METHODOLOGICAL AND EMPIRICAL INVESTIGATIONS IN QUANTIFICATION OF MODERN OPERATIONAL RISK MANAGEMENT

by

Sabyasachi Guharay

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Systems Engineering and Operations Research

Committee:

_____	Dr. KC Chang, Dissertation Director
_____	Dr. Jie Xu, Dissertation Co-Director
_____	Dr. John F. Shortle, Committee Member
_____	Dr. Fei Li, Committee Member
_____	Dr. Ariela Sofer, Department Chair
_____	Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering
Date: _____	Summer Semester 2016 George Mason University Fairfax, VA

Methodological and Empirical Investigations in Quantification of Modern Operational Risk Management

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Sabyasachi Guharay
Master of Arts
The Wharton School, University of Pennsylvania, 2006
Bachelor of Science
Princeton University, 2003

Director & Co-Director: KC Chang, Professor & J Xu, Professor
Department of Systems Engineering and Operations Research

Summer Semester 2016
George Mason University
Fairfax, VA

DEDICATION

I wish to dedicate this thesis to my parents, Dr. Samar K. Guharay and Mrs. Karabi Guharay, for their never-ending encouragement in my academics.

ACKNOWLEDGEMENTS

First, I would like to express my sincerest gratitude to Professor KC Chang of the Systems Engineering & Operations Research (SEOR) department for his critical role in developing this work. This work certainly would not have matured to this point without his extremely insightful help, patience, and guidance at many critical junctures. I remain indebted to Prof. Chang.

I'd like to also acknowledge a series of invaluable contributions that have helped me ask key questions, build knowledge, and drive towards rigorous solutions to challenging problems. To begin, when I first entered George Mason University (GMU) in 2012, Prof. John F. Shortle served as my advisor. I owe him a great deal of thanks and sincere gratitude for his guidance throughout the Ph.D. process in SEOR at GMU, along with agreeing to serve on my committee and giving me a comprehensive exam. I began working under the guidance of Prof. KC Chang starting from my second year at GMU. I wish to thank him for his guidance regarding the application of financial systems engineering methods to operational risk starting from my second year when I took his course in Financial Systems Engineering: OR 688. In my third year, I took a course with Prof. Jie Xu on Simulation Optimization: OR 750, and afterwards he became my Co-Director of the Ph.D. His invaluable guidance during the fourth year in this research has been crucial for its progress. In addition, I would like to thank Prof. Fei Li for serving as an outside SEOR committee member and his extremely valuable feedback. Besides members of my committee, I would like to thank Prof. C-H Chen and Prof. K.B. Laskey in the SEOR department for their continual support and valuable discussions throughout my time at GMU. In addition, I would like to thank Dr. Michael Mulhearn who as my first instructor at GMU for Systems Definition and Cost Modeling: SYST 510, offered valuable feedback and guidance throughout my Ph.D. program. Our departmental administrator, Mr. Angel Manzo's continuous help throughout my Ph.D. years has been essential for navigating through the administrative portions of the Ph.D. program. In addition the camaraderie of the present and former graduate students such as Bob Aarhus, Redouane Betrouni, Jeronymo Carvalho, John Checco, Todd Martin, Saba Neyshabouri, Ryan O'Neil, Nicholas Paul, Sajjad Taghiye and Zhenming Wang have provided the bedrock of a crucial learning experience for me during the Ph.D. process. Finally, I would like to thank Prof. Rao Mulpuri of the Electrical & Computer Engineering department at GMU. It is due to his initial encouragement that I applied to the SEOR Ph.D. program while being a full-time employee at the U.S. Federal Government.

Without his encouragement and support, I would not have applied to the SEOR Ph.D. program at George Mason.

Outside of GMU, I would like to thank Profs. James A. Yorke and Brian R. Hunt of the University of Maryland, College Park for sparking my interests in the mathematical sciences starting from my school years. I worked with them during my high-school and early undergraduate years on several projects which lead to my first peer-reviewed publications on mathematical genetics. From my undergraduate years, I would like to thank Prof. Robert Vanderbei and Prof. René Carmona of Princeton University for sparking my interests in operations research and applied statistics. From my current profession, I would first like to thank Mr. Ali Samad-Khan of Stamford Risk Analytics for being the first to introduce me to the field of modern ORM in 2008. In addition, I would like to thank Dr. Subhomoy Ghosh at the National Institutes of Standards and Technologies (NIST) for helpful discussions on certain statistical concepts. Currently, working in the U.S. Federal Government, I would like to thank my current supervisors, Mr. John D'Ambrosio and Ms. Tamara Powell, at the US Department of Treasury, for their continual support in my professional and academic development. Their encouragement and support allowed me the peace of mind to focus on the needed long-hours for the completion of this dissertation. I owe thanks to my brother, Mr. Antardeb Guharay at Capital One, for his patience in listening and thoroughly editing the dissertation for clarity of content. Last but certainly not at all the least; I would like to thank my father, Dr. Samar K. Guharay, for his life-long mentoring and support. Without his encouragement and support, I would have never had the zeal and tenacity to develop interests in pursuing scientific research. This thesis is dedicated in part to him and also to Mrs. Karabi Guharay, my mother who always shares my world with unbounded inspiration, encouragement, and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

AggLoss ..... Aggregate Loss distribution
Burr ..... Type XII Burr distribution
C ..... CDF of a Copula
cdf/CDF ..... Cumulative distribution function
DJIA ..... Dow Jones Industrial Average
erf ..... Error function
EC ..... Economic Capital
EL ..... Expected Loss
EVT ..... Extreme Value Theory
FFT ..... Fast Fourier Transform
GoF ..... Goodness-of-Fit
GMCM ..... Gaussian Mixture Copula Model
$\perp$ ..... Independent (Orthogonal)
GPD ..... Generalized Pareto distribution
i.i.d. ..... Independent and Identically Distributed
KRI ..... Key Risk Indicator
LDA ..... Loss Data Approach
LHS ..... Left-hand side
LN ..... Lognormal distribution
LNG ..... Lognormal-Gamma distribution
MCS ..... Monte-Carlo Simulation
MLE ..... Maximum Likelihood Estimation
$\mathcal{O}$ ..... Order of (Big O notation)
ORM ..... Operational Risk Management
NBD ..... Negative Binomial distribution
pdf/PDF ..... Probability distribution function
pmf/PMF ..... Probability mass function
Pois ..... Poisson distribution
POT ..... Peak over Threshold
Φ ..... CDF of a Standard Normal distribution
ϕ ..... PDF of a Standard Normal distribution
Ran ..... Range of a function
RHS ..... Right-hand side
S&P 500 ..... Standard & Poor's 500 Index
Σ ..... Correlation Matrix for the Multivariate Normal distribution
VaR ..... Value-at-Risk

ABSTRACT

METHODOLOGICAL AND EMPIRICAL INVESTIGATIONS IN QUANTIFICATION OF MODERN OPERATIONAL RISK MANAGEMENT

Sabyasachi Guharay, Ph.D.

George Mason University, 2016

Dissertation Director & Co-Director: Professors KC Chang and Professor J Xu

Establishing robust quantitative metrics which allow decision makers to determine the amount of risk in a system with extreme loss events is a problem of interest in many scientific fields. One of the fundamental metrics which is universally accepted in all fields of risk management is the quantity known as Value-at-Risk (VaR). Both academic researchers and industry practitioners are currently looking at ways to make this estimate more statistically robust and accurate with minimal assumption requirements. In particular, modern Operational Risk Management (ORM), a subfield of risk management, closely investigates methodologies to robustly estimate VaR. With this brief background in mind, this dissertation investigates two fundamental components of modern ORM: (1) Statistically modeling severity (magnitude) of losses and estimating corresponding Aggregate Loss distribution; (2) Robust Estimation of Value-at-Risk. One of the key

problems in the modeling of loss severity is that there is no currently known flexible severity loss distribution which can generalize to fit any type of severity data.

This dissertation finds two three-parameter statistical distributions, Type XII Burr distribution (Burr) and Lognormal-Gamma distribution (LNG), as flexible to fit both heavy and thin-tailed loss magnitude data. Also, in reference to the second fundamental component of modern ORM, this work examines the fundamental current assumption in Monte-Carlo Simulation (MCS) based estimation of VaR: the independence of loss severity and loss frequency (count). A theoretical argument is shown along with simulation evidence contrary to this fundamental assumption which provides the impetus for this work - this dissertation develops two new quantitative approaches for estimating VaR which do not rely on the assumption of independence between frequency and severity. These methods are known as the following: (1) Data Partition of Frequency and Severity (DPFS) through distribution-free method; (2) Distribution based partitioning (DBP) of frequency and severity using copulas. The DPFS involves using clustering analysis, specifically K-means algorithm, to partition the frequency and severity components of the loss data. The DBP approach using copulas is a parametric approach where a specific frequency and severity distribution along with a particular copula function is specified *a priori* and then is fit to the data. Verification and Validation (V&V) of the two new methodologies for computing VaR through both analytical argument and MCS are conducted. In both cases, this dissertation shows the new methodologies primarily perform superior to, and in the worst cases at least as well as, the current VaR best practices, or “classical” method. In addition, it is shown through a

mathematical justification how in two extreme instances, the classical methodology has a systematic bias in estimating the VaR. Finally, this thesis contributes to the field by implementing these new methodologies on five distinct publicly available datasets from four different and diverse domains: (1) Financial Indices data of Standard & Poor's (S&P) 500 and Dow Jones Industrial Average (DJIA); (2) Chemical Loss spills as tracked by the US National Coast Guard; (3) Australian automobile accidents; (4) US hurricane data. It is observed that the classical approach inaccurately estimates VaR for 80% of the simulated data cases studied and 60% of the real-world data cases studied. The new methodologies developed attain accurate VaR estimates which are within the 99% bootstrap confidence interval bounds for both simulated and real-world data. In summary, this thesis contributes to the overall field of risk management in providing new methodologies which better estimate VaR and does not require any critical and unnecessary assumptions. Academic researchers can use these methodologies and findings from the loss severity analysis to further improve upon more efficient methodologies in risk metric evaluations. Industry practitioners can directly apply these methodologies to other real-world data (both public and proprietary) as part of their toolkit when handling operational risk.

CHAPTER 1: INTRODUCTION

The science of studying rare events is a well-known problem of interest in various academic fields ranging from the purely theoretical, such as mathematics and mathematical statistics, to diverse real world disciplines like climatology, insurance, hydrology and financial systems engineering. Typically, most of these fields have focused on the study of events that are frequently occurring - those which occur in the body of a probability distribution - leaving a considerable gap in the accurate modeling and quantification of rarely occurring events. In most cases, events which occur rarely are classified as "outliers" and ignored (or sometimes even incorrectly discarded). It is in fact an innate/fundamental part of human nature as argued by Daniel Kahneman in Prospect Theory [1] (winner of the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel) to disregard rare events. In his work, Kahneman shows from psychological experiments that humans view near-zero probabilities (rarely occurring events) as identical to zero probability (events that simply cannot occur). This perspective from psychology drives home the significant importance of developing non-human based, accurate modeling of rare events. This is precisely what the field of modern quantitative risk management seeks to achieve. The framework for this field is narrated next.

1.1 Fundamentals of Quantitative Risk Management Framework

The overall purpose of the risk management framework is to develop a scientific basis for modeling both rarely occurring and common risks. Historically, the focus of this framework has been on modeling the more commonly occurring risks. However, more recent events such as the 2008 Financial Crisis, 2010 Liquidity Squeeze, 2013 US Government potential default on Treasury bonds etc. showed that the so-called Black Swan [2-8] events can occur in a short time span (i.e. high frequency) and potentially devastate the world economy. These events have been a natural catalyst for many academic researchers and industry practitioners to shift the focus of their work to analyzing rare events. This has been one of the new focus areas of modern quantitative risk management. Thus, while it may be human nature to ignore or neglect most low-probability outlier events, in the modern risk management context, it is crucial that these events are properly modeled and examined. While the mathematics behind low-probability events has been rigorously studied since the 1940's [9], applying it in rigorous and consistent manner in a modern quantitative risk management framework is still missing. Overall this process is considered somewhat of an art rather than a formal science. This is partially due to the difficulties borne from data coming from various correlated sources in some cases, and in others various independent sources. Properly combining all of these data sources is a fully challenging quest in the field of Information Fusion [10]. In the modern risk management practice, many simplifications and assumptions are made to the mathematics of this process which makes the risk management decision-making process incomplete or biased. The primary reason these

simplifications have been necessary is that there are multiple sources of data and the science of integrating them properly is not well understood and practiced [10].

Unlike in other academic disciplines, the modern quantitative risk management framework has largely been developed in the past decade by for-profit financial institutions and government institutions like the Federal Reserve and the Basel Committee on Banking Supervision (BCBS) [11] instead of traditional academia. This framework's principal goal is to analyze and quantify the risks associated with various events. In this dissertation, the focus is on the risks which are typically faced by any risk management institutions (such as insurance companies, federal agencies, banks, etc.). Broadly speaking, a financial institution (such as bank, insurance company, hedge fund, etc.) is exposed to three important types of risk: (1) Market risk; (2) Credit risk, and (3) Operational risk. Market risk can be broadly thought of as macro-economic changes to the overall financial landscape (e.g. stock prices, interest rates) which can adversely affect the portfolio value of a financial institution [10]. Credit risk can be broadly thought of the risk from a failing counterparty in a transaction [10]. These two risks have been extensively studied and there is a good confluence between theory and practice along with academics and industry practitioners [10]. However, the third risk, an equally important branch of risk management, Operational Risk Management (ORM) is a newer type of risk and is a much more active area of recent research.

Market and credit risk can be exploited to minimize loss or maximize profit; however, operational risk is unique in that it cannot be used to generate profit, but rather the sole goal is to minimize losses. In more precise terms, the goal is to manage the

operational risks so that losses are maintained within the financial institution's risk tolerance [12]. In the next section, a further detailed description of this risk is narrated.

1.2 Operational Risk Management

In 2001, the BCBS formally gave a definition of operational risk known in the BASEL I framework. There have been multiple iterations of the exact definition of operational risk. Currently however, BASEL III (2010) has defined operational risk as the following: "The risk of loss resulting from inadequate or failed internal processes, people and systems or from external events" [12]. A few relevant examples of this can include a rogue trader who performs unauthorized or illegal financial transactions, losses from naturally occurring events such as hurricane Sandy/Katrina, credit card fraud, identity theft of federal tax returns, rogue trader activity etc. The loss resulting from operational risk comes from multiple data sources and types. To manage the risk, the international regulatory agency of BCBS regulates and stipulates conditions for financial institutions. One of these is that financial institutions are required to mitigate themselves from all types of risk (including operational risk) by holding Economic Capital (EC) of an appropriate amount to absorb these losses. In other words, financial institutions are required to hold a "rainy day" fund (also known as a capital buffer) to absorb shocks which result from operational risk [10]. The key question that arises is how much should they hold to absorb what level (percentage) of loss and for how long? If they hold too little, a large shock can result in the institution getting wiped out. But if they hold too much capital, then they are losing out on opportunity costs of other business activities [10].

The aforementioned concept is one of the key practical questions that arise from operational risk. From a mathematical point of view, this concept can be described as the Value-at-Risk [13]. Conceptually speaking, a VaR figure of V dollars represents that one is X% sure of not losing more than V dollars in time T. So the practitioner sets the time T and probability X a priori¸ and computes V accordingly. Formally speaking, one sets an a priori confidence level of $\alpha \in (0, 1)$ and a time horizon T. The $VaR_{\alpha,T}$ is given by the smallest number l such that the probability that the loss of L exceeds l is at most (1- α) in time horizon T [10]. Thus, mathematically the $VaR_{\alpha,T}$ is defined as the following:

$$VaR_{\alpha,T}(L) = \inf\{l \in \mathbb{R} : P(L > l) \leq 1 - \alpha\} \text{ or}$$
$$VaR_{\alpha,T}(L) = \inf\{l \in \mathbb{R} : F_L(l) \geq \alpha\} \quad (1)$$

where $F_L(l)$ is a continuous and strictly increasing loss cumulative distribution function (CDF) [10]. In addition to $VaR_{\alpha,T}$, there are other quantities that financial institutions measure which is similar to Value-at-Risk, namely the conditional VaR, also known as Expected Shortfall. This is also given for a confidence level α and time horizon T, but it measures the average loss given the fact that the loss exceeds a VaR.

Now that the major conceptual framework is given, the next important feature in modern operational risk is organizing the loss data. The loss data are organized according to seven official Basel III defined event types and eight defined business lines [11]. This is usually given in a matrix format as shown below in table (1). Each block in the matrix for table (1) below shows that there are different numbers of losses in each cell.

Next, the BCBS has described the fundamental of the modeling of the operational risk which is known as the Loss Data Approach (LDA) [11]. When modeling operational

risk, there are two basic components: (1) Frequency of losses; (2) Severity of losses. The simplest explanation is that one is interested in how often losses will occur (frequency), and also how large will the losses be given that they have occurred (severity). Banks and other financial institutions obviously dread the instances where large losses (severity)

Table 1 Matrix for Typical ORM Loss Data

	Business Lines							
Events	Corporate Finance	Sales & Trading	Retail Bank	Commercial Bank	Payment & Settle	Agency Services	Asset Management	Retail Brokerage
Internal Fraud (IF)	$L_1^{1,1}, L_2^{1,1}, \cdots, L_{n_1}^{1,1}$	$L_1^{2,1}, \cdots, L_{n_2}^{2,1}$	$\cdots$					
External Fraud (EF)	$L_1^{1,2}, L_2^{1,2}, \cdots, L_{n_2}^{1,2}$	$\vdots$		$\ddots$				
Employee Practices & Workplace Safety (EPWS)	$\vdots$	$\vdots$			$\ddots$			
Clients, Products, & Business Practice (CPBP)	$\vdots$	$\vdots$				$\ddots$		
Damages to Physical Assets (DPS)	$\vdots$	$\vdots$					$\ddots$	
Business Disruption & Systems Failures (BDSF)	$\vdots$	$\vdots$						$\ddots$
Execution , Delivery, & Process Manage (EDPM)	$L_1^{1,7}, L_2^{1,7}, \cdots, L_{n_7}^{1,7}$	$L_1^{2,7}, \cdots, L_{n_{14}}^{2,7}$						$L_1^{8,7}, \cdots, L_{n_{56}}^{8,7}$

occur in a large number of instances (i.e. frequency). This is known as a high probability, high impact event. Contrary to the fears of many chief financial officers, these types of events almost never take place. The reason is that any reasonable financial institution will have proper risk management practices which would identify key risk indicators (KRIs) [12] that can prevent/mitigate frequent occurrences of large losses. In otherwords, any good financial institution will have checks in place to ensure that their employees cannot regularly steal billions of dollars. So if there is a rogue employee committing theft, it should be a rare event, and not a frequent event. However, there is an economic trade-off. To achieve close to zero losses, a financial institution can employ the strictest of KRI's. For example, they can mandate employees change their computer access passwords every minute. This practice for example, can ensure with near certainty that the company will not face any economic losses due to a hacked password. However, the trade-off for this procedure will be that most employees will face enormous frustration and will most likely be unable to be productive and thus lead the financial institution to loss in economic productivity. Thus, there comes the issue of trade-off. How high does the financial institution want the KRI's to be placed versus how much EC does the financial institution wish to hold as back-up reserves? This is a key unresolved question still. Overall, the modern ORM framework primarily is focused on low probability, high impact events, i.e. rare occurrences of large losses.

According to the guidelines from the BCBS, the aggregated operational loss process can be modeled as a random sum model [10]. The compound loss process (also

known as the Aggregate Loss) is assumed to follow a random sum {S} expressed as the following:

$$S = \sum_{k=0}^{N} L_k \ , L_k \overset{iid}{\sim} F_\gamma \qquad (2)$$

Here the loss magnitudes (severity) are described by the random independent and identically distributed (i.i.d.) sequence of $\{L_k\}$. This is assumed to follow the CDF F_γ which belongs to a parametric family of continuous probability functions and the counting process N is assumed to follow a discrete counting process. The key point here is that in equation (2) there is an inherent assumption of independence between severity and frequency distributions. Is this a reasonable assumption? This fundamental assumption will be investigated in this dissertation- the modern framework assumes complete independence.

In figure (1) below, it is graphically illustrated how the frequency and the severity process are traditionally thought as "independent" (silo) processes which come together to calculate the annualized aggregate loss [10]. The aggregate loss is used to find the final VaR estimate. In the modern ORM framework, the frequency of losses is estimated along with the severity of the losses using two different statistical distributions. The reason is the frequency comes from a discrete probability mass function (PMF), while the severity is assumed to be continuous and comes from a probability density function (pdf). The LDA approach then advocates combining the severity and frequency via MCS, to compute the annualized (or any other specified time unit) aggregate loss. Once the aggregate loss distribution has been determined, one can estimate the Expected Loss (EL) and also upper quantiles to obtain an estimate of the operational risk VaR. Most financial

institutions tend to estimate the VaR using a confidence level of at least 99.9% (if not 99.99%, which would hold for a 1 in 10,000 year event).

Throughout this dissertation, the term modern is used in front of operational risk. The reason is that there is a more qualitative branch of operational risk which looks at risk events based on the framework of likelihood and impact. The likelihood approach is based on subjectivity and thus is now classified as traditional operational risk management. The fundamental differences in the approach are shown in figure (2) below (taken directly from reference [14]).

Now that the overall processes of frequency and severity have been described, the next stage involves explaining in detail how to measure the frequency and the severity. In practice, most financial institutions have an internal loss data collection exercise which they conduct every year. They break down the loss severity and frequency for each cell in the matrix in table (1) above. As such, an operational risk modeler can fit the losses that were collected for event type *a* and business line *b* - say ($L_1^{a,b}$, $L_2^{a,b}$, ..., $L_N^{a,b}$) - to estimate the severity distribution. Likewise, a similar approach can be used to statistically estimate how often the losses are happening to get the frequency distribution. These are thought of as two distinct data sources. The idea is to combine these two approaches and derive one distinct estimate. An overview of the frequency and severity process is narrated next.

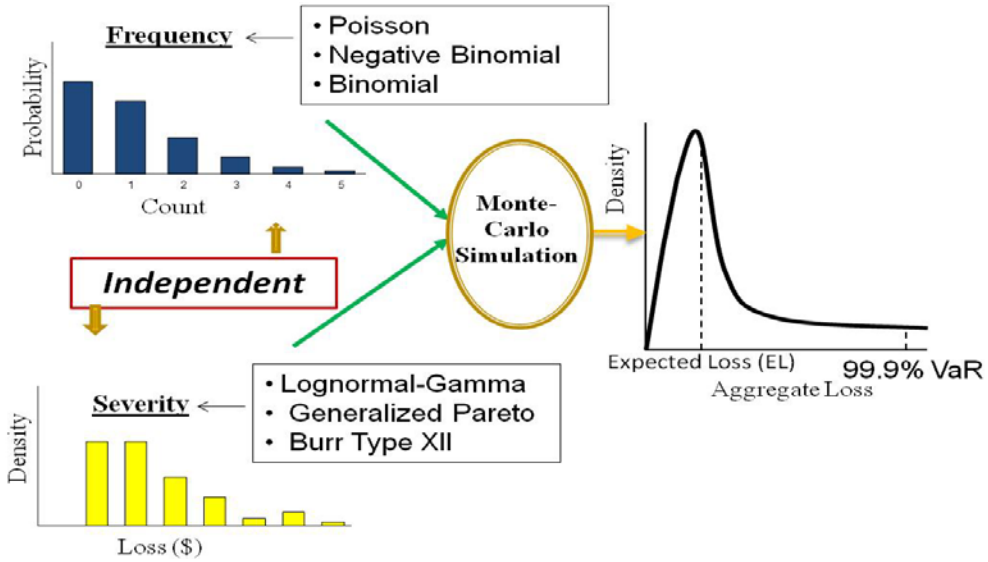


Figure 1 Frequency & Severity Process for VaR Estimation

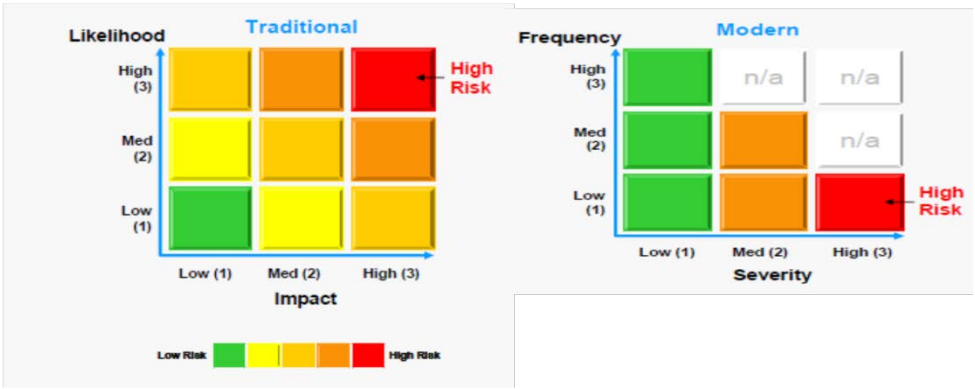


Figure 2 Traditional versus Modern ORM Framework (taken from reference [14])

Three types of discrete distributions, i.e. probability mass functions, are primarily used to model the frequency of losses: (1) Poisson; (2) Binomial; and (3) Negative Binomial distribution (NBD). The Poisson distribution has an interesting characteristic where μ (mean) is equal to the σ (standard deviation) and is characterized by a single parameter, λ. This distribution is the easiest one to model since it involves only statistically estimating a single parameter. The binomial distribution has two parameters,

n (sample size) and *p* (probability). Likewise, the negative binomial distribution also has two parameters, *r* (# failures till success) and *p* (probability). In terms of mean and variance, the binomial distribution is appropriate when $\mu > \sigma$, while the negative binomial distribution is appropriate when $\mu < \sigma$.

In most instances it is possible to determine the frequency distribution used by simply computing the relationship between sample mean and sample variance. Overall, there is not much difference when using different frequency distributions. In figure (3) [10], an overlap analysis of the similarity in the frequency distributions behavior is shown. It is clear to see the overlap when distinguishing between Poisson, Binomial and Negative Binomial distributions. It shows that in most cases there is not a great benefit in spending laborious effort to determine the ideal frequency distribution. A notable exception would be if historical loss data collection exercise of a financial institution clearly exhibits cases when say $\mu > \sigma$ in all cases (empirically). In this case, a binomial distribution should be chosen as a fit for the frequency. Likewise the same would be true if the reverse was observed and then the NBD could be used.

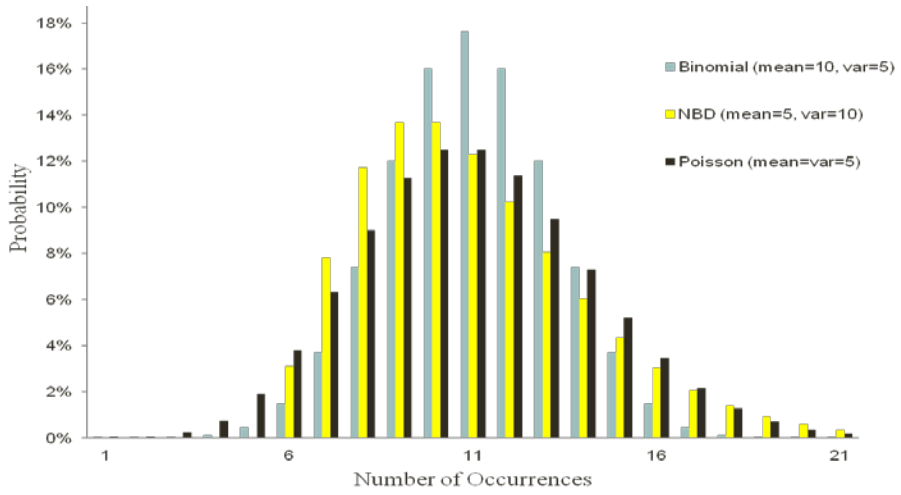


Figure 3 Comparison of different frequency PMFs

Thus the choice and usage of frequency distribution is not of utmost importance in the LDA modeling approach. The choices for the severity distribution types are narrated next.

Unlike in the case of the frequency, there are a plethora of valid statistical distributions that can be used to model the loss severity data. Listed below is a non-exhaustive list (for illustrative purposes only) of sample statistical distributions that are typically used to fit loss data severity: (1) Lognormal -- since losses are always strictly non-negative; (2) Type XII Burr (Burr) distribution; (3) Generalized Pareto distribution (GPD); (4) Weibull; (5) Pareto; (6) Lognormal-Gamma [14]. The current LDA approach suggests that users fit a variety of severity distributions and then use statistical goodness-of-fit (GoF) tests to determine the "optimal" severity distribution. This approach is highly time consuming and can lead to false positive instances.

In figure (4) [14], a typical operational loss data set for the severity of losses is shown. This figure shows that there exists in almost all cases a loss data collection threshold, say T. The reason is that financial institutions may regularly lose small amounts from pennies to even 100 dollars due to rounding or clerical mistakes (teller hands out an extra hundred-dollar bill for example, and this is discovered after customer picks up the deposit). Most institutions will not keep an inventory of these losses in the Loss Data Collection exercise that they undertake. The economic reason that is most often provided is that institutions are interested in larger losses and the ones that tend to occur more frequently. For their internal exercise, financial institutions keep track of the largest losses. That is why in figure (4), the loss severity histogram is shown starting from a finite positive loss and moving forward.

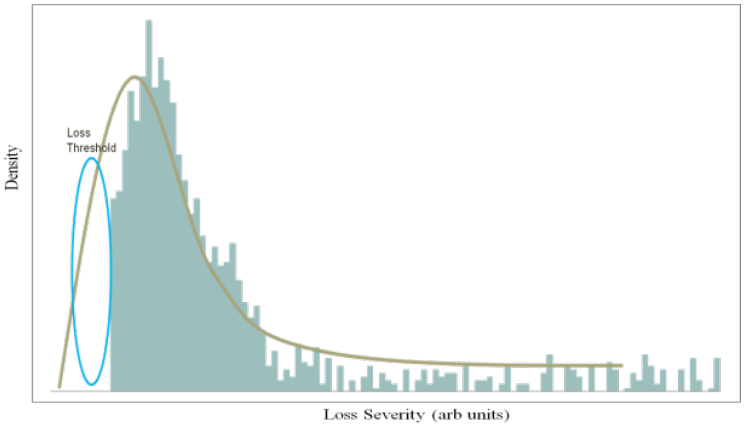


Figure 4 Typical Characteristics of Loss Severity Data

Now the exact approach to the LDA can be visualized accurately as shown below in figure (5) [14]:

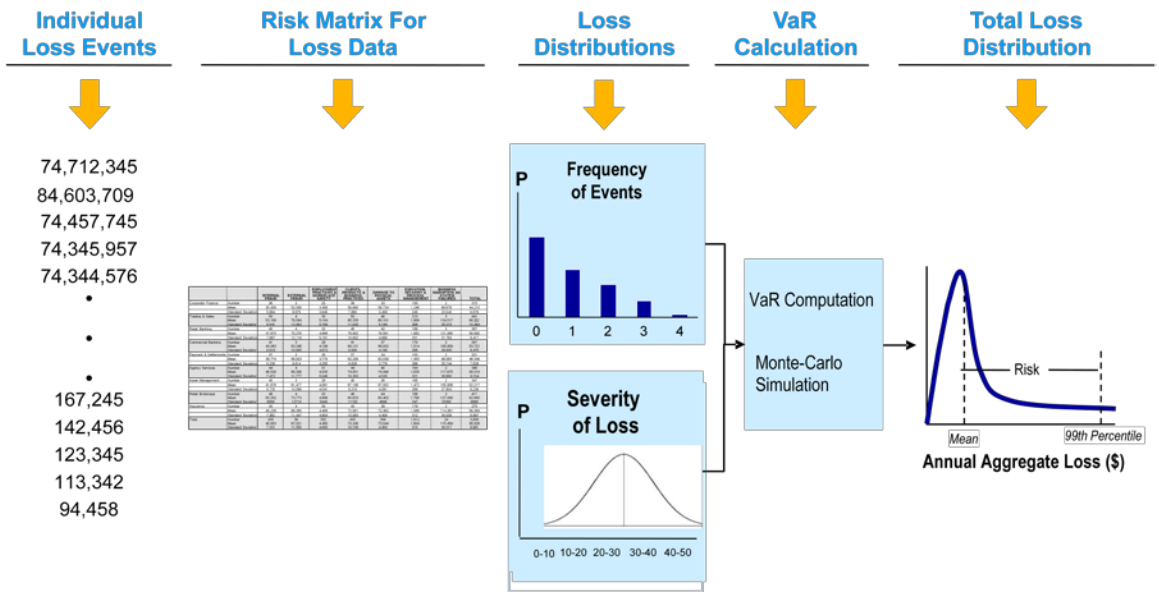


Figure 5 Visual Overview of modern ORM process

Thus so far, the overall modern ORM framework has been described. As mentioned in this section, one key component is the presence/absence of correlation. Are the cells in the matrix in table (1) correlated? How about severity and frequency of loss? These are important questions and thus the dependence structure is narrated next.

1.3 Dependence Structure

The simplest assumption for LDA would be to assume independence. However, from real-life studies [14-19], there is ample evidence that a dependence structure exists between the cells in the matrix in table (1). Thus, it is important to use measures of dependence. The classical measure of dependence is that of the Pearson's correlation coefficient [20].

This measure is a good quantity if it is known that there is a linear relationship between the two variables of interest. However, in most cases of modern ORM, it is well

known that the relationship is highly non-linear in nature. Thus, the classical Pearson's correlation coefficient is inadequate. To model the highly nonlinear dependence structure, one idea is to utilize the mathematical technique of copulas.

In the broad sense, a copula is a mathematical method for modeling the joint distribution (i.e. full dependence structure which includes linear & nonlinear dependencies) of multiple loss events. There are various types of copulas based on the statistical relationship between the joint distributions. Further details of the dependency structure and proposed research in this area are narrated in section 3.3. With this, the overall background for modern ORM has been narrated. Next, the goal is to proceed to describe the research objectives and the unique contributions that come out of this dissertation.

1.4 Overall Research Objectives and Unique Contributions

The research in this dissertation is two-fold: (1) Identify the potential flaws in the fundamental metrics for risk assessment, especially in instances of extreme loss types. Further, in addition to identification, the goal is to develop methodologies for improved risk assessment; (2) Investigate the performance of these methodologies across diverse scenarios (with sufficient uniqueness among them to capture the real-world possibilities) and perform V&V to demonstrate the merit of the newly developed methodologies on both simulated data and real-world datasets relevant to important national problems. For the methodological research, there are two major components: (1) Modeling the Severity and Aggregate Loss distribution; (2) Robust estimation of VaR through distribution-free and parametric based methodologies. For the first question, the current best practices

involve a trial-and-error based methodology of trying *n* number of severity and frequency distributions and then using statistical goodness-of-fit tests and choosing one. This is a time consuming procedure and usually does not yield optimal results. Instead of focusing on this approach, in this dissertation, the goal is to find flexible severity distributions which can fit the loss data and accurately estimate quantiles in the Aggregate Loss (AggLoss) distribution. The contribution of finding this approach is to give practitioners a quick way to estimate the quantiles of the Aggregate Loss distribution without having to rely on GoF tests.

Second, one of the primary questions that this dissertation addresses is the validity of the independence assumption between frequency and severity. This is one of the fundamental premises in modern ORM. My goal in this dissertation is to advance the risk-metric VaR calculation in such a manner where this assumption of independence between severity and frequency is not required. For example, suppose the loss data generation process has clear dependence or correlation between severity and frequency. A natural financial case is where the process exhibits that the higher the loss severity, the less frequent the loss occurs. In that case can a better method (which does not assume independence of frequency and severity) than the classical VaR estimation methodology via MCS be developed and be useful? One of the first goals here is to show through a mathematically rigorous argument that the current approach (known forward as the "classical" approach) has a bias in the estimation of VaR (under specific conditions). This argument is also demonstrated via large scale MCS. After showing the theoretical limitations in the classical methodology, the goal is to develop two types of quantitative

methods which account for cases where there is clear dependency between severity and frequency. The two types should be independent of each other, and one is non-parametric (distribution-free) while the second is a parametric approach. Both of these approaches are verified and validated using large scale MCS. For the empirical portion, the goal of this dissertation is to analyze publicly available data from different sectors which involves risk: (1) Financial Sector Losses; (2) Chemical Spills handled by US Coast Guard; (3) Automobile Accidents from Insurance domain; (4) US Hurricane losses based on natural calamities. One of the primary objectives in this dissertation is to use publicly available data so that the work can be easily peer-reviewed. One of the apparent flaws in most empirical analysis papers in modern ORM is that the data used is highly proprietary and not shared with the general public. This dissertation will only use publicly available data across different and diverse domains. The goal is to show how the new methodologies developed can be useful not just in the financial risk domain, but to demonstrate also in other diverse domains such as insurance, climatology, government loss, to name a few. This discovery process will be useful in exploring potential risks and anomalies of modern ORM which may have been ignored in traditional analysis and open new avenues for practitioners.

CHAPTER 2: LITERATURE REVIEW

While modern ORM is a relatively new academic discipline (compared to market and credit risk which has been extensively studied since the mid-20th century), there is already a plethora of academic and industry based literature in this field. In this dissertation, the literature review will be broken down into three distinct categories: (1) work on modern ORM estimation; (2) work on accurately modeling the dependency structure; (3) empirical applications of modern ORM and their results; and finally (4) the gaps in the current literature. There are several subsections for each of these topics which are narrated next. Overall, this chapter adds to the basic background provided in Chapter 1. This addition shows the current state of research in areas which have direct pertinence to the planned contribution of this dissertation.

2.1 Modern Operational Risk Management

In section 1.2 of the dissertation, there is detailed introductory background information regarding modern ORM based on the LDA approach. This approach is now considered “standard textbook approach” which is used by operational risk practitioners in any respected risk management institution. However, there has been considerable academic and industry based research on improving the LDA approach for modeling quantitative operational risk. For this portion there are three components where there is active research: (1) Modeling severity of losses; (2) Modeling frequency of losses; (3)

Modeling aggregate loss to estimate VaR. The next section begins by narrating the seminal papers in this subfield from which most of the academic research is based upon.

2.1.1 Modeling Severity of Losses

One of the first papers in this area is by de Fontnouvelle et al. (2003) [21] in a technical report where loss data is first used to quantify operational risk. In this paper three types of severity distributions are studied: (1) Generalized Pareto distribution (GPD); (2) Exponential; and (3) Log-logistic distributions. They estimate VaR using each of these distributions and determine that regardless of choice of severity distribution, the operational risk VaR figure is an important consideration for risk management practitioners in addition to market and credit risk VaR.

Later on, Rachev et al. (2006) [22] focus on Peak-over-Threshold (POT) distributions, namely Pareto class family for modeling operational risk. They study the general mathematical properties of these distribution classes and specifically investigate the power tail decay property. They apply their study to a loss data collection exercise from the 2002 BCBS and find estimates for the VaR and ES.

Shortly afterwards, one of the primary research contributions in severity modeling is done by Dutta and Perry (2007) [23] in their technical report. In this technical report, the authors study various severity statistical distributions and advocate the use of the four parameter g&h distribution (this is later described in section 3.1.1). They study the following other severity distributions: (1) Exponential; (2) Weibull; (3) Gamma; (4) Truncated Lognormal; (5) Log-logistic; and (6) GPD. They find in a small dataset that it

is difficult to determine any universal severity distribution which can work for all type of data (heavy, thin, and normal tailed).

Later on in 2008, Ergashev [24] discusses the benefits of using the Lognormal-Gamma (LNG) distribution for modeling operational risk. This paper focuses on methods to estimate the LNG distribution using Markov Chain Monte Carlo (MCMC) methods. In addition to their paper, a research report by Samad-Khan et al. [14] in the Society of Actuaries provided a first glimpse of using LNG for modeling risk.

More recently, there has not been a major push in severity modeling per say. However, Guillen et al. (2011) [25] did add on to the research presented in [22] by modeling the small, moderate and large losses with Pareto Positive Stable (PPS) distribution. The advantage of this distribution is that it can be quickly fit using the Method-of-Moments (MoM) as opposed to the classical Maximum Likelihood Estimation (MLE) based methods. They test their data on a synthetic operational risk data and find faster convergence of results as opposed to MLE.

Overall, the fundamental paper in [23] is primarily used for choosing severity distributions. A clear conclusion is not given by the above authors, but rather "ad-hoc" rules of when one can choose a particular one is described below. Practitioners in general tend to use lognormal distribution to model their financial institutions' severity losses (due to easy of convenience in computation). Even currently, there is no consensus on whether a "universal" severity can be found to fit general operational risk data.

2.1.2 Modeling Frequency of Losses

As shown in the previous section, the general consensus is that the choice of frequency distribution is not as crucial as the choice of severity when computing the final VaR estimate. As expected, this subfield of modeling the frequency has not received as much attention in academic and industry literature. However, there are some important papers for understanding this sub-field which is narrated next.

One of the first papers is by de Fontnouvelle et al. (2007) [26]. In this paper, different PMF's, namely Poisson, Negative Binomial and Binomial, are studied for a specific operational risk loss database from the BCBS. They conclude that the Poisson distribution is a reasonably robust model for their limited real dataset.

Later on, Dahen and Dionne (2010) [27] use external data from a non-standard operational risk consortium (called Algo OpData) and developed a geometric distribution regression model. This technique is commonly used in marketing research data when modeling count related data. They found that the geometric model proved to be working well (in terms of estimation) comparing to the regular PMFs that are used (such as binomial, NBD, Poisson).

Most recently, a couple of important papers on frequency modeling have been well received in the community. Gomes and Gzyl (2014) [28] argue for new statistical techniques to estimate the standard frequency PMFs. Using synthetic data, they have found that the K-Means algorithm and the Expectation Maximization (EM) algorithm do better at estimates for frequency than classical MLE or MoM.

