile of X. For example, Solvency II sets $p = 99.5\%$ for loss reserving purposes, where an insurer must be prepared to cover a loss at this level.

The second risk measure estimated is the Tail Value at Risk (TVaR), which is the expected loss given that a loss has exceeded the $100p$ quantile of X. This can be expressed as $TVaR_p = E(X|X > VaR_p)$. Thus, the TVaR focuses on the tail of X, as it measures the expectation of values superior to the VaR. If a variable has heavy tails, then there could be considerable differences between the VaR and TVaR, as will be shown later on.

Both risk measures are estimated for the right tails of the simulated series, as the purpose is to analyze risky scenarios, of high losses, for each claim type. The basic Python code for this simulation has been included in Appendix 5, but can be summarized in the following steps:

1. Use a given copula to simulate a vector ($length = 156$) of random conditional probabilities.
2. Revert these probabilities into claim severities using the inverse cdf of the respective marginal distribution.
3. Estimate the VaR and TVaR for this simulated series at a predetermined confidence level. This result is saved in a Python dictionary.
4. Reiterate this loop 10,000 times. A mean estimate is then taken of the VaR and TVaR.
5. Repeat this process for the next confidence level. The confidence levels tested ranged from 95% to 99.5% in increments of 0.5%.

Figure 20 compares the simulated VaR_p and $TVaR_p$ for property damage and property loss claims, through the use of Gumbel's copula. This plot illustrates that property loss has a heavier tail than property damage, as it has a higher simulated $TVaR_p$ and VaR_p, for all p values. This result is consistent with the data, as the property loss severity had a much higher maximum than the property damage severity.

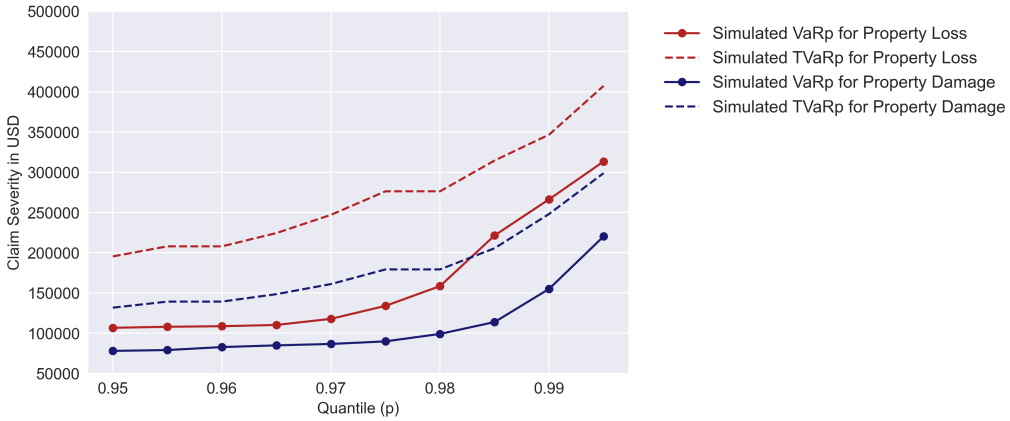


FIGURE 20: Simulated VaR_p and $TVaR_p$ for raw claim types using Gumbel's copula

Another interesting analysis is to compare the simulated risk measures for one series using different copulas. Figure 21 illustrates the simulated $TVaR_p$ for the property loss claims. This plot shows all of the copulas yield more conservative $TVaR_p$ values than the historical values after the 95% quantile.

Clayton's copula does not seem to be a good fit, as suggested by Figure 21, estimating an unrealistically high $TVaR_p$, compared to the other copulas. Furthermore, Gumbel's copula yields $TVaR_p$ values higher than the two elliptical copulas, starting from the 97% quantile. This may be due to the upper tail dependence of Gumbel's copula.

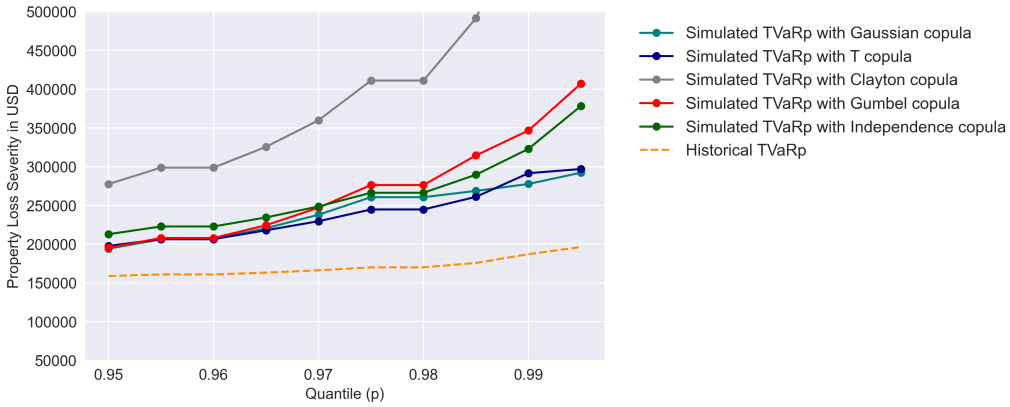


FIGURE 21: Simulated $TVaR_p$ for property loss claims using different copulas

These differences are not quite as pronounced when analyzing the simulated VaR_p for the property loss claims, illustrated in Figure 22. In this plot, the simulated VaR_p from the copulas only exceed the historical values starting at the 98% quantile. This finding illustrates a limitation of the VaR, as it undervalues extreme events for random variables with heavy tails, such as this one.