In addition, Badescu et al. (2014) [29] expand on the work in [28]. The authors use an Erlang-based multivariate mixed Poisson distribution to model the frequency.

They assume the loss severities follow the mixture of Erlang distributions and thus are able to successfully develop a closed-form formula for the aggregate loss distribution. They test the computational efficiency and accuracy of this approach using a modified loss data set and claim promising results.

Overall, the specific analysis of types of frequency distributions for modeling losses has not been a great academic focus in this field. The above limited work shows the relatively low interest in pursuing this avenue much further due to the primary fact that Poisson distribution acts as a good benchmark for modeling real-world frequency.

2.1.3 Modeling Distribution of Aggregate Losses and VaR

For this type of methodology, a series of mathematically rigorous papers have been developed. The goal of this subfield is to find optimal ways of computing the aggregate loss and then the VaR. One of the first significant researches conducted in this area is done by Panjer (1981) [30]. The paper (which arose out of interest in insurance applications) provides a recursive algorithm to compute the probability distribution of a compound random variable. The limitation of this approach is that it only works for (a, b, 0) class of distributions. A further detail of this is explained in the section 3.1.1.

After the landmark paper of Panjer [30], the next authoritative work comes from Böcker and Klüppelberg (2005) [31]. In the paper, the authors propose a simple closed form approximation for operational VaR. The key caveat here is that the severity distribution must be that of a heavy-tail type. They argue that mathematically this approximation is valid for large datasets. In order to do this, they compute error bounds for this approximation.

Böcker and Sprittulla (2006) [32] extended the above paper [31] with a minor extension. They assume that the operational loss severity has a finite mean (but no condition on higher moments), and they significantly reduce the approximation error from the paper [31]. In both of these papers [31-32], there are major assumptions made to the loss severity. The advantage of this methodology is that they find a closed-form approximation which can significantly reduce computational time. However, this formula is known to have errors for small-samples.

The work of Jin and Ren (2010) [33] use a different methodology rather than MCS to compute the aggregate loss. Like Panjer's method, they do not rely on Monte Carlo to compute the VaR. Instead they use the Fast Fourier Transform (FFT) and extend the concept of exponential tilting for univariate FFT as previously shown. They describe how to attack several numerical issues such as aliasing, and floating-point representation error. They argue that the FFT method is as reliable as using Panjer's method. However, they only show that it works for certain cases of severity distributions.

Opdyke et al. (2012) [34] describe the limitations of using MLE as an approach for severity estimation and the VaR estimation. This paper argues that if there is even a modest violation of the i.i.d. assumption in the data, the MLE estimates are highly non-robust. They provide rigorous simulation studies to show their argument.

Extending the above paper, Opdyke (2013) [35] mathematically argues that the current MC method of estimating VaR overestimates the VaR due to Jensen's Inequality. Specifically, they show that when using LDA approaches for any of the heavy-tailed or skewed severity distributions, all unbiased estimators of the severity distribution

parameters generate biased capital estimates due to Jensen's Inequality. In this paper it is argued that VaR always appears to be a convex function of these severities' parameter estimates because the severity quantile being estimated is usually large and heavy-tailed. They show for a class of severity functions, the reduced-Bias Capital Estimator can correct for this. They show its applicability for several severity functions such as lognormal, Burr, and Gamma.

2.2 Dependency Structure Analysis

In all of the aforementioned papers, there is an implied assumption of independence between the severity and the frequency component. However, when modeling real life operational risk losses, in almost all cases there is some correlation or dependency structure. There have been many papers which have modeled this structure. In this literature review, the most promising and relevant papers are described below. These fall under two classes: (1) Copula analysis; (2) Global Correlation structure. This is narrated next.

2.2.1 Distribution Dependent Analysis via Copula

As previously described in section 1.3, there are several standard methods of measuring correlation or dependency structure. One of the most popular methods in industry since the 1990's is the copula approach [36]. It has been widely used in market and credit risk modeling. However, it isn't until the early 2000's that this methodology is first introduced for modern ORM.

Frachot et al. (2004) [37] provide a first glimpse in the academic literature on addressing both the frequency and severity of losses individual correlations across the

matrix cell (as shown in table (1)). They calculate an upper bound of the aggregate loss correlation for both high frequency low severity data, and low frequency high severity data in a limited operational risk dataset. They showed that the correlation between two aggregate losses is typically below 5%, and thus open a wide score for large diversification effects. This is much larger than anticipated by the BCBS.

Di Clemente and Romano (2004) [38] did some of the primary work on Extreme Value Theory (EVT) along with copula analysis for insurance loss data. They utilize a MCS in order to determine the loss distribution and calculate both VaR and ES. They assume a severity distribution of lognormal in the center and left tail, and then model the right tail using POT type distributions. They use both a *t*-Copula and Gaussian copula to model across three business lines. They find that the Gaussian copula is more stable for the VaR estimates.

The use of non-standard copulae (i.e. Gaussian or Student's *t* distribution based) is first introduced by Böcker and Klüppelberg (2008) [39]. The authors here invoke the concept of Lévy copula to model the dependence structure of operational risk loss events. They derive first order approximations for the correlations with this copula assuming a heavy-tailed GPD severity distribution. They conclude from simulations that it is not worthwhile to estimate precise frequency correlations between different cells (i.e. in table (1)), but all effort should be made for accurate modeling of the severity. This was based on large MCS based studies of their methodology.

Fantazzini et al. (2008) [40] study observed correlations in operational losses in the Operational Risk Exchange database. They implement a Gaussian and *t*-Copula to

model the correlations across three business lines testing several severity distributions along with several different PMFs for the frequency. They observe that the most stable estimates for VaR came from using the Gamma distribution for severity and Poisson distribution for frequency. They also reported some computational efficiency issues in using MLE based methods.

Abate et al. (2009) [41] have extended the work in [34] to generalize how using EVT for severity and standard copulas can be used to model operational risk across several business lines. They assume that the severity data can be modeled using POT distribution (such as the GPD). Then they show that using VaR as a risk measure may lead to an inaccurate estimation (based on MCS study). They show that there are stability issues (in terms of computation) when using the GPD based distribution. They find that the *t*-Copula is useful in modeling their synthetic dataset which they obtained.

The aforementioned works all use a frequentist approach to finding and estimating the optimal copula, but the work of Valle (2009) [42] first introduces in the operational risk context the usage of the Bayesian copulae. They develop a MCMC model to estimate both the Bayesian Gaussian and Bayesian *t*-Copula. They use uninformative priors and then use MCMC method of Gibbs sampler to compute the posterior distribution for each case. They apply this methodology to a sample financial institutions' loss data across five business lines. They find evidence of good MCMC convergence and argue that the Bayesian approach can be superior due to pitfalls in frequentist MLE procedures.

Böcker and Klüppelberg (2010) [43] extend their work from [31] to derive approximations for a closed form VaR for a multivariate operational risk context. They add the use of Lévy copula to model the dependence structure. They work out a first order approximation to this VaR estimate and compute error bounds. They try several different types of severity distributions for a fixed Poisson frequency and compute the error approximations in each case. Overall they argue that the Lévy copula is useful (theoretically speaking) for multivariate VaR analytical approximations.

Finally, Brechmann et al. (2013) [44] introduces the idea of pair copula constructions for modeling both the frequency and severity components together. Using this approach, the authors are able to model the dependence of the seven-dimensional distribution of the losses per event type (i.e. in table (1)), in terms of pairwise dependence and tail dependence. They analyze a specific financial institution's operational loss data. Their results show that there is a significant decrease in the required economic capital comparing to the standard BCBS approach of summing up the VaRs across each cell in the matrix. The bivariate copulas used involve the Gaussian and Clayton based copulas.

Overall, there is a rich literature in using copulas in modern ORM framework. In most cases, there has been the use of standard Gaussian and or *t*-Copula with some exceptions to using Lévy copula. In almost all cases this has been conducted from a frequentist rather than a Bayesian approach. Besides the copula, there has been some work on developing a non-linear global correlation coefficient. This is narrated next.

2.2.2 Global Correlation Coefficient

One of the strengths of using a correlation coefficient is that one can give a simple metric to describe the dependency structure. For copulas, one is given a full distribution. However, there has been work since the late 1990's on developing a global correlation coefficient. This measure has some technical issues when computing from a real-life dataset.

The work of Darbelley (1998) [45] in a technical report explains the concept of global correlation coefficient which can be defined from the mutual information between two random variables. The global correlation coefficient, λ, is a function of two random variables X and Y is defined as $\lambda(X, Y) = (1- \exp(-2I(X, Y)))^{1/2}$ where I(X, Y) is the mutual information between two random variables X and Y. Some interesting mathematical derivations from mutual information are shown in the paper.

Next, the question is how to measure the mutual information from numerical data. Moddemeijer (1999) [46] wrote an interesting paper for this application. Specifically, in the case of two signals with dependent observations, he derives a statistic to estimate the variance of the histogram based mutual information estimator. There are several statistical flaws in this methodology such as high bias and variance. To alleviate this, he proposes some corrections based on histogram modeling.

Dionisio et al. (2004) [47] describe how mutual information can be used to measure dependency in nonlinear time series. This paper has some theoretical first order approximation estimates for the estimator in question. They study two different nonlinear time series for an empirical application and find good convergence property.

The same authors extend the above paper in 2006 [48] to develop an entropy based independence test. They develop a non-parametric approach by conducting a new test of independence among distributions based on Shannon's entropy. They compute the critical values of their new test through simulation. In addition, they apply their new metric to two time series data from market risk, and find good convergence properties.

Kaskov et al. (2004) [49] extend the work in [47-48]. The authors here develop two classes of improved estimators for mutual information from samples of random points with a bivariate joint distribution function. They argue that the histogram binning method has computational weakness, and instead advocates the use of entropy estimates from K-nearest neighbor distances. They prove that this measure is data efficient adaptive and have minimal bias. They perform a test on two times series market risk data sets, and find good convergence with known results.

These methodologies were first introduced to operational risk very recently by Li et al. (2014) [50]. They apply the global correlation coefficient using the variance-covariance matrix approach for computing VaR. They apply this approach to operational risk losses from three major Chinese banks and compute the VaR of each bank. They find that the VaR estimates for the overall sector for the Chinese regulators are lower than required based on the global correlation coefficient. The dataset is proprietary and not shared to the public.

Overall, significant work has been developed for the global correlation coefficient. However, there are still some unresolved questions on how to accurately

compute it for estimating VaR. The variance-covariance approach has a few assumptions which may not be true for many datasets.

2.3 Application of Modern ORM to Real-World data: Empirical Research

The focus of most of the previous papers has been on extending theory of modern ORM and then potentially applying them to a relevant dataset in the proprietary loss exchange database. However, there have been several papers which focus on simply the empirical applications of modern ORM. The vast majority have been in financial application although there is an interesting application to chemical spills which is narrated at the end.

One of the first relevant papers which simply focus on the applications for financial institutions modern ORM is that of Aue and Kalkbrener (2006) [51]. This paper presents the LDA model for the Deutsche Bank for their internal data. They specifically work on scenario analysis along with the traditional VaR estimation using a lognormal severity and Poisson frequency. The interesting aspect of this paper is that they perform sensitivity analysis on how the LDA approach estimates can be different from scenario analysis.

Next, Chapelle et al. (2008) [52] analyze the implication of the LDA approach through a study of four categories of two business lines and two event types from real-life data from a large financial institution (name is withheld). They analyze the data using a mixed model by calibrating one distribution for describing "normal" losses (namely under $100 million) and another POT distribution for larger losses. They use a NBD for modeling the frequency of these losses across all matrix cells. They also uniquely

perform some sensitivity analysis to see how the impact of modern ORM on bank profitability.

Cope and Antonini (2008) [53] perform extensive empirical research on the observed correlations among operational losses in the loss exchange database. They use Spearman's rho and Pearson's correlation coefficient, ρ, to estimate these correlations. They do not employ any copula based approach. They find important implications for diversification benefits when aggregating losses across different operational risk categories.

Cope et al. (2009) [54] extend the previous work but study the property of data sufficiency in internal and external operational risk data. Here they investigate what minimum thresholds are necessary for collected internal operational risk data and their consequences in severity and frequency modeling. They find that a minimum threshold of 500 data points is necessary for good MLE convergence.

Colombo and Desando (2008) [55] use a purely scenario based approach for computing VaR for the Italian bank Intesa Sanpaolo. This is based on finding a loss distribution from expert opinions of five institutional managers. They compare their findings from a scenario based approach to that of the standard LDA approach assuming lognormal severity and Poisson frequency. They find that for this particular bank, the scenario approach is more efficient (in time) in estimating their economic capital requirements.

Finally, a recent paper in 2014 by Liu and Cortes [56] shows an interesting application of modern ORM for the five top banks in Taiwan. This paper specifically

demonstrates that by applying risk managerial strategies, banks can improve their performance based on a risk-adjusted return on capital (RAROC). They apply a stochastic frontier approach and perform shock absorption to these systems and observe the potential operational risk loss fallout. This approach is very different from LDA but it uses concepts form market risk to analyze operational risk fallout in Taiwanese banks.

The previous papers all focus on empirical applications to financial institutions. There is one notable exception in that there have been applications to chemical spills. The idea here is how much economic capital should an industry hold to mitigate itself from the worst type of spills? A 2007 paper by Meel et al. [57] extensively analyze the accident database up till 2006 from the National Response Center (NRC) of the US Coast Guard. This paper is the first to perform an operational risk assessment using frequency and severity to compute an ES. Instead of using MC approach, they use the FFT to estimate the VaR and ES figures.

Overall, there have been a plethora of application based papers in modern ORM. The vast majority consist of applications to banking and financial losses. However, a notable exception is found in modeling losses from chemical spills along with looking at aviation data, and market risk from a loss perspective.

2.4 Gaps in the Literature & Primary Motivation for Dissertation

There are several areas in the literature where contributions can be made. The fundamental over-arching question that is not answered in the literature is the following: Is it fair to assume the independence of severity and frequency? In the financial world, it is well-known that in any capitalist system, severe losses should not happen with high

frequency. The reason is that any competent institution will develop KRIs to ensure that large losses do not occur frequently which would wipe them out. This similar logic should exist for government based loss models. The current best practices assume the independence of severity and frequency, and then moves forward in the research. Can methodologies be developed which are agnostic to the relationship between severity and frequency and robustly estimate the VaR? This dissertation challenges the aforementioned fundamental premise by developing two new quantitative methodologies to address this issue. Two additional specific areas where there is a potential gap in the literature which this dissertation addresses are narrated next.

For the severity modeling portion there are still the following outstanding questions:

(1) Can a "flexible" severity distribution be used to model severity losses?

(2) How well does the flexible severity distribution play in accurately estimating the VaR?

With regards to the dependency structure (between frequency and severity) there are several unresolved questions:

(1) There is still no agreement on whether some copula functions are more powerful than others in modeling dependence between frequency and severity? Can a mixture distribution be used?

(2) Can distribution-free approaches be used to model this correlation?

(3) Do these new approaches with dependency structure perform at least as well (if not hopefully better) as the classical approach to estimating VaR?

Finally for the empirical application of modern ORM methodology section, almost all current applications have been either to financial institutions' individual losses or data from a proprietary exchange. The problem with this approach is that the data is always proprietary and cannot be validated through peer-review. What is missing in the literature is the usage of publicly available loss datasets across different domains. To address this gap, one of the fundamental components in the empirical analysis of this dissertation is using publicly available data from different domains such as financial losses, spills monitored by government agency, insurance losses and natural calamities. This way the work can be peer-reviewed by anyone with internet access. In addition, there is potential for showing the pertinence of the new methodologies across different disciplines.

CHAPTER 3: METHODOLOGY

This chapter describes the methodology used in the study of quantification of modern ORM. In this chapter and each subsection, the current state-of-the-art methodology is narrated at first. Afterwards, the contribution that this dissertation makes is described. This way there is a delineation of what the contribution of this dissertation is in direct tandem to what is currently used as the state-of-the-art.

3.1 Modeling Frequency & Severity Component

3.1.1 Current State-of-the-Art

This section expands on the introductory materials described in chapter 1. To begin, the fundamental definition of frequency refers to the number of events that occur within a given time period (this is defined from physics). In the modern ORM context, frequency is a stochastic parameter and thus is expressed through a probability mass function. Therefore, the domain for the frequency distribution is a subset of all non-negative integers, i.e. $\mathbb{Z}$. Therefore, in theory, any discrete statistical distribution on non-negative integers can be potentially used to model frequency. The details of the three most commonly used distributions are given next.

The most commonly used PMF is that of the Poisson. The Poisson probability distribution is given by the following list of probabilities on non-negative integers:

$$P(N = k \mid \lambda) = \frac{\lambda^k}{k!} e^{-\lambda},\ k = 0, 1, 2, \ldots$$

where ($\lambda > 0$) is a parameter of the distribution. Thus the Poisson distribution is a single parameter distribution. It is straightforward to show that the mean and the variance of the Poisson distribution are both equal to λ. The estimation of the parameter λ is taken in the form of either method of moments or using maximum likelihood estimation, both of which yield the same estimator.

The next popular one is the negative binomial distribution. The negative binomial distribution is given by the following probabilities on the set of $\mathbb{Z}$:

$$P(N = k \mid r, p) = \frac{\Gamma(r+k)}{k!\Gamma(r)} p^r (1-p)^k,\ k = 0, 1, 2, \ldots,\ r > 0,\ 0<p<1$$

where $\Gamma(r) = (r - 1)!$ and r (# of successes till failure) and p (probability) are two parameters of the distribution.

The NBD is more flexible than the Poisson distribution, since it has two parameters. Hence, it is possible to compute the mean and the variance of this distribution as the following:

$$E(N) = r(1-p)/p,\ Var(N) = r(1-p)/p^2$$

Therefore, the variance is greater than the mean, a distinctive feature that differentiates it from the Poisson distribution. Thus, if empirical evidence suggests that there is excessive variability in the frequency, then the natural conclusion is using the negative binomial distribution rather than the Poisson distribution should be more fruitful.

An interesting fact about the NBD is that it can be modeled as the Poisson distribution mixed with a prior distribution on its mean parameter, i.e. λ. Thus, if one

introduces volatility to the mean parameter (λ) of the Poisson distribution, one can statistically obtain the negative binomial distribution [14]. This can be done using a Gibbs sampler methodology [20]. Suppose there is a prior distribution of a gamma pdf with Gamma(mean=1, variance=β). Then suppose one draws a random value, say g, from the aforementioned gamma probability distribution. Then, conditional on the given g (which was just drawn), draw an integer N from the Poisson distribution with mean of λg. The resultant integer value, will follow the NBD with mean of λ and variance of $\lambda(1+\beta)$ [14].

This way of analyzing the NBD has a clear implication for MCS. In other words, the NBD data can be simulated by first drawing a number g from the Gamma distribution and then drawing a random number from the Poisson distribution with mean of λg, where λ is the mean frequency. In this case, it is more convenient (for interpretability) to re-parameterize the NBD in terms of mean (λ) and the variance of the frequency and to determine the parameter β of the gamma distribution accordingly. The mean and the variance of the frequency can be estimated by calculating the sample mean and sample variance from internal data [14].

Next, the binomial distribution is of importance for modeling frequency. The binomial probability distribution is given by the following equation:

$$P(N = k \mid p, n) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k},\ k = 0, 1, 2, \ldots, n, \text{ and } 0 \le p \le 1,$$

where n (sample size) and p (probability) are the two parameters of this distribution.

The mean and variance for the binomial distribution is the following:

$$E(N) = np, \text{Var}(N) = np(1-p)$$

Thus, the binomial distribution is applicable when frequency has a variance smaller than the mean. Next, the details of the severity component of the losses are discussed.

Operational loss data is almost always collected above a certain dollar (loss amount) threshold. This makes it difficult to model loss severity, because, except for a very few well-defined distributions that are developed to model truncated data sets (such as the Pareto distribution), most loss severity distributions are developed to model data sets where there was no threshold (where the data was collected from the ground up). These distributions cannot be used in their original form to fit truncated data. A discussion on fitting truncated data is presented next.

The primary method of estimation is maximum likelihood estimation for fitting empirical data. This is primarily because of the statistical property of consistency, asymptotic normality and efficiency [20]. MLE is a process used to fit empirical data to a theoretical distribution which is selected *a priori*. If one is given a pre-specified theoretical distribution, MLE is used to find the set of parameters that have the maximum probability (i.e. likelihood) of describing the empirical data set. Generally, the likelihood function is the PDF, but where loss data are censored or truncated at a threshold, say T, an adjustment is required to incorporate the missing mass of data. In order to accommodate truncated data, the likelihood function must be modified to describe the conditional likelihood (conditioned on the threshold T). For left truncated data, one can

define a normalized PDF by taking the original PDF, and dividing it by the probability of the data being drawn from above the threshold, T, as shown likelihood (LH) below [14]:

$$\mathrm{LH}(\theta \mid T, x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n \mid T, \theta) = \frac{\prod_{i=1}^{n} \mathrm{PDF}(x_i|\theta)}{[1-\mathrm{CDF}(T|\theta)]^n}$$
$$\hat{\theta}_{\mathrm{MLE}} = \mathrm{argmax}_{\theta \in \Theta} \mathrm{LH}(\theta \mid T, \mathrm{data}) \quad (3)$$

where θ is the parameter vector, the x_i refers to the actual empirical data (usually transformed into a log scale, i.e. in logarithmic values), and T is the threshold value above which the data is collected. The goal is to use numerical constrained optimization techniques (since the parameter space is almost always bounded) to estimate the value of $\hat{\theta}_{\mathrm{MLE}}$ from the entire theoretically possible parameter space.

In general, the values of a probability density function are typically small (less than 1), and the product of many such terms will quickly render the joint LH value to a level that is computationally indistinguishable from 0. Thus a better objective function to be maximized is the log-likelihood function (LLH) for data set (x_1,..., x_n), which transforms the above LH into the following equation (4):

$$\mathrm{LLH} = \log(\mathrm{LH}) = -n \times \log[1 - \mathrm{CDF}(T|\theta)] + \sum_i \log[\mathrm{PDF}(x_i|\theta)] \quad (4)$$

Next, the details of fitting the severity distribution are provided. There is a plethora of statistical distribution functions which can fit the severity. Ideally one requires that the distribution contain positive support. The following non-exhaustive list, which is presented next, provides some commonly used distributions and their PDFs. The work in [23] includes four parameter distributions which are not used by industry due to lack of consistency in MLE fits. Either the PDF or CDF (based on mathematical simplicity) for each of these distributions are mathematically described below.

- Lognormal (LN)

The PDF is the following:

$$f(x|\,\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right) = \frac{\phi(z)}{\sigma x}\text{, where } z = \frac{\log(x)-\mu}{\sigma}$$

for all $x \geq 0$, where ($\mu \in \mathbb{R}$, $\sigma > 0$) are the domain of the parameters of the distribution and $\phi(\bullet)$ is the density function for the normal distribution with mean μ and variance σ^2.

- Lognormal-Gamma (LNG)

The PDF is the following:

$$f(x \mid \mu,\sigma,k) = \int_0^\infty \gamma(y|\,k)\phi(x|\mu,\sigma^2 \times y)dy$$

where $\gamma(\bullet)$ is the density function for the Gamma distribution. Similar to lognormal, the data consists of $x \geq 0$, and where ($\mu, k \in \mathbb{R}$, $\sigma > 0$) is constant parameter of the distribution. The value of k is the kurtosis of the distribution (where normal has a value of 3).

A more intuitive way to write the LNG distribution is that X is distributed as LNG given the following:

$$Y \equiv \log(X) = \mu + \sigma\sqrt{\gamma}Z$$

where μ, σ are the mean and standard deviation, and Z is a standard normal distribution and γ is a random variable from a Gamma(1/k, k) distribution.

- Weibull

The PDF is the following:

$$f(x \mid a, b) = \frac{b}{a}\left(\frac{x}{a}\right)^{b-1} \exp\left(-\left[\frac{x}{a}\right]^{b}\right)$$

for all $x \geq 0$, where ($a > 0$ is scale parameter and $b > 0$ is shape parameter).

- Three parameter type XII Burr distribution (Burr)

The PDF is the following:

$$f(x \mid \phi, \tau, \alpha) = \left(\frac{\alpha\tau}{\phi}\right)\left(\frac{x}{\phi}\right)^{\tau-1}\left\{1 + \left(\frac{x}{\phi}\right)^{\tau}\right\}^{-(\alpha+1)}$$

for all $x \geq 0$, where ($\phi, \tau, \alpha > 0$). It is a very stable distribution family that can express a wide range of distribution shapes. The Burr distribution subsumes cases of many commonly known distributions such as gamma, lognormal, beta distributions (excluding U-shaped instances). Also, there are some instances where several compound distributions can be algebraically shown to match a Burr distribution. An example of the compounding process is compounding a Weibull distribution with a gamma distribution for its scale parameter which will result in algebraic equivalence to a Burr distribution [20]. An alternate example of this is compounding an exponential distribution with a gamma distribution for its rate parameter, $1/\mu$, can be shown to have the same characteristics as a Burr distribution [59]. Taking two asymptotic limiting cases of the first two parameters of the Burr distribution yields a Pareto and a Weibull distribution [20].

The aforementioned Burr distribution is well-known to fit a wide range of empirical data from diverse areas such as hydrology, failure modeling, climate change, to financial risk loss modeling [59]. The power of this distribution is that it can capture a broad range of data (heavy to thin tails) across its parameter set.

- Generalized Pareto distribution

The CDF is the following:

$$F(x\,|a,b) = \begin{cases} 1 - \exp\left(-\frac{x}{a}\right), \text{if } b = 0 \\ 1 - \left(1 + \frac{bx}{a}\right)^{-1/b}, \text{otherwise} \end{cases}$$

where a > 0 (also known as scale), and b (sometimes called shape) is a real number.

- Pareto

The CDF is the following:

$$F(x \mid a, T) = \begin{cases} 1 - \left(\frac{T}{x}\right)^{a}, \text{if } x > T \\ 0, \text{otherwise} \end{cases}$$

where a > 0 is a constant parameter of the distribution and T is the threshold.

- Gamma

The PDF is the following:

$$f(x \mid a, b) = \frac{1}{b^{a}\Gamma(a)} x^{a-1} \exp\left(\frac{-x}{b}\right)$$

where (Γ is the usual gamma function, i.e., $\Gamma(r) = (r-1)!$, $b > 0$ is the scale parameter, and $a > 0$ is the shape parameter). Now with this knowledge in mind, there are several research questions. This is narrated next.

3.1.2 Research Contribution in this Area

There is a consensus in that Poisson is a reasonable choice for frequency. However, there is absolutely no consensus on whether there is a flexible severity distribution which can be applied to reliably model severity. Obviously it is not possible to find a single distribution which can best fit a universal severity data set. However, it is well-known statistically that higher parameterized distributions are more "flexible" in being able to fit more complex data types. However that does not imply that using distributions with the highest number of input parameters will be better, because then there is the issue of the “curse of dimensionality” and possibility of over-fitting. The higher the number of parameters to estimate, the higher the amount of data is needed for accuracy. Thus a practical question is the following: Can a "flexible" severity distribution be found which can fit various types of loss severity data?

As mentioned in the literature review section 2.1.1, the only paper (technical non-peer reviewed work) is in [23] where a serious investigation on what severity distribution should be used is discussed. The aforementioned paper introduces the four-parameter statistical distribution, g&h distribution, to the modern ORM framework. However, this paper does emphasis the weakness of using this distribution for limited size datasets. This distribution is defined as the following for the four non-negative parameters (a, b, g, h):

$$f(Z \mid a, b, g, h) = a + b*(\exp(gZ)-1)*(\exp(hZ^2/2)/g$$

The goal for this part of the research will be to test the robustness of high parameter distributions along with extreme-value type distributions of GPD, Pareto (for example) for their robustness when modeling different types of losses. Thus the research goal will be to use robust large-scale MCS studies and test the sensitivity of the final VaR, EL figures (given a known frequency distribution). Can a single distribution model data reasonably well for different types of loss severity? This dissertation successfully answers this question in chapter 4.

3.2 Modeling Aggregate Loss distribution

3.2.1 Current State-of-the-Art

One of the main concepts that is crucial in modern ORM is computing not only the VaR, but also the Expected Loss (EL) and Unexpected Loss (UL) from the aggregate loss distribution (i.e. combining frequency and severity). The plot in figure (6) [14] illustrates this pictorially.

The VaR here can be computed and then subtracting the EL gives the important concept of UL. The UL is one of the crucial practical points for any operational risk manager. A natural question that arises next is the following: Is it possible to easily compute these properties for the Aggregate Loss distribution? In general, it is very difficult if not impossible to exactly describe the aggregate loss distribution, in closed form (e.g., in terms of the distribution function), even when the frequency and severity distributions are both well known. In general, numerical methods such as MCS, Panjer and FFT are used to calculate the aggregate distribution and its relevant statistics (such as the EL, UL and VaR).

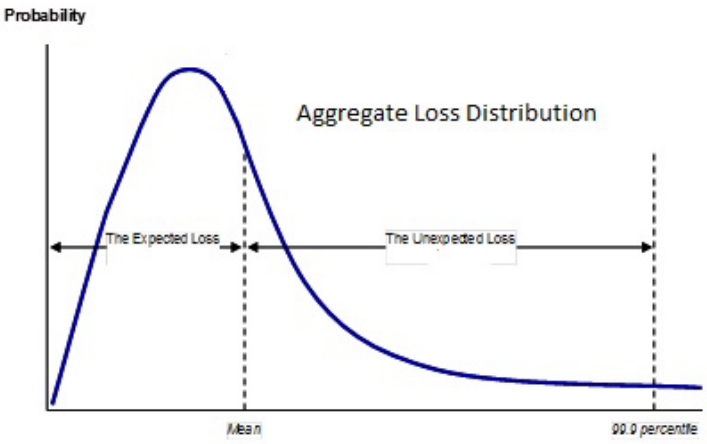


Figure 6 Pictorial Representation of EL and UL from Aggregate Loss

There is a relatively simple relation between the mean of Aggregate Loss (S from equation (1)) and the frequency (N) and severity (L), as mathematically shown below.

$E(S) = E(N) \times E(L)$, where $E(\bullet)$ represents the Expectation operator.

This is due to the fact that one can apply the law of iterated expectations [20] as follows: $E(S) = E[E(S|N)]$. However, for each N fixed, one can adjust the previous to the following:

$$E(S|N) = \sum E(L) = NE(L).$$

$$\rightarrow \; E(S) = E[NE(L)] = E(L) \times E(N)$$

Similarly, it is possible to show that Variance (Var) of the Aggregate Loss can be expressed as the following:

$Var(S) = Var(N)*E(L)^2 + E(N)*Var(L)$, where $Var(\bullet)$ represents Variance.

Now the above can be shown to be true from the following argument invoking at first the fundamental definition of variance, i.e. $Var(S) = E(S^2) - E^2(S)$.

On the other-hand, $E(S^2) = E[E(S^2|N)]$ from the law of iterated expectations.

Now, $E(S^2|N) = Var(S|N) + E^2(S|N) = N*Var(L) + N^2*E^2(L)$.

Hence, $E(S^2) = E[N*Var(L) + N^2*E^2(L)]$

$= E(N)*Var(L) + E(N^2)*E^2(L)$

$= E(N)*Var(L) + [Var(N) + E^2(N)] *E^2(L)$

$= E(N)*Var(L) + Var(N)*E^2(L) + E^2(N) *E^2(L)$.

Therefore,

$Var(S) = E(N)*Var(L) + Var(N) *E^2(L) + E^2(N) *E^2(L) - E^2(S)$

$= E(N) *Var(L) + Var(N)*E^2(L)$

The last statement is true since $E^2(S) = E^2(N) *E^2(L)$ (this is just squaring the mean). Note one of the key assumptions here is the independence between frequency (N) and severity (L).

The next portion shows to compute the curve in figure (6). There are several ways to perform this calculation: (1) Monte Carlo; (2) Panjer's Algorithm; (3) Fourier Transform (FT).

Monte-Carlo Simulation

In almost all cases, an *exact* analytical solution is impossible to obtain for the Aggregate Loss distribution. Therefore, the conceptually best way to move forward is to use simulation. This is done by applying the calculation of the aggregate loss distribution

modeled within the modern ORM framework via two steps: (1) event frequency, and (2) individual loss severity. The algorithm is presented below next [10]:

1. Determine the severity distribution and optimal parameters from MLE fits.
2. Determine optimal frequency distribution parameters.
    2.1 Set a high simulation threshold value N (minimum of 10,000).
3. Set the iteration counter t = 1.
4. Draw a random number of losses from the Frequency distribution, n.
5. Given the number n, draw n losses, L_1, L_2,...,L_n from the severity distribution.
6. Sum all n of the severity losses and call that value S_t (Aggregate Loss for time t).
7. Increment iteration counter t = t+1, and go to step 4.
8. Iterate till t hits the maximum iteration threshold, N.
9. $\{S_1, S_2, ...,S_N\}$ is the Aggregate Loss distribution. Next, empirically compute the mean, and 99.9 percentiles to get EL and VaR.

Simulation is also flexible to incorporate new logical steps in loss generation that could easily change the resultant distribution. For example, one may want to know what impact the purchase of new insurance contracts (i.e. insuring for extreme losses) would have on the aggregate loss distribution and subsequently the VaR estimate. Instead of having to refit the after-insurance severity distribution (likely to cause model inconsistency and inaccuracy compared to the severity without insurance), the simulation process can incorporate an additional step for insurance coverage after each event loss is generated (i.e. using a logical check if insurance is applied or not). The resultant aggregate loss distribution would automatically have the insurance component built-in,

and the result can be compared to the case where insurance is not purchased. Also in this case the difference due to model inconsistency caused by severity distribution fitting is minimized since the insurance component is checked as a logical step during the simulation [14].

Panjer's Algorithm

Panjer's recursion is currently widely used in the insurance industry but not much in the modern ORM and risk management community. Panjer's algorithm [30] has the following goal to compute the distribution of S, the aggregate loss distribution:

$$S = \sum_{i=1}^{N} L_i$$

where *N* is a discrete random variable distributed on non-negative integers and $\{L_i\}_i^N$ is a sequence of i.i.d. random variables and *L* represents the individual losses. One of the requirements for this approach is that *L* is independent of *N*, and that *L* comes from a *discrete* distribution. Now most severity distributions are continuous (like lognormal, Burr, Weibull, etc.). Therefore, the algorithm first discretizes the entire loss region, namely, the half interval [0, +∞), into loss buckets with equal bin width so that they can be alternatively discretely numbered such as 0, 1, 2, 3, …, n and so on (similar to how a histogram is created). Depending on the multiplier that one chooses to associate with each bucket, each integer number may refer to a specific monetary value (this is purely for labeling purposes). As an example, if the multiplier is $1,000, then bucket 1 would correspond to $1,000 and bucket 2 would correspond to $2,000, etc. Therefore, the numerical implementation of Panjer's algorithm first requires that the severity

distribution be assigned as discrete probabilities for each loss bucket 0, 1, 2, ..., n. For practical purposes, there is a maximum bucket number n, after which all the buckets will have zero probability. So in order to use Panjer's algorithm one must discretize the values of the severity distribution (round to the nearest integer). Next the mathematical formulation is described.

To begin this process it is important to define the following:

$p_n = P[N = n]$ for n=0,1,2,...

$f_k = P[L_i = k]$ for k=0,1,2,...,

$f_k^n = P[L_1 + L_2 + \cdots + L_n = k]$ for n=1,2,3,... and k=0,1,2...

$g_k = P[S = k]$ for k=0,1,2....

Now, Panjer's recursion is defined as the following for (*a*, *b*, 0) class of distributions such as the following:

$$p_k = P[N = k] = \left(a + \frac{b}{k}\right) * p_{k-1}, \quad k \geq 1$$

for some $a + b \geq 0$ and p_0 is determined by the fact that $\sum_{i=0}^{\infty} p_i = 1$.

Now define the Probability Generating Function (PGF) of a random variable *N* as $PGF_N(z)$. Mathematically this is known to be the following:

$PGF_N(z) = P_N(z) = E[z^N]$ (for which the expectation exists)

Now,

$$g_0 = PGF_N(f_0)$$

$$g_0 = p_0 * \exp(f_0 b) \text{ if } a = 0,$$

$$g_k = \frac{1}{1 - af_0} \sum_{j=1}^{k} \left(a + \frac{b * j}{k} \right) * f_j * g_{k-j}$$

For a Poisson random variable for example, a = 0, b = λ, and $p_0 = \exp\{-\lambda\}$. Using the above recursion formula, it is possible to obtain the distribution of the aggregate loss function. But the question that naturally arises is how large is it necessary for *k* to be from the above equation? In general, it is necessary to choose a *k* such that the following will hold true [30]:

$$\sum_{i=0}^{k} g_k > 99.99\%$$

Since this algorithm has nested loop, the complexity of calculation for Panjer's recursion is of $\mathcal{O}(n^2)$, where n is the number of buckets for which aggregate probabilities are desired. Therefore, the higher the percentile, the more computational time is required for calculation. The actual time required may also depend on the granularity of loss buckets, mean frequency, and the overall computational speed [14].

On the surface, Panjer's algorithm does not require the fitting of severity distribution, since empirical data can be directly turned into loss bucket probabilities (equivalent to bin ranges in histogram modeling). This is very problematic, especially for operational loss events where loss data is not detailed enough (granular) for a fine lattice partition (based on data collection). Moreover, a lack of detailed tail-end descriptions (for example, probabilities are all zero for buckets beyond the maximum loss that has been so far collected) almost always leads to serious underestimation of the high percentile capital (i.e. upper tail) for aggregate losses [14]. Thus, in reality an operational risk

modeler may have to go back and use MLE to fit a theoretical distribution for severity, which is then discretized into loss bucket probabilities. In this case, one still needs to tackle the issues of how to fit a distribution to truncated loss data (same as in standard MLE modeling). As an alternative, Panjer's algorithm adapted to absolutely continuous severity distributions is also available; but its numerical implementation still needs discretized integration [14].

Perhaps the more serious disadvantage of Panjer's method is that it is inappropriate for calculating diversified total risk exposure based on multi-unit loss distributions (even if correlations are assumed to be zero among business units), since it was not originally designed [30] to deal with multivariate distributions (i.e. correlations). One way to get around this problem is to lump all the data from multiple units into one single/large pool and apply Panjer's method accordingly, to get a diversified aggregate loss/VaR estimate [14]. However, this requires re-specification of the severity and frequency distributions for the pooled data set and model inconsistency may well dominate the true benefit of diversification when compared to the VaR estimates for individual business lines [14].

Fourier Transform (FT)

This method allows the density of a probability distribution to be turned into its associated Characteristic Function (CF). To explain this concept, it is important to begin expand equation (1) as the following for the random sum S (from equation (2)):

$$S = L_1 + \cdots + L_N$$

$$\Rightarrow F_S(l) = P(S \leq l)$$

$$= \sum_{n=0}^{\infty} P(N = n) * P(S \leq l \,|N = n)$$

Now, it is important to calculate the PGF of S, $P_S(z)$, as the following:

$$P_S(z) = E[z^S] = \sum_{n=0}^{\infty} E[z^{L_1+\cdots+L_N}|N = n]P[N = n]$$

$$= \sum_{n=0}^{\infty} E\left[\prod_{j=1}^{n} z^{L_j}\right] P[N = n]$$

$$= \sum_{n=0}^{\infty}[P_L(z)]^n P[N = n]$$

$$= E\left[P_L^N(z)\right] = P_N[P_L(z)]$$

The above is true assuming the independence of L_1, ..., L_N for a fixed n. Now the CF always exists for any random variable (unlike the moment generating function). The CF of a random variable S, $\varphi_S(z)$ is defined as the following [58]:

$$\varphi_S(z) = E[\exp\{izS\}] = P_N[\varphi_L(z)], \text{ where } i = \sqrt{-1} \text{ (imaginary number).}$$

Now for any continuous function f(l), the FT is the mapping:

$$\tilde{f}(z) \propto \int_{-\infty}^{\infty} f(l)\exp(izl)dl$$

While the Inverse Fourier Transform (IFT) can be used to recover the original function as the following:

$$f(l) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{f}(z)\exp(izl)dz$$

Now for the distribution of losses, f(l) is the PDF of the severity losses. The FT can be optimized by using Fast Fourier Transform (FFT) instead of a discrete FT. The algorithm is the following in the modern ORM context [58]:

(1) First discretize the loss severity into the following steps:

$f_L(0), f_L(1), ..., f_L(n-1)$, where $n = 2^r$ for some integer r, and *n* is the number of points desired in the distribution of aggregate loss, S, $f_S(l)$.

(2) After discretization of the losses, the FFT is computed to obtain $\varphi_L(z)$, i.e. the CF of the discretized distribution.

(3) Using the PGF transformation of loss frequency distribution, transform the values obtained in Step (2) to calculate the value of $\varphi_S(z) = P_N[\varphi_L(z)]$.

(4) At this stage if one computes the Inverse FFT, the aggregate losses, S is obtained.

(5) Sort the aggregate losses, S, and compute VaR.

Note that the FFT requires the loss region to be discretized into 2^n lattice points with corresponding (discrete) probabilities given for any type of loss severity. For the same reasons as mentioned in the previous section on Panjer's algorithm, it is highly undesirable to use the empirical distribution directly out of the actual loss data for severity. This is especially true when the purpose of this exercise is to calculate the VaR estimate for the aggregate distribution subject to a very high percentile level (e.g. 99^{th} quantile or above). Therefore, a continuous distribution has to be fitted (using MLE) and then discretized for the lattice points [14].

One important note is that the speed of FFT procedure is typically faster (depending the precision magnitude necessary) than the Panjer's recursion. For example, the computational complexity of FFT is of $\mathcal{O}(\log_2 n)$, where 2^n is the number of lattice points (i.e., loss buckets for the modern ORM context) for the loss region [14]. Notice

that unlike Panjer's algorithm; here the set of lattice points always represents the full loss severity spectrum (rather than the region up to a certain percentile). Since the actual number of lattice points required is only a function of desired level of precision, the percentile level of VaR to be calculated does not directly impact the computational speed of the final VaR estimation. However, it does have an indirect impact of the speed, to the extent that finer lattice partition may be necessary for a robust estimate of the upper tail VaR regions.

Single-Loss Approximation Formula

There is a closed form formula for the VaR for certain special cases of frequency and severity in the classical case [31-32]. This formulation assumes that the severity follows a heavy-tailed distribution. Statistically speaking, this means that the severity comes from a class of subexponential distributions, meaning that their tails decay slower than any exponential tail. The precise mathematical definition of a subexponential distribution is that the tail of the sum of *n* subexponential random variables has the same order of magnitude as the tail of the maximum random variable among them [31]:

$$\lim_{x \to \infty} \left(\frac{P(X_1 + X_2 + \cdots + X_n > x)}{P(\max(X_1, X_2, \cdots, X_n) > x)} \right) = 1 \text{ for all } n \geq 2$$

In otherwords, the sum of *n* i.i.d. severities is most likely to be large because of one of the terms being large. Another way to look at this phenomenon is that severe overall losses are due to a single large loss rather than the consequence of accumulated small independent losses [31]. Now assuming that the frequency distribution is either Poisson or NBD with distribution f(N), then under specific weak regularity conditions [32], it has

been shown that the following holds true for the VaR as the quantile (κ), $\kappa \to 1$ given severity distribution is a subexponential with cdf of F:

$$\mathrm{VaR}(\kappa) = \mathrm{F}^{-1}\left\{1 - \frac{1-\kappa}{\mathrm{E[N]}}(1 + o(1))\right\} \qquad (5)$$

Specifically if, F is LN(μ, σ) and N is Pois(λ) then the single-loss approximation following equation (5) from [31] above is the following (where Φ^{-1} is the inverse of a standard normal cdf):

$$\mathrm{VaR}(\kappa) = \exp\left\{\mu - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\lambda}\right)\right\} \qquad (6)$$

A more refined approximation to the above equation is calculated in [31]. This approximation does a refinement by mean correction as the following if F is LN(μ, σ) and N is Pois(λ):

$$\mathrm{VaR}(\kappa) = \exp\left\{\mu - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\lambda}\right)\right\} + (\lambda - 1)\exp\left[\mu + \frac{\sigma^2}{2}\right] \qquad (7)$$

These approximations are used in the following section to give a mathematical argument as to why the classical methodology is biased in extreme cases of disparity between frequency and severity.

3.2.2 Research Contribution in this Area

One of the major areas of contribution in this dissertation is using a distribution free methodology to better estimate the VaR quantity. The classical methodology (the current state-of-the-art) makes a fundamental assumption: independence of frequency and severity. This dissertation starts by addressing this question: Can one come up with a universal method which can address cases where frequency is independent of severity

and the reverse case, i.e. frequency is dependent of severity? The current classical methodology assumes the case where frequency is independent of severity. To motivate this question, a theoretical argument is presented which demonstrates that in certain circumstances the classical methodology systematically either underestimates or overestimates the VaR. To further buttress the mathematical argument, large scale MCS studies are performed to show the weaknesses in the classical methodology. After demonstrating the shortcomings of the classical methodology both theoretically and computationally, a new approach is developed using distribution free methods from cluster analysis. Afterwards, a parametric-based approach is shown using copulas to estimate VaR without using the assumption of independence between severity and frequency. These are narrated next.

3.2.2.1 Mathematical Argument Showing Bias in Classical Methodology

One of the first questions that this dissertation asks is if there is any need to modify the classical approach. Recall again from the previous subsection that the classical approach posits that frequency is independent of severity. The limits of the classical approach are tested next using a mathematical argument.

Suppose that there are two independent compound Poisson processes (S_X, S_Y) which have two components: (1) Loss Severity (X, Y) and (2) Loss Frequency (N, M) such that:

$$S_Y = \sum_{i=1}^{N} Y_i$$

$$S_X = \sum_{j=1}^{M} X_j$$

where $X \sim LN(\mu, \sigma)$, $Y \sim LN(\mu/k, \sigma)$, while $N \sim Poisson(\lambda)$ and $M \sim Poisson(\lambda/l)$, where $k, l > 1$.

Then a compound Poisson process, S_Z, is defined such that:

$$Z = \begin{cases} Y_i & \text{with } p = \dfrac{m}{m+n} \\ X_i & \text{with } p = \dfrac{n}{m+n} \end{cases}$$

$$S_Z = \sum_{i=1}^{N} Y_i + \sum_{j=1}^{M} X_j$$

where S_Y denotes the random sum (i.e. aggregate loss) for Y, and S_X denotes the random sum for X, and S_Z denotes the random sum for X+Y. Note that $P\{S_Z \leq z\} = F(S_z)$ is by definition the Aggregate Loss distribution and $VaR(S_Z \mid \kappa) = F^{-1}(Sz \mid \kappa)$ is defined for a right tail quantile, κ. S_x represents the "High Severity/Low Frequency" Regime and S_Y represents the "Low Severity/High Frequency" regime. For this analysis, the operational risk is divided into two distinct cases. In Case (I), the mean severity parameter (M) is much larger than the mean frequency parameter (Λ), while the variance parameter of the severity (Σ) is fixed and very small. In Case (II), the mean severity (M) parameter is much smaller than the mean frequency parameter (Λ) (and the variance (Σ) is fixed and small). This is formalized in the following manner.

<u>Case (I): Mean Severity >> Mean Frequency</u>

In this case, the average severity is significantly larger than the mean frequency. From an analytic point of view, this means that $M \sim \mathcal{O}(\exp(\Lambda))$, and fix Σ to be small (low variance model) such that $0 < \Sigma < 1$, and M to be large, i.e. $M >> 1$ and the mean

frequency is at least one event per time unit, i.e. $\Lambda \geq 1$. For simplicity sake, Σ is set to be 0.8. The severity is assumed to follow a lognormal distribution while the frequency follows a Poisson distribution.

Case (II): Mean Frequency >> Mean Severity

In this case, the average frequency is significantly higher than the mean severity. Specifically, suppose that $\Lambda \sim \mathcal{O}(\exp(\exp(M)))$, and let fix Σ to be small (low variance model) such that $0 < \Sigma < 1$ and Λ to be large, i.e. $\Lambda >> 1$. The severity is assumed to follow a lognormal distribution while the frequency follows a Poisson distribution.

Now how can one approximate the truth? It is argued that in Case I, the severity dominates the frequency, the tail region of S_X will approximate the $VaR(S_Z \mid k)$, where $\kappa \rightarrow 1$ in Case I. Likewise for Case II, the tail region of S_Y approximates the $VaR(S_Z \mid \kappa)$, where $\kappa \rightarrow 1$ since frequency dominates severity. This argument is shown formally and then validated through simulation in the next section.

To begin the theoretical argument for Case (I), the Single Loss Approximation [31] as shown in equation (6) is implemented. This gives the following set of equations:

$$\mathrm{VaR}\left(S_X \mid \kappa, \mu, \sigma, \frac{\lambda}{L}\right) \approx \exp\left[\mu - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\frac{\lambda}{L}}\right)\right] \tag{8}$$

$$\mathrm{VaR}\left(S_Y \mid \kappa, \frac{\mu}{k}, \sigma, \lambda\right) \approx \exp\left[\frac{\mu}{k} - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\lambda}\right)\right] \tag{9}$$

$\mu \sim \mathcal{O}(\exp(\lambda))$, so $\exp(\lambda)$ is substituted for μ in the RHS of (8).

$$\mathrm{VaR}\left(S_X \mid \kappa, \mu, \sigma, \frac{\lambda}{L}\right) \approx \exp\left[e^{\lambda} - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\frac{\lambda}{L}}\right)\right] \tag{10}$$

$$\mathrm{VaR}\left(S_Y \mid \kappa, \frac{\mu}{k}, \sigma, \lambda\right) \approx \exp\left[e^{\frac{\lambda}{k}} - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\lambda}\right)\right] \tag{11}$$

$$\Phi^{-1}(x) \propto \text{erf}^{-1}(x)$$

The Maclaurin series expansion of $\text{erf}^{-1}(x)$ (this is the error function from approximating the normal integral) centered on x = 0 has the following property:

$\text{erf}^{-1}(x) \propto x + \mathcal{O}(x^3)$.

$\text{erf}^{-1}(1/\lambda) \propto 1/\lambda + \mathcal{O}(1/\lambda^3)$.

Therefore, the RHS of (10) and (11) can be approximated as the following:

$$\text{VaR}\left(S_X \mid \kappa, \mu, \sigma, \frac{\lambda}{L}\right) \approx \exp\left[\underbrace{e^{\lambda} - \sigma\frac{L(1-\kappa)}{\lambda}}_{A}\right] \qquad (12)$$

$$\text{VaR}\left(S_Y \mid \kappa, \frac{\mu}{k}, \sigma, \lambda\right) \approx \exp\left[\underbrace{e^{\frac{\lambda}{k}} - \sigma\left(\frac{1-\kappa}{\lambda}\right)}_{B}\right] \qquad (13)$$

Next, examining the rate of change of A and B (from equations (12)-(13) from above) with respect to λ:

$$\frac{\partial A}{\partial \lambda} = e^{\lambda} + \sigma\frac{L(1-\kappa)}{\lambda^2}$$

$$\frac{\partial B}{\partial \lambda} = \frac{e^{\frac{\lambda}{k}}}{k} + \sigma\frac{(1-\kappa)}{\lambda^2}$$

Now observing the region of the extreme right tail, i.e. $\kappa \to 1$, then $\lim_{\kappa\to 1} \sigma\frac{L(1-\kappa)}{\lambda} = 0$ for $\lambda > 1$ and L fixed and finite. Likewise, $\lim_{\kappa\to 1} \sigma\frac{(1-\kappa)}{\lambda} = 0$ for $\lambda > 1$ can be computed. So, the rate of change of A is larger than B with respect to λ because k > 1. Also the dominant term in both A and B is their first term. Since exp(λ) always dominates over exp(λ /k) given that k > 1 (easily seen through Taylor expansion of exp), the $\text{VaR}(S_X \mid \kappa)$ > $\text{VaR}(S_Y \mid \kappa)$ as one approaches the upper right quantile. Case (II) is proceeded to next

and that usage of the updated VaR approximation [32] from equation (7) is adapted since frequency portion dominates in this instance.

Case (II) using Single Loss Approximation with Mean Correction

$$\mathrm{VaR}\left(S_X \mid \kappa, \mu, \sigma, \frac{\lambda}{L}\right) \approx \exp\left[\mu - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\frac{\lambda}{L}}\right)\right] + \left\{\frac{\lambda}{L} - 1\right\}\exp(\mu)\exp\left(\frac{\sigma^2}{2}\right) \quad (14)$$

$$\mathrm{VaR}\left(S_Y \mid \kappa, \frac{\mu}{k}, \sigma, \lambda\right) \approx \exp\left[\frac{\mu}{k} - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\lambda}\right)\right] + \{\lambda - 1\}\exp\left(\frac{\mu}{K}\right)\exp\left(\frac{\sigma^2}{2}\right) \quad (15)$$

For Case (II), $\lambda \sim \mathcal{O}(\exp(\exp(\mu)))$, so $\exp(e^{\mu})$ is substituted for λ in the RHS of (14)-(15). In addition, the approximation of Φ^{-1} is also used next (as used in the above Case (I)).

$$\mathrm{VaR}\left(S_X \mid \kappa, \mu, \sigma, \frac{\lambda}{L}\right) \approx \exp\left[\mu - \sigma\left(\frac{L(1-\kappa)}{\exp(e^{\mu})}\right)\right] + \left\{\frac{\exp(e^{\mu})}{L} - 1\right\}\exp(\mu)\exp\left(\frac{\sigma^2}{2}\right) \quad (16)$$

$$\mathrm{VaR}\left(S_Y \mid \kappa, \frac{\mu}{k}, \sigma, \lambda\right) \approx \exp\left[\frac{\mu}{k} - \sigma\left(\frac{1-\kappa}{\exp(e^{\mu})}\right)\right] + \{\exp(e^{\mu}) - 1\}\exp\left(\frac{\mu}{K}\right)\exp\left(\frac{\sigma^2}{2}\right) \quad (17)$$

Now the 2nd terms of the RHS of (16) and (17) are important terms since they are much larger in magnitude. Since $\exp(e^{\mu}) >> \exp(\mu)$, the RHS of (17) is larger than the RHS of (16) (since L > 1). So the $\mathrm{VaR}(S_Y \mid \kappa) > \mathrm{VaR}(S_X \mid \kappa)$.

Thus, it has been shown that for Case (I) the "True" VaR, i.e. $\mathrm{VaR}(S_Z \mid \kappa)$ is approximated as the following:

$$\lim_{\kappa \to 1} \mathrm{VaR}(S_Z \mid \kappa) \simeq \mathrm{VaR}\left(S_X \mid \kappa, \mu, \sigma, \frac{\lambda}{I}\right) \approx \exp\left[\mu - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\frac{\lambda}{I}}\right)\right]$$

Likewise, for Case (II) the "True" VaR, i.e. $\mathrm{VaR}(S_Z \mid \kappa)$ is approximated as the following:

$$\lim_{\kappa\to 1} \mathrm{VaR}(S_Z \mid \kappa) \simeq \mathrm{VaR}\left(S_Y \mid \kappa, \frac{\mu}{k}, \sigma, \lambda\right) \approx \exp\left[\frac{\mu}{k} - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\lambda}\right)\right]$$

Next, the focus of the analysis is on the approximation for $\mathrm{VaR}(S_Z \mid \kappa)$ from the classical method. The VaR estimation from the classical method will be compared to both Case (I) and Case (II). To begin, the frequency component is analyzed for the classical method. The classical method for the frequency component can be described as the following:

$$E[N] + E[M] = \lambda + \lambda/l = (l+1)\lambda/l$$

Next, severity component is examined.

In this case, the following can be expressed mathematically:

Z = *p**X+*(1-p)**Y, and the goal is to express (μ_Z, σ_Z^2) in terms of (μ, σ).

E[Z] = *p**E[X]+(*1*-*p*)*E[Y]

Var[Z] = p^2*Var[X] + *(1-p)*^{2}*Var[Y]

Suppose the proportion *p* is the same for both X and Y, (examining the simplest case with equal proportions and scaling factor consisting of k, l are set to 2). By Fenton-Wilkinson Approximation, the central moment matching [60] is computed next for estimating a single lognormal distribution for Z as the following:

E[Z]= $\exp\{\mu_Z + 0.5*\sigma_Z^2\}$

Var[Z]= $(\exp\{\sigma_Z^2\}-1)*\exp(2\mu_Z + \sigma_Z^2)$

$$\exp\{\mu_Z + 0.5*\sigma_Z^2\} = p * \exp\left\{\frac{2\mu+k\sigma^2}{2k}\right\} + (1-p) * \exp\left\{\frac{2\mu+\sigma^2}{2}\right\}$$

$$\exp\{\mu_Z + 0.5*\sigma_Z^2\} = \exp\left\{\frac{2\mu+\sigma^2}{2}\right\} * \left\{(1-p) + p * \exp\left[\frac{\mu}{k} - \mu\right]\right\}$$

$$\Rightarrow \mu_Z = \frac{2\mu+\sigma^2}{2} + \log\left\{(1-p) + p * \exp\left[\frac{\mu}{k} - \mu\right]\right\} - 0.5\sigma_Z^2$$

Looking at the simplest (algebraically speaking) case where the scaling factors are fixed (i.e. k,l=2), then

$$\mu_Z \propto \mu + 0.5\sigma^2 + \log\left\{1 + \exp\left[\frac{-\mu}{2}\right]\right\} - 0.5\sigma_Z^2 \quad (18)$$

Next, variances (equal proportions) are computed:

$$(\exp\{\sigma_Z^2\}\text{-}1)*\exp(2\mu_Z + \sigma_Z^2) \propto$$
$$\underbrace{[\exp\{\sigma^2\}\exp\{2\mu + \sigma^2\} - \exp\{2\mu + \sigma^2\}]}_{\text{Var}[X]} + \underbrace{[\exp\{\sigma^2\}\exp\left\{\frac{2\mu}{2} + \sigma^2\right\} - \exp\left\{\frac{2\mu}{2} + \sigma^2\right\}]}_{\text{Var}[Y]}$$

$$\Rightarrow \text{LHS} = [\exp\{\sigma^2 + \mu\}] * [\overbrace{\exp\{\sigma^2 + \mu\} - \exp\{\mu\} + \exp\{\sigma^2\} - 1}^{\gamma}]$$

Taking logarithm on both sides yields the following:

$$\Rightarrow 2\mu_Z + \sigma_Z^2 + \log\{\exp[\sigma_Z^2] - 1\} = [\sigma^2 + \mu] + \log(\gamma) \quad (19)$$

Now substitute RHS of (18) into LHS of (19):

$$\Rightarrow 2\mu + \sigma^2 + 2\log\left\{\underbrace{1 + \exp\left[\frac{-\mu}{2}\right]}_{\rho}\right\} + \log\{\exp[\sigma_Z^2] - 1\} = [\sigma^2 + \mu] + \log(\gamma)$$

$$\Rightarrow \log\{\exp[\sigma_Z^2] - 1\} = -\mu - \log(\rho^2) + \log(\gamma)$$

$$\Rightarrow \log\{\exp[\sigma_Z^2] - 1\} = \log\left(\frac{\exp\{-\mu\}*\gamma}{\rho^2}\right)$$

$$\Rightarrow \sigma_Z^2 = \log\left[\frac{\exp\{-\mu\}*\gamma}{\rho^2} + 1\right]$$

where $\gamma = [\exp\{\sigma^2 + \mu\} - \exp\{\mu\} + \exp\{\sigma^2\} - 1]$ and $\rho = \left\{1 + \exp\left[\frac{-\mu}{2}\right]\right\}$

$$\mu_Z = \mu + 0.5\sigma^2 + \log\left\{1 + \exp\left[\frac{-\mu}{2}\right]\right\} - 0.5 * \log\left[\frac{\exp\{-\mu\}*\gamma}{\rho^2} + 1\right]$$

For the classical estimate of VaR, the work of [31] is used, namely equation (6) below:

$$\text{VaR}\left(S_Z^{\text{Classical}} \mid \kappa, \mu_z, \sigma_Z, \frac{3\lambda}{2}\right) \approx \exp\left[\mu_z - \sigma_z \Phi^{-1}\left(\frac{1-\kappa}{\frac{3\lambda}{2}}\right)\right]$$

where

$$\mu_Z = \mu + 0.5\sigma^2 + \log\left\{1 + \exp\left[\frac{-\mu}{2}\right]\right\} - 0.5 * \log\left[\frac{\exp\{-\mu\}*\gamma}{\rho^2} + 1\right]$$

$\sigma_Z^2 = \log\left[\frac{\exp\{-\mu\}*\gamma}{\rho^2} + 1\right]$, where $\gamma = [\exp\{\sigma^2 + \mu\} - \exp\{\mu\} + \exp\{\sigma^2\} - 1]$ and

$$\rho = \left\{1 + \exp\left[\frac{-\mu}{2}\right]\right\}$$

Now that the classical VaR estimation is complete, it is important to compare this with both Case (I) and Case (II). For Case (I) where Mean Severity >> Mean Frequency,

$$\overbrace{\exp\left[\mu_z - \sigma_z\Phi^{-1}\left(\frac{1-\kappa}{\frac{3\lambda}{2}}\right)\right]}^{\text{"Classical"}} \quad ? \quad \overbrace{\exp\left[\mu - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\frac{\lambda}{2}}\right)\right]}^{\text{"Truth"}}$$

For Case (II) where Mean Frequency >> Mean Severity

$$\overbrace{\exp\left[\mu_z - \sigma_z\Phi^{-1}\left(\frac{1-\kappa}{\frac{3\lambda}{2}}\right)\right]}^{\text{"Classical"}} \quad ? \quad \overbrace{\exp\left[\mu - \sigma\Phi^{-1}\left(\frac{2(1-\kappa)}{\lambda}\right)\right]}^{\text{"Truth"}}$$

The procedure for comparison between the classical method and Case (I) is shown next.

$$\mu_Z = \mu + 0.5\sigma^2 + \log\left\{1 + \exp\left[\frac{-\mu}{2}\right]\right\} - 0.5 * \log\left[\frac{\exp\{-\mu\}*\gamma}{\rho^2} + 1\right] \qquad (20)$$

where $\gamma = [\exp\{\sigma^2 + \mu\} - \exp\{\mu\} + \exp\{\sigma^2\} - 1]$ and

$$\rho = \left\{1 + \exp\left[\frac{-\mu}{2}\right]\right\}$$

$$\log\left[\frac{\exp\{-\mu\} * \gamma}{\rho^2} + 1\right] = \log\left[\frac{\exp\{-\mu\}\,[\exp\{\sigma^2 + \mu\} - \exp\{\mu\} + \exp\{\sigma^2\} - 1]}{\left\{1 + \exp\left[\frac{-\mu}{2}\right]\right\}^2} + 1\right]$$

$$\Rightarrow \text{ LHS} = \log\left[\frac{[1+\exp\{-\mu\}][\exp\{\sigma^2\}-1]}{\left\{1+\exp\left[\frac{-\mu}{2}\right]\right\}\left\{1+\exp\left[\frac{-\mu}{2}\right]\right\}} + 1\right]$$

Now as μ is large, $\frac{[1+\exp\{-\mu\}]}{[1+\exp\{-\frac{\mu}{2}\}]} \to 1$ & $\frac{[\exp\{\sigma^2\}-1]}{\left\{1+\exp\left[\frac{-\mu}{2}\right]\right\}} \to [\exp\{\sigma^2\} - 1]$.

$$\log\left[\frac{\exp\{-\mu\} * \gamma}{\rho^2} + 1\right] \approx \log\left[\underbrace{\{\exp\{\sigma^2\} - 1\}}_{x} + 1\right]$$

Next, the Taylor approximations for log(1+x) and exp{x} are used for simplification.

$$\log(1 + x) = x - \mathcal{O}(x^2) \quad \text{for } |x| < 1$$

$$\sigma_Z^2 = \log\left[\frac{\exp\{-\mu\}*\gamma}{\rho^2} + 1\right] \approx \log\left[\underbrace{\{\exp\{\sigma^2\} - 1\}}_{x} + 1\right] \approx \exp\{\sigma^2\} - 1 \approx \sigma^2 \qquad (21)$$

$$\log\left\{1 + \underbrace{\exp\left[\frac{-\mu}{2}\right]}_{x}\right\} \approx \exp\left[\frac{-\mu}{2}\right] \approx 1 - \frac{\mu}{2} \qquad (22)$$

Substituting (21) and (22) into (20), the following is obtained

$$\mu_Z = \mu + 0.5\sigma^2 + \log\left\{1 + \exp\left[\frac{-\mu}{2}\right]\right\} - 0.5 * \log\left[\frac{\exp\{-\mu\}*\gamma}{\rho^2} + 1\right]$$

$$\Rightarrow \mu_Z \approx \mu + \frac{\sigma^2}{2} + 1 - \frac{\mu}{2} - \frac{1}{2}(\sigma^2) = 1 + \frac{\mu}{2}$$

$$\sigma_Z^2 \approx \sigma^2$$

Therefore,

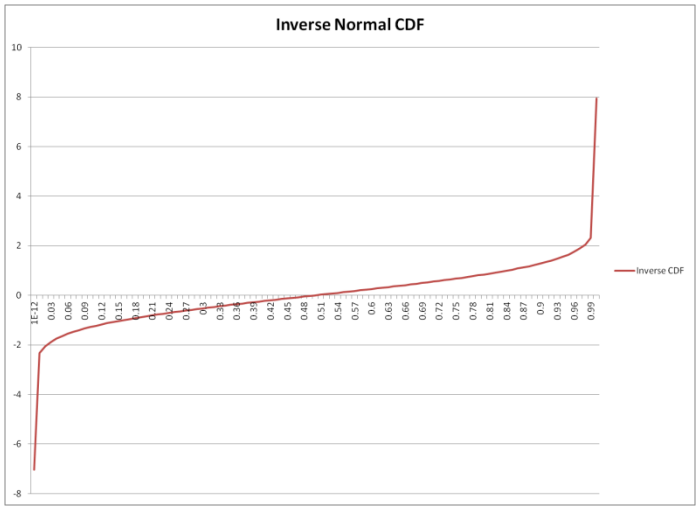
$$\overbrace{\exp\left[\mu_z - \sigma_z\Phi^{-1}\left(\frac{1-\kappa}{\frac{3\lambda}{2}}\right)\right]}^{\text{"Classical"}} \approx \exp\left[\frac{\mu}{2} + 1 - \sigma\Phi^{-1}\left(\frac{2(1-\kappa)}{3\lambda}\right)\right]$$

$$\exp\left[\frac{\mu}{2} + 1 - \sigma\Phi^{-1}\left(\frac{(1-\kappa)}{\frac{3\lambda}{2}}\right)\right] \text{ ? } \exp\left[\mu - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\frac{\lambda}{2}}\right)\right]$$

Now in this scenario, since $\mu >> \lambda$, and $\kappa \rightarrow 1$, $\Phi^{-1}(\cdot)$ decreases much more slowly than linearly, and thus $\frac{\mu}{2} + 1 - \sigma\Phi^{-1}\left(\frac{(1-\kappa)}{\frac{3\lambda}{2}}\right) \lesssim \mu - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\frac{\lambda}{2}}\right)$

$$\Rightarrow \exp\left[\frac{\mu}{2} + 1 - \sigma\Phi^{-1}\left(\frac{(1-\kappa)}{\frac{3\lambda}{2}}\right)\right] \lesssim \exp\left[\mu - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\frac{\lambda}{2}}\right)\right]$$

Why is it argued that $\Phi^{-1}(\cdot)$ component doesn't matter? The graphical representation of the classical inverse normal cdf is shown below.



Sample plot of Inverse Normal CDF

$\frac{dy}{dx}[\Phi^{-1}(x)] = \frac{1}{\phi(\Phi^{-1}(x))}$, where recall that ϕ is the standard normal pdf.

$$\Phi^{-1}(x) \propto \text{erf}^{-1}(x)$$

The Maclaurin series expansion of $\text{erf}^{-1}(x) \propto x + \mathcal{O}(x^3)$.

So the rate of increase of $\left[\frac{\mu}{2} + 1 - \sigma\Phi^{-1}\left(\frac{(1-\kappa)}{\frac{3\lambda}{2}}\right)\right]$ is determined by $\frac{\mu}{2}$ versus

that of $\left[\mu - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\frac{\lambda}{2}}\right)\right]$ is determined by μ.

Therefore, the VaR obtained from Classical Method asymptotically is smaller than the "True" Method in Case (I). This analysis is easily extendible for non-equal probabilities (p) and other higher values of scaling factor (l, k).

The procedure for comparison between the classical method and Case (II) is shown next. Using the approximation in [32],the adjusted VaR analysis is implemented from equation (7):

$$\mathrm{VaR}\left(S_Q^{\mathrm{Mean}} \mid \kappa, \mu, \sigma, \lambda\right) \approx \overbrace{\exp\left[\mu - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\lambda}\right)\right] + (\mathrm{E[Frequency]} - 1)\mathrm{E[Severity]}}^{\text{"Classical"}}$$

$$\exp\left[\mu_z - \sigma_z\Phi^{-1}\left(\frac{1-\kappa}{\frac{3}{2}\lambda}\right)\right] + \left(\frac{3}{2}\lambda - 1\right)\exp\left\{\mu_z + \frac{{\sigma_z}^2}{2}\right\} \approx$$

$$\exp\left[\frac{\mu}{2} + 1 - \sigma\Phi^{-1}\left(\frac{2(1-\kappa)}{3\lambda}\right)\right] + \left(\frac{3}{2}\lambda - 1\right)\exp\left[1 + \frac{\mu+\sigma^2}{2}\right] \qquad (24)$$

"True" VaR, i.e. $\mathrm{VaR}(S_Z \mid \kappa)$ is approximately the following:

$$\lim_{\kappa\to 1}\mathrm{VaR}(S_Z \mid \kappa) \simeq \mathrm{VaR}\left(S_Y \mid \kappa, \frac{\mu}{k}, \sigma, \lambda\right)$$

$$\mathrm{VaR}\left(S_Y \mid \kappa, \frac{\mu}{k}, \sigma, \lambda\right) \approx \exp\left[\frac{\mu}{k} - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\lambda}\right)\right] + (\lambda - 1)\exp\left(\frac{\mu}{k} + \frac{\sigma^2}{2}\right)$$

Next examining the simplest case for scaling factor (l, k = 2) as in Case (I), the RHS of the above is simplied to the following:

$$\exp\left[\frac{\mu}{2} - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\lambda}\right)\right] + (\lambda - 1)\exp\left(\frac{\mu+\sigma^2}{2}\right) \qquad (25)$$

Now, the first term has already been calculated in Case (I). The analysis in Case (II) requires the second term for the classical case, namely, $\left(\frac{3}{2}\lambda - 1\right)\exp\left[1 + \frac{\mu+\sigma^2}{2}\right]$. It is important to note that the inverse-normal cdf, $\Phi^{-1}(\cdot)$, decreases much more slowly than linearly (it decreases approximately $\mathcal{O}\left[\frac{1}{\exp\{x^2\}}\right]$), and thus there is not much difference between the first term in the RHS of equation (24) and the first term in the RHS of equation (25), namely:

$\exp\left[\frac{\mu}{2} + 1 - \sigma\Phi^{-1}\left(\frac{2(1-\kappa)}{3\lambda}\right)\right] \simeq \exp\left[\frac{\mu}{2} - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\lambda}\right)\right]$ (since μ is relatively small)

However since λ >> μ,

$\left(\frac{3}{2}\lambda - 1\right)\exp\left[1 + \frac{\mu+\sigma^2}{2}\right] \gtrsim (\lambda - 1)\exp\left(\mu + \frac{\sigma^2}{2}\right)$ since the λ term will dominate.

Therefore,

$$\exp\left[\frac{\mu}{2} + 1 - \sigma\Phi^{-1}\left(\frac{2(1-\kappa)}{3\lambda}\right)\right] + \left(\frac{3}{2}\lambda - 1\right)\exp\left[1 + \frac{\mu+\sigma^2}{2}\right]$$

$$\gtrsim \exp\left[\frac{\mu}{2} - \sigma\Phi^{-1}\left(\frac{1-\kappa}{\lambda}\right)\right] + (\lambda - 1)\exp\left(\mu + \frac{\sigma^2}{2}\right)$$

So the Classical Case is asymptotically approximately greater than the "Truth" in Case (II). ■

Thus, it has been shown that in the asymptotic case for the compound Poisson, the Classical method is "on average" close to the truth, but in extreme cases, it has been shown that it either underestimates or overestimates the true VaR estimate. Next, this

phenomenon is further demonstrated using a large scale Monte-Carlo simulation study which validates the above argument through evidence from numerical experiments.

3.2.2.2 Simulation Study Showing Bias in Classical Methodology

I test both Case (I) and Case (II) (as narrated from the previous section) using large scale MCS study. Recall that in Case (I), the severity of losses is the dominant component while the variance parameter of the severity (Σ) is fixed to be small 0.8. Now for the simulation study for Case (I), the following simulation parameters are used:

- Two independent compound Poisson processes (S_X, S_Y) which have two components: (1) Loss Severity (X, Y) and (2) Loss Frequency (N, M).

$$S_Y = \sum_{i=1}^{N} Y_i$$

$$S_X = \sum_{j=1}^{M} X_j$$

$$Z = \begin{cases} Y_i & \text{with } p = \dfrac{m}{m+n} \\ X_i & \text{with } p = \dfrac{n}{m+n} \end{cases}$$

$$S_Z = \sum_{i=1}^{N} Y_i + \sum_{j=1}^{M} X_j$$

where X ~ LN(μ=25, σ=0.8), Y ~ LN(μ' = 2.5, σ'=0.8), while N ~ Pois(λ=5) and M ~ Pois(λ' = 0.2). Daily loss data is generated from X and Y above with a monthly frequency. So for S_x, this means that there is a monthly average frequency of 0.2 losses per month, while for S_y this means that there is a monthly average frequency of 5 losses per month. The simulation number is large, namely, 10,000,000 months in order to

compute the extreme upper quantiles of the Aggregate Loss distributions of S_Z and S_X. The results are shown below in table (2).

Table 2 Results of Large scale MCS for Case I: Severity is Dominant Component

Case I				
Variable	**Y**		**X**	
Frequency Parameter	λ	5	0.125	λ
Severity Parameters	μ	2.5	25.0	μ
	σ	0.8	0.8	σ
	Classical Estimate of S_Z (combines S_X & S_Y)		**Truth ($S_Z \rightarrow S_X$)**	
Quantile (κ)	**Simulation**		**Simulation**	
99.999%	$606,726,980		$1,489,089,590,514	
99.995%	$100,104,807		$928,928,623,979	
99.99%	$23,773,049		$627,836,613,418	
99.95%	$12,213,875		$518,474,639,417	
99.9%	$2,229,758		$308,110,208,494	
99.5%	$1,001,221		$233,319,036,817	
99%	$431,986		$166,490,267,229	

Notice how in table (2) above, the classical methodology consistently underestimates the VaR(S_z) compared to the estimation for the VaR(S_X). The reason that one can justify that the true estimation for the upper tail region of VaR(S_z) only consists of VaR(S_X), is that in this case, mean severity is much greater than mean frequency. Thus for this case, for large mean severity loss (μ), high individual loss values are generated from severity distribution albeit less frequently (on average 0.2 vs. 5 times). The tail region of S_Z will consist of S_X in this Case (I) since S_X will solely generate the heavy losses and since $\mu \gg \lambda$, there will not be a large enough frequency (count) to make up through adding the small severity losses. Therefore the tail region of S_Z will consist from S_X only (this is mathematically shown in the previous section).

Next, simulation for Case (II) is shown. Recall that in Case (II), the frequency of the losses is the dominant component while the variance parameter of the severity (Σ) is fixed to be small 0.8. For this case, the following situation holds true: $\lambda \sim \mathcal{O}(\exp(\exp(\mu)))$. The specific paradigm for the simulation experiment is narrated next.

- Two independent compound Poisson processes (S_X, S_Y) which have two components: (1) Loss Severity (X, Y) and (2) Loss Frequency (N, M).

$$S_Y = \sum_{i=1}^{N} Y_i$$

$$S_X = \sum_{j=1}^{M} X_j$$

$$Z = \begin{cases} Y_i & \text{with } p = \dfrac{m}{m+n} \\ X_i & \text{with } p = \dfrac{n}{m+n} \end{cases}$$

$$S_Z = \sum_{i=1}^{N} Y_i + \sum_{j=1}^{M} X_j$$

where X ~ LN(μ=1, σ=0.8), Y ~ LN(μ' = 0.5, σ'=0.8), while N ~ Pois(λ=10,000) and M ~ Pois(λ' = 5,000). Daily loss data is generated from X and Y above with a half-century (every fifty years) frequency. So for S_x this means that there is on average 5,000 losses every fifty years, while for S_y this means that there is 10,000 losses every fifty years. Similar to the previous case, ten million half-century time periods are simulated in order to compute the extreme upper quantiles of the Aggregate Loss distributions of S_Z and S_X. The results are shown below in table (3).

Table 3 Results of Large scale MCS for Case II: Frequency is Dominant Component

Case II				
Variable	X		Y	
Frequency Parameter	λ	5000	10000	λ
Severity Parameters	μ	1.0	0.5	μ
	σ	0.8	0.8	σ
	Classical Estimate of S_Z (combines S_X & S_Y)		**Truth ($S_Z \rightarrow S_Y$)**	
Quantile (κ)	**Simulation**		**Simulation**	
99.999%	$45,374		$24,064	
99.995%	$45,094		$23,881	
99.99%	$44,878		$23,744	
99.95%	$44,778		$23,681	
99.9%	$44,520		$23,514	
99.5%	$44,394		$23,435	
99%	$44,258		$23,348	

Notice how in table (3) above, the classical methodology consistently overestimates the VaR(S_z) compared to the true estimation for the VaR(S_Y). The reason that one can justify that the true estimation for the upper tail region of VaR(S_z) only consists of VaR(S_Y), is that in this case, mean frequency >> mean severity. For large λ, high aggregate loss values are generated from combining the high frequency and low severity distributions. The tail region of S_Z will be dominated by S_Y in this case since S_X will have larger individual losses, but they will occur much more rarely and thus the combined sum will be much less than the lower *individual* losses. Since λ >> μ, there won't be enough number of individual large severity to make up for adding the large number of small losses. Therefore the tail region of S_Z will consist from S_Y only (this is mathematically shown in the previous section).

Now the approximation of saying that the tail region of S_z will solely consist of either S_X or S_Y (depending on whether one is looking at Case (I) or Case (II)) is conceptually accurate, however, the VaR quantile will not be the same. To explain this

phenomenon, a simple example in figure (7) is shown below which illustrates the concept.

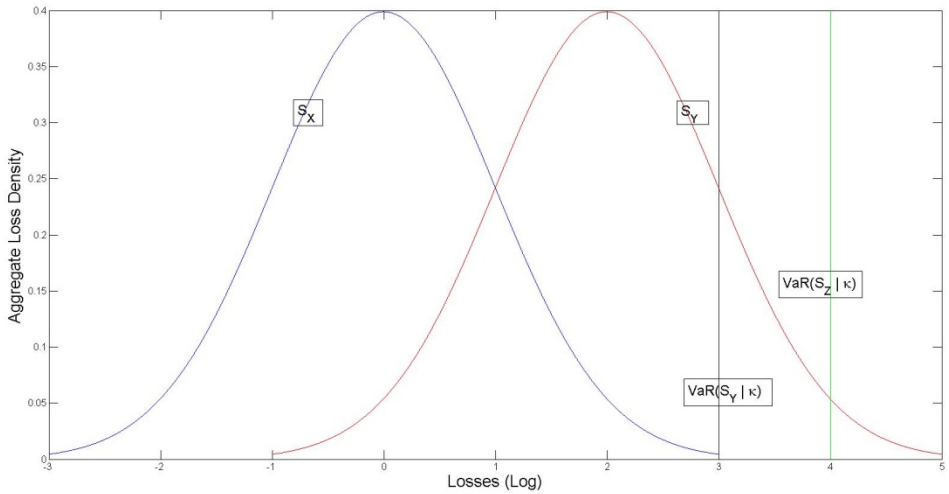


Figure 7 VaR Quantile change plot for two AggLoss distributions

Figure (7) represents a graphical representation of the two independent compound Poisson processes. Here $S_Y >> S_X$ as $\kappa \rightarrow 1$ (κ is the right tail). Now the right tail of S_Z (which consists of a random sum of X and Y as defined in previous section), will solely be determined by S_Y, but since there is probability mass that comes from S_x, the κ will be shifted further as shown in the above figure for S_Z. The exact calculation for the κ shift is done next, and large-scale MCS is performed to adjust the values from tables (3)-(4) above.

To begin, S_Z has a different corresponding quantile κ' for a given quantile κ for S_X or S_Y (depending on whether Case (I) or Case (II) is analyzed). In the situation arising

in Case (I), the true method would compute the VaR(S_Z) using VaR(S_X) while for Case (II), the true method would compute the VaR(S_Z) using VaR(S_Y). However, there is a probability mass for S_Z which comes from S_Y in Case (I) and likewise a mass coming from S_X in Case (II). So the adjusted quantile κ' is calculated for S_Z a given quantile κ for S_X (in the situation arising from Case (I)) and S_Y (in the situation arising from Case (II)) next.

The analysis for Case (I) with properly adjusted quantile, κ' is presented next.

Given the following information:

S_X has X ~ LN(μ, σ) and M ~ Pois(λ/l),

S_Y has Y ~ LN(μ/k, σ), while N ~ Pois(λ) where k, l > 1

Then, the adjusted quantile, κ', is the following:

κ' = (proportion of mass of Y) + (proportion of mass of X)*κ

proportion of mass of Y $= \frac{l\lambda}{\lambda(l+1)}$, proportion of mass of X $= 1 - \frac{l\lambda}{\lambda(l+1)}$

$\kappa' = \frac{l\lambda}{\lambda(l+1)} + \left[1 - \frac{l\lambda}{\lambda(l+1)}\right]\kappa$

The analysis for Case (II) with properly adjusted quantile, κ' is presented next.