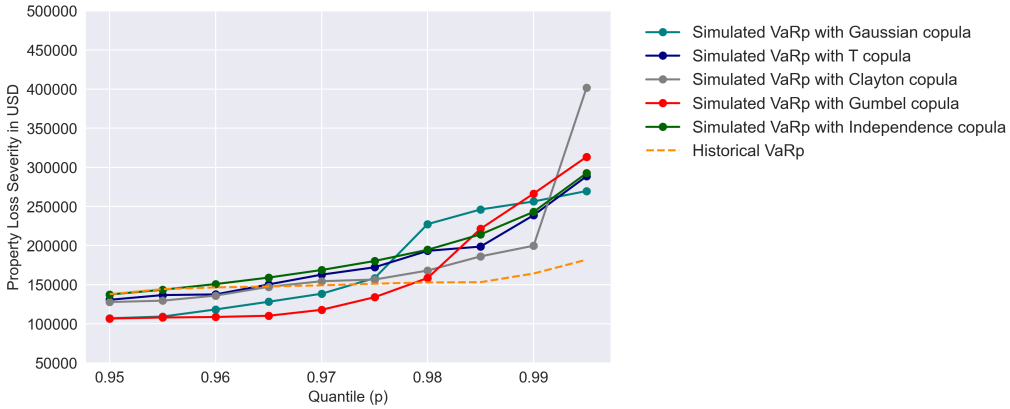


FIGURE 22: Simulated VaR_p for property loss claims using different copulas

Additional analysis can quantify the impact the marginal distributions, used in the copulas, have on the simulated risk measures. This is done by generating the Empirical Cumulative Distribution Function (ECDF) for all the claim types. These empirical distributions are then fit into a given copula, to generate simulations with the same procedure as before. If the copula type is the same, but one includes parametric and the other empirical marginals, the resulting simulated risk measures provide insights on the tail of a random variable.

This is best illustrated in Figure 23. The plot includes simulated $TVaR_p$ from two Gumbel's copula. The first Gumbel's copula contains parametric marginals and has been used in Figures 19-21. The second Gumbel's copula contains the empirical marginals, and its $TVaR_p$ estimates are plotted with dotted lines.

For both severity series, the copula $TVaR_p$ estimates using parametric marginals are significantly higher than those obtained with empirical marginals. This highlights the heavy-tail nature of the two distributions used to model the property loss and property damage severities. The choice of the copula used can further accentuate these differences, as previously highlighted in Figure 20.

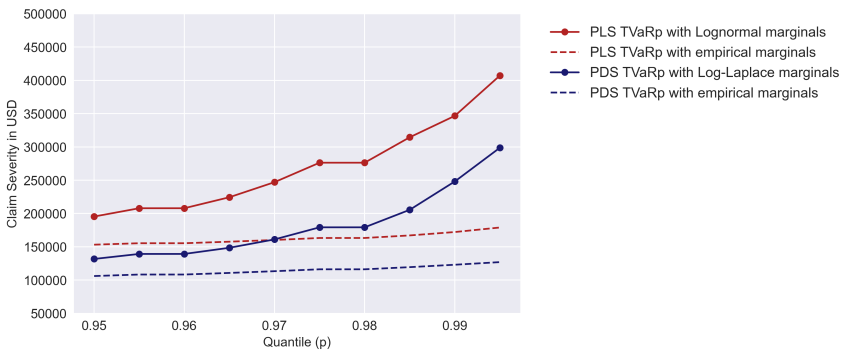


FIGURE 23: Simulated $TVaR_p$ using Gumbel's copula with different marginals

Table XVII summarizes these results and presents all the monthly risk measures obtained at the 99.5% confidence level. The table confirms Clayton's copula is a bad fit for these marginals, as it overestimates the risk measures for property loss, and underestimates the risk measures for property damage, compared to the historical values. In contrast, the risk measures of the other copulas exceed the historical values of the VaR_p and $TVaR_p$.

For example, Gumbel's copula yields a VaR_p estimate for PLS that is 172% higher than the historical, and a $TVaR_p$ 207% higher than the historical. Both of the elliptical copulas are more conservative in this regard. The Student t copula has a $TVaR_p$ for PLS that is 151% higher than the historical. For the Gaussian copula, this $TVaR_p$ estimate is 149% higher than the historical value.

The results for the independence copula are also interesting, though perhaps a larger number of iterations is needed to see if they diverge more strongly from the other copulas. What is significant, is that the independence case also yields higher risk measures than the historical values. This can be attributed to the heavy-tailed nature of the marginals included in the independence copula, as shown in Figure 23.

Measure	Historical	Gaussian cop	T cop	Clayton cop	Gumbel cop	Indep. cop
PLS VaR	181897.39	269656.82	288564.97	401826.30	313373.99	292540.70
PDS VaR	130443.56	155413.86	217871.75	125964.10	220262.34	226264.98
PLS TVaR	196328.29	292260.45	296987.34	1074281.08	407208.38	378359.62
PDS TVaR	133835.07	162481.01	274806.54	128396.17	298815.09	328110.80

TABLE XVII: Monthly risk measures obtained at 99.5% for raw claim types (USD)

Overall, the results suggest the differences between the historical risk measures and the simulated estimates using copulas are due to two factors. The first factor is the choice of the copula and its Goodness of Fit (GOF) with the marginals. Clayton's copula seems to be a poor fit and yields unreliable risk measures. The second factor is related to the heaviness of the tail in the marginals used for each random variable. The final section attempts to confirm these findings by simulating risk measures for the claim sites.

5.5 Simulating risk measures for claim sites with copulas

The first step is fitting different copulas for the raw claim sites. Table XVIII summarizes the results from the copula fits for the raw claim sites. This table shows Gumbel's copula is also a better fit than Clayton's, according to log-likelihood. Additionally, the two

elliptical copulas have higher likelihoods than the Archimedean copulas, and although not directly comparable just based on this measure, they can be considered better fits.

Type of copula	Parameter(s)	Log-likelihood
Gaussian	$\Sigma = \begin{pmatrix} 1 & 0.86 & 0.93 & 0.72 \\ 0.86 & 1 & 0.86 & 0.87 \\ 0.93 & 0.86 & 1 & 0.71 \\ 0.72 & 0.87 & 0.71 & 1 \end{pmatrix}$	376.84
Student t	$\Sigma = \begin{pmatrix} 1 & 0.86 & 0.98 & 0.69 \\ 0.86 & 1 & 0.86 & 0.85 \\ 0.98 & 0.86 & 1 & 0.68 \\ 0.69 & 0.85 & 0.68 & 1 \end{pmatrix}$; $df = 3.66$	298.81
Clayton	$\theta = 2.01$	247.85
Gumbel	$\theta = 2.25$	271.37

TABLE XVIII: Summary of multivariate copulas for marginals of raw claim sites

Figure 24 plots the simulated VaR_p and $TVaR_p$ for the two claim sites using Gumbel's copula. This plot shows the checked baggage claims have a much higher simulated VaR_p and $TVaR_p$ than the security checkpoints, for all p levels. This is to be expected, as there are also large differences in the claim severities of the two sites. Both variables were modelled with Log-Laplace distributions that appear to be heavy-tailed, as there are large differences in the VaR_p and $TVaR_p$ of each series.

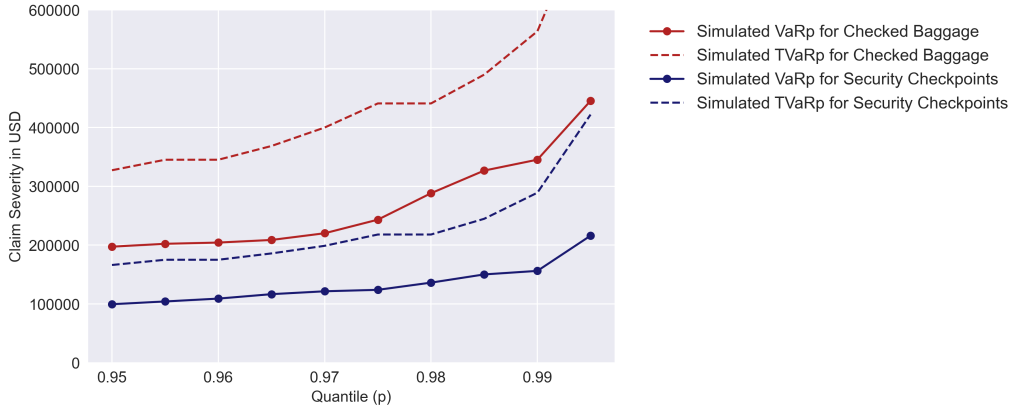


FIGURE 24: Simulated VaR_p and $TVaR_p$ for raw claim sites using Gumbel's copula

Figure 25 compares the VaR_p obtained for the security checkpoint claims using different copulas. This plot shows all of the copula estimates exceed the historical VaR. The

two elliptical copulas also yield lower VaR_p results than the other three copulas. In order to verify this result, it is necessary to also focus on the tail of the security checkpoint claims.

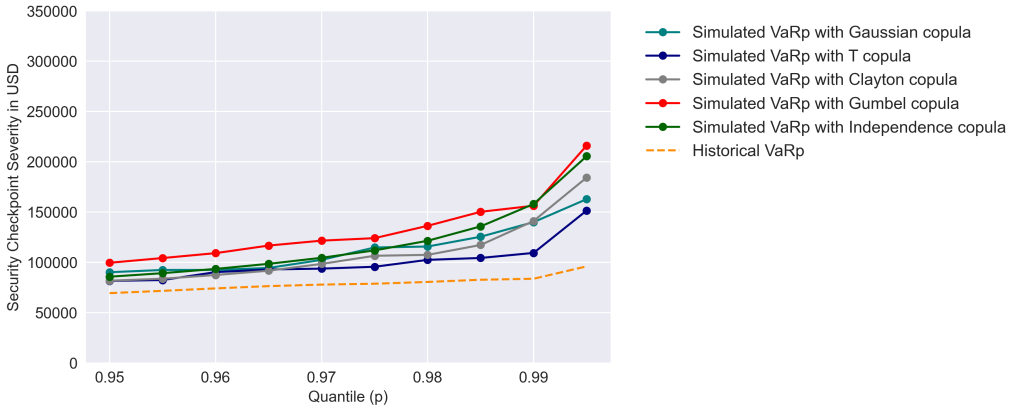


FIGURE 25: Simulated VaR_p for security checkpoint claims using different copulas

Figure 26 focuses on the tail of the security checkpoint claims and plots the simulated $TVaR_p$ with different copulas. This plot confirms Gumbel's copula provides the most conservative risk measures than the rest of the copulas. In this sense, the Gaussian copula yields the lowest $TVaR_p$ for quantiles greater than 98%.

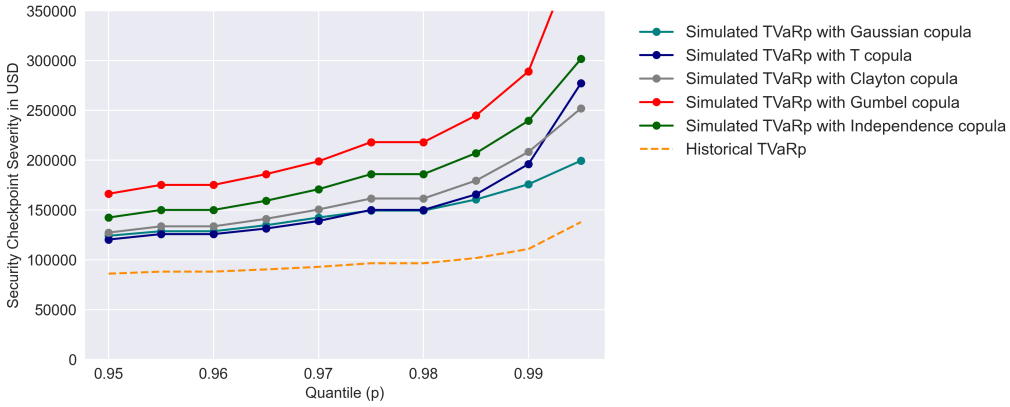


FIGURE 26: Simulated $TVaR_p$ for security checkpoint claims using different copulas

Figure 27 meanwhile, compares the $TVaR_p$ estimates from two Gumbel's copulas for the security checkpoint claims. The first copula includes the Log-Laplace distribution used to model CPS, and the second copula includes the empirical marginals of all the claim sites. This plot illustrates the copula containing parametric marginals generates higher $TVaR$ estimates for all p values, than the copula with empirical marginals. The plot also suggest the Log-Laplace distribution selected for CPS is heavy-tailed, as its

estimates are much higher than the historical values and those obtained with empirical marginals.