Given the following information:

S_X has X ~ LN(μ, σ) and M ~ Pois(λ/l), and

S_Y has Y ~ LN(μ/k, σ), while N ~ Pois(λ) where k, l > 1

Then, the adjusted quantile, κ', is the following:

κ' = (proportion of mass of X) + (proportion of mass of Y)*κ

$\kappa' = \left[1 - \frac{l\lambda}{\lambda(l+1)}\right] + \left[\frac{l\lambda}{\lambda(l+1)}\right]\kappa$

Table 4 VaR Results of Large scale MCS for Case (I) with scaled quantile shifting

Case I				
Variable	Y		X	
Frequency Parameter	λ	5	0.125	λ
Severity Parameters	μ	2.5	25.0	μ
	σ	0.8	0.8	σ
	Classical Estimate of S_Z (combines S_X & S_Y)		Truth ($S_Z \rightarrow S_X$)	
Shifted Quantile of Z (κ')	Simulation		Simulation	Quantile X (κ)
99.9999756%	$9,556,149,565		$1,489,089,590,514	99.999%
99.999878%	$3,158,683,916		$928,928,623,979	99.995%
99.999756%	$1,962,093,430		$627,836,613,418	99.99%
99.998780%	$532,225,663		$518,474,639,417	99.95%
99.997561%	$311,845,292		$308,110,208,494	99.9%
99.987805%	$83,001,435		$233,319,036,817	99.5%
99.975610%	$45,894,227		$166,490,267,229	99%

Table 5 VaR Results of Large scale MCS for Case (II) with scaled quantile shifting

Case II				
Variable	X		Y	
Frequency Parameter	λ	5000	10000	λ
Severity Parameters	μ	1.0	0.5	μ
	σ	0.8	0.8	σ
	Classical Estimate of S_Z (combines S_X & S_Y)		Truth ($S_Z \rightarrow S_Y$)	
Quantile (κ)	Simulation		Simulation	
99.999%	$45,374		$24,064	
99.995%	$45,094		$23,881	
99.99%	$44,878		$23,744	
99.95%	$44,778		$23,681	
99.9%	$44,520		$23,514	
99.5%	$44,394		$23,435	
99%	$44,258		$23,348	

Notice in both tables (4)-(5) above, the results are consistent with before: (1) In Case (I), the classical method underestimates the true VaR in the tail region; (2) In Case (II), the classical method overestimates the true VaR in the tail region. Thus, the dissertation has shown so far through a mathematical argument and through large scale MCS that in specific extreme cases, the classical methodology forms a bias in the estimate of VaR. Therefore, it would be interesting and pertintent to develop distribution-free (non-parametric) and parametric approaches which calculate VaR without making the strong assumption of independence between severity and frequency. The methodology for the

non-parametric case is called Data Partition of Frequency and Severity (DPFS) via K-means algorithm (distribution free method). The next subsection narrates this approach next.

3.2.2.3 Cluster Analysis

In the field of data mining, the overall purpose of clustering (an unsupervised learning technique) is to group data objects based solely on information characteristics found in the data that describes objects and their relationships [61]. The overall goal is that the objects within each group be similar to one another, and different from objects in other groups. Thus, the greater the homogeneity within a group and the greater the heterogeneity between groups, then the clustering algorithm is working the best. Now in the modern ORM context, this dissertation argues (and later on shows) the importance of clustering/correlating the frequency and severity objects of the loss data. The reason is that if frequency and severity are *truly* independent, then the clustering should show one single group. If they are correlated and or dependent, then the clustering should show more than one distinct group. The advantage is that clustering framework allows for both paradigms of independence and dependence between frequency and severity. The next question is what type of cluster analysis should this research use? This is narrated next through description of the K-means algorithm.

The K-means algorithm is an algorithm for putting N data points in an *I*-dimensional space into *K* clusters. Each cluster is parameterized by a vector $\mathbf{m}^{(k)}$ called its mean (average) [62]. The data points can be denoted by $\{\mathbf{x}^{(n)}\}$ where the superscript *n* runs from 1 to the data size *N*. Each vector comprises of *I* components, x_i. Next a metric is

used which defines the distances between points, for example, the Euclidean distance (L_2 norm):

$$\mathrm{d}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{x_i}-\mathrm{y_i})^2}$$

With this metric, the K-means algorithm is defined as the following [62]:

Initialization. Set K means $\{\mathbf{m}^{(k)}\}$ to random values (K must be set *a priori*).

Step 1. Each data point n is assigned to its nearest mean (known as centroid). In this notation, the guess for the cluster $k^{(n)}$ that the point $\mathbf{x}^{(n)}$ belongs to by $\hat{\mathrm{k}}^{(n)}$.

$$\hat{\mathrm{k}}^{(n)} = \underset{k}{\operatorname{argmin}}\{\mathbf{d}(\mathbf{m}^{(k)},\mathbf{x}^{(n)})\}.$$

Step 2. Now take each $\mathbf{x}_\mathrm{i}$ in the k^{th} cluster, n_k is the number of points in the k^{th} cluster, and $\mathbf{d_{i,k}^2}$ is the distance metric (Euclidean for example) between $\mathbf{x}_\mathrm{i}$ and the centroid of cluster k. If there exists a group s such that

$$\frac{n_k}{n_k-1}\mathbf{d_{i,k}^2} > \frac{n_s}{n_s-1}\mathbf{d_{i,k}^2}$$

then move $\mathbf{x}_\mathrm{i}$ to the cluster s.

Step 3. If there exists several clusters that satisfy the above inequality then move $\mathbf{x}_\mathrm{i}$ into the group that has the smallest value for

$$\frac{n_s}{n_s-1}\mathbf{d_{i,s}^2}$$

Step 4. Repeats Steps 1-3 until convergence criterion is met (usually a tolerance limit which the user can set).

With the description of the K-means algorithm defined above [62], the next natural question is how does one choose the correct value of K? In simulation experiments and certain types of real-life datasets, one may know *a priori* what the correct value should be for K in the K-means algorithm. However, in the majority of real-life datasets, it is not well-know how to select the appropriate K in advance. The brute-force trial and error method involves trying different values of K and then calculating the overall minimum distance. However, there is a rich literature from the data mining field regarding procedures which speed up finding the "optimal" K. This is narrated next.

There are several methods such as Mojena upper tail rule [63], Silhouette Statistic (SS) by Kaufman and Rousseeuw [64], Tibshirani gap statistic [65], Calinski & Harabasz index [66] to name a few. Similar to GoF measures, each of the above metrics has their own strengths and weaknesses. In this dissertation, the SS is utilized because this has both a numerical and a visualization component. This is narrated next.

The SS is a measure to estimate the optimal number of groups in a given dataset. Given a data point $\mathbf{x_i}$, denote the average distance to all other points in its own cluster as a_i. For any other cluster C, let $\bar{d}$(i, C) represent the average distance of $\mathbf{x_i}$ to all observations in cluster C. Finally, let b_i denote the minimum of the average distances $\bar{d}$(i, C), where the minimum is taken as C ranges over all clusters except the observation's own cluster. Then silhouette width (SW) for the i^{th} observation of **x** is defined as the following [62]:

$$-1 \leq SW_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \leq 1$$

If an observation has a value close to 1, then the data point is closer to its own cluster than a neighboring one. If a silhouette width is close to -1, then it is not very well-clustered. A width of 0 indicates that the observation could just as well belong to its current cluster or another one that is near to it. With this notation in mind, one uses the average silhouette width as an indicator of best grouping defined as the following (which averages over all observations) [62]:

$$\overline{sw} = \frac{1}{n}\sum_{i=1}^{n} SW_i$$

where the average silhouette width greater than 0.5 indicates good partition of the data, while a value of less than 0.2 indicates no partition is necessary. Visually, all of the clusters should have roughly the same thickness. If in a particular partition, a specific cluster has a larger thickness than the others, this graphically indicates poor choice of *K* for the clustering. Two sample choices of K for a sample dataset is shown next in figures (8) - (9).

It is interesting to notice in figure (8) that the average SW value is close to 0.66 (the green line) while in figure (9) for the same dataset, the average SW is closer to 0.69. This indicates that for this particular dataset there is clear clustering structure in the data and 5 groups is marginally better than choosing 3 groups.

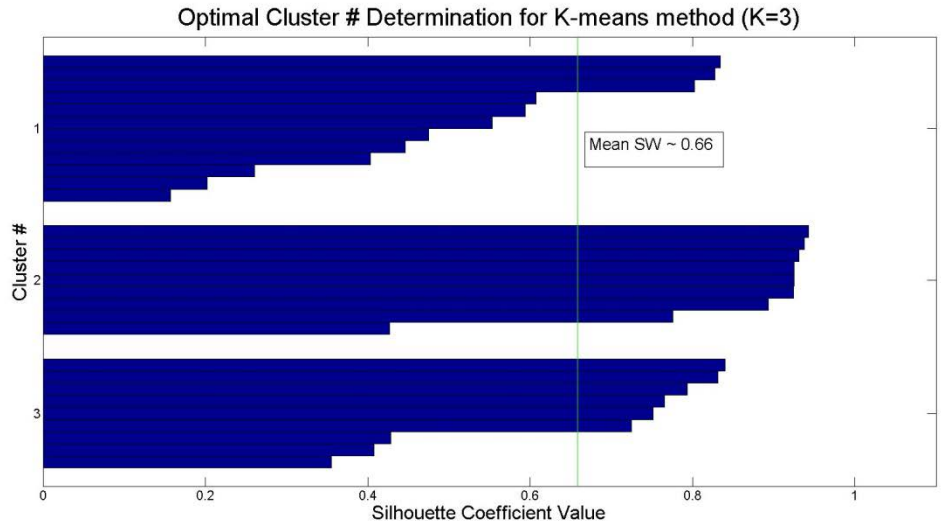


Figure 8 Sample Silhouette Statistic technique for K-means Algorithm (3 clusters)

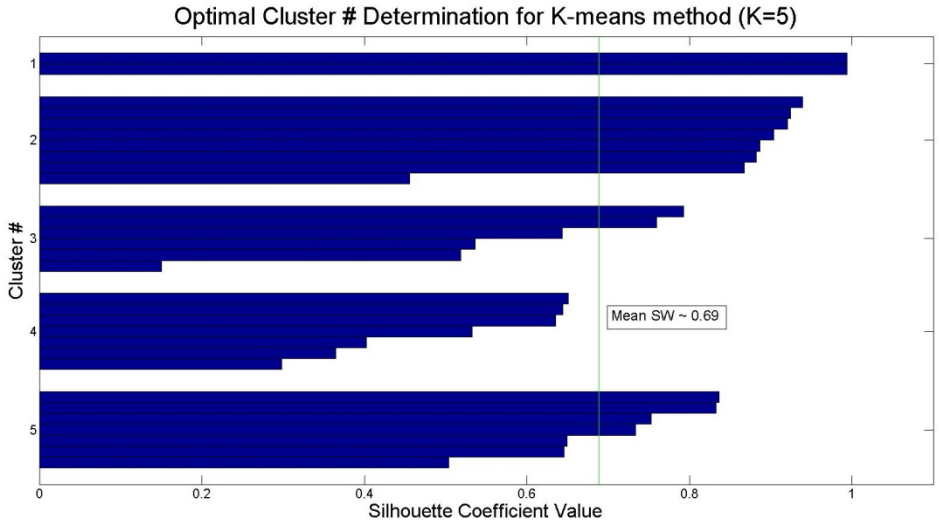


Figure 9 Sample Silhouette Statistic technique for K-means Algorithm (5 clusters)

Application to modern ORM VaR Calculation

Now with the basics of the K-means clustering algorithm described above, one can proceed to defining different ways to partition the frequency and severity component

in the modern ORM setting. This dissertation tests out three distinct possibilities which are narrated next.

The first possibility is using no partition at all, i.e., the classical method. This is known as the null method (Method 0) where frequency is assumed to be independent of severity. The VaR through MLE estimates of the parameters for the severity and the frequency distributions are the same as described in previous methodology section.

Next, a case of 2-D K-means methodology is tested. The 2-D notation here is used to denote using both the frequency and severity for each dimension. Specifically, the mean severity in one dimension and frequency in the second dimension are used in the K-means. There is a pragmatic reason why for any 2-D K-means methodology a summary statistic usually must be applied for the severity. In almost all cases, the loss magnitudes are collected at a *different* time interval than the frequency of interest. For example, daily losses in the financial market happen once per day. When calculating the VaR, in most cases, institutions are interesting in how much they can expect to lose (EL) and the VaR for a different time horizon, say monthly or even annually. Thus in this example, the frequency is every 30 days or 365 days (monthly versus annually respectively). However, the severity is collected daily. Thus mathematically speaking there is not a *one-to-one* correspondence between frequency and severity. In fact, there is a *one-to-many* correspondence between frequency and severity. So in order to force a one-to-one relationship, a summary statistic which best describes the severity during the frequency time unit has to be used. Standard statistical theory argues that either the mean or median suffices as a sufficient statistic for estimating the central tendency of a dataset. So for

example suppose that the severity consists of daily loss data while the frequency consists of count of losses for the monthly time unit. If the original dataset consisted of *N* total daily losses for *m* months, then the new dataset consists of a *m*x2 matrix of data where the first column represents the mean daily loss magnitude for month *i* and the 2nd column represents the corresponding monthly loss count for month *i*. This method, which is called K-means: Method I or K-means: Mean Severity/Frequency, applies K-means algorithm to the average severity and frequency. This method calculates the mean severity for each frequency time unit (say monthly), so average loss/month and its corresponding count of losses in that respective month. In addition to looking at severity and frequency simultaneously, a second method, i.e. K-means: Method II is examined, where the K-means partitioning is done solely on the severity of the loss data. So this is called K-means: Severity only and Implied Frequency. This algorithm applies K-means algorithm for the loss severity only and then calculates the corresponding implied frequency. I illustrate using an example. Suppose again that the severity consists of daily loss data while the frequency consists of monthly loss count. Assume that there are *N* total daily loss data and *m* total months (where by definition $m < N$). Suppose the severity only data (i.e. losses L_1, ..., L_N) are split into *k* groups. Then for each month from 1 to *m*, compute the frequency in each group say n_i for i=1...k. The implied frequency for group *i* is simply the average of n_i taken over all *m* months. The reason that the term implied frequency is applied here is due to the fact that this frequency comes after K-means is computed on the severity (i.e. this does not exist in the original data). So this frequency is *implied* from the K-means severity only portion. With this new implied frequency, one

can use the classical methodology and compute the VaR through MLE estimates of severity distributions for each group *i* and their corresponding frequency in each group *i*.

In this work, several other ad-hoc partitioning methods for the K-means have been tried. In the end, the above two methods are determined to be the most successful methods among them. The simulation study and the results are shown in section 4. In summary, the algorithm for computing VaR [67] is described in the following based on the clustering analysis:

1. Use K-means algorithm to split the severity/frequency data into K components.
2. For each of the K components, compute the severity loss parameters and frequency parameter.
3. For a large integer *N*, a minimum of 10,000 iterations, do the following:
   (a) Draw a random number, n_i of loss frequency from frequency distribution for month *i*.
   (b) Draw n_i losses from severity distribution for region K.
   (c) Sum the losses to compute the Aggregate Loss distribution.
   (d) Perform steps (a)-(c) for each region of the K-means split.
4. Generate the Aggregate Loss distribution, and then compute VaR.

3.3 Modeling Distribution based Partitioning (DBP) for Estimating VaR

3.3.1 Current State-of-the-Art

The simplest method to include correlations is to use the linear correlation coefficient in the aggregation process. Currently this is done for severity and frequency separately. For the aggregation across j Business Lines (BL^s), one may use the normal

assumption of the VaR for confidence level α, and time period 1 year, and the aggregated VaR, $\mathrm{VaR}_{\alpha}^{\mathrm{Agg}}$, can be determined as the following [14]:

$$\mathrm{VaR}_{\alpha}^{\mathrm{Agg}} = \sqrt{\sum_{i=1}^{n} \mathrm{VaR}^2(i|\alpha) + 2\sum_{i=1}^{I}\sum_{j=1}^{J}\left[(\mathrm{VaR}(j|\alpha)) * (\mathrm{VaR}(i|\alpha)) * \rho_{ij}\right]^2}$$

where the following holds true:

- VaR_i is the Value-at-Risk of the BL_i for a certain percentile α
- $\rho_{i,j}$ is the correlation coefficient between the BL_i and the BL_j

The above approach only has a point estimate for the correlation across different BL's. However, the current best practices include using parametric based approach to model a distribution. This is done through the methods of parametric copulas. The basic mathematical foundation for this methodology is narrated next.

A *k*-dimensional copula C: $[0, 1]^k \rightarrow [0, 1]$ is a function which is a CDF with uniform distributions as the marginals. Thus, copulas are the functions *C* satisfying the following property [68]:

1. $C(u_1, u_2, ..., u_k)$ is an increasing function in each component u_i.
2. $C(u_1, u_2, ..., u_k) = u_i$ if $u_j = 1$, for all $j \neq i$ and $u_i \in [0, 1]$.
3. For all $(a_1, ..., a_k)$, $(b_1, ..., b_k) \in [0, 1]$ with $a_i \leq b_i$ for all then the following:

$\sum_{i_1}^{2}\sum_{i_k}^{2}(-1)^{[i_1+\cdots+i_k]}\, C\left(u_{1,i_1}, \cdots, u_{1,i_k}\right) \geq 0$, where $u_{j,1} = a_j$ and $u_{j,2} = b_j$.

It is rather obvious to note that the aforementioned condition 1 above holds true for any cumulative distribution function. The second aforementioned condition ensures that the marginals are uniforms. Finally the third condition above ensures that the mass of any *k*-dimensional rectangle $x_{i=1}^{k}[a_i, b_i]$ is non-negative.

The most important (from a risk management perspective) is to find a connection between copulas and multivariate random variables. This is formalized by Sklar's Theorem [69]. Overall, this theorem states that any *k*-dimensional random variable adopts a copula representation. In otherwords, this theorem provides a relationship between copulas and multivariate random variables with any given marginals. Specifically, Sklar's Theorem states the following [69]:

1. Let F be a joint cdf with margins F_1,..., F_k. Then there exists a copula C:$[0,1]^k \rightarrow [0,1]$ such that $F(x_1, ..., x_k) = C(F_1(x_1), F_2(x_2), ..., F_k(x_k))$, for all $x_j \in [-\infty, \infty]$. If the marginals are continuous, then the copula is unique. Otherwise the copula is uniquely determined only on Ran $F_1 \times$ Ran $F_2 \times \cdots \times$ Ran F_k, where Ran F_i denotes the range of the cdf F_i.

2. Conversely, if C is a copula and F_1,..., F_k are univariate CDFs, then F is a multivariate cdf with margins F_1,..., F_k and copula C.

An immediate consequence of Sklar's Theorem is the copula for a set of *k* random variables X_1, ..., X_k can be expressed as the following [68]:

$$C(u_1,...,u_k) = F(F^{-1}{}_1(u_1), ..., F^{-1}{}_k(u_k)),$$

where F is the joint CDF of (X_1, ..., X_k) whose quantile functions are $F_1^{-1}, \cdots, F_k^{-1}$. Now there are multiple types of copulas which are well-known. Various classes of parametric copulas such Gaussian copula, Student *t*-copula, Archimedean copulas etc. have all been used to model the relationship between the cells of the matrix in Table 1. In this dissertation the Gaussian and Student *t* copula is examined and is thus defined next.

To define a Gaussian copula, there is a need to specify the correlation matrix, Σ. With that, the Gaussian copula $C^{\text{Gaussian}}(u_1, u_2, ..., u_k \mid \Sigma) = \Phi_\Sigma(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_k))$, where Φ is the CDF of the Standard Normal distribution for the k-dimensional multivariate function [68]. Next the Student *t*-distribution is well-known with ν degrees of freedom, t_ν. The multivariate *t*-distribution in *k*-dimensions with ν degrees of freedom, $t_{\nu,\Sigma}$, is defined from $\mathbf{X} = X_1,...,X_k \sim \text{Normal}(0, \Sigma)$ with the following [68]:

$$(\eta_1, \ldots, \eta_k) = \left(\frac{X_1}{\sqrt{\xi/\nu}}, \ldots, \frac{X_k}{\sqrt{\xi/\nu}}\right) \text{ where } \xi \sim \chi^2(\nu) \perp \mathbf{X}$$

$$C^t(u_1, \ldots, u_k \mid \Sigma, \nu) = t_{\nu,\Sigma}(t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_k))$$

In figure (10) below, a sample plot of two random variables X_1 and X_2 where $X_1 \sim$ Gamma(2,1) and $X_2 \sim t_5$ with a Gaussian copula (correlation matrix of $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ where $\rho = 0.7$) is shown. The first variable could theoretically be considered as a frequency and the second a severity (as an example). The research question is what type of correlation can best model this dynamic relationship? Should a Gaussian copula be used with say Poisson frequency and lognormal severity? How about *t*-copula? In particular, the focus will be to use non-parametric copulas so that a parametric dependency does not need to be established. The estimation for the copulas is done through the log-likelihood process and using MLE to fit the best parameter estimates. Afterwards, a bootstrap based goodness-of-fit test is used to calculate a non-parametric *p-value* of the likelihood of the model being a good fit. The basics of the bootstrap procedure for estimating the p-value of the model is based on GoF [70-87] and is narrated next.

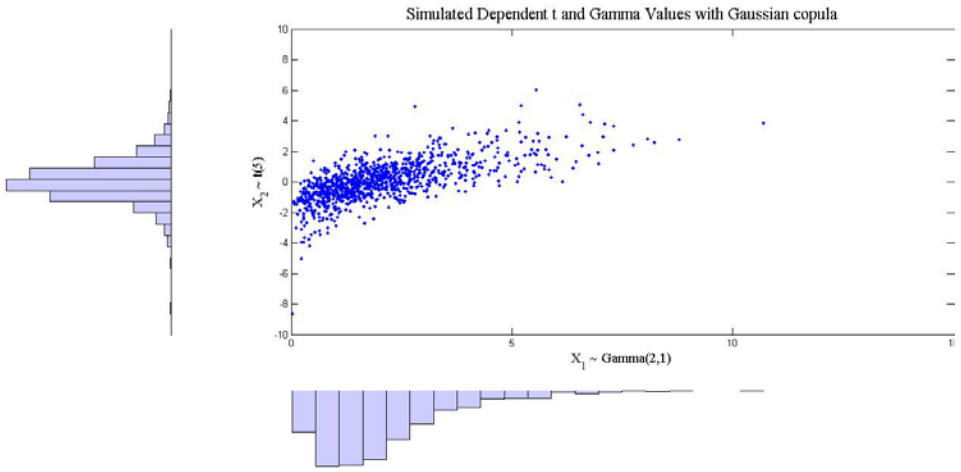


Figure 10 Example of Gaussian Copula used for Correlation Analysis

The GoF test is used following the work of [70]. This is based on the empirical copula of the data (defined shortly below) using a consistent estimator of the unknown copula *C*. Let $\widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_n$ be the pseudo-observations from *C* and let there be a vector of ranks $R_1, \ldots, R_n$ of the data X_{ij}. Now the pseudo-observations are defined as the following: $\widehat{\mathbf{U}}_i = R_i/(n+1)$, $i \in \{\mathbf{1}, \mathbf{2}, \ldots, n\}$. The empirical copula [70] is then classically defined as the empirical cdf computed from the pseudo-observations, i.e.

$$C_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(\widehat{U}_i \leq \mathbf{u})} \quad \text{where } \mathbf{u} \in [0,1]^d$$

Then the empirical copula process, $\mathbb{C}_n$, is defined as the following:

$$\mathbb{C}_{\mathbf{n}} = \sqrt{n}\{C_n(\mathbf{u}) - C_{\theta_n}(u)\}, \text{ where } \mathbf{u} \in [0,1]^d \qquad (23)$$

where C_n is the empirical copulas as defined in equation (23) above and C_{θ_n} is an estimator of *C* under the null hypothesis, H_0: $C \in \{C_\theta\}$ holds while H_A: $C \notin \{C_\theta\}$. Based on large scale Monte-Carlo numerical experiments by Berg [88], the statistic of interest is S_n which has an analytical formulation as shown below:

$$S_{\mathrm{n}} = \int_{[0,1]^{\mathrm{d}}}^{\infty} \mathbb{C}_{\mathbf{n}}(\mathbf{u})^2 \, \mathrm{d}C_{\mathrm{n}}(\mathrm{u}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \{C_{\mathrm{n}}(\widehat{\mathbf{U}}_{\boldsymbol{i}}) - C_{\theta_{\mathrm{n}}}(\widehat{\mathbf{U}}_{\boldsymbol{i}})\}^2 \qquad (24)$$

An approximate *p*-value for S_{n} can be attained computationally via a parametric bootstrap approach. This is used to assess the GoF and is narrated next.

Following the work of Genest *et al.* [89] an approximate p-value for the test can be calculated using the statistic for S_{n}. This is based on the following procedure below [89]:

1. Compute C_{n} explicitly from the pseudo-observations $\widehat{\mathbf{U}}_{\mathbf{1}}, \ldots, \widehat{\mathbf{U}}_{\mathbf{n}}$ and estimate θ from $\widehat{\mathbf{U}}_{\mathbf{1}}, \ldots, \widehat{\mathbf{U}}_{\mathbf{n}}$ by means of a rank-based estimator of θ_{n}.
2. Compute the test-statistic, S_{n} as defined above (in Eq. 24).
3. For a large integer *N*, a minimum of 100 (bootstrap replicate number), repeat the following steps for every $k \in \{1, 2, \ldots, N\}$:

   (a) Generate a random sample of $\mathbf{X}_1^{(\mathrm{k})}, \ldots, \mathbf{X}_{\mathrm{n}}^{(\mathrm{k})}$ from copula $\mathbf{C}_{\boldsymbol{\theta}_{\mathbf{n}}}$ and compute the associated *sample* pseudo-observations $\widehat{\mathrm{U}}_1^{(k)}, \ldots, \widehat{\mathrm{U}}_{\mathrm{n}}^{(k)}$.

   (b) Next define the quantity $C_{\mathrm{n}}^{(k)}(\mathbf{u})$ such as the following:

   $$C_{\mathrm{n}}^{(k)}(\mathbf{u}) = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{1}_{\left(\widehat{\mathrm{U}}_1^{(k)} \leq \mathbf{u}\right)} \qquad \mathbf{u} \in [0,1]^{\mathrm{d}}$$

   and compute an estimate of $\theta_{\mathrm{n}}^{(\mathrm{k})}$ of θ from $\widehat{\mathrm{U}}_1^{(k)}, \ldots, \widehat{\mathrm{U}}_{\mathrm{n}}^{(k)}$ using the rank-based estimator as shown in Step 1.

   (c) Compute an approximate independent realization of S_{n} under the null hypothesis using the following:

$$S_{\mathrm{n}}^{(\mathrm{k})} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \{C_{\mathrm{n}}^{(\mathrm{k})}(\widehat{\mathbf{U}}_{\mathrm{i}}^{(\mathrm{k})}) - C_{\theta_n^{(k)}}(\widehat{\mathbf{U}}_{\mathrm{i}}^{(\mathrm{k})})\}^2$$

4. With the above, a *p*-value (*p*) can be approximated for the test as the following:

$$p = \frac{1}{\mathrm{N}} \sum_{k=1}^{N} \mathbf{1}_{\left(S_{\mathrm{n}}^{(k)} \geq S_{\mathrm{n}}\right)}$$

With the previously discussed background, I proceed forward to discussing the area of work where this dissertation makes contributions. In particular, the usage of Gaussian Mixture Copula Model (GMCM) in the setting of modern ORM is first used in this dissertation (to the best of my knowledge). The basic formalism is narrated next and then the application to modern ORM through severity and frequency is narrated next.

3.3.2 Research Contribution in this Area

There is recent work which combines the area of unsupervised learning and copulas: Gaussian Mixture Copula Models (GMCM) [90]. The unsupervised learning portion comes in because it is unknown how many components will be in the mixture. The basic idea is that one would use unsupervised learning to determine the number of Gaussian components of the copula. The main strength of this procedure is that since with multiple components, the correlation is no longer linear (as in the case of the classical Gaussian copula), but rather non-linear correlations can be modeled. A simple example figure is shown below in figure (11). Notice how in the figure (11), there is a clear non-linear correlation trend between the simulated x and y-axis. Also, three clear clusters can be seen in the simulated data in figure (11). The mathematical framework for the GMCM is narrated next.

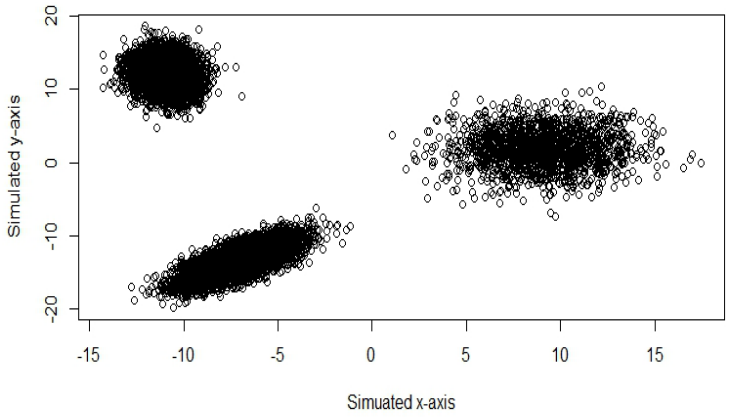


Figure 11 Example of Simulated Mixture Copula Process

The framework follows from the seminal paper in 2011 [90]. For the general framework, suppose one can observe a large *p x d* matrix of observed values, x_d^p, which are naturally clustered into say *m* groups (this is known *a priori*). The general GMCM assumes a m-component Gaussian mixture model as a latent process, $\mathbf{Z} = (Z_1, ..., Z_d)^T$ with the following distribution [90]:

$$\begin{cases} H \sim \text{Categorical}(\alpha_1, \dots, \alpha_m) \\ \mathbf{Z}|H = h \sim N_d(\mu_h, \Sigma_h) \end{cases}$$

where the Categorical distribution is the generalized Bernoulli distribution, H is a dummy variable where $H \in \{1, 2, \dots, m\}$ corresponds to the class and $(\alpha_1, \dots, \alpha_m)$ are the mixture probabilities for h = 1,...,m with $\sum_{h=1}^{m} \alpha_h = 1$. The parameters of interest (which will be estimated) are the following:

$$\boldsymbol{\theta} = \underbrace{(\alpha_1, \dots, \alpha_m, \mu_1, \dots, \mu_m, \Sigma_1, \dots, \Sigma_m)}_{3m \text{ terms}}$$

Since this is a copula process with Gaussian mixtures, the marginal cdf is denoted as the following:

$$\Omega_k(z|\alpha_1, \ldots, \alpha_m, \mu_1, \ldots, \mu_m, \Sigma_1, \ldots, \Sigma_m) = \sum_{h=1}^{m} \alpha_h \Phi_k(z|\mu_h, \Sigma_h)$$

Next the above framework needs to be extended to the observable data (i.e. moving out from the latest phase). Let $\mathbf{X} = (X_1, ..., X_d)^T$ be an observation with *known* marginal cdfs $(F_1, ..., F_d)$, where the following relationship is assumed [91]:

$$X_k = F_k^{-1}(\Omega_k(Z_k|\alpha_1, \ldots, \alpha_m, \mu_1, \ldots, \mu_m, \Sigma_1, \ldots, \Sigma_m)), \qquad k \in \{1 \ldots d\}$$

Then using the standard probability integral transformation of the above equation, the vector $\mathbf{U} = (U_1, ..., U_d)^T$ where $U_k = \Omega_k(Z_k) = F_k(X_k)$ have uniform marginals (as required by the theory of copulas). Overall, the GMCM operates on three distinct levels: (1) latent level **Z**; (2) copula level **U** which is distributed according to the Gaussian mixture copula density *c*; and (3) observed data level **X** based on the marginal transformation of **U** [91]. Next, a description of the estimation process is given, by briefly narrating the likelihood calculation in the model.

Given that $(F_1,..., F_d)$ are known, one can derive an expression for the log-likelihood for the aforementioned model. Since this is based on vectors, following the work in [91] the introduction of the notation of vector functions is used as, $\Omega_0: \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}^d$ and $F_0: \mathbb{R}^d \rightarrow \mathbb{R}^d$ where Θ is the parameter space. Note that the vector function Ω_0 applies the k^{th} marginal transformation, Ω_k on the k^{th} entry of the observation and the vector function F_0 applies the k^{th} marginal transformation F_k on the k^{th} entry of the observation [91]. Using vector notation this can expressed as the following:

$$\Omega_0(\mathbf{Z} \mid \boldsymbol{\theta}) = \begin{pmatrix} \Omega_1(Z_1 \mid \boldsymbol{\theta}) \\ \vdots \\ \Omega_d(Z_d \mid \boldsymbol{\theta}) \end{pmatrix} \quad \text{and} \quad F_0(\mathbf{X}) = \begin{pmatrix} F_1(X_1) \\ \vdots \\ F_d(X_d) \end{pmatrix}$$

Now using the probability integral transformation, $\boldsymbol{Z}$ is transformed by Ω_0 into the marginally uniformed distributed random vector $\boldsymbol{U}$ with cdf and pdf as the following respectively [91]:

$$C(\mathrm{u} \mid \boldsymbol{\theta}) = \Omega(\Omega_0^{-1}(\mathrm{u} \mid \boldsymbol{\theta}) \mid \boldsymbol{\theta})$$

$$c(\mathrm{u} \mid \boldsymbol{\theta}) = \frac{\omega(\Omega_0^{-1}(\mathbf{u}))}{\prod_{\mathrm{k}=1}^{\mathrm{d}} \omega_{\mathrm{k}}(\Omega_{\mathrm{k}}^{-1}(\mathrm{u}_{\mathrm{k}}))}$$

With the above framework, one can fit the GMCM model. The usage of this copula approach in the context of modern ORM is narrated next.

The current literature does not use copula based methodology to directly compute the VaR from the severity and frequency. One of the major contributions that this dissertation makes is that the copula based methodology can be used to obtain an estimate for the frequency and severity parameters per time unit. For example, suppose that the loss severity are daily distributed as LN(μ, σ) while the loss frequency is monthly following a Poisson distribution, Pois(λ). Then the goal of this method is to obtain a parameter estimate for severity and frequency on a monthly basis. Thus for each month *i*, there will be a 3-tuple estimate of $(\hat{\lambda}_i, \hat{\mu}_i, \hat{\sigma}_i)$. The best way to show this is through figure (12). As one can see in the aforementioned figure, the severity parameters and the frequency parameters are estimated for a time unit and their surface shows their relationship. The strength of this approach is that the VaR is computed by incorporating the relationship between severity and frequency for each time unit. So if indeed there is a correlation between frequency and severity, the copula surface approach will capture this dependency. The algorithm for computing VaR is summarized in the following:

1. Compute the surface of $(\hat{\lambda}_i, \hat{\mu}_i, \hat{\sigma}_i)$ for i=1...m months using copula fit.

2. Create a scatterplot (for visualization purposes) for the severity and frequency parameters.

3. For a large integer *N*, a minimum of 1,000,000 iterations, do the following:

   (a) Generate a random draw of $(\lambda_k, \mu_k, \sigma_k)$ from the 3-D surface as a realization for month k.

   (b) Compute the Aggregate Loss for the realization of $(\lambda_k, \mu_k, \sigma_k)$ as the Monte-Carlo estimate for month k.

4. Generate the AggLoss distribution, and then compute VaR.

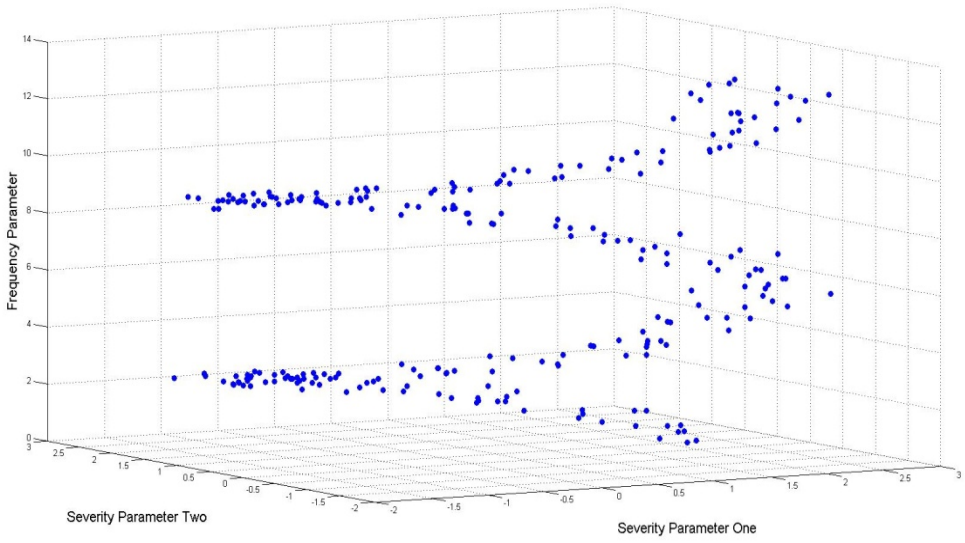


Figure 12 Sample 3-D Scatterplot for Estimating VaR through Copula

3.4 Application of Modern ORM to Real-World Data: Empirical Research

3.4.1 Current State-of-the-Art

All of the modern ORM applications have been almost exclusively on financial proprietary data. The main problem is that the current empirical papers rely on data which are not availably publicly. In almost all cases, the research papers do not share the data for public usage or peer review. In the rare cases, where the data is cited, it is available in very expensive loss data exchanges where users have to pay a hefty fee to access the data. To counteract this trend, all of the empirical analysis in this dissertation consists of data which is freely available for the general public. One of the goals of this dissertation is to form a more generalized consensus on robust risk-metrics, and thus datasets are studied across many practical domains. Specifically, financial loss datasets, along with insurance, natural calamities, and chemical spills as monitored by US Coast Guard are studied. The details of the data are narrated next.

3.4.2 Research Contributions in this Area

In addition to using real-world data for testing the new methodologies developed, verification is conducted of the methodologies by testing on known simulated scenario data. The purpose of using simulated data is that five distinct and diverse scenarios can be created where the true VaR value is known *a priori*. Then a comparison of the new methodology with the classical methodology can be determined to see which performs better and in what conditions. This simulated data through five distinct scenarios is narrated next. Afterwards, the real-life datasets which are used for testing the new methodology is discussed.

3.4.2.1 Simulated Data Analysis through Five Distinct Scenarios

Five distinct simulated scenarios are created to test the new methodologies. One of the scenarios involves generating data based on the classical methodology assumption, i.e. independence of severity and frequency. The remaining four scenarios will have correlations between the frequency and severity, albeit in different types of manner. The goal is to verify and validate each of the methodologies strengths' and weaknesses on simulated data where the true estimate of VaR is known. To begin, a description using a flow-chart format for each scenario and their sample severity and frequency distributions from large MCS is shown. The parameters for the scenario study are shown in tables (6)-(7) below [67].

Table 6 Severity Parameters for Simulation Study

	Severity		
	Low	Medium	High
μ	0.5	4.5	9.3
σ	0.25	0.5	0.6
Mean	$1.70	$102.00	$13,095
Median	$1.65	$90.02	$10,938
Standard Dev	$0.43	$54.36	$8,620

Table 7 Frequency Parameters for Simulation Study

	Frequency		
	Low	Medium	High
λ	2	10	30

Scenario (I) replicates the diagram from [14] as shown in figure (2). There are three types of severity (low/medium/high) arising from low frequency, two types of

severity (low/medium) for medium frequency and low severity in cases of high frequency. The algorithm for Scenario (I) with corresponding simulation parameters is listed next along with the simulated data histograms as shown below in figure (13).

Scenario I: Hi/Med/Low Severity mapping to Hi/Med/Low Frequency (one-to-many)

(1) For i=1:*n* months (set *n* to a large number of iterations, which is 10,000)

(2) Generate a uniform random number, $u \sim U[0,1]$.

(2a) If $u \leq 0.5$, then Low Frequency Region

(2a.1) Generate a $\lambda_{Low} \sim Poisson(\Lambda = 2)$

(2a.2) Generate a uniform discrete random number $k \in \{1,2,3\}$

(2a.3) If $k = 1$, then sample λ_{Low} cases from LN(0.5, 0.25)

(2a.3) If $k = 2$, then sample λ_{Low} cases from LN(4.5, 0.5)

Else, sample λ_{Low} cases from LN(9.3, 0.6).

(2b) If $1/2 \leq u < 5/6$, then Medium Frequency Region

(2b.1) Generate a $\lambda_{Med} \sim Poisson(\Lambda = 10)$

(2b.2) Generate a uniform discrete random number $k \in \{1,2\}$

(2b.3) If $k = 1$, then sample λ_{Med} cases from LN(0.5, 0.25)

If $k = 2$, then sample λ_{Med} cases from LN(4.5, 0.5)

(2c) If $u \geq 5/6$, then High Frequency Region

(2c.1) Generate a $\lambda_{Hi} \sim Poisson(\Lambda = 30)$

(2c.2) Sample λ_{Hi} cases from LN(0.5, 0.25)

(3) End Loop.

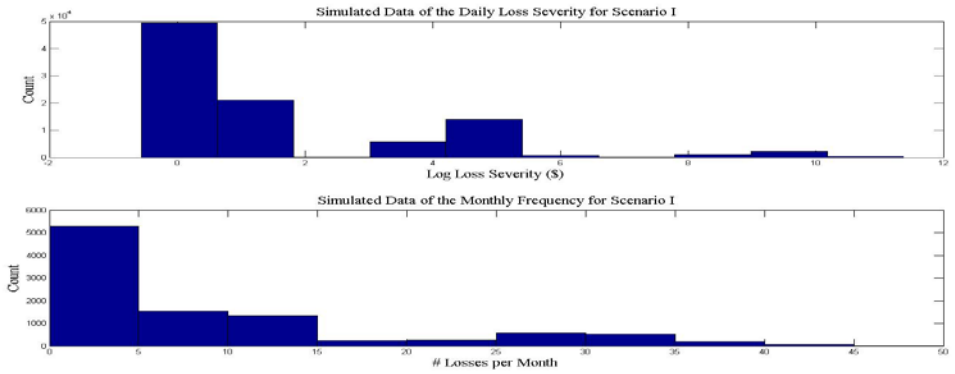


Figure 13 Severity and Frequency distribution of Scenario (I) simulated data

Scenario (II) maps one frequency type to one severity type. Specifically, Scenario (II) simplifies Scenario (I) by having high frequency with low severity, medium frequency with medium severity and finally low frequency with high severity. The algorithm for Scenario (II) with corresponding simulation parameters is listed next along with the simulated data histograms as shown below in figure (14) [67].