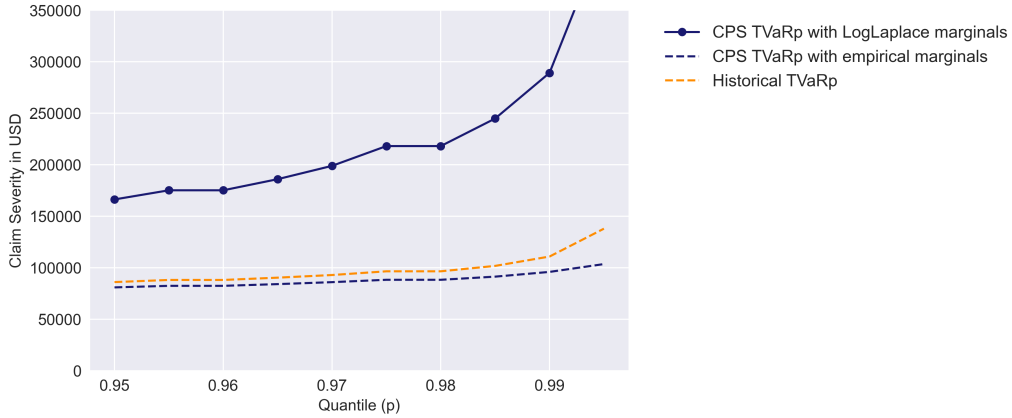


FIGURE 27: Simulated $TVaR_p$ using Gumbel's copula with different marginals

Table XIX summarizes these results and presents all the monthly risk measures obtained at the 99.5% confidence level. The first result that stands out is the $TVaR_p$ obtained for CBS using the independence copula, which is nearly 2 million. This value seems unrealistic and is not comparable to those obtained with parametric copulas. A similar conclusion could be made for the Gaussian copula, which yields high risk measures for CBS but relatively low estimates for CPS, in comparison to Gumbel's copula.

Measure	Historical	Gaussian cop	T cop	Clayton cop	Gumbel cop	Indep. cop
CBS VaR	259019.58	639220.33	482656.16	436257.96	445350.78	909156.41
CPS VaR	95929.58	162753.20	162753.20	184214.74	216060.15	205430.66
CBS TVaR	261424.43	910176.52	913644.30	834543.75	778463.95	1948296.78
CPS TVaR	137850.99	199462.25	277221.58	251982.85	421945.04	301612.57

TABLE XIX: Monthly risk measures obtained at 99.5% for raw claim sites (USD)

These results confirm the fitted claim sites have heavy tails, as all the simulated risk measures exceed the historical values. In some cases, like those obtained with the independence copula, the estimates seem too conservative and unrealistic. Thus, the choice of which copula should be used has a qualitative element, as there has to be a certain balance. The "ideal" copula should yield risk estimates that are higher than the historical values, but not unrealistic for loss reserving purposes. Future research could perform these simulations with additional copulas, or select heavy-tailed marginals for the claim severities *a priori*.

6 CONCLUSIONS

This dissertation highlights the usefulness of modelling dependencies with copulas using the TSA claims database. The initial aggregation of claims in months evidences that most claims concentrate in two types (property damage and property loss), and also take place in mostly two airport sites (checked baggage and security checkpoints). The presence of non-stationary time series is also detected, and corrected through a polynomial detrending process.

Furthermore, by finding the best parametric distributions for the aggregated series in Chapter 4, it is also possible to appreciate differences between each risk. In the case of the claim counts, the Negative Binomial distribution is a better fit than the traditionally used Poisson. The inclusion of detrended series also allows for the use of distributions with positive and negative supports, with better goodness-of-fit than those distributions used for the raw series.

The application of bivariate copulas demonstrates the existence of tail dependence between the different claim counts and severities. This result challenges the traditional risk modelling approach of assuming independence between claim counts and amounts. The Archimedean copulas applied to the data helped in visualizing non-linear dependence structures.

Lastly, the results from the multivariate copula simulations show that modelling claims with copulas can produce more conservative risk estimates, for both the claim sites and claim types. These results are influenced by the heaviness of the tails of the parametric marginal distributions included in the copulas. All of the severity series are best modelled with heavy-tailed distributions, which yield VaR_p and $TVaR_p$ estimates that exceed the historical values.

These results also show sensitivity to the type of copula used, and evidences how different assumptions on claim dependencies can lead to different risk measures for loss reserving purposes. Additionally, knowing these dependence structures could also be useful for insurance pricing purposes, such as estimating premiums, among other aspects of risk management.

Future research can focus on using alternative distributions to model the claim counts, such as fitting discrete mixtures, in order to improve the goodness of fit. With more computational power it could also be useful to perform more iterations to confirm if these results are consistent. Furthermore, the use of vine copulas could also be analyzed, to contrast its simulated risk measures with the parametric approach adopted in this dissertation.

REFERENCES

Alvarez, M., Sala, C., Sun, Y., Pérez, J., Zhang, K., Montanez, A., Bonomi, G., Veeramachaneni, K., Ramírez, I., Hofman, F., Lima, P. & Ivantsiv, N. (2018), 'Copulas python library'.
URL: *https://sdv.dev/Copulas/index.html*

Aussenegg, W. & Cech, C. (2012), 'A new copula approach for high-dimensional real world portfolios', *Working Paper Series* (68).

Bok, D. (2019), 'Copulae python library'.
URL: *https://copulae.readthedocs.io/en/latest/index.html*

Brechmann, E., Hendrich, K. & Czado, C. (2013), 'Conditional copula simulation for systemic risk stress testing', *Insurance: Mathematics and Economics* **53**(3), 722–732.

Caillault, C. & Guegan, D. (2005), 'Empirical estimation of tail dependence using copulas. application to asian markets', *Quantitative Finance* **5**, 489–501.

Cheng, G., Ping, L. & Shi, P. (2007), 'A new algorithm based on copulas for var valuaion with empirical calculations', *Theoretical Computer Science* **378**, 190–197.

Choi, S. & Yoon, J. (2020), 'Modeling and risk analysis using parametric distributions with an application in equity-linked securities', *Mathematical Problems in Engineering* **2020**.

Czado, C. (2018), *Analyzing Dependent Data with Vine Copulas*, Springer, New York, NY.

Department of Homeland Security (2019), 'Tsa claims data'.
URL: *https://www.dhs.gov/tsa-claims-data*

Ding, W. (2015), Copula Regression Models for the Analysis of Correlated Data with Missing Values, PhD thesis, University of Michigan.

Dorey, M. & Joubert, P. (2005), 'Modelling dependencies: An overview', *Finance and Investment Conference* .

Ferreira, M. (2013), 'Non-parametric estimation of the tail-dependence coefficient', *REVSTAT - Statistical Journal* **11**(1), 1–16.

Frahm, G., Junker, M. & Schmidt, R. (2005), 'Estimating the tail-dependence coefficient: Properties and pitfalls', *Insurance: Mathematics and Economics* **37**(1), 80–100.

Frees, E., Lee, G. & Yang, L. (2016), 'Multivariate frequency-severity regression models in insurance', *Risks* **4**(4).

Haberland, M. (2020), 'Scipy issue 11948'.
URL: *https://github.com/scipy/scipy/issues/11948*

Haff, I., Aas, K. & Frigessi, A. (2010), 'On the simplified pair-copula construction - simply useful or too simplistic?', *Journal of Multivariate Analysis* **101**, 1296–1310.

Hasebe, T. (2013), 'Copula-based maximum-likelihood estimation of sample-selection models', *The Stata Journal* **13**(3), 547–573.

Hernández, J., Hammoudeh, S., Nguyen, D., Al Janabi, M. & Reboredo, J. (2016), 'Global financial crisis and dependence risk analysis of sector portfolios: A vine copula approach', *MPRA Paper* (7399).

Hochrainer-Stigler, S., Pflu, G., Dieckmann, U., Rovenskaya, E., Thurner, S., S., P., G., B., Linnerooth-Bayer, J. & Brannstrom, A. (2018), 'Integrating systemic risk and risk analysis using copulas', *Int J Disaster Risk Sci* **9**, 561–567.

Hofert, M., Machler, M. & McNeil, A. (2012), 'Likelihood inference for archimedean copulas in high dimensions under known margins', *Journal of Multivariate Analysis* **110**, 133–150.

Klugman, S., Panjer, H. & Wilmot, G. (2008), *Loss Models: From Data to Solutions, 3rd. Edition*, Wiley, Hoboken, NJ.

Manner, E. (2007), 'Estimation and model selection of copulas with an application to exchange rates', *METEOR* **56**.

Masarotto, G. & Varin, C. (2017), 'Gaussian copula regression in r', *Journal of Statistical Software* **77**(8).

Nelsen, R. (1997), 'Dependence and order in families of archimedean copulas', *Journal of Multivariate Analysis* **60**, 111–122.

Nelsen, R. (2006), *An Introduction to Copulas*, Springer, Cham, Switzerland.

Nicolas, M. (2021), 'Pycop python library'.
URL: *https://pypi.org/project/pycop/*

Oh, D. (2014), Copulas for High Dimensions: Models, Estimation, Inference, and Applications, PhD thesis, Duke University.

Omari, C., Nyambura, S. & Mwangi, J. (2018), 'Modeling the frequency and severity of auto insurance claims using statistical distributions', *Journal of Mathematical Finance* **8**, 137–160.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011), 'Scikit-learn: Machine learning in python', *Journal of machine learning research* **12**(Oct), 2825–2830.

Safari-Katesari, H. & Zaroudi, S. (2020), 'Count copula regression model using generalized beta distribution of the second kind', *Statistics in Transition* **21**(2).

Shemyakin, A., Huan, Z., Benson, S., Burroughs, R. & Mohr, J. (2019), 'Copula models of economic capital for life insurance companies', *Society of Actuaries* .

Shi, P., Feng, X. & Boucher, J. (2016), 'Multilevel modeling of insurance claims using copulas', *The Annals of Applied Statistics* **10**(2), 834–863.

Sklar, A. (1973), 'Random variables, joint distribution functions and copulas', *Kybernetika* **9**(6), 450–460.

Valdez, E. (2014), 'Empirical investigation of insurance claim dependencies', *European Actuarial Journal* **4**(1), 155–179.

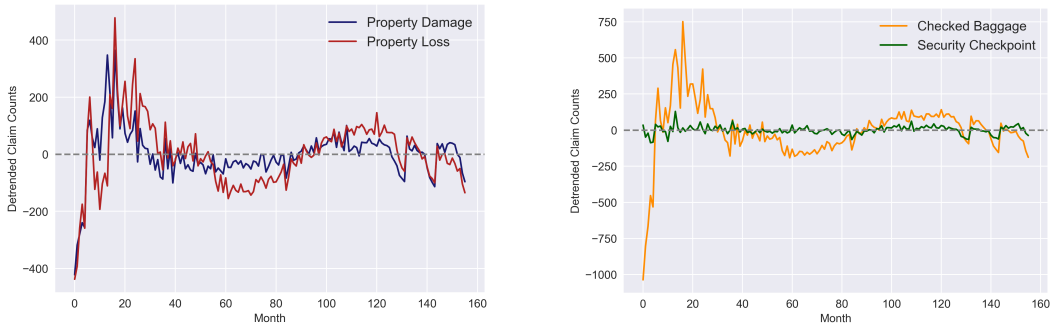
Vandenberghe, S., Verhoest, N. & De Baets, B. (2010), 'Fitting bivariate copulas to the dependence structure between storm characteristics: A detailed analysis based on 105 year 10 min rainfall', *Water Resources Research* **46**(W01512).