Scenario II: Hi/Med/Low Severity mapping to Hi/Med/Low Frequency (one-to-one)

(1) For i=1:*n* months (set *n* to a large number of iterations, which is 10,000)

(2) Generate a $\lambda_{Low} \sim Poisson(\Lambda = 2)$

(3) Sample λ_{Low} samples from LN(9.3, 0.6)

(4) Generate a $\lambda_{Med} \sim Poisson(\Lambda = 10)$

(5) Sample λ_{Med} samples from LN(4.5, 0.5)

(6) Generate a $\lambda_{Hi} \sim Poisson(\Lambda = 30)$

(7) Sample λ_{Hi} samples from LN(0.5, 0.25)

(8) Frequency for Month i = $\lambda_{Low} + \lambda_{Med} + \lambda_{Hi}$

(9) Aggregate Loss for Month i = Sum of Hi, Med, Low Frequency Severities

(10) End Loop.

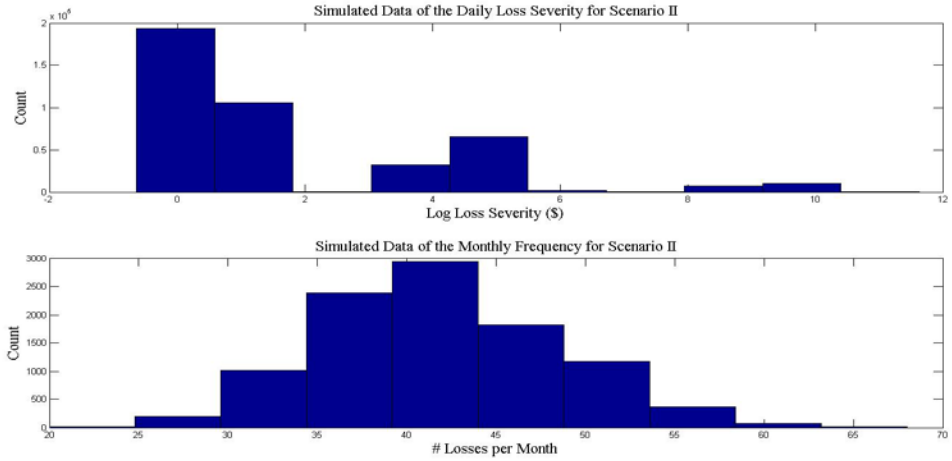


Figure 14 Severity and Frequency distribution of Scenario (II) simulated data

Scenario (III) generates data assuming the classical framework: independence of frequency and severity. The algorithm for Scenario (III) with corresponding simulation parameters is listed next along with the simulated data histograms as shown below in figure (15) [67].

Scenario III: Independence of Severity and Frequency

(1) For i=1:*n* months (set *n* to a large number of iterations, which is 10,000)

(2) Generate a $\lambda_{Fixed} \sim$ Poisson($\Lambda = 14$)

(3) Sample λ_{Fixed} samples from LN(5, 2)

(4) Sum to get the Aggregate Loss for Month i

(5) End Loop.

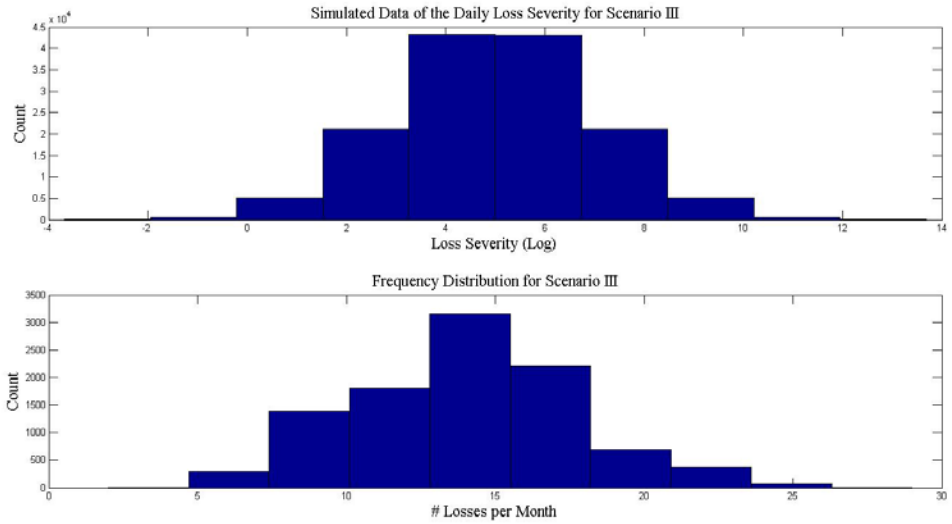


Figure 15 Severity and Frequency distribution of Scenario (III) simulated data

Scenario (IV) generates data which mimics a mixture process for low and high frequency/severity. Specifically, low frequency corresponds to high severity with a low probability, while typically high frequency corresponds to low severity with a high probability. For severity component, the variance is fixed to be the same for both regions and only the mean parameter is changed. The algorithm for Scenario (IV) with corresponding simulation parameters is listed next along with the simulated data histograms as shown below in figure (16).

Scenario IV: Mixture process for Hi/Low Severity to Frequency

(1) For i=1:n months (set *n* to a large number of iterations, which is 10,000)

(2) Generate a u ~ U[0,1]

(2a) If u < 0.3 then

(2b) Generate a λ_{Low} ~ Poisson($\Lambda = 5$) & sample from LN(10, 0.5)

(2c) else; Generate a λ_{Hi} ~ Poisson($\Lambda = 50$) & sample from LN(1, 0.5)

(3) Aggregate Loss for Month i = Sum of Low + Hi Severities

(4) End Loop.

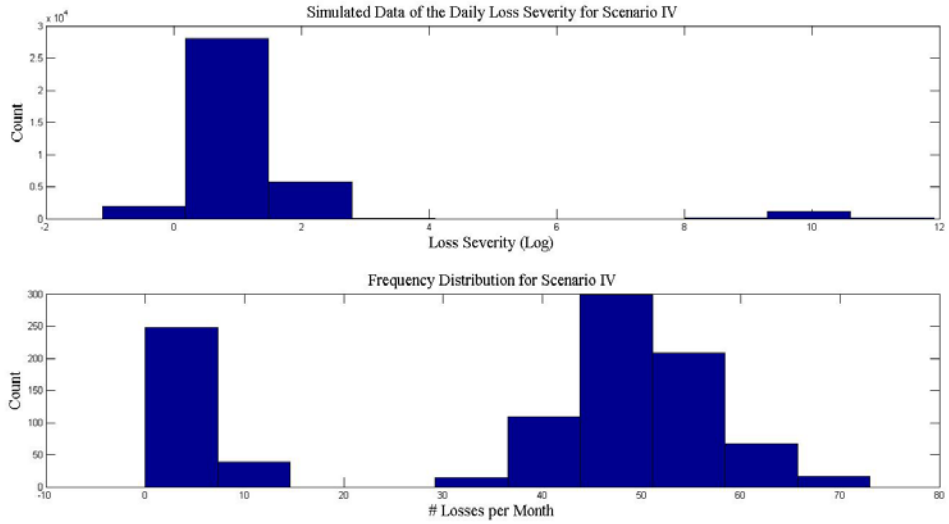


Figure 16 Severity and Frequency distribution of Scenario (IV) simulated data

Scenario (V) generates data which has a perfect correlation between frequency and severity. Specifically, lower frequency corresponds to higher severity while higher frequency corresponds to lower severity. For severity component, the variance is fixed to be the same for both regions and only the mean parameter is changed. The algorithm for Scenario (V) with corresponding simulation parameters is listed next along with the simulated data histograms as shown below in figure (17).

Scenario V: Perfect correlation between Severity and Frequency

- Assume $\lambda \sim$ Discrete Uniform[10, 210] (represents annual losses)
- Generate $\mu \mid \lambda$ from the following table (8) below:

Table 8 Perfectly correlated frequency & severity for Scenario (V) study

μ	λ
10	10
9.9	11
9.8	12
9.7	13
9.6	14
9.5	15
9.4	16
⋮	⋮
1	210

- Given μ, λ then generate losses from LN(μ, 0.1) with count of λ losses.

(1) For i=1:n years (set *n* to a large number of iterations, which is 1,000)

(2) Generate a λ ~ Discrete U[10, 210]

(2a) Given λ, find corresponding μ from table (8) above.

(2b) Draw λ samples from LN(μ, 0.1)

(3) Aggregate Loss for Year i = Sum λ samples from LN(μ, 0.1)

(4) End Loop.

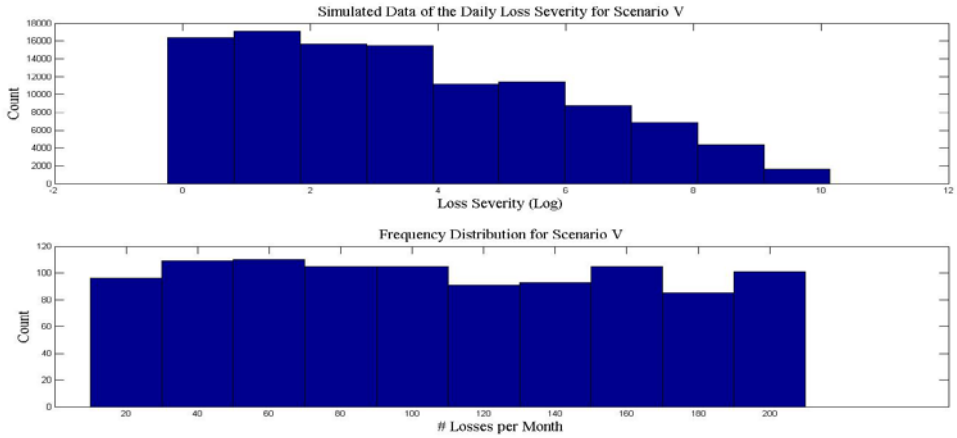


Figure 17 Severity and Frequency distribution of Scenario (IV) simulated data

These five scenarios are created based on the following motivation. It is expected that in Scenarios (I) and (II), the K-means methodology should outperform the classical methodology. In Scenario (III), classical methodology should do well, and the K-means methodology should show that K=1 is best partition of data. Finally for Scenarios (IV) and (V), the copula based methodologies should outperform the classical methodologies. The results for verification are presented in chapter 4. Next, narration of the real-world datasets studied in this dissertation is given.

3.4.2.2 Real-World Datasets across Multiple Domains for VaR Analysis

There are four areas where empirical analysis is conducted to validate the new methodologies for estimating the VaR. The categories are the following: (1) Financial losses; (2) Government insured losses; (3) Natural Losses; (4) Insurance based losses. The basics characteristics of these datasets are described next.

The first category that this dissertation investigates is financial losses. Note that most losses from financial firms are proprietary and can't be publicly obtained. However, there is a plethora of stock market based data which is publicly available. Usually for stock market data, most people are interested in gains and losses. However, from a modern ORM context, this dissertation looks only at the losses. So the natural application of this methodology would be to a risk-averse investor who is highly conscious of minimizing his/her losses. Another practical scenario for this is a new hedge-fund manager who wants to prove that he/she does not have severe drawdown in order to build his/her reputation. Thus, the key idea is to minimize the downside risk. In figures (18)-(19), the severity and frequency histogram for the following two indices are shown: (1) Standard & Poor's 500 (S&P 500); and (2) Dow Jones Industrial Average (DJIA) [91]. For this data, daily log return, i.e. $\log(P_{t+1}/P_t)$(where P_t represents the daily closing price at time t), is computed and only analyzed in instances when this return is negative. In order to have finite positive values, this return is scaled by $10,000 to indicate a portfolio loss. For S&P 500, the data is used from 1928 - 2015, while for DJIA the data is used from 1950-2015 [67]. The severity time unit is daily losses, while the frequency time unit is monthly loss. There are approximately 10,000 loss data points for the S&P 500 and approximately 7,500 daily loss data points for the DJIA. There are approximately 1,000 months of frequency data for S&P 500 and 780 months for DJIA. Therefore, a monthly Aggregate VaR is computed.

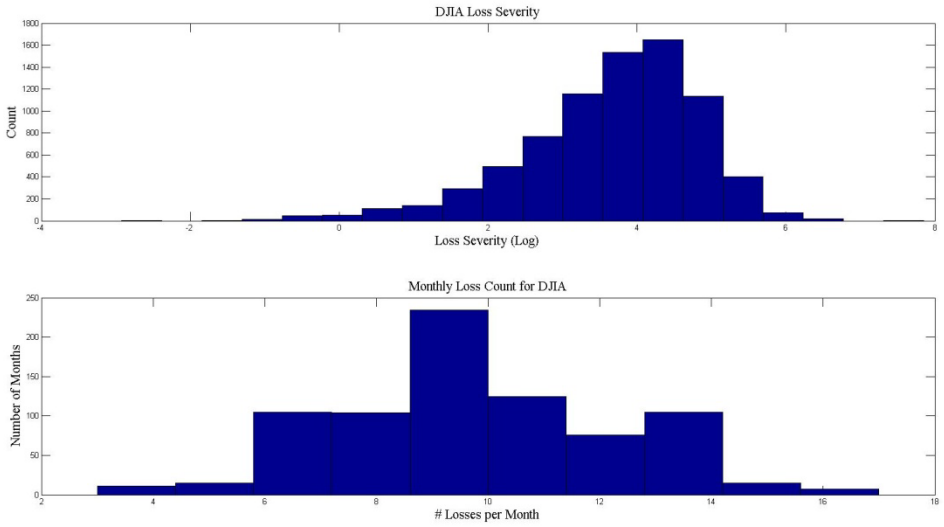


Figure 18 Data Characteristics of DJIA

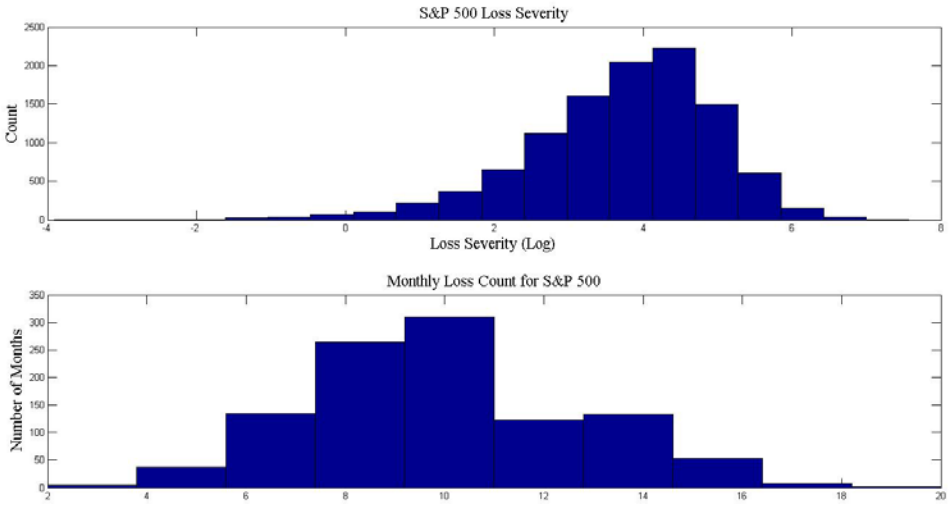


Figure 19 Data Characteristics of S&P 500

The next dataset that is investigated falls in the government tracked losses. The publicly available data in this domain comes from US Coast Guard's Chemical Spills loss

database [92]. From 1990 till 2015, there is approximately 300 months worth of data. The spills occur anytime during the day. In figure (20) below, the loss data characteristics are shown [67]. Overall there are approximately 5,000 individual losses which span the approximately 300 months from 1990 - 2015. Thus, a monthly Aggregate VaR is computed.

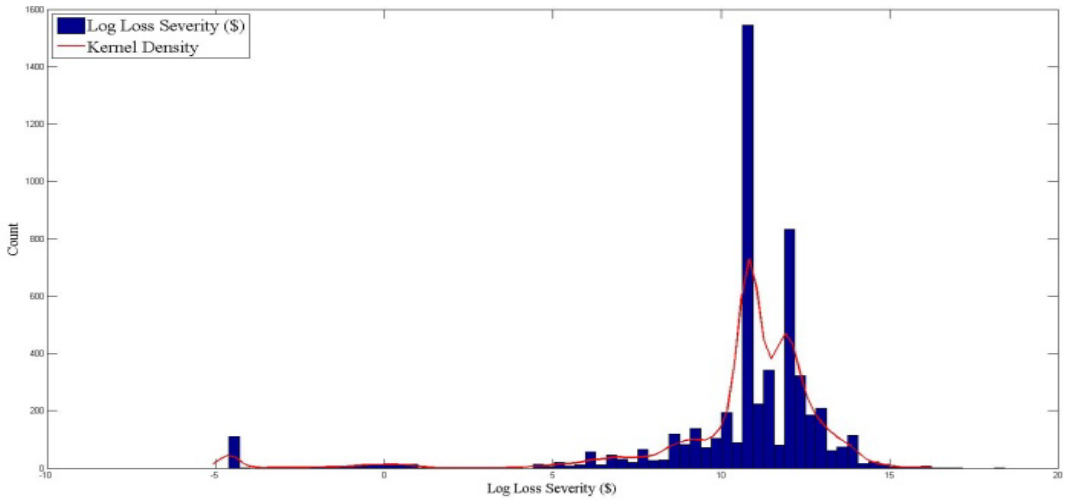


Figure 20 Data Characteristics of Chemical Spills US Coast Guard

The next dataset that is investigated falls in the insurance world. The publicly available data in this domain comes from automobile accidents in Australia from 1989-1999 [93]. There is approximately 120 months worth of frequency data. The accident loss severity occurs anytime during the day. In figure (21) below, the data characteristics are shown. Overall there are approximately 22,000 individual losses which span over approximately 120 months. Thus, a monthly Aggregate VaR is computed.

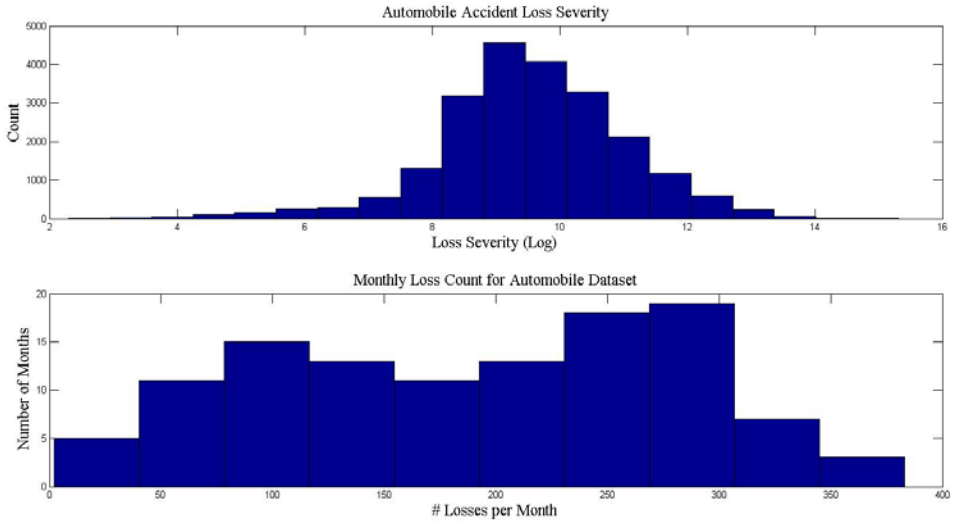


Figure 21 Data Characteristics of Australian Automobile accidents

The final dataset that is investigated falls in the natural calamities. The publicly available data in this domain comes from US hurricane losses from 1900-2005 [94]. There is 105 years worth of frequency data. The accident loss severity occurs anytime during the year. In figure (22) below the data is shown for approximately 200 individual losses which span 105 years. Thus, an annual Aggregate VaR is computed.

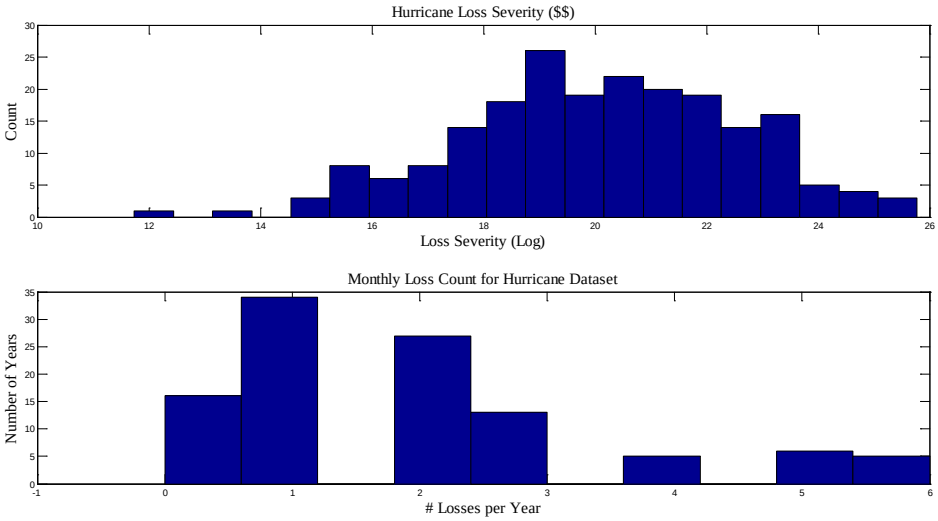


Figure 22 Data Characteristics of US hurricanes

CHAPTER 4 RESULTS

This chapter shows the results which form the basis of the contributions of this dissertation. This section is organized by first showing the contribution this dissertation provides in determining a flexible severity distribution which can fit operational risk loss data and its impact on estimating the Aggregate Loss distribution. Next, the results and the discussions of the findings for the contributions regarding robust estimation of VaR through DPFS via distribution-free methods and DBP via copulas both of which do not assume the independence assumption of severity and frequency are shown.

4.1 Flexible Severity Distribution and Impact on VaR

Determining flexible severity distribution is largely a Monte-Carlo simulation based study. Therefore, the flexibility is tested using high parameter distributions, i.e. three parameter distributions such as Burr and LNG with heavy tails. A large dataset consisting of 10,000,000 daily loss severities from a Lognormal-Gamma distribution is generated with the following parameters: mean (μ)=9; standard deviation (σ)=2; and kurtosis (k)=5 [10]. Then this empirical loss severity dataset is fit to the following severity type distributions: (1) Weibull, (2) lognormal, (3) Lognormal-Gamma, (4) GPD, and (5) Burr. Instead of performing graphical/statistical tests of goodness of fits (such as χ^2 test or Anderson-Darling tail tests, etc), a numerical comparison using empirical percentiles is shown below in table (9) [10].

Table 9 Fitting random high severity data; using LNG as base

True Severity Distribution		Fitted Severity Distributions				
Percentile	Lognormal-Gamma	Weibull	Lognormal	LNG	GPD	Burr
99.95	62,358,321	19	1,523	62,358,325	408,234,509	4,560,456
99.9	25,789,098	8	521	25,789,100	157,892,234	1,740,731
99.5	3,451,989	-	22	3,451,990	21,456,897	226,475
99	1,345,897	-	3	1,345,890	8,567,987	91,234
95	214,589	-	-	214,591	1,082,321	8,759
90	92,345	-	-	92,340	450,232	2,845
50	7,528	-	-	7,530	51,234	43
25	3,214	-	-	3,233	11,357	9
Min	0	-	-	0	0	0

LNG	
Parameters	Theoretical
Mean	9
Standard Deviation	2
Kurtosis	5

.

Notice how one can get a quick estimate of the fit by just looking at the percentile comparisons in tabular form. For example, at the 99.9 percentile, the theoretical (true) value is approximately $25 million, and the GPD does an overestimate of approximately $160 Million, while the Burr does an underestimate of $1.8 million [10]. Notice how the Weibull and lognormal fail completely to fit this leptokurtic type of data. This result can perhaps be explained by the fact that Weibull is primarily characterized as a thin-tailed distribution, and similarly lognormal has an exact kurtosis of 3. As expected, the LNG fits itself quite well. However, as mentioned in section 3, since the LNG does not have a *closed form* density, a "good" MLE fit is not guaranteed [10].

Next, with the same framework it is important to test if the results translate forward when performing MCS for Aggregate Loss distribution. For this simulation study, the frequency distribution is set to a Poisson frequency distribution with parameter

of λ=10 losses per month and the corresponding VaR is calculated as shown in table (10) below [10].

Table 10 Simulation results for Aggregate Loss distribution using LNG as the base

True Severity Distribution		Fitted Severity Distributions				
Percentile	Lognormal-Gamma	Weibull	Lognormal	LNG	GPD	Burr
99.95	62,358,321	19	1,523	62,358,325	408,234,509	4,560,456
99.9	25,789,098	8	521	25,789,100	157,892,234	1,740,731
99.5	3,451,989	-	22	3,451,990	21,456,897	226,475
99	1,345,897	-	3	1,345,890	8,567,987	91,234
95	214,589	-	-	214,591	1,082,321	8,759
90	92,345	-	-	92,340	450,232	2,845
50	7,528	-	-	7,530	51,234	43
25	3,214	-	-	3,233	11,357	9
Min	0	-	-	0	0	0

Aggregate Loss Distribution		Fitted Aggregate Loss				
Percentile	Lognormal-Gamma	Weibull	Lognormal	LNG	GPD	Burr
99.95	19,139,142,100	1,756,897,234	6,812,234,232	18,947,760,669	49,893,497,789	18,568,806,451
99.9	8,823,025,340	1,000,000,982	3,098,124,521	8,734,805,108	22,044,926,382	2,876,897,121
99.5	1,332,466,429	271,897,787	675,098,123	1,319,151,766	2,894,958,587	1,102,232,923
99	414,756,364	160,324,345	309,987,323	410,618,811	1,445,367,061	402,407,431
95	78,282,735	40,897,232	61,098,123	77,509,906	130,723,020	75,960,716
90	34,835,875	25,092,109	30,235,897	34,497,486	76,246,227	33,808,528
50	7,863,879	7,876,909	6,109,232	7,795,245	13,854,942	7,640,333
25	3,110,585	3,123,451	4,098,123	3,089,469	8,758,838	3,028,684
Min	456,173	200,321	259,879	461,606	389,599	370,279
						985
EL	98,621,642	17,109,091	43,098,123	97,645,419	229,049,322	95,693,486

LNG	
Parameters	Theoretical
Mean	9
Standard Deviation	2
Kurtosis	5

It is interesting to note that while the MLE fit has failed for the Burr distribution, the Aggregate Loss distribution estimates are very reasonable. The true EL is actually around $98 million while the Burr distribution estimated through MCS the VaR around $96 million [10]. For the 99.95% quantile, the Burr distribution estimated an $18.6 billion value for the VaR, while the actual VaR value is near $19 billion [10]. This shows an error rate of about 0.2% which is quite reasonable by any standard.

The next simulation is using the base (i.e. theoretical) distribution as the GPD. The reason is that this is a classic EVT class of POT distribution. Particularly, this distribution can be used to model data consisting solely of heavy tails. Therefore, it is interesting to see how the results vary from the previous case where a heavy tail and a

large body (i.e. LNG distribution) is used as the base. As shown in table (11) below, the GPD fails to fit itself at the $0 threshold [10]. It can only fit itself from a certain finite threshold ($20K in this example) as shown in the figure below. This is not surprising, since GPD comes from the EVT class of POT distributions. It is also important to notice from table (11) that for the MLE portion only, the Lognormal-Gamma does a remarkable job in the fit. For the lower ends of the distribution, like at the 25th percentile, the Lognormal-Gamma is showing a severity value of $7,700 while the actual severity value is $7,500 [10]. For the higher ends of the tail, the 99.95% actual value is around $145 million (for the severity) while the Lognormal-Gamma is showing $130 million [10]. Also the Burr distribution does very well in this regards. When one moves to the Monte Carlo results for the Aggregate Loss distribution in table (11), the Lognormal-Gamma does a reasonable job fitting this distribution. In reality, the GPD is not used that often to fit the operational risk loss data due to its stability issues. However, the figure below shows that even if GPD is the true distribution, the three parameter distribution of Lognormal-Gamma can do a reasonable job to fit and simulate Aggregate Loss distribution [10]. While the three-parameter Burr distribution performs marginally better amongst all distributions, it is not at all intuitive to interpret the meaning of the parameter estimates. On the other hand, the LNG distribution has clear and intuitive statistical meaning for each of its three parameters, namely, mean, variance and kurtosis. Thus based on the above simulation studies, the LNG distribution is chosen as a good "flexible" distribution for fitting both the severity and compute VaR from the aggregate loss distribution.

Table 11 Simulation results for Aggregate Loss distribution using GPD as the base

True Severity Distribution		Fitted Severity Distributions					
Percentile	Generalized Pareto (GPD)	GPD	Pareto	Burr	Weibull	Log-Normal	LogNormal Gamma
99.95	144,796,533	364,879,590	2,527,142,976	145,378,707	605,699	14,437,250	130,100,754
99.9	63,013,402	157,100,095	865,766,761	63,295,855	242,680	7,554,730	56,838,765
99.5	9,121,221	22,192,718	71,973,536	9,170,229	19,193	1,431,737	9,189,903
99	3,976,520	9,546,900	24,657,210	3,984,361	5,020	639,038	4,194,104
95	561,239	1,336,205	2,049,821	564,694	82	70,556	629,080
90	245,608	566,559	702,242	237,006	7	21,795	262,738
50	23,543	66,842	58,379	20,813	0	346	24,158
25	7,528	34,830	31,197	6,641	0	39	7,713
Min	0	20,000	20,000	0	0	0	0

Aggregate Loss Distribution		Fitted Aggregate Loss					
Percentile	Generalized Pareto (GPD)	GPD	Pareto	Burr	Weibull	Log-Normal	LogNormal Gamma
99.95	35,139,142,100	55,437,218,661	1,000,008,212,620	36,378,190,759	1,254,760,110	1,649,039,896	26,390,136,271
99.9	15,823,025,340	24,494,361,551	379,078,656,551	16,019,872,405	857,129,394	1,053,398,400	7,892,311,833
99.5	2,346,966,429	3,618,696,992	31,536,012,354	2,342,102,756	330,293,322	357,854,139	1,423,073,123
99	1,014,756,364	1,605,962,288	10,833,119,786	1,033,727,750	216,012,691	223,948,081	684,915,818
95	168,282,735	261,444,081	934,832,496	168,027,860	80,203,322	77,335,901	135,310,321
90	81,835,875	127,075,362	333,501,838	81,443,369	52,338,203	49,736,322	70,840,306
50	17,865,369	27,707,922	32,169,281	17,525,310	17,621,573	16,820,061	17,297,126
25	11,311,585	17,515,681	15,881,642	10,970,930	11,611,139	11,217,838	11,008,223
Min	1,456,173	2,590,765	1,356,906	1,134,607	1,159,222	1,416,721	1,170,332
EL	426,621,642	699,049,732	1,506,450,045	179,554,385	30,172,394	30,478,464	140,943,017

Generalized Pareto	
Parameters	Theoretical
Scale parameter	19,000
Shape parameter	-1.2
Fit Threshold	20,000
Sample size	10,000,000

4.2 VaR Estimation via DPFS with Distribution-Free & DBP using Parametric Approaches

This section begins by verifying the distribution-free and parametric copula based methodology on the simulated data from Scenarios (I) - (V). Afterwards, the validation of the analysis is done using the real-world data results (as described in the previous section) are shown. Then, a summary of the overall results found and discussion of the implications of the results from all of the analysis are provided at the end.

4.2.1 Distribution-Free Approach using Clustering

The K-means algorithm is analyzed using Method I and Method II as mentioned in chapter 3. Results begin with the simulated data from the five scenarios (from chapter 3) and then follow up with the real-world data. Afterwards, a summary of the results are discussed.

4.2.1.1 Simulation Scenarios (I) - (V) Results

Next, the K-means algorithm is analyzed using Method I and Method II for Scenario (I). The 2-D K-means clustering for Scenario (I) is shown figure (23). Figure (26) shows the partitioning using K-means Severity only Implied Frequency (Method II) clustering. Notice how the K-means fails to separate the data perfectly for the 2-D case, while it does a good job for the case in Method II.

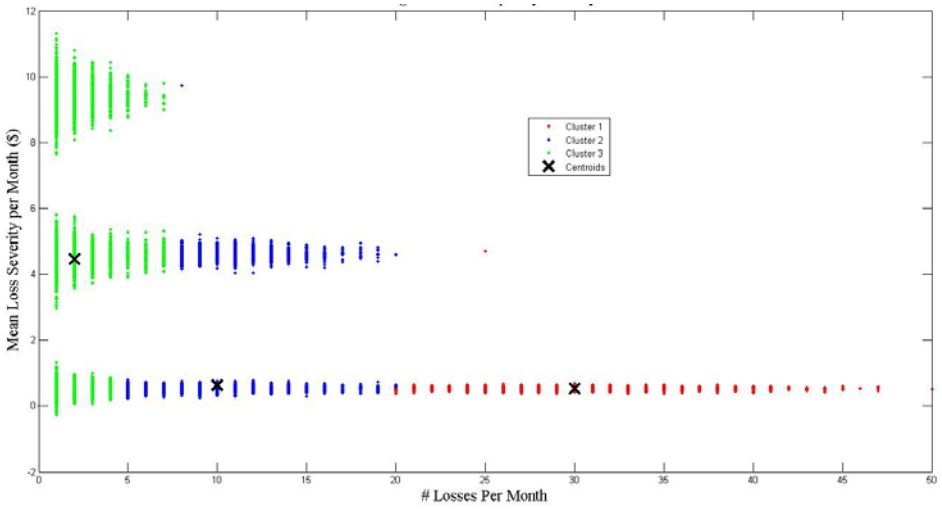


Figure 23 K-Means: 2-D (Method I) for Scenario (I)

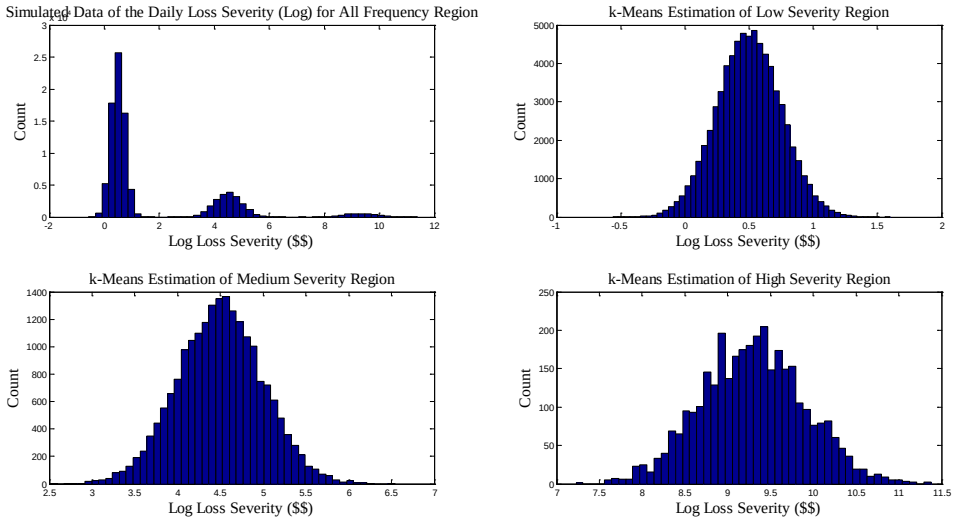


Figure 24 K-Means: Severity Only Implied Frequency (Method II) for Scenario (I)

Next, the analysis of the K-means algorithm using Method I and Method II for Scenario (II) is described. The 2-D K-means clustering for Scenario (II) is shown in figure (25) below. The partitioning using K-means: Method II is given in figure (26) [67]. Notice how the K-means fails to separate the data perfectly for the 2-D case, while it does a good job for Method II.

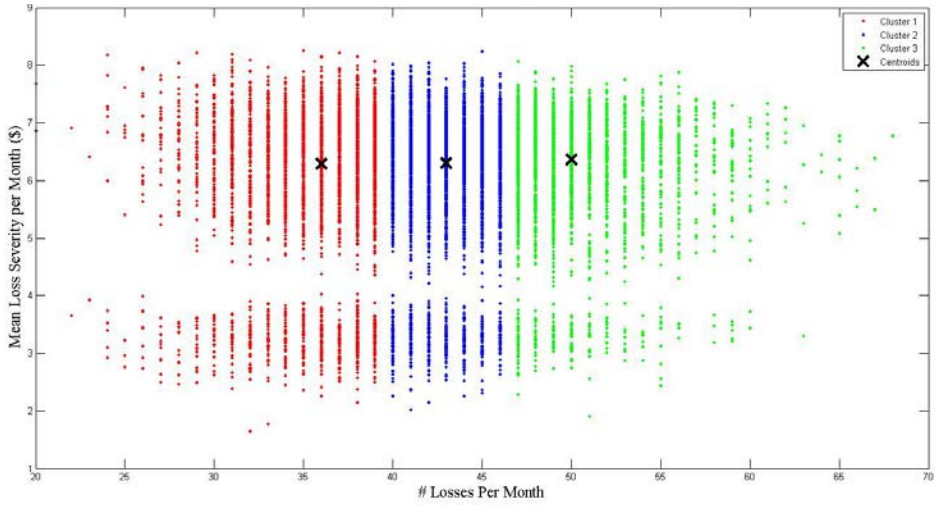


Figure 25 K-Means: 2-D for Scenario (II)

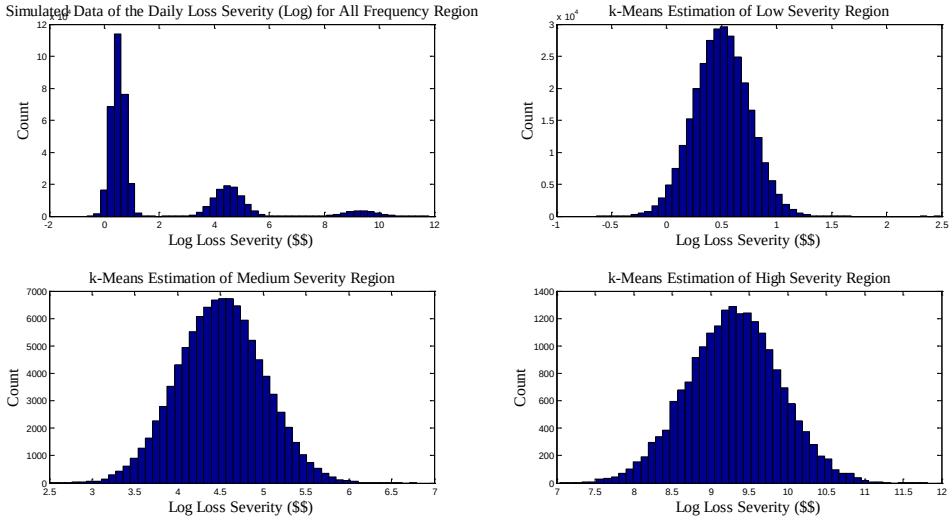


Figure 26 K-Means: Method II for Scenario (II)

Then the K-means algorithm using Method I and Method II for Scenario (III) are analyzed. For this particular scenario, the strength of using the silhouette technique to compute the optimal value of K is shown. This is presented in figure (27) below [67]. The average silhouette value is around 1, and thus a single cluster is appropriate. So the 2-D

K-means from Method I and K-means: Severity only to Implied Frequency (Method II) is meaningless since K=1, implies that the classical method is sufficient.

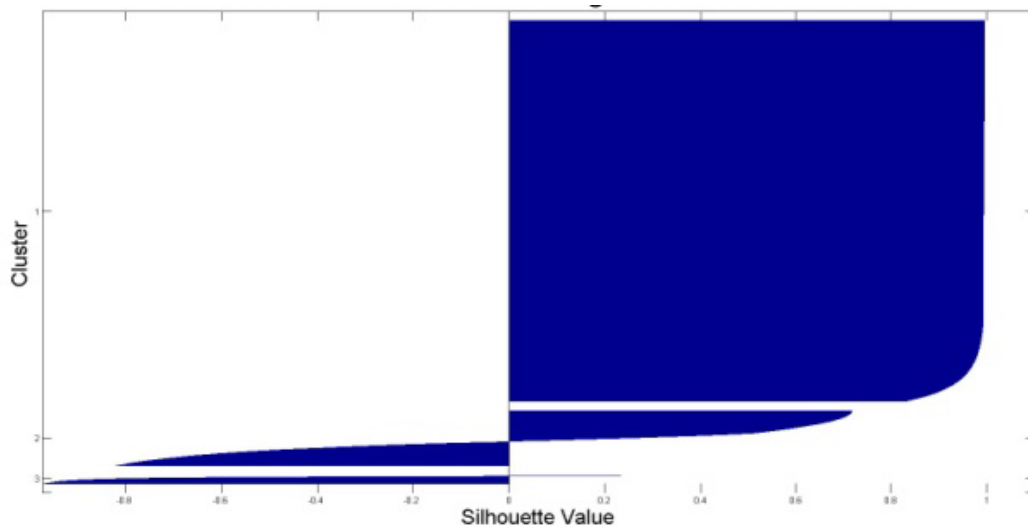


Figure 27 Silhouette plot for determining optimal K for Scenario (III)

Next, an analysis of the K-means algorithm using Method I and Method II for Scenario (IV) is described. The 2-D K-means clustering for Scenario (IV) is shown in figure (28). In addition, the partitioning of the data using K-means: Method II clustering is shown in figure (29) below. Notice how both the K-means methodology properly identify the clusters.