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors (2020), 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python', *Nature Methods* **17**, 261–272.

A APPENDICES

A.1 Descriptive statistics of all the detrended series

Appendix A.1 goes into further detail of the detrended series presented at the end of Chapter 3. The graphs in figure A.1 show the detrended claim counts cycle around the y axis at zero.



(a) Detrended counts by type

(b) Detrended counts by site

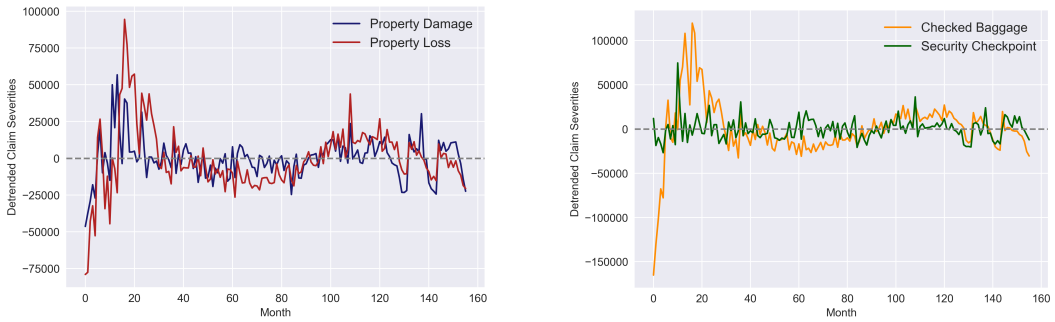
FIGURE A.1: Monthly detrended TSA claim counts from 2003-2015

To confirm their mean is zero and the detrending process was successful, Table A.1 presents some descriptive statistics obtained using Python. This table shows all of the series have a mean of zero, indicating the third order polynomial detrending was successful in eliminating the negative trend component.

Statistic	DPDX	DPLX	DCBX	DCPX
Minimum	−422.36	−437.84	−1036.66	−87.47
Maximum	364.01	477.95	750.79	128.99
Mean	0.00	0.00	0.00	0.00
Standard Deviation	89.81	116.70	198.10	28.76
Skewness	−0.34	0.08	−0.95	−0.02
Kurtosis	6.89	2.91	7.94	2.88

TABLE A.1: Descriptive statistics for detrended TSA claim counts

The issue here is that the detrended series have negative values, which cannot be readily interpreted. This also occurs with the detrended severities, as demonstrated in Figure A.2. These graphs both show how the detrended claim severities cycle around the y axis at zero.



(a) Detrended severities by type (b) Detrended severities by site

FIGURE A.2: Monthly detrended TSA claim severities from 2003-2015

The ranges for all series are much more pronounced, due to the scale, as reflected in Table A.2. This table displays the high ranges of the claim severities, especially for CBS.

Statistic	DPDS	DPLS	DCBS	DCPS
Minimum	-46348.60	-79007.97	-165131.60	-26816.67
Maximum	56741.45	94453.69	119657.70	74654.35
Mean	0.00	0.00	0.00	0.00
Standard Deviation	13940.64	22305.52	32757.34	12515.92
Skewness	0.50	0.54	-0.48	1.55
Kurtosis	3.19	4.06	7.48	7.41

TABLE A.2: Descriptive statistics for detrended TSA claim severities

Thus, it is difficult to use a measure like VaR, as it is a tail based measure which will take into account positive severity deviations from the trend.

A.2 Details of the bin selection process for the chi-squared tests

As stated in Chapter 4, 20 bins were selected to perform the χ^2 tests, using a quantile approach. This means that each bin has a 5% probability of occurring for the distribution being tested. The upper and lower bounds for each bin are generated using the corresponding inverse cumulative distribution function; for ranges from 0-5%, 5-10%, and so on. Thus, each expected observation $E_{\{j=1,2,...,k\}}$ has approximately uniform values across all bins, but the bins themselves depend on the distribution being tested.

For example, the first distribution tested for the property damage counts is a Poisson with $\lambda = 526.32$. Table A.3 illustrates the first 10 bins obtained for this distribution by quantile, along with the observed values O_j, and the expected values from this Poisson distribution E_j. .

Quantile	Lower bound	Upper bound	O_j	E_j
$0-5\%$	0	489	136	8.2638
$5-10\%$	489	497	1	7.9218
$10-15\%$	497	503	0	8.7721
$15-20\%$	503	507	1	7.2924
$20-25\%$	507	511	0	8.3996
$25-30\%$	511	514	0	6.9510
$30-35\%$	514	517	0	7.4188
$35-40\%$	517	520	0	7.7815
$40-45\%$	520	523	0	8.0218
$45-50\%$	523	526	1	8.1286

TABLE A.3: Sample bins estimated for property damage counts using a Poisson

The *SciPy* library includes a function to estimate the χ^2 statistic and its p-value. Recall that in this case, the critical value has $(20-1-1=18)$ degrees of freedom, as the Poisson has a single parameter. The results from the test are: $\chi^2 = 2105.36$ with $p = 0.00$. The null hypothesis is rejected and the Poisson distribution cannot be considered a good fit for the property damage counts. This is the basic process used to test the remaining distributions used throughout this dissertation.

A.3 Notation for SciPy probability distribution functions

This appendix is dedicated to outlining the SciPy notation for the continuous distribution functions used throughout the dissertation, including how location (L) and scale (S) parameters are incorporated. The probability density functions ($f(x)$) are as follows:

1. Log-Laplace distribution:

$$f(x,c) = \begin{cases} \frac{c}{2}x^{c-1} & \text{for } 0 < x < 1 \\ \frac{c}{2}x^{-c-1} & \text{for } x \geq 1 \end{cases}$$

where c is a shape parameter and $c > 0$, $f(x,c,L,S)$ is equal to $f(y,c)/S$ with $y = \frac{(x-L)}{S}$

2. Johnson SU distribution:

$f(x,a,b) = \frac{b}{\sqrt{x^2-1}}\phi(a + b\log(x + \sqrt{x^2+1}))$

where x, a, and b are real scalars, a and b are shape parameters, $b > 0$, ϕ is the standard normal, and $f(x,a,b,L,S)$ is equal to $f(y,a,b)/S$ with $y = \frac{(x-L)}{S}$

3. Generalized Logistic distribution:

$f(x,c) = c\frac{\exp(-x}{)}(1+\exp(-x))^{c+1}$

for $x \geq 0$, where c is a shape parameter and $c > 0$, and $f(x,c,L,S)$ is equal to $f(y,c)/S$ with $y = \frac{(x-L)}{S}$

4. Lognormal distribution:

$f(x,s) = \frac{1}{sx\sqrt{2\pi}}\exp\left(-\frac{\log^2(x)}{2s^2}\right)$

for $x > 0$, where s is a shape parameter and $s > 0$, and $f(x,s,L,S)$ is equal to $f(y,s)/S$ with $y = \frac{(x-L)}{S}$

5. Student t:

$f(x,df) = \frac{\Gamma((df+1)/2)}{\sqrt{\pi df}\Gamma(df/2)}(1+x^2/df)^{-(df+1)/2}$;

where x is a real number, df are the degrees of freedom and $df > 0$, Γ is the gamma function, and $f(x,df,L,S)$ is equal to $f(y,df)/S$ with $y = \frac{(x-L)}{S}$

6. Non-central t (NCT):

If Y is a standard normal variable and V is an independent χ^2 random variable with df degrees of freedom, then:

$X = \frac{Y+nc}{\sqrt{V/df}}$ has a non-central Student t distribution

where $df > 0$, nc is the non-centrality parameter and must be a real number, and $f(x,df,nc,L,S)$ is equal to $f(y,df,nc)/S$ with $y = \frac{(x-L)}{S}$

7. Hyperbolic Secant:

$f(x) = \frac{1}{\pi}\text{sech}(x)$

where x is a real number, sech is the hyperbolic secant function, and $f(x, L, S)$ is equal to $f(y)/S$ with $y = \frac{(x-L)}{S}$

8. Double Gamma:

$f(x, a) = \frac{1}{2\Gamma(a)}|x|^{a-1}\exp(-|x|)$

for a real number x, where a is a shape parameter and $a > 0$, Γ is the gamma function, and $f(x, a, L, S)$ is equal to $f(y, a)/S$ with $y = \frac{(x-L)}{S}$

A.4 Definitions of the copulas used and their tail dependence coefficients

This appendix formulates the copulas used in chapter 5, along with their tail dependence coefficients. The definitions for the empirical copula are retrieved from Caillault & Guegan (2005) and Frahm et al. (2005). Alvarez et al. (2018) is the reference for the three Archimedean copulas. The two elliptical copulas considered for the final two sections of Chapter 5 are retrieved from Bok (2019).

1. Empirical copula:

If $\mathbf{z} = \{(z_{1_k}, z_{2_k})\}_{k=1}^{N}$ denotes a sample of size N from a continuous bivariate distribution, the empirical copula is the function $\tilde{C}$ given by:

$$\tilde{C}(\frac{i}{N}, \frac{j}{N}) = \frac{\#((z_1, z_2), z_1 \leq z_1^{(i)} and\, z_2 \leq z_2^{(j)})}{N} \tag{A.1}$$

where # is used for cardinal; $z_1^{(i)}$, $z_2^{(j)}$ for $1 \leq i, j \leq N$ represent the order statistics obtained from the sample.

The upper tail dependence coefficient for the empirical copula is then given by:

$$\tilde{\lambda}_U = 2 - \frac{1 - \tilde{C}(\frac{N-k}{N}, \frac{N-k}{N})}{1 - \frac{N-k}{N}} \tag{A.2}$$

for $0 < k \leq N$, where $\frac{k}{N}$ represents the threshold or quantile for the estimation.

2. Clayton's copula:

Clayton's copula, with parameter θ, is a member of the Archimedean family and has the

following expression:

$$C(u,v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}, \text{ for } \theta > 0 \tag{A.3}$$

Along with the copula density function:

$$c(U,V) = (\theta+1)(uv)^{-\theta-1}(u^{-\theta} + v^{-\theta} - 1)^{\frac{2\theta+1}{\theta}}, \text{ for } \theta > 0 \tag{A.4}$$

Clayton's copula only has a lower tail dependence coefficient, given by:

$$\tilde{\lambda}_L = 2^{-\frac{1}{\theta}}, \text{ for } \theta > 0 \tag{A.5}$$

3. Gumbel's copula:

Gumbel's copula, with parameter θ, is a member of the Archimedean family and has the following expression:

$$C(u,v) = e^{-((-\ln u)^{\theta} + (-\ln v)^{\theta})^{\frac{1}{\theta}}}, \text{ for } \theta \geq 1 \tag{A.6}$$

Along with the copula density function:

$$c(U,V) = \frac{C(u,v)}{uv} \frac{((-\ln u)^{\theta} + (-\ln v)^{\theta})^{\frac{2}{\theta}-2}}{(\ln u \ln v)^{1-\theta}} (1 + (\theta-1)((-\ln u)^{\theta} + (-\ln v)^{\theta})^{-\frac{1}{\theta}}) \tag{A.7}$$

for $\theta \geq 1$

Gumbel's copula only has an upper tail dependence coefficient, given by:

$$\tilde{\lambda}_U = 2 - 2^{-\frac{1}{\theta}}, \text{ for } \theta \geq 1 \tag{A.8}$$