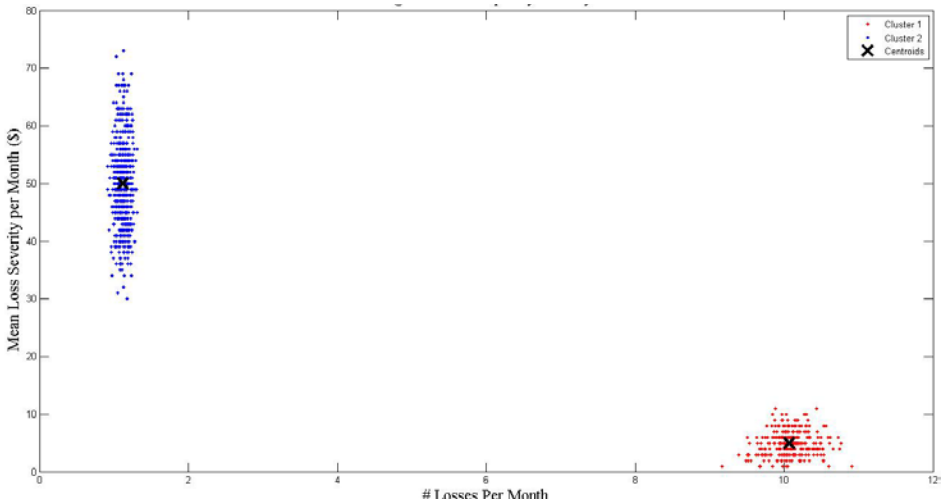


Figure 28 K-Means: 2-D for Scenario (IV)

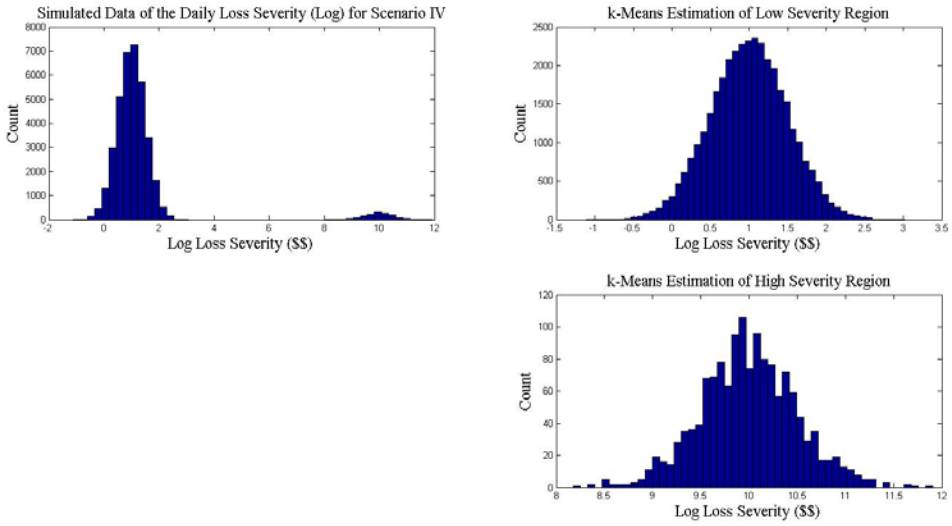


Figure 29 K-Means: Method II for Scenario (IV)

Finally, an analysis of the K-means algorithm using Method I and Method II for Scenario (V) is described. The 2-D K-means clustering for Scenario (V) is given in figure (30) below. Figure (31) shows the partitioning using K-means: Severity only Implied

Frequency clustering. Notice how the 2-D K-means fails to separate the data along with the K-means: Method I.

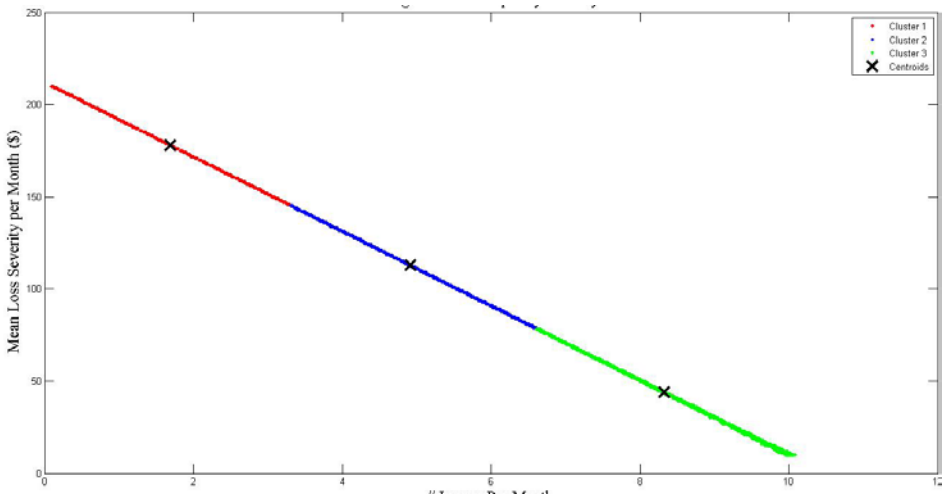


Figure 30 K-Means: 2-D for Scenario (V)

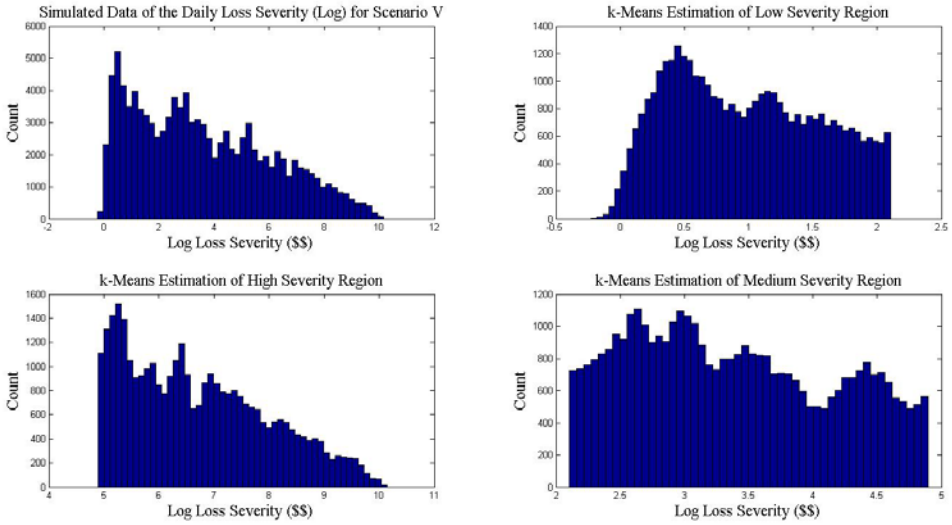


Figure 31 K-Means: Method II for Scenario (V)

Overall, the results from the dissertation show that the K-means methodology works well on most types of the scenarios. For Scenarios (I) - (III), the Method II works well. For the correlated cases, Method I (i.e. mean severity with frequency, also known as 2-D K-means) works well. For Scenario (V), the K-means algorithm does not work well, because there is perfect correlation and the partition requires a large value of K which is not selected. The VaR results for Scenarios (I) - (V) are shown next in tables (12) - (16) below.

Table 12 VaR results for Scenario I using DPFS K-means: Methods I and II

Scenario I			
		DPFS K-Means	
Historical		Method I	Method II
VaR		VaR	VaR
0%	$0	$0	$0
25%	$13	$5	$14
50%	$55	$23	$57
90%	$16,659	$21,844	$20,964
95%	$33,966	$39,285	$37,375
98%	$52,560	$56,327	$54,427
99%	$65,165	$70,248	$66,828
99.5%	$77,042	$85,470	$78,008

Table 13 VaR results for Scenario II using DPFS K-means: Methods I and II

Scenario II			
		DPFS K-Means	
Historical		Method I	Method II
VaR		VaR	VaR
0%	$161	$2,413	$152
25%	$10,514	$10,820	$10,722
50%	$22,815	$16,629	$23,159
90%	$57,130	$53,861	$56,196
95%	$69,590	$86,086	$69,155
98%	$85,204	$161,826	$82,845
99%	$96,556	$240,311	$94,040
99.5%	$107,539	$376,988	$105,457

Table 14 VaR results for Scenario III using DPFS K-means: Methods I and II

Scenario III			
		DPFS K-Means	
Historical		Method I	Method II
VaR		VaR	VaR
0%	$125	$4	$108
25%	$5,143	$5,190	$5,137
50%	$9,268	$9,315	$9,372
90%	$30,377	$30,408	$31,450
95%	$44,748	$44,704	$45,185
98%	$72,481	$72,299	$74,414
99%	$103,036	$103,087	$107,022
99.5%	$145,555	$145,573	$146,053

Table 15 VaR results for Scenario IV using DPFS K-means: Methods I and II

Scenario IV			
		DPFS K-Means	
Historical		Method I	Method II
VaR		VaR	VaR
0%	$0	$40	$43
25%	$144	$640	$8,656
50%	$167	$1,157	$28,646
90%	$146,025	$4,299	$82,321
95%	$185,097	$6,638	$101,008
98%	$227,096	$11,856	$122,585
99%	$255,899	$17,324	$141,709
99.5%	$282,384	$24,684	$158,573

Table 16 VaR results for Scenario V using DPFS K-means: Methods I and II

Scenario V			
		DPFS K-Means	
Historical		Method I	Method II
VaR		VaR	VaR
0%	$230	$73,831	$8,433
25%	$2,090	$248,248	$42,501
50%	$17,247	$304,640	$61,651
90%	$238,502	$462,573	$155,336
95%	$260,502	$527,759	$219,225
98%	$268,995	$617,283	$355,505
99%	$272,649	$687,479	$525,180
99.5%	$275,452	$803,500	$805,360

4.2.1.2 Real-World Data Results

To begin, the results for the financial loss datasets are shown next. Figure (32) below shows the analysis where K-means: 2-D clustering is calculated for the S&P 500 data [67]. For the K-means: Severity only with Implied Frequency, the best fit is found to be K=2, so a corresponding plot is shown in figure (33).

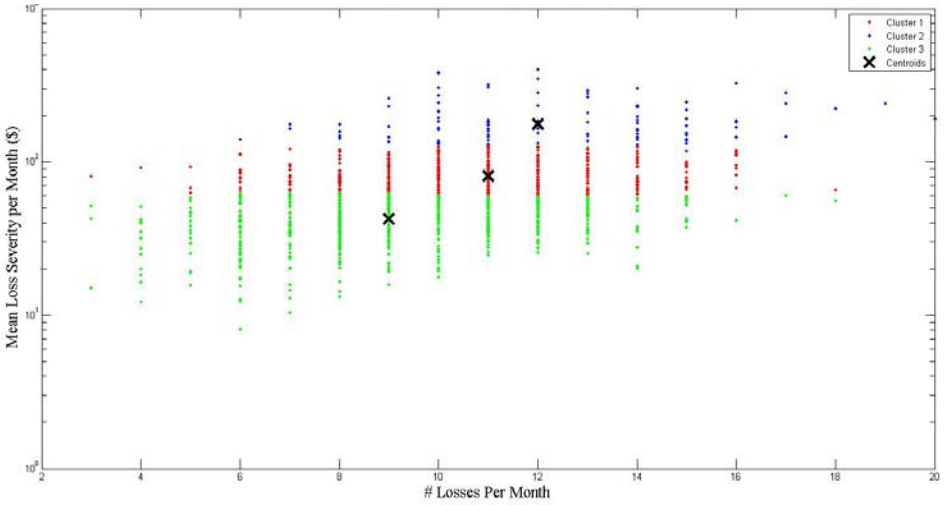


Figure 32 K-Means: Method I (2-D) for S&P 500

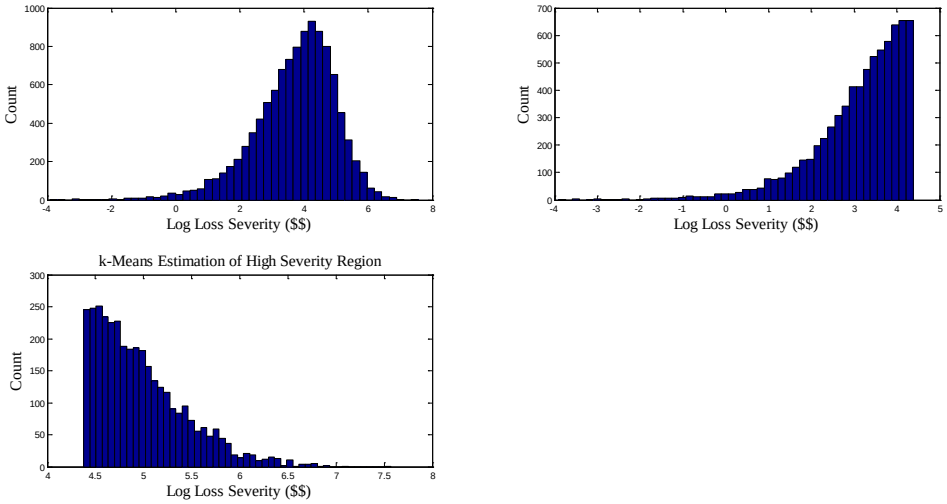


Figure 33 K-Means: Method II (Severity Only Implied Frequency) for S&P 500

The VaR results are computed next using the K-means methodology and compare with historical data and classical methodology. The results are shown in table (17) [67]. Notice how Method I does a very good job in estimating the VaR.

Table 17 VaR results for S&P 500 using K-means Methods I and II

			k-Means Split	
Historical		Classical	Method I	Method II
VaR		VaR	VaR	VaR
Mean	$766	$764	$937	$1,606
0%	$40	$0	$0	$0
25%	$369	$508	$499	$1,160
50%	$586	$723	$783	$1,554
90%	$1,431	$1,234	$1,710	$2,432
95%	$2,027	$1,404	$2,132	$2,712
98%	$3,143	$1,611	$2,772	$3,043
99%	$3,808	$1,756	$3,337	$3,272
99.5%	$4,149	$1,895	$3,990	$3,488
99.9%	$4,825	$2,190	$5,981	$3,943

Next the results for DJIA in figure (32) are shown below where the 2-D K-means clustering for the DJIA data is given. For the K-means clustering using Method II, the best fit is found at K=1, so the K-means plot is the same as the standard severity histogram as shown in previous section (figure 18).

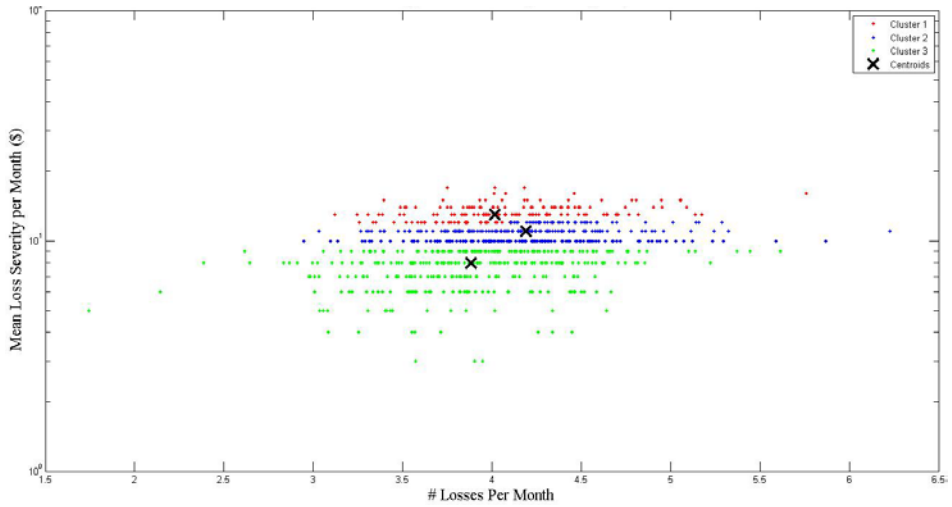


Figure 34 K-Means: Method I (2-D) for DJIA

The VaR results are computed using the K-means methodology and compare with historical data and classical methodology. The results are shown below in table (18). Notice how both Method I and Method II perform well in estimating the VaR.

Table 18 VaR results for DJIA using K-means Methods I and II

DJIA				
			k-Means Split	
Historical		Classical	Method I	Method II
VaR		VaR	VaR	VaR
Mean	$287	$341	$850	$339
0%	$29	$6	$412	$7
25%	$364	$448	$1,098	$441
50%	$566	$683	$1,316	$678
90%	$1,122	$1,375	$1,773	$1,390
95%	$1,465	$1,666	$1,910	$1,702
98%	$2,070	$2,075	$2,088	$2,150
99%	$2,325	$2,454	$2,192	$2,516

Next the results for chemical spills in figures (35)-(36) are shown where the 2-D K-means clustering for the data is given [67]. For the K-means: Severity only Implied Frequency, the best fit is found at K=3, so the K-means plot is shown below in figure (36) [67].

After the K-means Method I and Method II are determined, the corresponding VaR results from these methodologies are computed and compared with historical data and classical methodology. The results are shown below in table (19) [67]. Notice how Method II does a reasonable job in estimating the VaR.

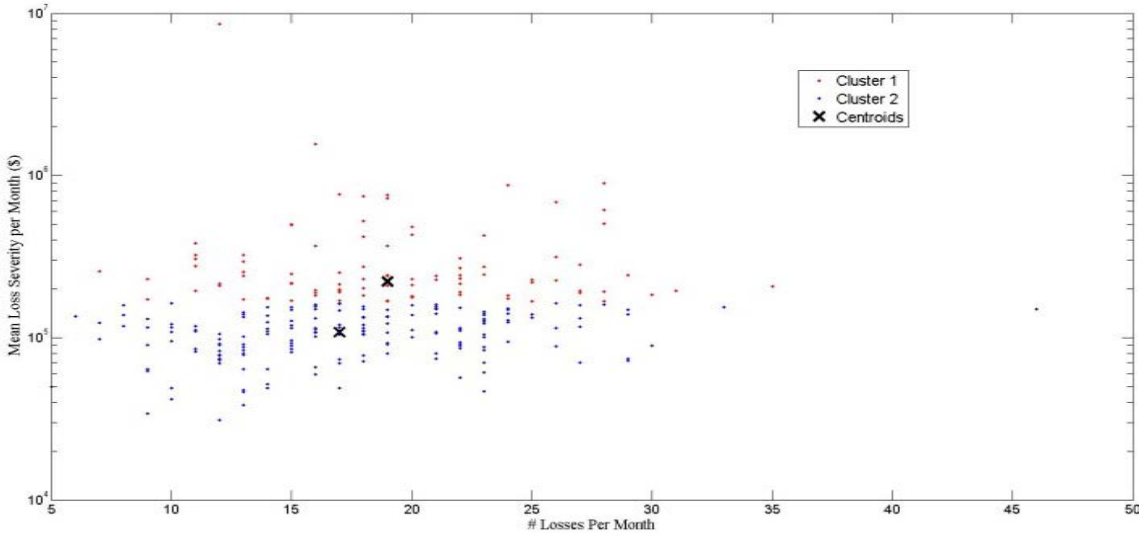


Figure 35 K-Means: Method I (2-D) for Chemical Spills

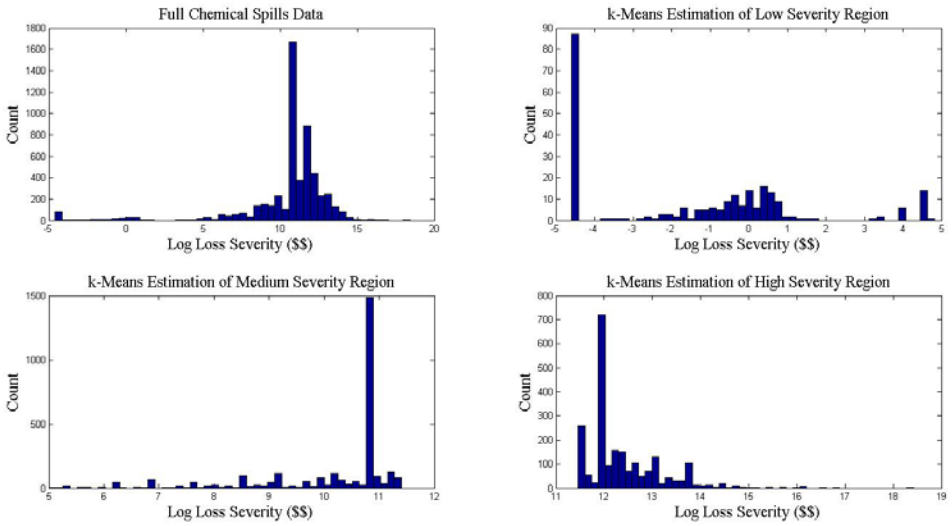


Figure 36 K-Means: Method II for Chemical Spills

Table 19 VaR results for Chemical Spills using K-means Methods I and II

Chemical Spills				
			k-Means Split	
Historical		Classical	Method I	Method II
VaR		VaR	VaR	VaR
Mean	$3,704,235	$20,509,595	$78,656,988	$6,724,650
0%	$250,000	$0	$10,058	$0
25%	$1,579,058	$0	$6,439,424	$0
50%	$2,528,113	$0	$15,803,952	$0
90%	$6,070,004	$23,174,443	$114,902,912	$5,414,227
95%	$8,223,224	$55,905,371	$226,507,895	$9,211,239
98%	$14,428,712	$146,956,339	$519,706,874	$23,887,867
99%	$21,006,921	$284,977,729	$939,953,129	$51,923,390

Afterwards, the results for Automobile accidents are shown in figure (37) below. Here the 2-D K-means clustering is given. For the K-means clustering using Method II, the best fit is found at K=1, so the K-means plot is the same as the standard severity histogram as shown in figure (25).

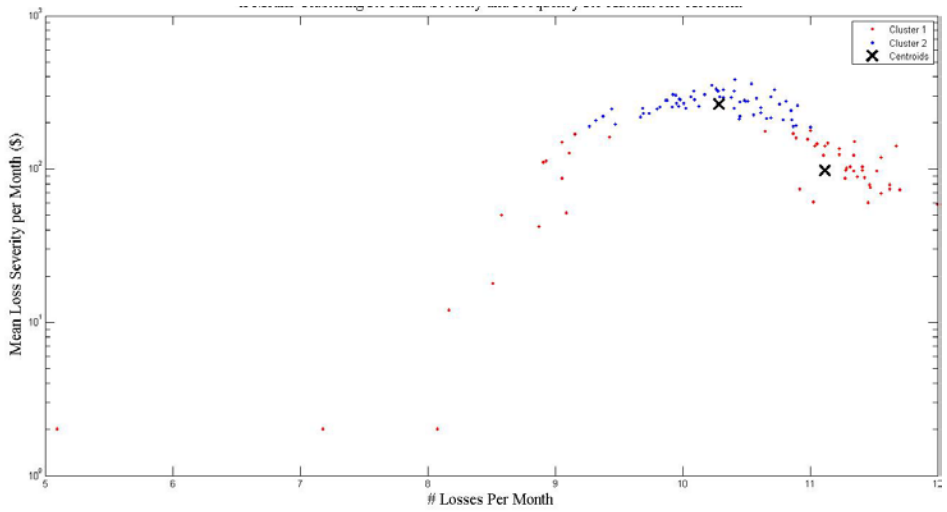


Figure 37 K-Means: Method I (2-D) for Automobile Accidents

The VaR results are computed next using the K-means methodology are compared with historical data and classical methodology. The results are shown below in table (20). Notice how Method II does a reasonable job in estimating the VaR.

Table 20 VaR results for Automobile Accidents using K-means Methods I and II

Automobile				
			k-Means Split	
Historical		Classical	Method I	Method II
VaR		VaR	VaR	VaR
Mean	$7,351,826	$7,903,965	$9,909,820	$7,902,410
0%	$327	$3,652,871	$6,193,134	$3,506,035
25%	$5,279,309	$6,764,007	$9,097,156	$6,780,871
50%	$8,179,051	$7,693,910	$9,859,308	$7,678,260
90%	$11,382,899	$9,935,066	$11,442,192	$9,950,250
95%	$12,841,319	$10,802,019	$11,906,385	$10,820,026
98%	$14,045,680	$12,037,425	$12,461,354	$11,991,467
99%	$14,717,942	$12,988,392	$12,871,537	$12,900,951

Finally, the results for US hurricanes are shown. For both the 2-D K-means (Method I) and the K-means: Method II, it is observed that the best fit occurs at K=1, so

the K-means plot is the same as the standard severity histogram as shown in figure (22). The VaR results are shown in table (21) below. The results show that K-means: Method II is the only one that performs reasonably well in VaR estimation.

Table 21 VaR results for US Hurricanes using K-means Methods I and II

US Hurricanes				
			k-Means Split	
Historical		Classical	Method I	Method II
VaR		VaR	VaR	VaR
Mean	$10,304,351,516	$25,005,170,479	$11,966,093,532,725,000	$15,856,419,244
0%	$0	$0	$0	$0
25%	$145,567,136	$143,064,352	$984,718	$185,902,032
50%	$1,546,550,238	$1,636,020,964	$891,143,303	$1,889,066,337
90%	$26,622,376,781	$32,421,416,186	$8,797,128,358,329	$27,635,698,982
95%	$51,396,924,874	$73,341,304,140	$106,917,479,563,651	$53,991,567,167
98%	$76,583,740,780	$186,681,225,463	$1,509,316,668,391,140	$123,786,358,452
99%	$113,860,560,854	$378,170,152,946	$8,990,749,087,773,140	$230,825,029,288

4.2.2 Parametric Approach using Copula

To begin this subsection, an analysis of the copula based VaR computation as mentioned in the previous chapter is described. As mentioned in the previous section, the first step is to assume a distribution based structure. In this section, the severity is lognormal and the frequency is Poisson distribution for the parametric structure. Therefore, the estimation involves the triplet (λ, μ, σ) for each frequency time period. Note that there are closed form expressions for these quantities from MLE as shown below:

- $\hat{\mu} = \frac{\sum_{i=1}^{n} \text{Log}(x_i)}{n}$ for a sample of $x_1, x_2, ..., x_n \sim LN(\mu, \sigma)$ distribution

- $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left[\text{Log}(x_i) - \frac{\sum_{i=1}^{n}\text{Log}(x_i)}{n}\right]^2}{n}$ for a sample of $x_1, x_2, ..., x_n \sim LN(\mu, \sigma)$

- $\hat{\lambda}$ = count in each time period

Afterwards, the results for the simulated data are shown and then followed up with the results for the real-world data. To conclude, a summary of the results are discussed.

4.2.2.1 Simulation Scenarios (I) - (V) Results

To begin, an analysis of the copula methodology by using Gaussian/*t*-copula along with GMCM is described. Scenario (I) is used as a starting example. Figure (38) shows the frequency and severity parameters for the data on a per month basis.

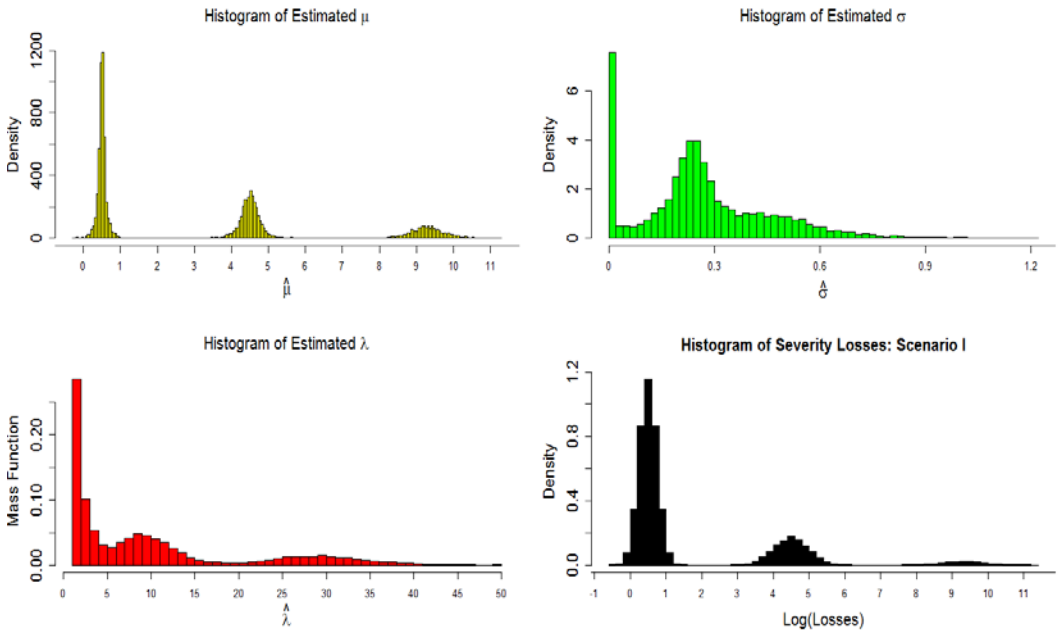


Figure 38 Severity/Frequency Parameter Estimates for Scenario (I)

The corresponding Pearson correlation matrix among the frequency and severity parameters is shown in figure (39) below.

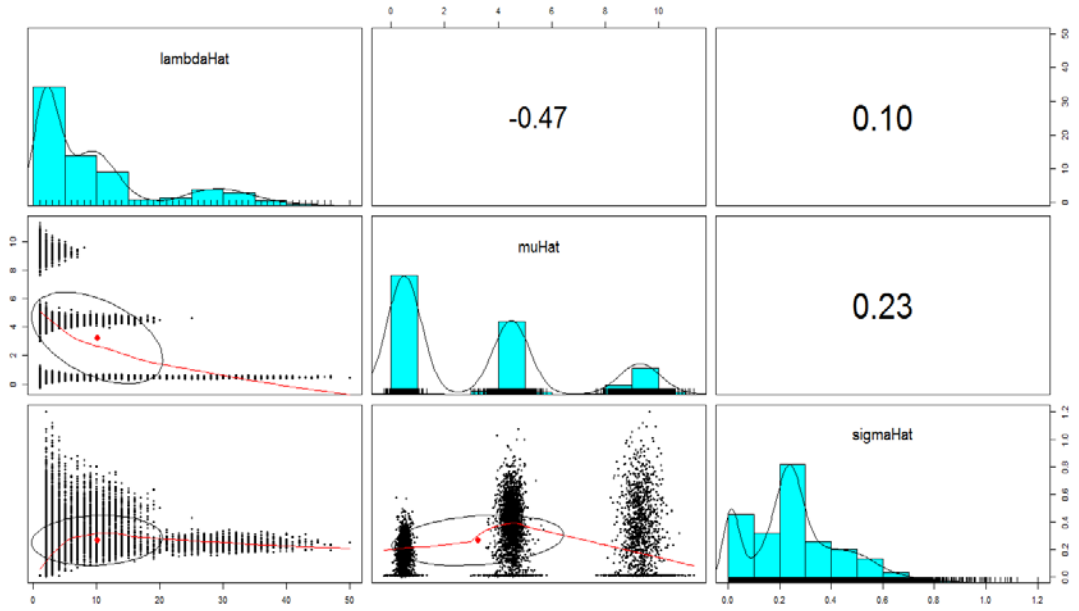


Figure 39 Severity/Frequency Pearson Correlation Estimates for Scenario (I)

The results from copula fitting the empirical surface for Gaussian/*t*/GMCM copulas are given in figures (40)-(42) below.

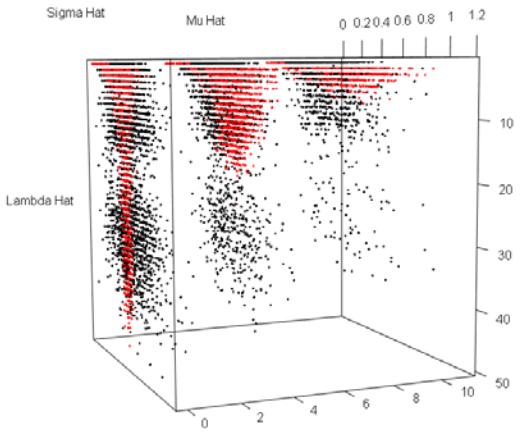


Figure 40 Surface Plot for Scenario (I) using Gaussian Copula (red is data; black is copula)

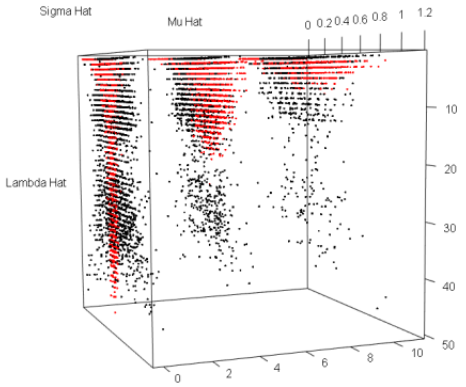


Figure 41 Surface Plot for Scenario (I) using *t*-Copula (red is data; black is copula)

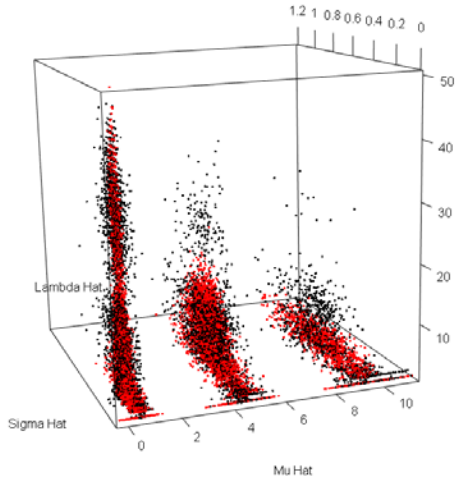


Figure 42 Surface Plot for Scenario (I) using GMCM Copula (red is data; black is copula)

Scenario (II) results are analyzed next. Figure (43) shows the frequency and severity parameters per month time unit.

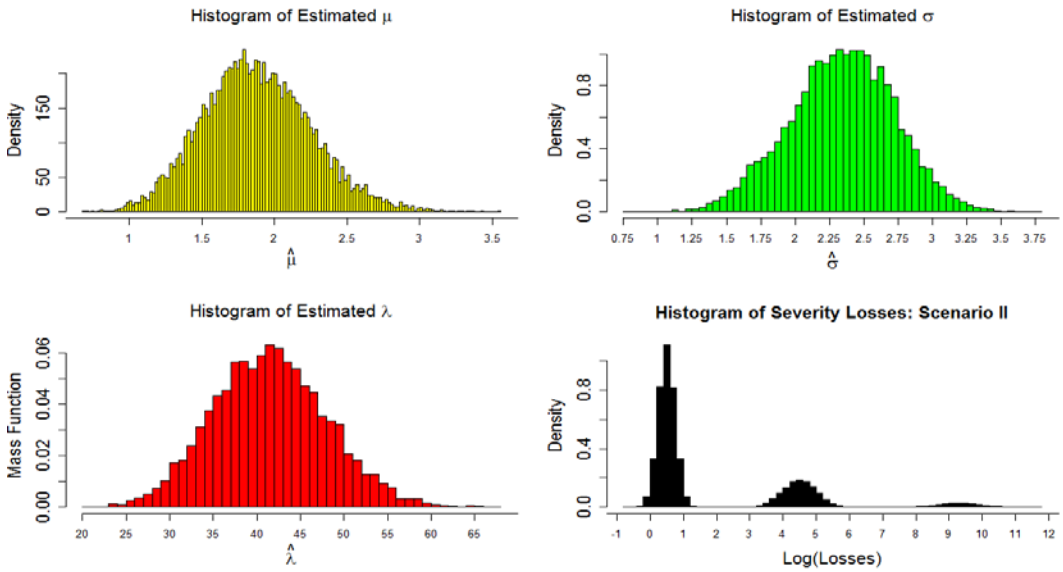


Figure 43 Severity/Frequency Parameter Estimates for Scenario (II)

The resulting Pearson correlation matrix among the frequency and severity parameters is as shown in figure (44) below.

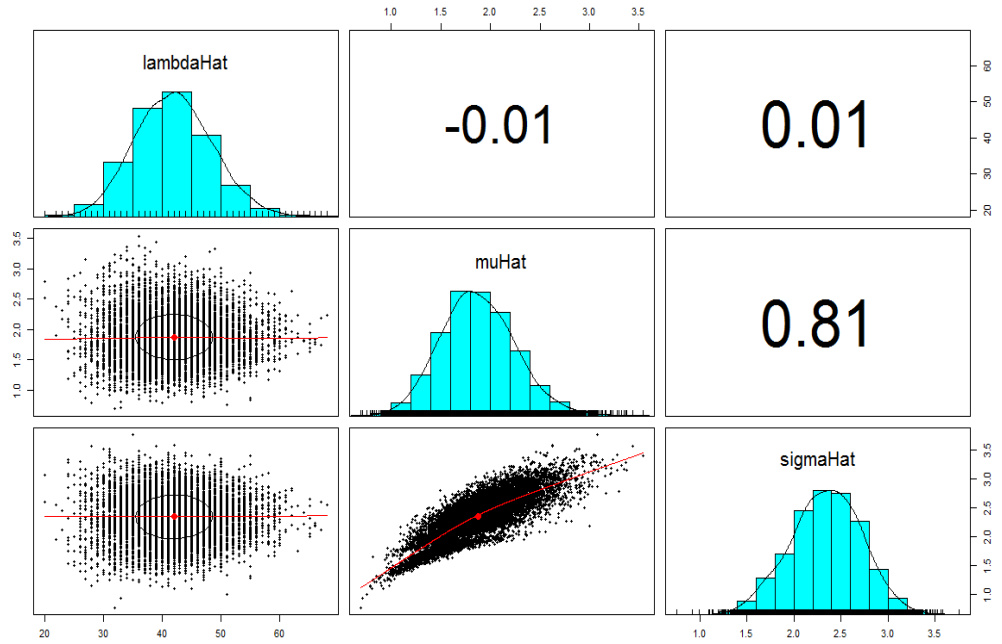


Figure 44 Severity/Frequency Pearson Correlation Estimates for Scenario (II)

The final results from copula fitting the empirical surface for Gaussian/t/GMCM copulas is shown in figures (45)-(47) below.

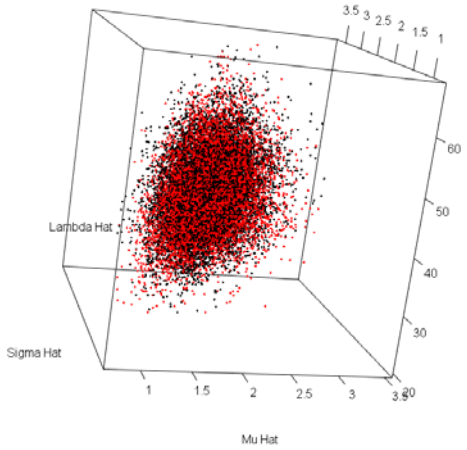


Figure 45 Surface Plot for Scenario (II) using Gaussian Copula (red is data; black is copula)

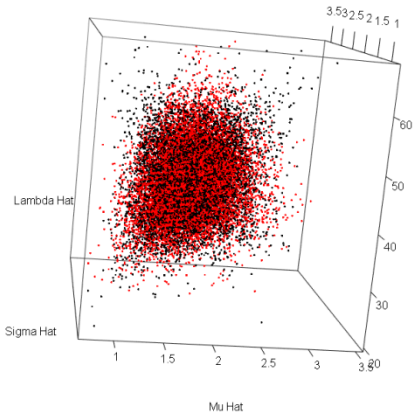


Figure 46 Surface Plot for Scenario (II) using *t*-Copula (red is data; black is copula)

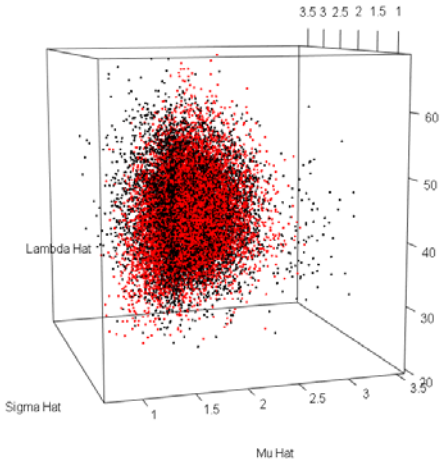


Figure 47 Surface Plot for Scenario (II) using GMCM Copula (red is data; black is copula)

Next, the results from Scenario (III) are analyzed. Figure (48) shows the frequency and severity parameters per month time period.

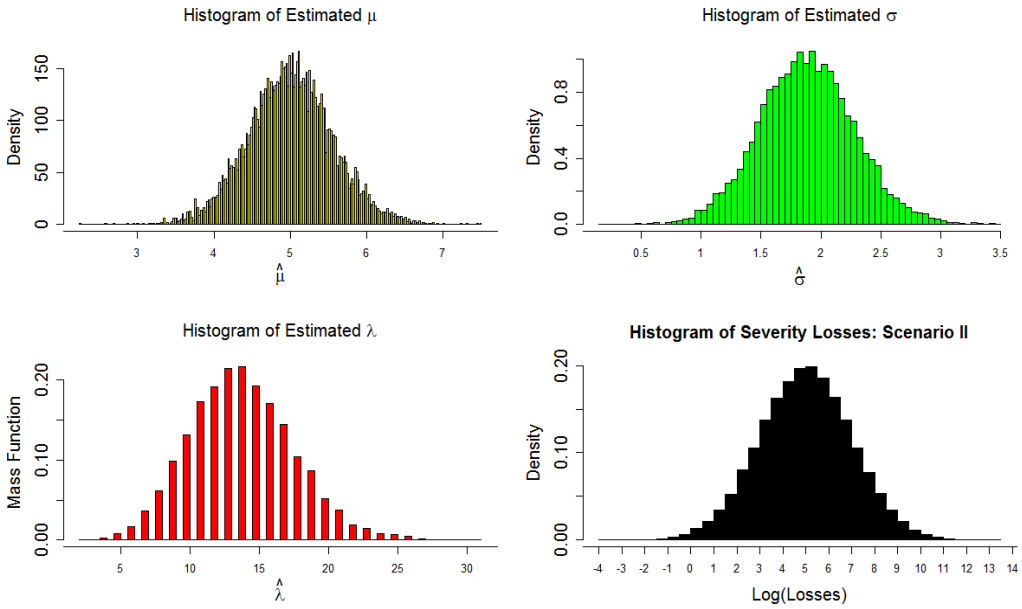


Figure 48 Severity/Frequency Parameter Estimates for Scenario (III)

The Pearson correlation matrix among the frequency and severity parameters is shown in the figure (49) below.