4. Frank's copula:

Frank's copula, with parameter θ, is a member of the Archimedean family with no tail dependence, and has the following expression:

$$C(u,v) = -\frac{\ln\left(\frac{1+g(u)g(v)}{g(1)}\right)}{\theta}, \text{ for } -\infty < \theta < +\infty \tag{A.9}$$

where $g(x) = e^{-\theta x} - 1$. The copula density function is given by:

$$c(U,V) = \frac{-\theta g(1)(1+g(u+v))}{(g(u)g(v)+g(1))^2}, \text{ for } -\infty < \theta < +\infty \tag{A.10}$$

5. Gaussian copula:

The Gaussian copula is an elliptical copula with equal weight placed on each tail. The Gaussian copula is also determined exclusively by the correlation matrix of its marginals:

$$C_\Sigma(u_1, u_2, ..., u_n) = \Phi_\Sigma(N^{-1}(u_1), N^{-1}(u_2), ..., N^{-1}(u_n)) \tag{A.11}$$

where Σ is the correlation matrix and must be non-negative, and N^{-1} is the quantile function or inverse cumulative distribution function.

6. Student t copula:

The Student t copula is another elliptical copula, with heavier tails than the Gaussian copula. This copula is determined by the correlation matrix, Σ, and the degrees of freedom, v, parameter:

$$C_{\Sigma,v}(u_1, u_2, ..., u_n) = t_{\Sigma,v}(t_v^{-1}(u_1), t_v^{-1}(u_2), ..., t_v^{-1}(u_n)), \text{ for } v > 0 \tag{A.12}$$

where t_v^{-1} is the quantile or inverse cumulative distribution function. Although they were not estimated for this dissertation, the tail coefficients for the t copula are equivalent and given by:

$$\tilde{\lambda}_U = \tilde{\lambda}_L = 2\bar{t}_{v+1}\left(\sqrt{\frac{(v+1)(1_\rho)}{1+\rho}}\right), \text{ for } v > 0 \tag{A.13}$$

where ρ is the correlation coefficient between the variables, $\bar{t}_{v+1} = 1 - t_{v+1}(u)$, and t_{v+1} is the Student t distribution with $v+1$ degrees of freedom.

A.5 Basic Python code for simulation of risk measures

The first step of the simulation is to generate the four marginal distributions for the series and to fit them into a given copula. For this example, the simulation is shown using a Gaussian copula:

```
#loading the raw claims dataset:
df = pd.read_csv("aggregated_series.csv")

#adding the marginals to the existing dataframe
df['Fpdx'] = ss.nbinom.cdf(df['pdx'], 2.139883173181995, 0.00922814420430195)
df['Fplx'] = ss.nbinom.cdf(df['plx'], 1.7579089737306162, 0.006834721956499432)
df['Fpds'] = ss.loglaplace.cdf(df['pds'], 2.481704232311386, 0, 37891.93289315458)
df['Fpls'] = ss.lognorm.cdf(df['pls'],  0.9086819314521254, 0.0, 31402.11730335455)

from copulae import GaussianCopula

#setting the dimensions:
_, ndim = df.iloc[:,9:].shape

#initializing the copula:
g_cop = GaussianCopula(dim=ndim)

#fitting the copula to the marginals:
g_cop.fit(df.iloc[:,9:], to_pobs=False)
```

Additionally, two functions were also created to make the estimation of the VaR and TVaR more direct in the simulation:

```
#making a VaR function:
def VAR(series, CL):
    return series.quantile(CL, axis=0)

#Tail VaR function:
def TVAR(series, ci):
    var = VAR(series, ci)
    return series[series.gt(var, axis=1)].mean()
```

Both functions take dataframes, or tables in Python format, and confidence levels as arguments. This allows VaR and TVaR estimates for all series in a given dataframe. In this case, the risk measures are applied only to the two severity variables, PLS and PDS.

The simulation starts with the construction of a list containing the confidence intervals tested. Afterwards, the conditional random probabilities are generated through the copula. These vectors are converted back to the severities through the inverse cdf of the respective marginals:

```
########################## GAUSSIAN COPULA SIMULATION#########
pi = [0.95, 0.955, 0.96, 0.965, 0.97, 0.975, 0.98, 0.985, 0.99, 0.995]

#generating 10000 simulations:
sims = {}
for it in range(0,10000):
    sims[it] = g_cop.random(len(df))

#finding the inverse of the simulated series:
for key in sims:
    sims[key]['pls'] =  ss.lognorm.ppf(sims[key]['Fpls'], 0.9086819314521254, 0.0, 31402.1173033545)
    sims[key]['pds'] =  ss.loglaplace.ppf(sims[key]['Fpds'], 2.481704232311386, 0, 37891.93289315458)
```

The VaR and TVaR for the two series are then obtained by looping through this dictionary containing the simulated series:

```
#creating VaR and TVaR dictionaries from the simulations:
vpl1 = []
tvpl1 = []
vpd1 = []
tvpd1 = []
for p in tqdm(pi):
    vi = {}
    vi2 = {}
    tvi = {}
    tvi2 = {}
    for key in sims:
        vi = list(VAR(sims[key].iloc[:,[4]], p))
        vi2 = list(VAR(sims[key].iloc[:,[5]], p))
        tvi = list(TVAR(sims[key].iloc[:,[4]], p))
        tvi2 = list(TVAR(sims[key].iloc[:,[5]], p))
    vpl1.append(mean(vi))
    vpd1.append(mean(vi2))
    tvpl1.append(mean(tvi))
    tvpd1.append(mean(tvi2))
```

This process is then repeated using the different copulas shown in Chapter 5. The Frank copula yielded highly volatile simulation results, so it is omitted from the analysis. This copula seems to consistently generate conditional probabilities close to zero, so the simulated risk measures were much lower than those generated by the other copulas.