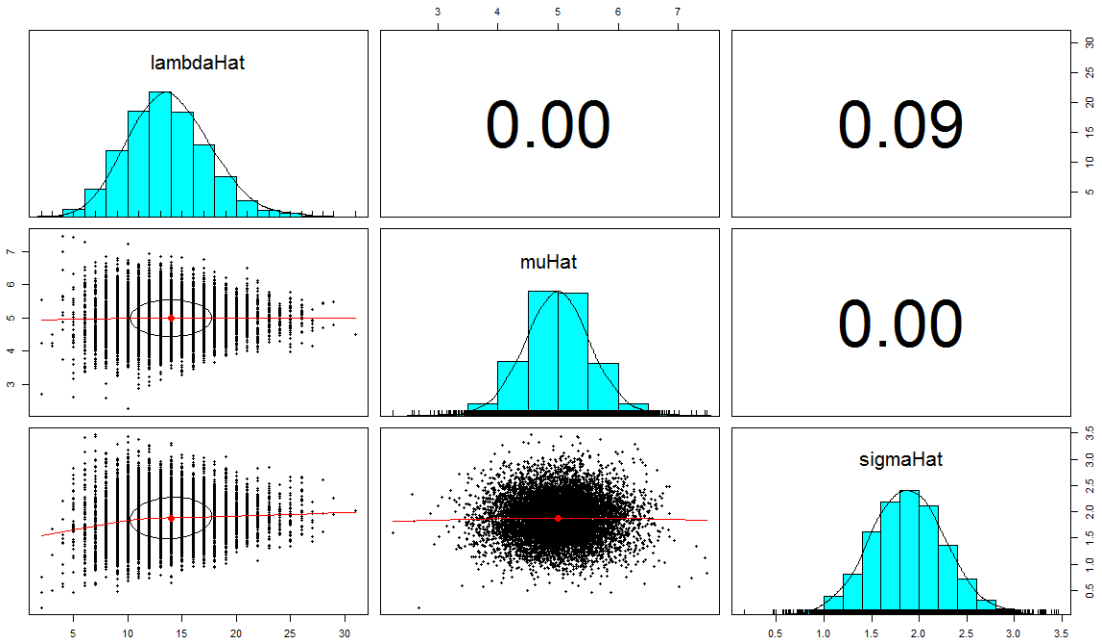


Figure 49 Severity/Frequency Pearson Correlation Estimates for Scenario (III)

Next the computation of the mixture surface for Gaussian/*t* copulas as shown in figures (50)-(52) below. Since the GMCM found K=1, the GMCM becomes the same as the Gaussian copula, and thus its surface plot is not shown in this case.

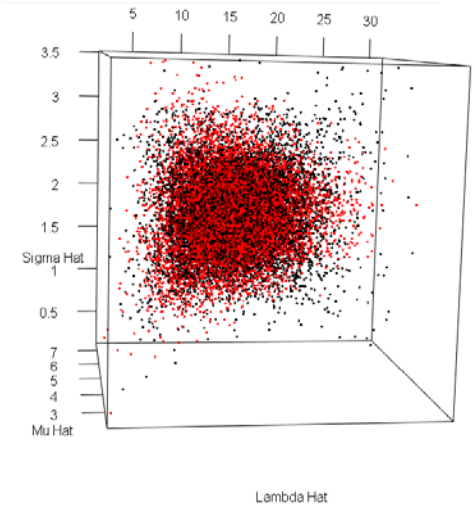


Figure 50 Surface Plot for Scenario (III) using Gaussian Copula (red is data; black is copula)

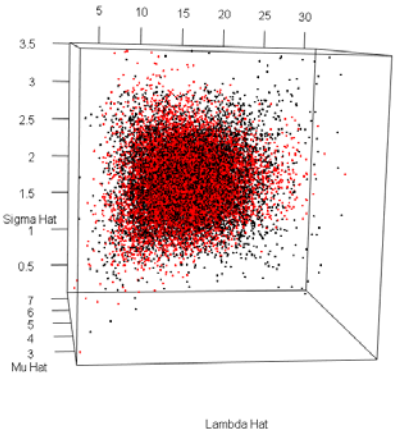


Figure 51 Surface Plot for Scenario (III) using *t*-Copula (red is data; black is copula)

Next the analysis moves to Scenario (IV). Figure (52) shows the frequency and severity parameters per month time unit.

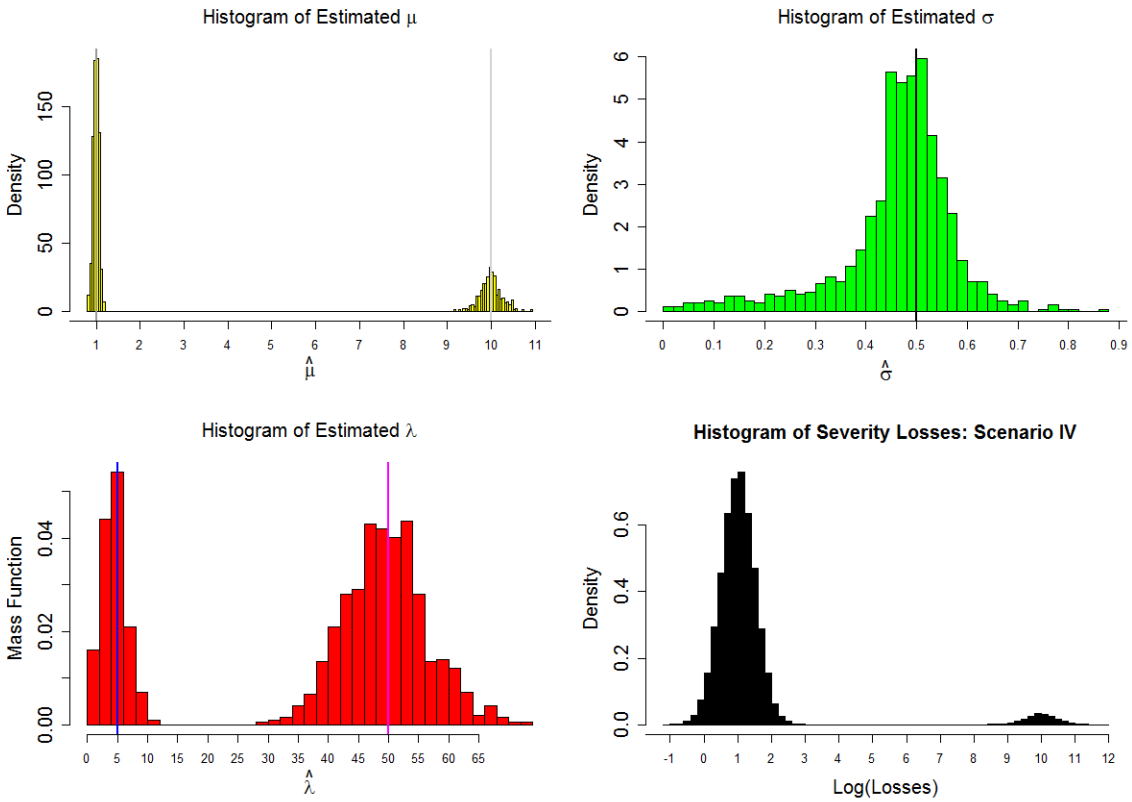


Figure 52 Severity/Frequency Parameter Estimates for Scenario (IV)

The Pearson correlation matrix among the frequency and severity parameters is shown in figure (53) below.

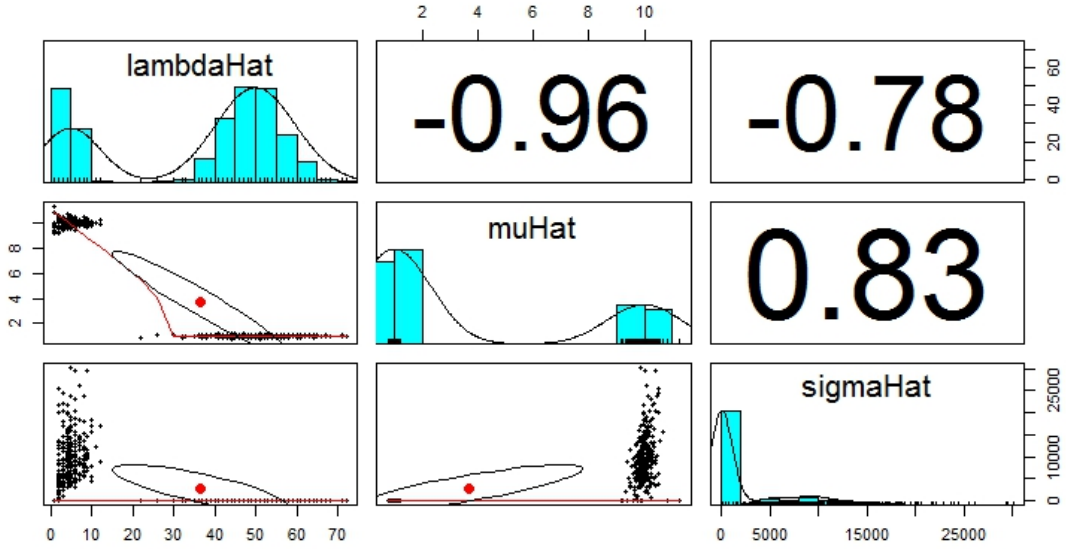


Figure 53 Severity/Frequency Pearson Correlation Estimates for Scenario (IV)

The mixture surface for Gaussian/GMCM copulas are computed and shown in figures (54)-(55) below. The *t* is identical to Normal in shape and not shown.

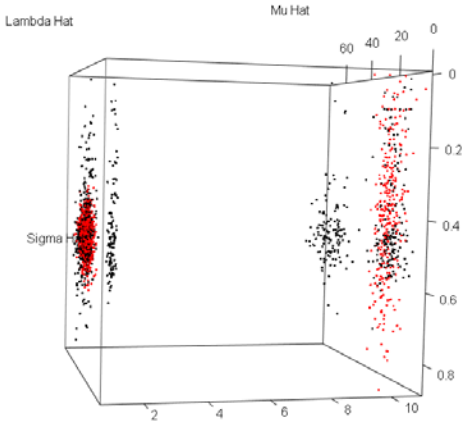


Figure 54 Surface Plot for Scenario (IV) using Gaussian Copula (red is data; black is copula)

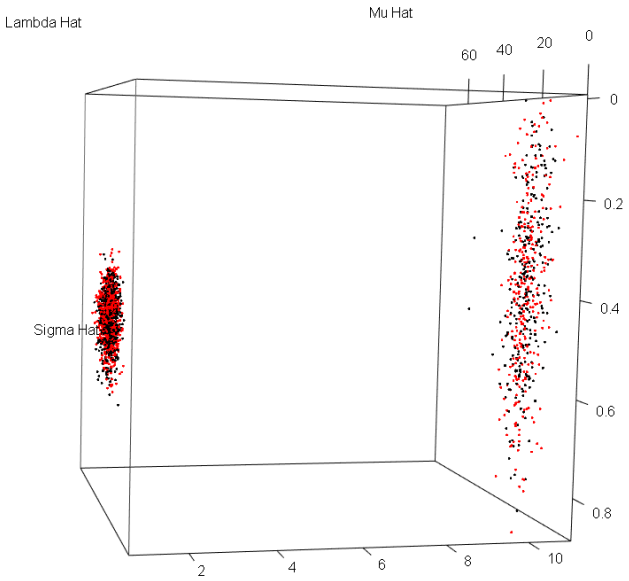


Figure 55 Surface Plot for Scenario (IV) using GMCM Copula (red is data; black is copula)

The final Scenario (V) is analyzed here. Figure (56) shows the frequency and severity parameters per month time unit.

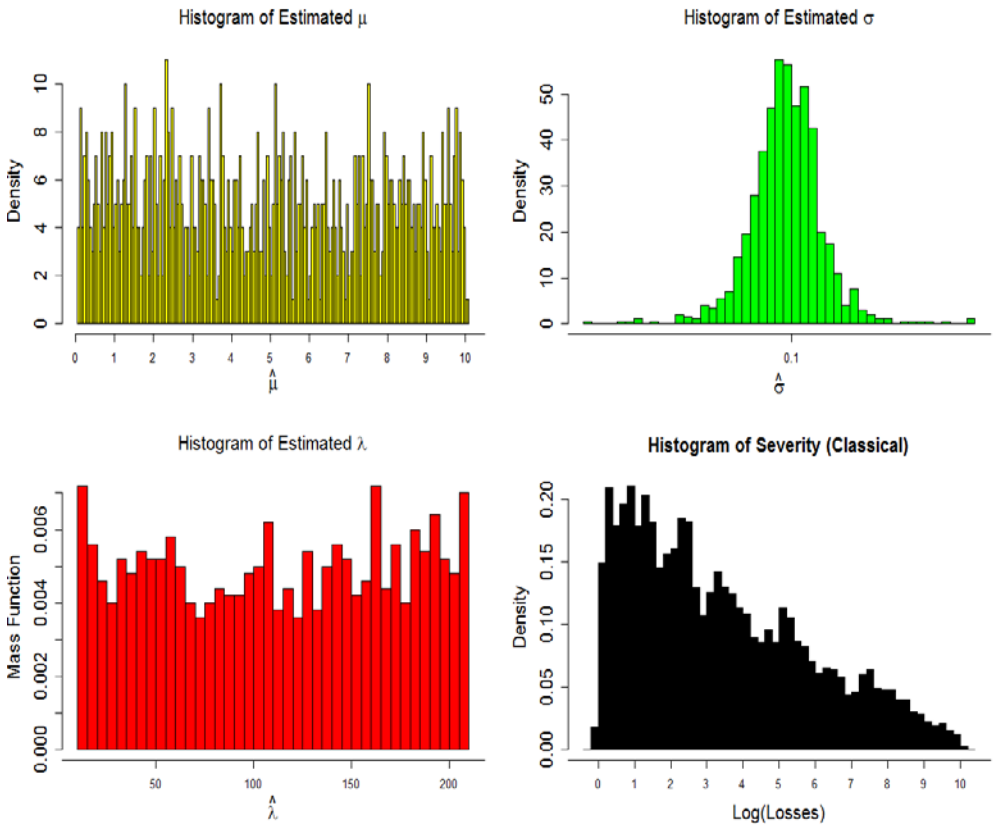


Figure 56 Severity/Frequency Parameter Estimates for Scenario (V)

The Pearson correlation matrix among the frequency and severity parameters is shown in figure (57) below.

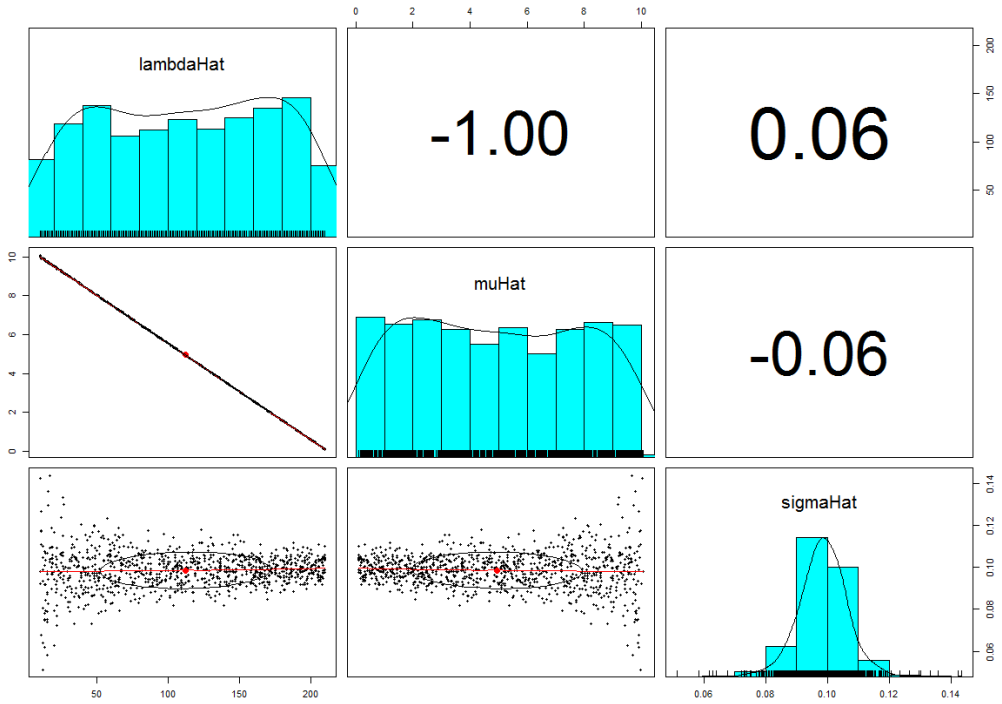


Figure 57 Severity/Frequency Pearson Correlation Estimates for Scenario (V)

Next, the mixture surface for Gaussian/*t* copulas are computed and shown in figures (58)-(59). Since the GMCM found K=1, the GMCM becomes the same as the Gaussian copula, and thus its surface plot is not shown in this case.

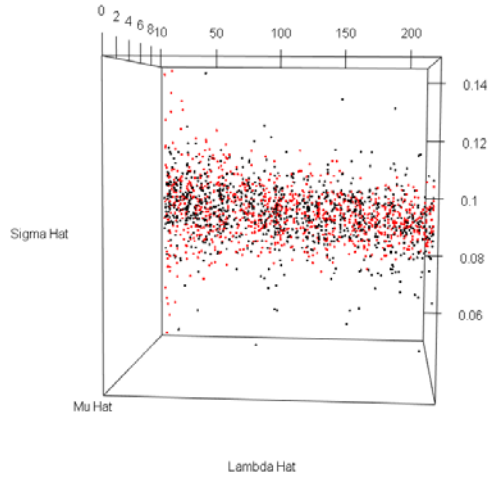


Figure 58 Surface Plot for Scenario (V) using Gaussian Copula (red is data; black is copula)

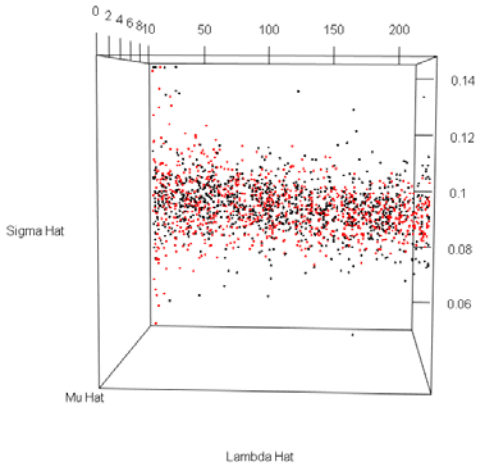


Figure 59 Surface Plot for Scenario (V) using *t*-Copula (red is data; black is copula)

Now that all of the copula calculations and fits are shown, the VaR estimates from this methodology are shown. This is given in tables (22)-(26). Afterwards, the analysis of real-world datasets is presented in the next section.

Table 22 VaR results for Scenario I using Parametric Copula methodology

Scenario I Results					
			Copula		
Historical VaR		Classical Method VaR	Normal VaR	Student T VaR	Gaussian Mixture Copula VaR
0%	$0	$0	$0	$0	$0
25%	$13	$118	$9	$10	$13
50%	$55	$262	$49	$51	$57
90%	$16,659	$1,253	$18,468	$16,112	$19,901
95%	$33,966	$2,067	$58,118	$55,966	$49,573
98%	$52,560	$3,805	$133,194	$150,503	$96,300
99%	$65,165	$5,920	$209,223	$271,317	$136,675
99.5%	$77,042	$9,028	$325,273	$387,643	$167,253

Table 23 VaR results for Scenario II using Parametric Copula methodology

Scenario II Results					
			Copula		
Historical VaR		Classical Method VaR	Normal VaR	Student T VaR	Gaussian Mixture Copula VaR
0%	$161	$183	$69	$55	$60
25%	$10,514	$1,872	$1,168	$1,165	$1,156
50%	$22,815	$3,080	$2,544	$2,464	$2,487
90%	$57,130	$9,522	$15,307	$14,527	$15,070
95%	$69,590	$14,302	$28,186	$26,001	$28,202
98%	$85,204	$25,504	$65,391	$57,718	$63,784
99%	$96,556	$38,792	$112,870	$100,300	$104,598
99.5%	$107,539	$61,201	$202,611	$171,972	$198,425

Table 24 VaR results for Scenario III using Parametric Copula methodology

Scenario III Results					
			Copula		
Historical VaR		Classical Method VaR	Normal VaR	Student T VaR	Gaussian Mixture Copula VaR
0%	$125	$0	$0	$0	$0
25%	$5,143	$5,127	$3,553	$3,490	$3,553
50%	$9,268	$9,234	$7,584	$7,488	$7,584
90%	$30,377	$30,280	$38,859	$37,090	$38,859
95%	$44,748	$44,682	$64,416	$63,991	$64,416
98%	$72,481	$72,090	$128,535	$128,767	$128,535
99%	$103,036	$102,021	$232,959	$220,152	$232,959
99.5%	$145,555	$143,970	$369,283	$407,959	$369,283

Table 25 VaR results for Scenario IV using Parametric Copula methodology

Scenario IV Results					
			Copula		
Historical VaR		Classical Method VaR	Normal VaR	Student T VaR	Gaussian Mixture Copula VaR
0%	$0	$78	$0	$103	$87
25%	$144	$421	$128	$2,063	$147
50%	$167	$591	$163	$6,088	$165
90%	$146,025	$1,223	$796,317	$53,156	$74,433
95%	$185,097	$1,571	$118,263	$107,196	$156,484
98%	$227,096	$2,196	$1,377,574	$374,233	$228,040
99%	$255,899	$2,749	$1,550,531	$562,645	$257,474
99.5%	$282,384	$3,378	$1,675,394	$1,177,644	$282,252

Table 26 VaR results for Scenario V using Parametric Copula methodology

Scenario V Results					
			Copula		
Historical VaR		Classical Method VaR	Normal VaR	Student T VaR	Gaussian Mixture Copula VaR
0%	$230	$9,034	$245	$231	$245
25%	$2,090	$36,517	$879	$1,713	$879
50%	$17,247	$52,223	$6,315	$14,936	$6,315
90%	$238,502	$125,085	$222,966	$234,803	$222,966
95%	$260,502	$174,906	$261,706	$257,916	$261,706
98%	$268,995	$265,873	$276,293	$266,826	$276,293
99%	$272,649	$379,968	$281,689	$271,631	$281,689
99.5%	$275,452	$541,806	$282,797	$274,750	$282,797

4.2.2.2 Real-World Data: Financial, Government, Insurance and Hurricanes

The results for the financial loss datasets are shown next. Figure (60) shows the frequency and severity parameters per month time unit for S&P 500.

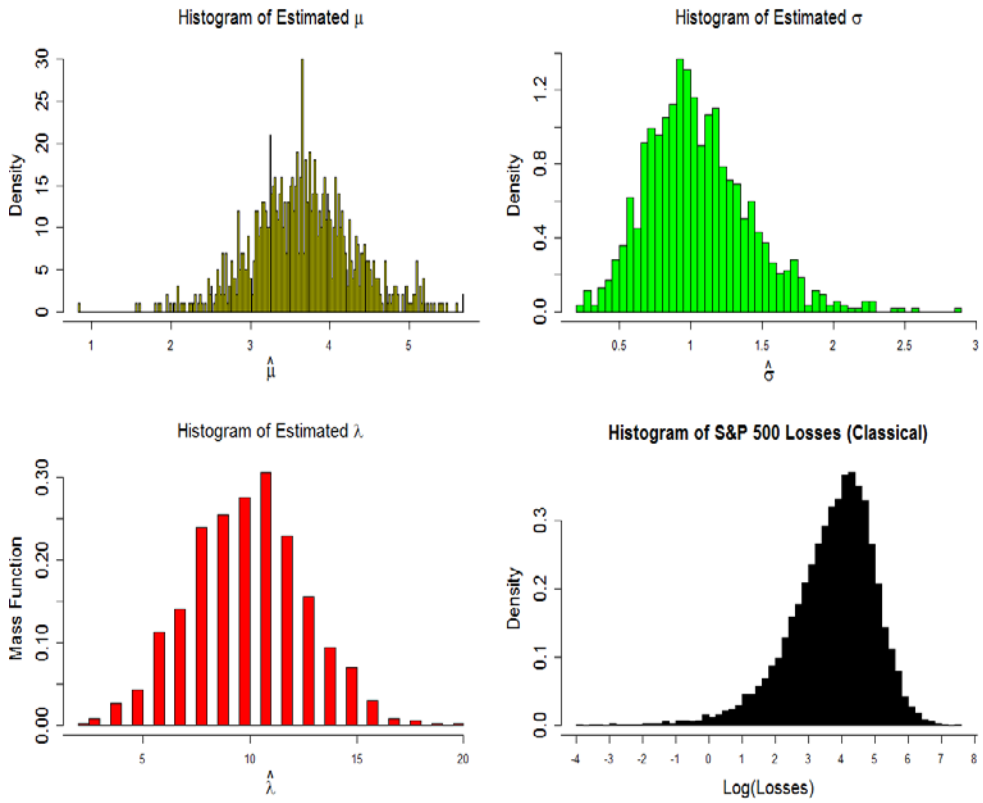


Figure 60 Severity/Frequency Parameter Estimates for S&P 500

The Pearson correlation matrix among the frequency and severity parameters is shown in figure (61) below.

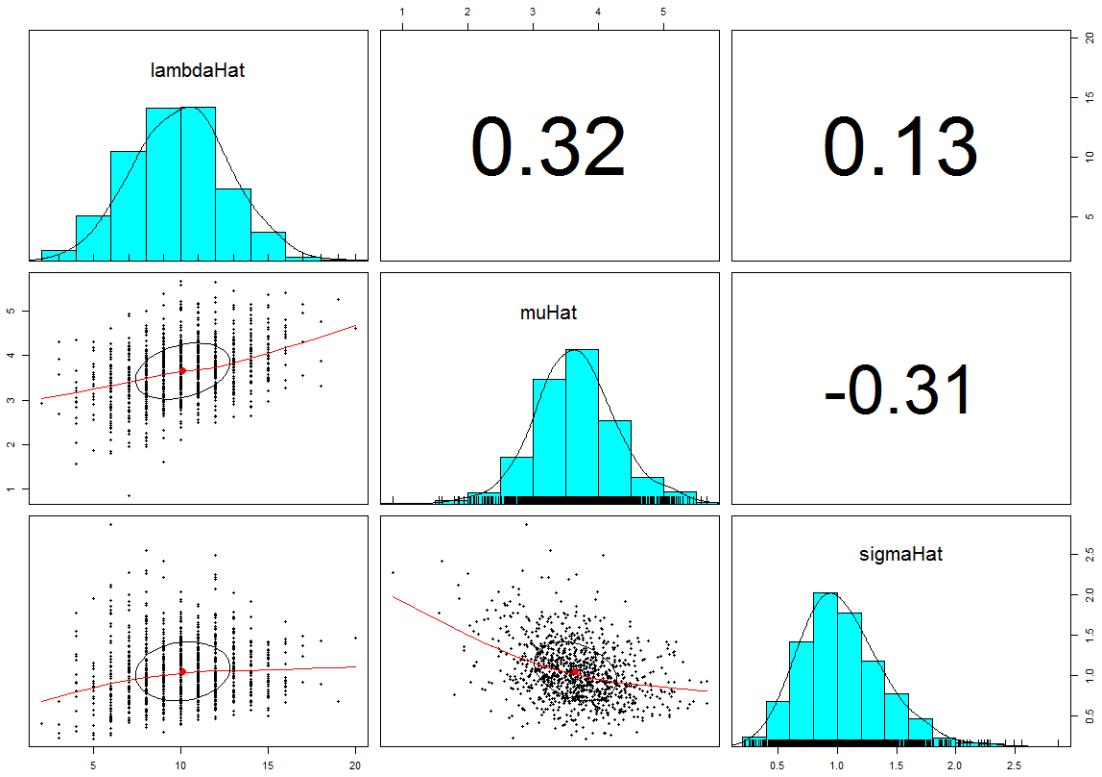


Figure 61 Severity/Frequency Pearson Correlation Estimates for S&P 500

Next, the mixture surface for Gaussian/*t* copulas are computed and shown in figures (62)-(63) below. The *t*-Copula is identical in surface plot and is thus not shown.

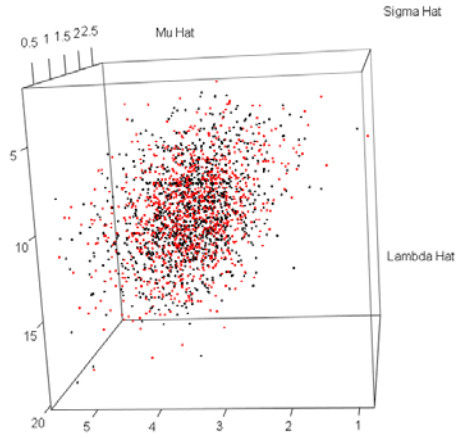


Figure 62 Surface Plot for S&P 500 using Gaussian Copula (red is data; black is copula)

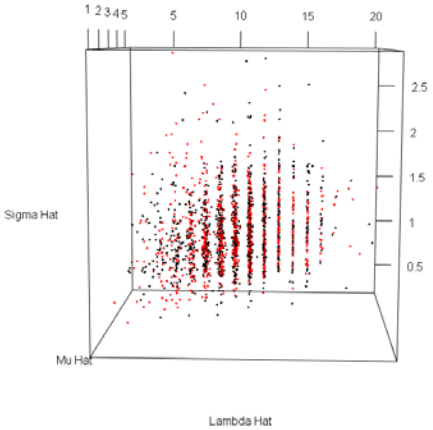


Figure 63 Surface Plot for S&P 500 using GMCM (red is data; black is copula)

The analysis of the DJIA dataset is shown next. Figure (64) shows the frequency and severity parameters per month time unit for this financial dataset.

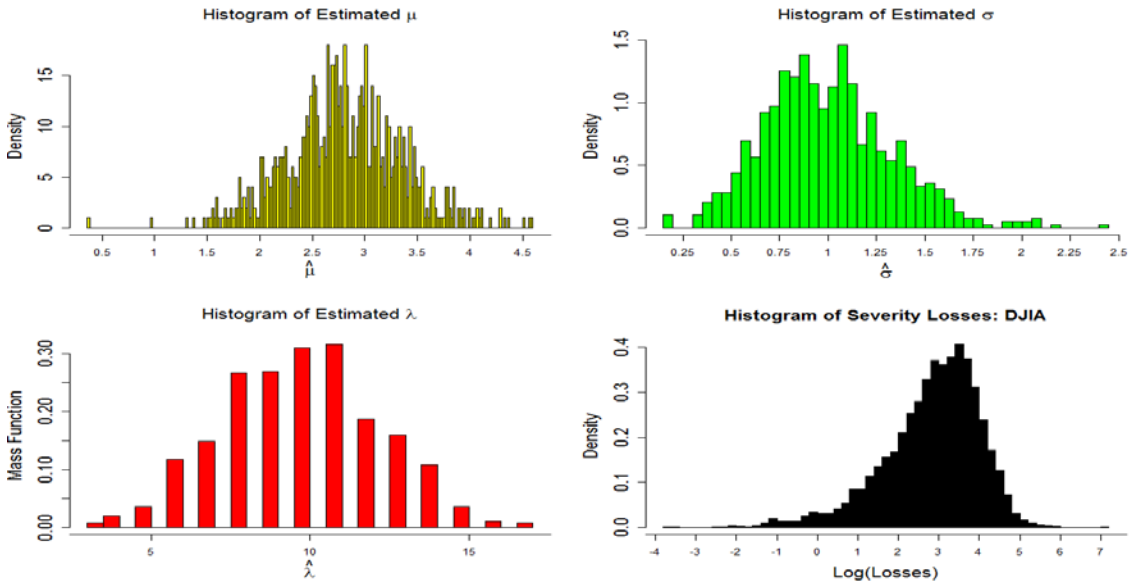


Figure 64 Severity/Frequency Parameter Estimates for DJIA

The resulting Pearson correlation matrix among the frequency and severity parameters is shown in figure (65) below.

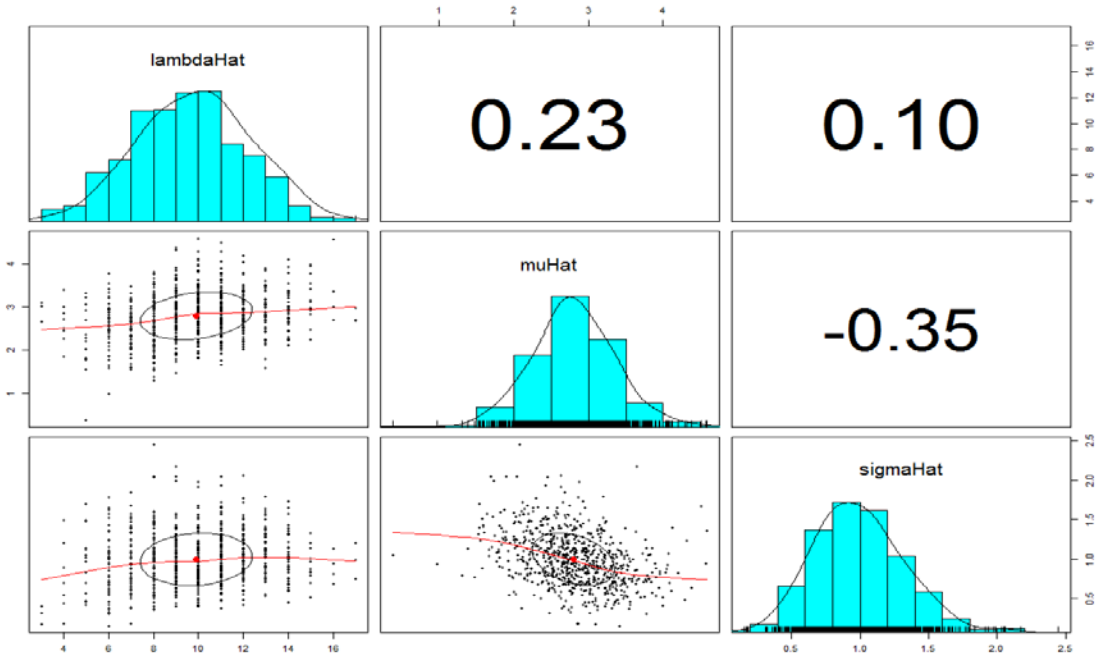


Figure 65 Severity/Frequency Pearson Correlation Estimates for DJIA

The surfaces for Gaussian/*t*/GMCM copulas are computed and shown in figures (66)-(68) below.

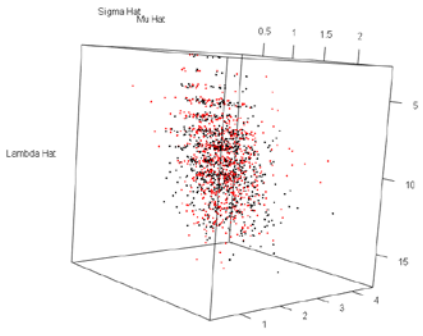


Figure 66 Surface Plot for DJIA using Gaussian Copula (red is data; black is copula)

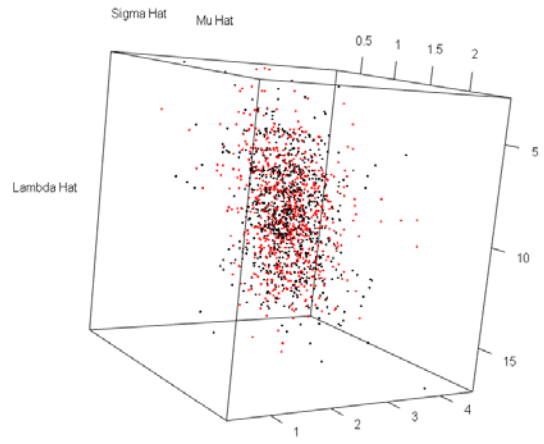


Figure 67 Surface Plot for DJIA using *t*-Copula (red is data; black is copula)

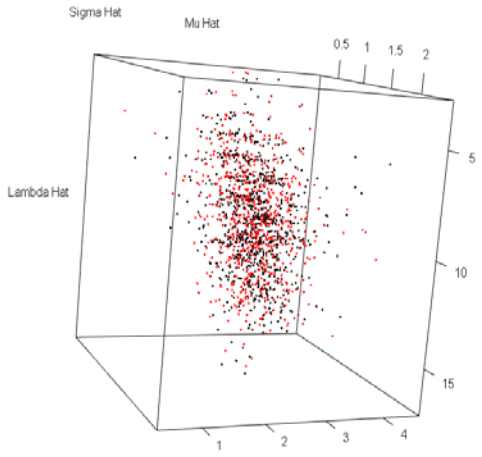


Figure 68 Surface Plot for DJIA using GMCM Copula (red is data; black is copula)

The chemical spills dataset analysis is shown next. Figure (69) shows the frequency and severity parameters per month time unit for this government monitored loss dataset.

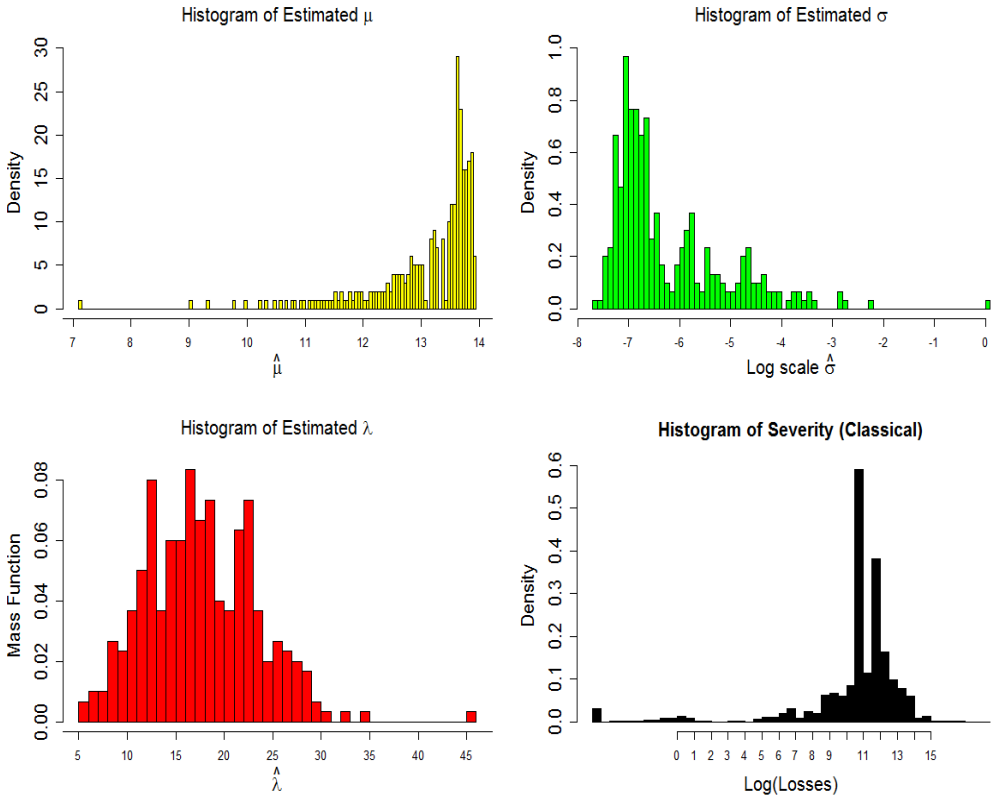


Figure 69 Severity/Frequency Parameter Estimates for Chemical Spills

The corresponding Pearson correlation matrix among the frequency and severity parameters is shown in figure (70) below.

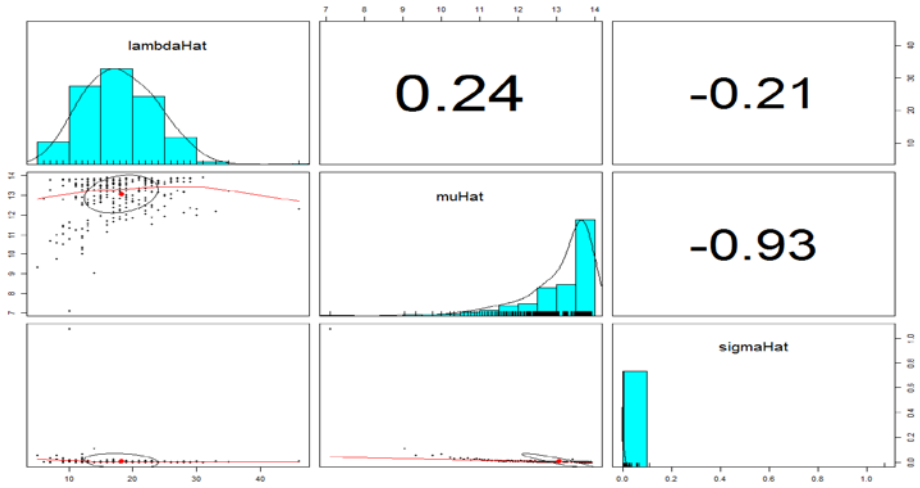


Figure 70 Severity/Frequency Pearson Correlation Estimates for Chemical Spills

The surfaces for Gaussian/*t*/GMCM copulas are computed and shown in figures (71)-(73) below.

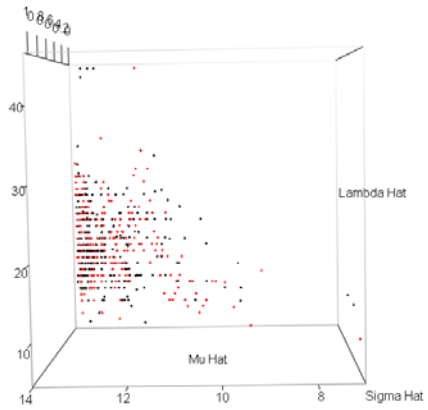


Figure 71 Surface Plot for Chemical Spills using Gaussian Copula (red is data; black is copula)

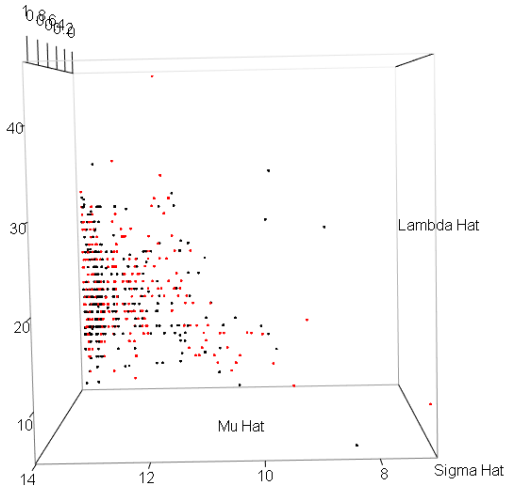


Figure 72 Surface Plot for Chemical Spills using *t*-Copula (red is data; black is copula)

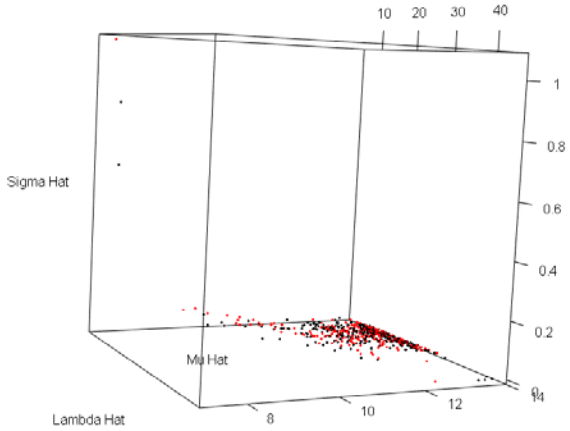


Figure 73 Surface Plot for Chemical Spills using GMCM Copula (red is data; black is copula)

Next, the automobile accident dataset is shown. Figure (74) shows the frequency and severity parameters per month time unit for this insurance dataset.

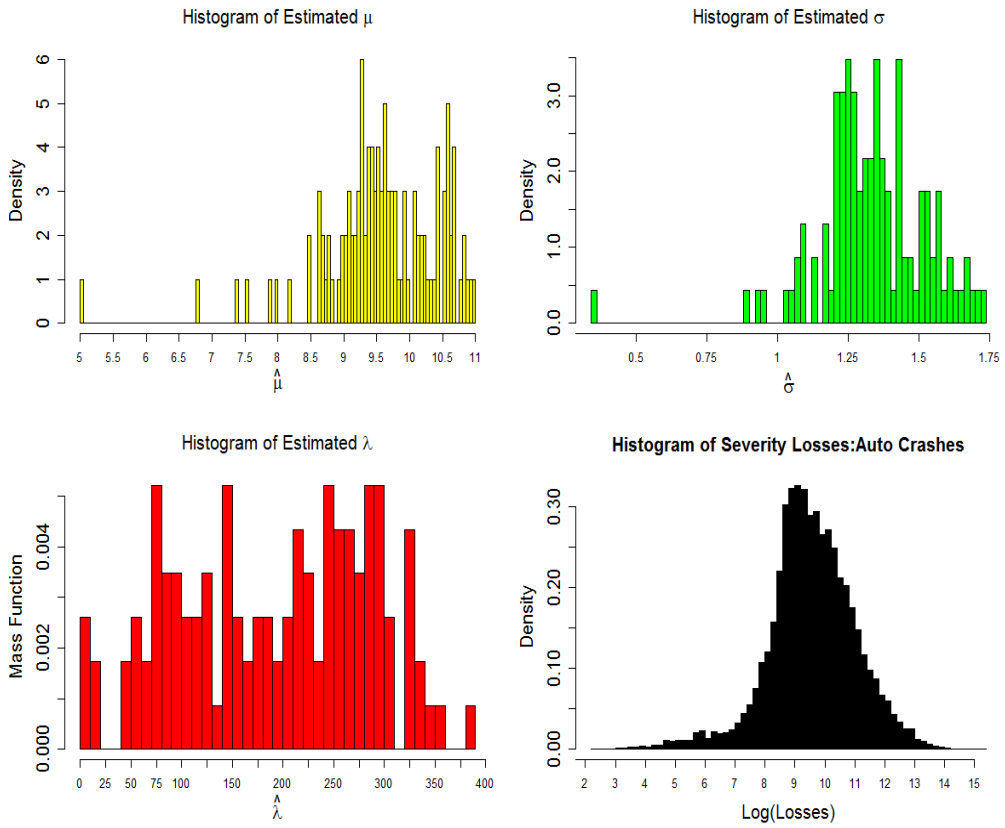


Figure 74 Severity/Frequency Parameter Estimates for Automobile Accident

The Pearson correlation matrix among the frequency and severity parameters is shown in figure (75) below.

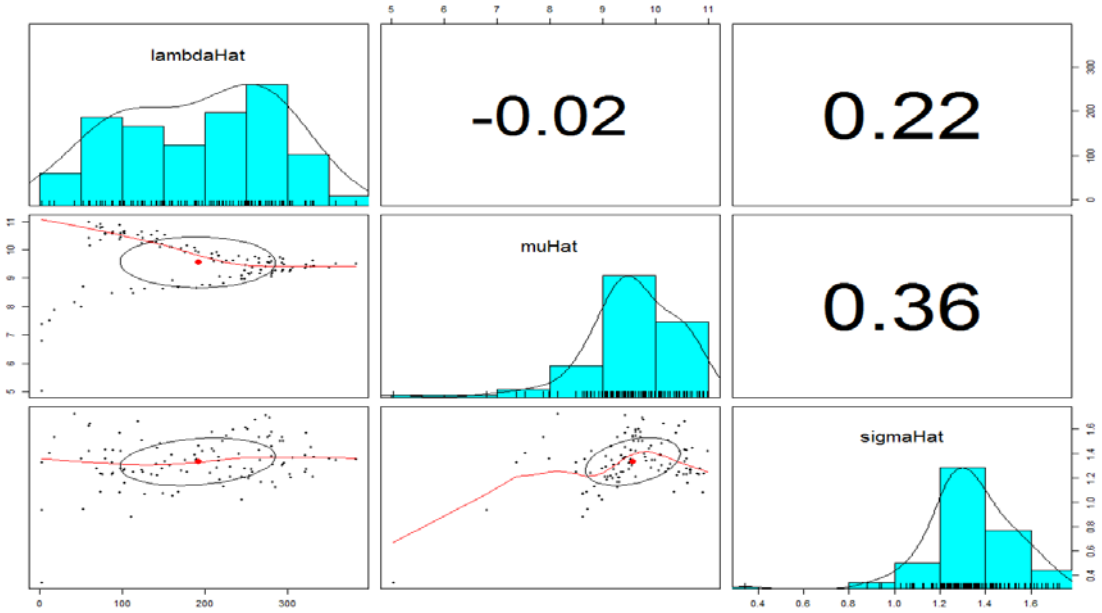


Figure 75 Severity/Frequency Pearson Correlation Estimates for Automobile crashes

The surfaces for Gaussian/*t*/GMCM copulas are computed and shown in figures (76)-(78) below.

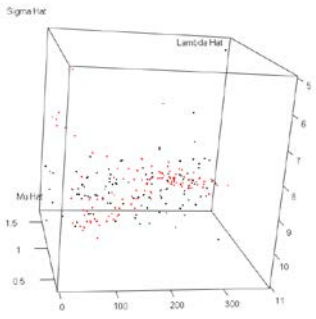


Figure 76 Surface Plot for Automobile using Gaussian Copula (red is data; black is copula)

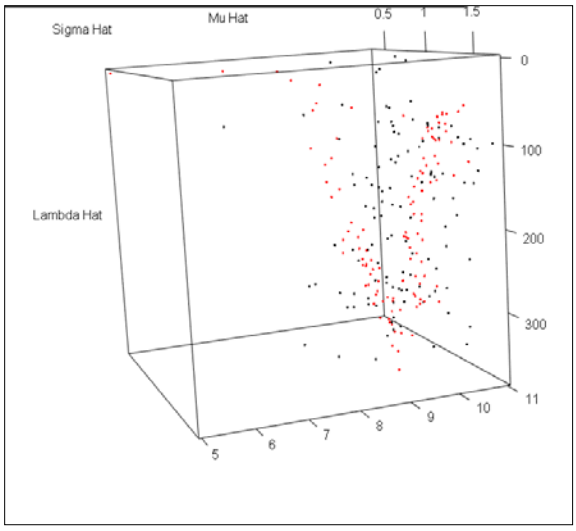


Figure 77 Surface Plot for Automobile Accidents using *t*-Copula (red is data; black is copula)

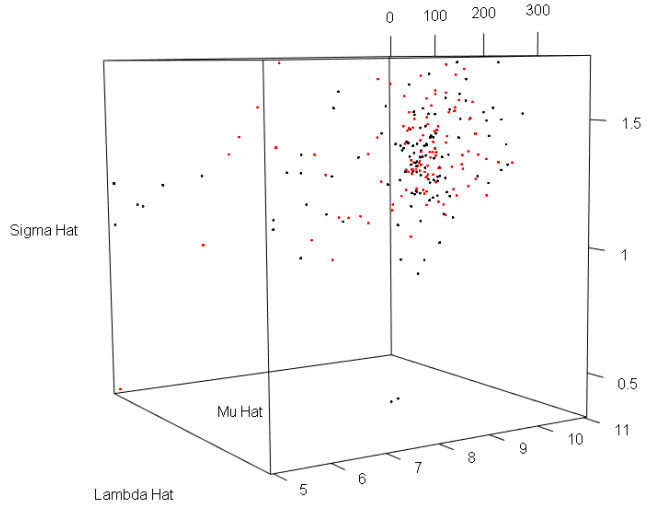


Figure 78 Surface Plot for Automobile Accidents using GMCM Copula (red is data; black is copula)

Finally the results from the US hurricane dataset are shown. In figure (79) below, the frequency and severity parameters are shown per month time unit for this natural disaster dataset.

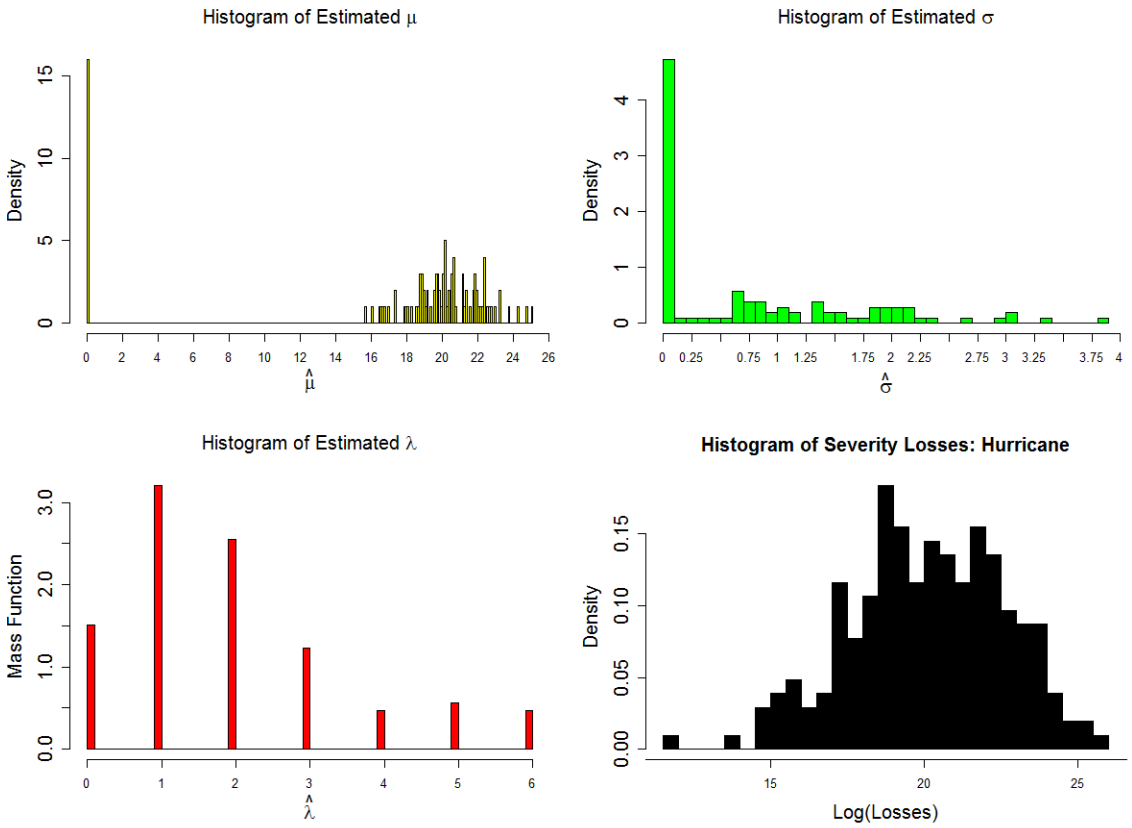


Figure 79 Severity/Frequency Parameter Estimates for US Hurricanes

The Pearson correlation matrix among the frequency and severity parameters is shown in figure (80) below.

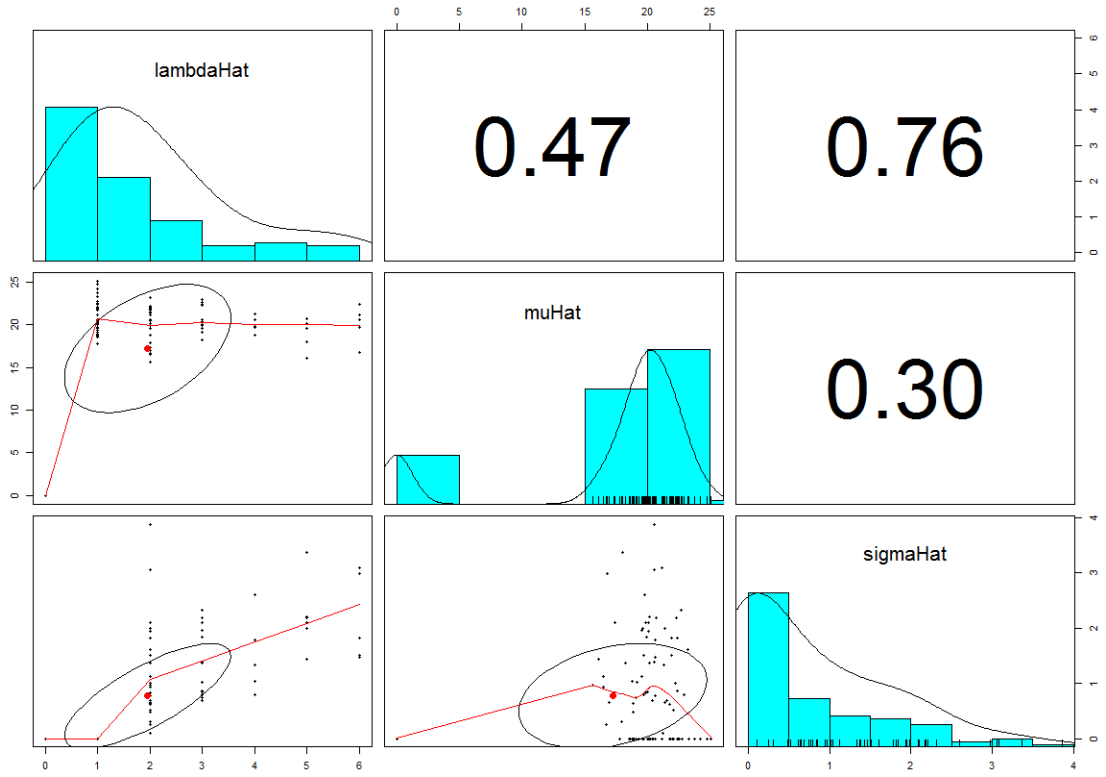


Figure 80 Severity/Frequency Pearson Correlation Estimates for US Hurricanes

The surfaces for Gaussian/*t*/GMCM copulas are computed and shown in figures (81)-(83) below.

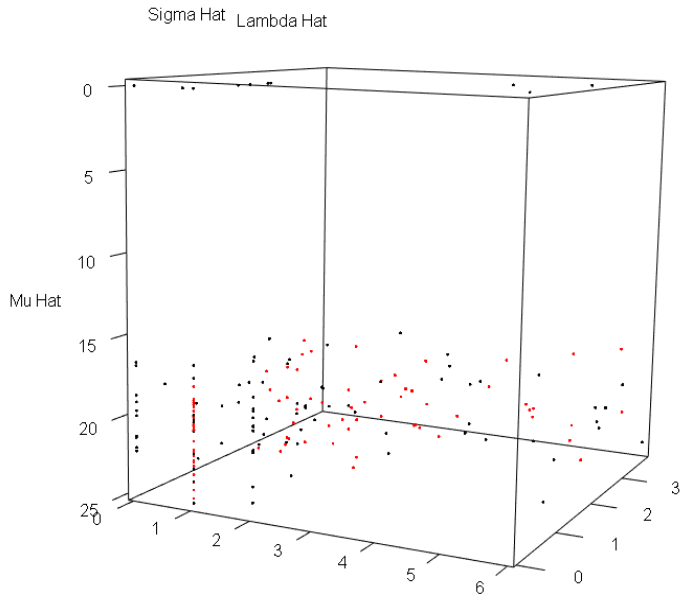


Figure 81 Surface Plot for US Hurricanes using Gaussian Copula (red is data; black is copula)

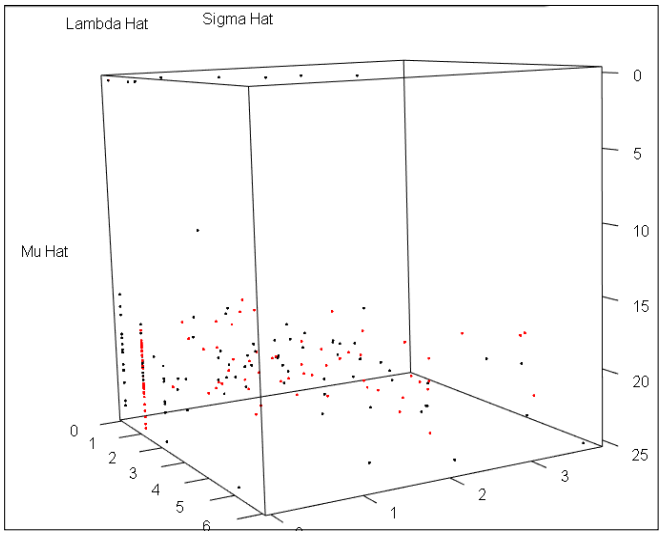


Figure 82 Surface Plot for US Hurricanes using *t*-Copula (red is data; black is copula)

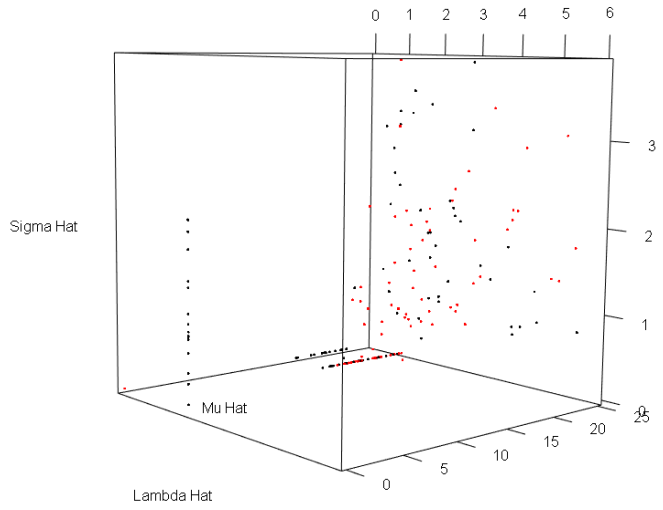


Figure 83 Surface Plot for US Hurricanes using GMCM Copula (red is data; black is copula)

Next, the copula based VaR results are shown for the real-world data. In this case, the comparison is shown relative to historical VaR estimates. This is shown next in tables (27) - (31). For the automobile crashes and hurricane data, the GMCM performs the best with respect to matching the historical data based VaR. The normal/*t* copula works best for the financial loss data and the chemical spills dataset.

Table 27 VaR Results for S&P 500 using Parametric Copula methodology

SP 500				
		Copula		
Historical		Normal	Student T	Gaussian Mixture Copula
VaR		VaR	VaR	VaR
0%	$40	$0	$0	$5
25%	$369	$311	$309	$322
50%	$586	$594	$588	$607
90%	$1,431	$2,049	$1,961	$1,943
95%	$2,027	$2,876	$2,956	$2,978
98%	$3,143	$4,382	$4,801	$4,644
99%	$3,808	$5,947	$7,160	$6,714
99.5%	$4,149	$8,043	$10,834	$9,721

Table 28 VaR Results for DJIA using Parametric Copula methodology

DJIA				
		Copula		
Historical		Normal	Student T	Gaussian Mixture Copula
VaR		VaR	VaR	VaR
0%	$29	$21	$3	$15
25%	$364	$246	$320	$307
50%	$566	$503	$590	$602
90%	$1,122	$1,538	$1,448	$1,651
95%	$1,465	$1,884	$2,238	$2,521
98%	$2,070	$2,447	$2,547	$3,534
99%	$2,325	$3,071	$3,320	$3,860

Table 29 VaR Results for Chemical Spills using Parametric Copula methodology

Chemical Spills				
		Copula		
Historical		Normal	Student T	Gaussian Mixture Copula
VaR		VaR	VaR	VaR
0%	$250,000	$3,767	$1,786	$48,229
25%	$1,579,058	$4,959,969	$4,983,645	$4,973,240
50%	$2,528,113	$10,832,908	$10,495,439	$9,833,337
90%	$6,070,004	$23,568,699	$23,275,790	$25,691,621
95%	$8,223,224	$27,670,975	$27,283,755	$29,239,610
98%	$14,428,712	$31,950,010	$31,506,144	$33,737,442
99%	$21,006,921	$34,874,776	$34,383,831	$37,529,789

Table 30 VaR Results for Automobile Crashes using Parametric Copula methodology

Automobile				
		Copula		
Historical		Normal	Student T	Gaussian Mixture Copula
VaR		VaR	VaR	VaR
0%	$327	$1,440	$0	$0
25%	$5,279,309	$2,921,533	$3,281,798	$5,101,715
50%	$8,179,051	$6,662,785	$6,066,204	$8,177,328
90%	$11,382,899	$23,932,125	$20,024,615	$14,390,955
95%	$12,841,319	$33,947,505	$25,378,522	$17,128,886
98%	$14,045,680	$43,604,744	$42,101,120	$21,260,755
99%	$14,717,942	$50,816,258	$50,234,097	$25,302,068

Table 31 VaR Results for US Hurricanes using Parametric Copula methodology

US Hurricanes				
		Copula		
Historical		Normal	t	Gaussian Mixture
VaR		VaR	VaR	VaR
0%	$0	$0	$0	$0
25%	$145,567,136	$0	$0	$0
50%	$1,546,550,238	$43,626,519	$83,492,715	$64,194,926
90%	$26,622,376,781	$24,771,773,034	$23,468,203,215	$14,479,751,412
95%	$51,396,924,874	$71,113,326,546	$96,668,651,534	$56,725,726,195
98%	$76,583,740,780	$277,118,598,958	$671,464,063,908	$99,273,860,617
99%	$113,860,560,854	$523,265,834,638	$1,885,740,861,357	$160,029,640,458

4.2.3 Summary of Results

Now, a comparison of the results from all of the methodologies to the classical methodology and historical VaR are discussed. The results are summarized in tables (32) - (41). In order to have statistical robustness, a bootstrap 99% confidence interval is computed for the historical data. Then, one can check whether the new and classical methodologies VaR estimates fall within the interval or not. To begin, the simulated data

through the five scenarios are discussed. Afterwards, the dissertation proceeds to discuss the real-world data. In order to provide a clearer understanding (rather than just looking at numbers), graphical representations of the VaR estimates are provided in figures (84) - (93) below after the tabular representations (in tables (32)-(41)).

Table 32 Comparison of New & Classical Methodology VaR for Scenario (I)

Scenario I									
					DPFS K-Means		DBP Copula		
Lower 99%	Historical (Ground Truth)		Upper 99%	Classical	Method I	Method II	Normal	*t*	Gaussian Mixture
	VaR			VaR	VaR	VaR	VaR	VaR	VaR
$0	0%	$0	$0	$0	$0	$0	$0	$0	$0
$12	25%	$13	$14	$118	$5	$14	$9	$10	$13
$54	50%	$55	$57	$262	$23	$57	$49	$51	$57
$14,596	90%	$16,659	$18,727	$1,253	$21,844	$20,964	$18,468	$16,112	$19,901
$31,696	95%	$33,966	$36,241	$2,067	$39,285	$37,375	$58,118	$55,966	$49,573
$49,535	98%	$52,560	$55,700	$3,805	$56,327	$54,427	$133,194	$150,503	$96,300
$61,116	99%	$65,165	$69,317	$5,920	$70,248	$66,828	$209,223	$271,317	$136,675
$71,623	99.5%	$77,042	$82,852	$9,028	$85,470	$78,008	$325,273	$387,643	$167,253

Table 33 Comparison of New & Classical Methodology VaR for Scenario (II)

Scenario II									
					DPFS K-Means		DBP Copula		
Lower 99%	Historical (Ground Truth)		Upper 99%	Classical	Method I	Method II	Normal	*t*	Gaussian Mixture
	VaR			VaR	VaR	VaR	VaR	VaR	VaR
$43	0%	$161	$272	$183	$2,413	$152	$69	$55	$60
$10,049	25%	$10,514	$10,982	$1,872	$10,820	$10,722	$1,168	$1,165	$1,156
$22,202	50%	$22,815	$23,448	$3,080	$16,629	$23,159	$2,544	$2,464	$2,487
$55,867	90%	$57,130	$58,439	$9,522	$53,861	$56,196	$15,307	$14,527	$15,070
$67,844	95%	$69,590	$71,384	$14,302	$86,086	$69,155	$28,186	$26,001	$28,202
$82,609	98%	$85,204	$87,893	$25,504	$161,826	$82,845	$65,391	$57,718	$63,784
$92,905	99%	$96,556	$100,448	$38,792	$240,311	$94,040	$112,870	$100,300	$104,598
$102,635	99.5%	$107,539	$113,191	$61,201	$376,988	$105,457	$202,611	$171,972	$198,425

Table 34 Comparison of New & Classical Methodology VaR for Scenario (III)

Scenario III									
					DPFS K-Means		DBP Copula		
Lower 99%	Historical (Ground Truth)		Upper 99%	Classical	Method I	Method II	Normal	*t*	Gaussian Mixture
	VaR			VaR	VaR	VaR	VaR	VaR	VaR
$1	0%	$125	$295	$0	$4	$108	$0	$0	$0
$5,001	25%	$5,143	$5,288	$5,127	$5,190	$5,137	$3,553	$3,490	$3,553
$9,036	50%	$9,268	$9,509	$9,234	$9,315	$9,372	$7,584	$7,488	$7,584
$29,172	90%	$30,377	$31,627	$30,280	$30,408	$31,450	$38,859	$37,090	$38,859
$42,364	95%	$44,748	$47,276	$44,682	$44,704	$45,185	$64,416	$63,991	$64,416
$66,761	98%	$72,481	$78,875	$72,090	$72,299	$74,414	$128,535	$128,767	$128,535
$91,757	99%	$103,036	$116,180	$102,021	$103,087	$107,022	$232,959	$220,152	$232,959
$124,353	99.5%	$145,555	$171,729	$143,970	$145,573	$146,053	$369,283	$407,959	$369,283

Table 35 Comparison of New & Classical Methodology VaR for Scenario (IV)

Scenario IV									
					DPFS K-Means		DBP Copula		
Lower 99%	Historical (Ground Truth) VaR		Upper 99%	Classical	Method I	Method II	Normal	*t*	Gaussian Mixture
				VaR	VaR	VaR	VaR	VaR	VaR
$0	0%	$0	$0	$78	$40	$43	$0	$103	$87
$143	25%	$144	$145	$421	$640	$8,656	$128	$2,063	$147
$166	50%	$167	$168	$591	$1,157	$28,646	$163	$6,088	$165
$141,624	90%	$146,025	$150,458	$1,223	$4,299	$82,321	$796,317	$53,156	$74,433
$179,958	95%	$185,097	$190,261	$1,571	$6,638	$101,008	$118,263	$107,196	$156,484
$220,317	98%	$227,096	$234,175	$2,196	$11,856	$122,585	$1,377,574	$374,233	$228,040
$246,722	99%	$255,899	$265,264	$2,749	$17,324	$141,709	$1,550,531	$562,645	$257,474
$270,610	99.5%	$282,384	$295,100	$3,378	$24,684	$158,573	$1,675,394	$1,177,644	$282,252

Table 36 Comparison of New & Classical Methodology VaR for Scenario (V)

Scenario V									
					DPFS K-Means		DBP Copula		
Lower 99%	Historical (Ground Truth) VaR		Upper 99%	Classical	Method I	Method II	Normal	*t*	Gaussian Mixture
				VaR	VaR	VaR	VaR	VaR	VaR
$228	0%	$230	$231	$9,034	$73,831	$8,433	$245	$231	$245
$1,925	25%	$2,090	$2,292	$36,517	$248,248	$42,501	$879	$1,713	$879
$15,738	50%	$17,247	$18,862	$52,223	$304,640	$61,651	$6,315	$14,936	$6,315
$234,360	90%	$238,502	$242,341	$125,085	$462,573	$155,336	$222,966	$234,803	$222,966
$258,871	95%	$260,502	$261,966	$174,906	$527,759	$219,225	$261,706	$257,916	$261,706
$267,960	98%	$268,995	$269,994	$265,873	$617,283	$355,505	$276,293	$266,826	$276,293
$271,586	99%	$272,649	$273,688	$379,968	$687,479	$525,180	$281,689	$271,631	$281,689
$274,238	99.5%	$275,452	$276,676	$541,806	$803,500	$805,360	$282,797	$274,750	$282,797

Table 37 Comparison of New & Classical Methodology VaR: Chemical Spills

Chemical Spills									
				Old	DPFS K-Means		DBP Copula		
Lower 99%	Historical (Ground Truth) VaR		Upper 99%	Classical	Method I	Method II	Normal	*t*	Gaussian Mixture
				VaR	VaR	VaR	VaR	VaR	VaR
$250,000	0%	$250,000	$558,655	$0	$10,058	$0	$3,767	$1,786	$48,229
$1,252,850	25%	$1,579,058	$1,892,100	$0	$6,439,424	$0	$4,959,969	$4,983,645	$4,973,240
$2,235,025	50%	$2,528,113	$3,054,550	$0	$15,803,952	$0	$10,832,908	$10,495,439	$9,833,337
$4,932,000	90%	$6,070,004	$7,921,420	$23,174,443	$114,902,912	$5,414,227	$23,568,699	$23,275,790	$25,691,621
$6,556,489	95%	$8,223,224	$14,371,808	$55,905,371	$226,507,895	$9,211,239	$27,670,975	$27,283,755	$29,239,610
$8,397,040	98%	$14,428,712	$24,979,120	$146,956,339	$519,706,874	$23,887,867	$31,950,010	$31,506,144	$33,737,442
$11,594,676	99%	$21,006,921	$102,667,700	$284,977,729	$939,953,129	$51,923,390	$34,874,776	$34,383,831	$37,529,789

Table 38 Comparison of New & Classical Methodology VaR: S&P 500

SP 500									
				Old	DPFS K-Means		DBP Copula		
Lower 99%	Historical (Ground Truth) VaR		Upper 99%	Classical	Method I	Method II	Normal	*t*	Gaussian Mixture
				VaR	VaR	VaR	VaR	VaR	VaR
$40	0%	$40	$66	$0	$0	$0	$0	$0	$5
$344	25%	$369	$397	$508	$499	$1,160	$311	$309	$322
$545	50%	$586	$617	$723	$783	$1,554	$594	$588	$607
$1,314	90%	$1,431	$1,608	$1,234	$1,710	$2,432	$2,049	$1,961	$1,943
$1,799	95%	$2,027	$2,326	$1,404	$2,132	$2,712	$2,876	$2,956	$2,978
$2,484	98%	$3,143	$3,734	$1,611	$2,772	$3,043	$4,382	$4,801	$4,644
$3,222	99%	$3,808	$4,210	$1,756	$3,337	$3,272	$5,947	$7,160	$6,714
$3,667	99.5%	$4,149	$4,826	$1,895	$3,990	$3,488	$8,043	$10,834	$9,721

Table 39 Comparison of New & Classical Methodology VaR: DJIA

DJIA									
				Old	DPFS K-Means		DBP Copula		
Lower 99%	Historical (Ground Truth) VaR		Upper 99%	Classical VaR	Method I VaR	Method II VaR	Normal VaR	*t* VaR	Gaussian Mixture VaR
$29	0%	$29	$107	$6	$412	$7	$21	$3	$15
$332	25%	$364	$406	$448	$1,098	$441	$246	$320	$307
$526	50%	$566	$615	$683	$1,316	$678	$503	$590	$602
$1,040	90%	$1,122	$1,306	$1,375	$1,773	$1,390	$1,538	$1,448	$1,651
$1,251	95%	$1,465	$1,803	$1,666	$1,910	$1,702	$1,884	$2,238	$2,521
$1,571	98%	$2,070	$2,352	$2,075	$2,088	$2,150	$2,447	$2,547	$3,534
$1,938	99%	$2,325	$3,539	$2,454	$2,192	$2,516	$3,071	$3,320	$3,860

Table 40 Comparison of New & Classical Methodology VaR: Automobile Crashes

Automobile									
				Old	DPFS K-Means		DBP Copula		
Lower 99%	Historical (Ground Truth) VaR		Upper 99%	Classical VaR	Method I VaR	Method II VaR	Normal VaR	*t* VaR	Gaussian Mixture VaR
$327	0%	$327	$265,962	$3,652,871	$6,193,134	$3,506,035	$1,440	$0	$0
$2,618,809	25%	$5,279,309	$6,632,639	$6,764,007	$9,097,156	$6,780,871	$2,921,533	$3,281,798	$5,101,715
$7,076,471	50%	$8,179,051	$9,004,338	$7,693,910	$9,859,308	$7,678,260	$6,662,785	$6,066,204	$8,177,328
$10,147,760	90%	$11,382,899	$13,182,960	$9,935,066	$11,442,192	$9,950,250	$23,932,125	$20,024,615	$14,390,955
$11,001,852	95%	$12,841,319	$14,481,365	$10,802,019	$11,906,385	$10,820,026	$33,947,505	$25,378,522	$17,128,886
$12,587,402	98%	$14,045,680	$16,722,353	$12,037,425	$12,461,354	$11,991,467	$43,604,744	$42,101,120	$21,260,755
$12,983,397	99%	$14,717,942	$16,722,353	$12,988,392	$12,871,537	$12,900,951	$50,816,258	$50,234,097	$25,302,068

Table 41 Comparison of New & Classical Methodology VaR: US Hurricanes

US Hurricanes									
				Old	DPFS K-Means		DBP Copula		
Lower 99%	Historical (Ground Truth) VaR		Upper 99%	Classical VaR	Method I VaR	Method II VaR	Normal VaR	*t* VaR	Gaussian Mixture VaR
$0	0%	$0	$0	$0	$0	$0	$0	$0	$0
$32,147,184	25%	$145,567,136	$358,069,353	$143,064,352	$984,718	$185,902,032	$0	$0	$0
$577,494,764	50%	$1,546,550,238	$3,841,822,355	$1,636,020,964	$891,143,303	$1,889,066,337	$43,626,519	$83,492,715	$64,194,926
$14,468,733,186	90%	$26,622,376,781	$57,663,865,630	$32,421,416,186	$8,797,128,358,329	$27,635,698,982	$24,771,773,034	$23,468,203,215	$14,479,751,412
$22,102,340,936	95%	$51,396,924,874	$115,750,000,000	$73,341,304,140	$106,917,479,563,651	$53,991,567,167	$71,113,326,546	$96,668,651,534	$56,725,726,195
$42,263,864,537	98%	$76,583,740,780	$161,331,092,065	$186,681,225,463	$1,509,316,668,391,140	$123,786,358,452	$277,118,598,958	$671,464,063,908	$99,273,860,617
$54,632,606,422	99%	$113,860,560,854	$161,331,092,065	$378,170,152,946	$8,990,749,087,773,140	$230,825,029,288	$523,265,834,638	$1,885,740,861,357	$160,029,640,458

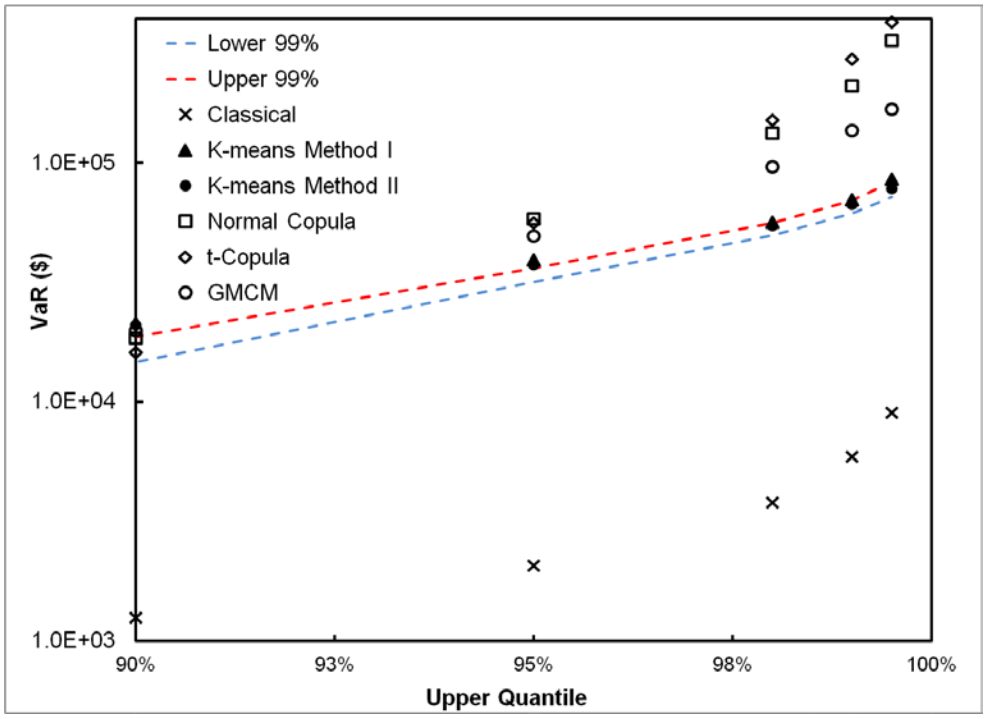


Figure 84 VaR Results for DPFS, DBP & Classical: Scenario (I)

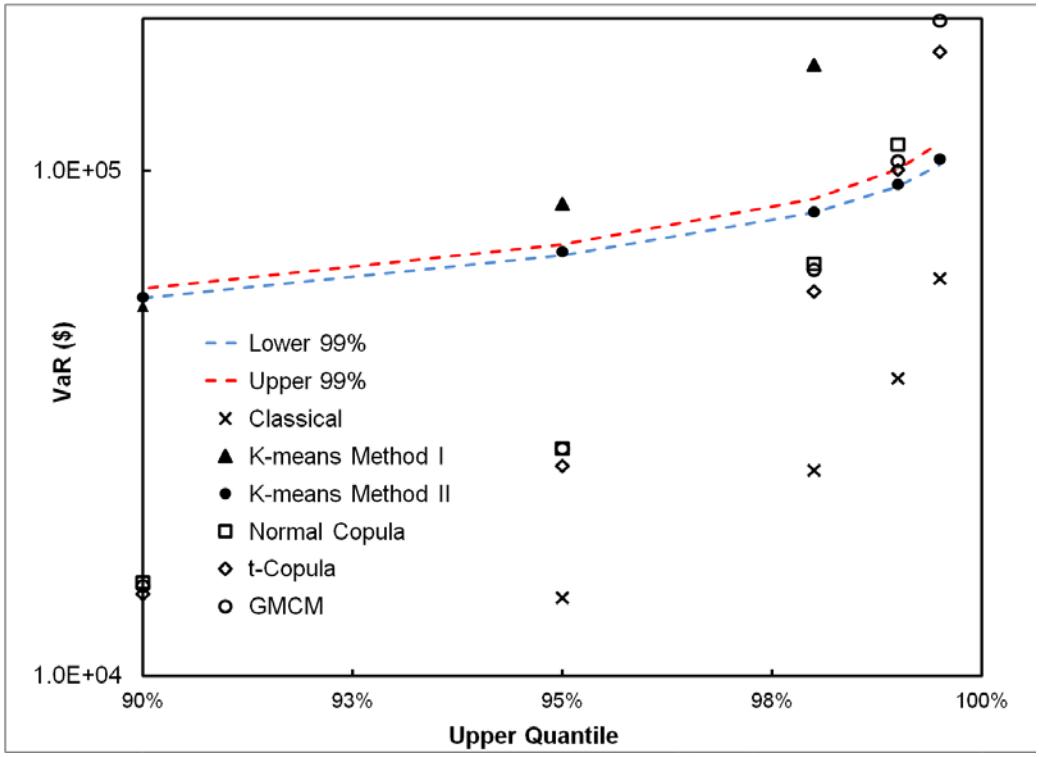


Figure 85 VaR Results for DPFS, DBP & Classical: Scenario (II)

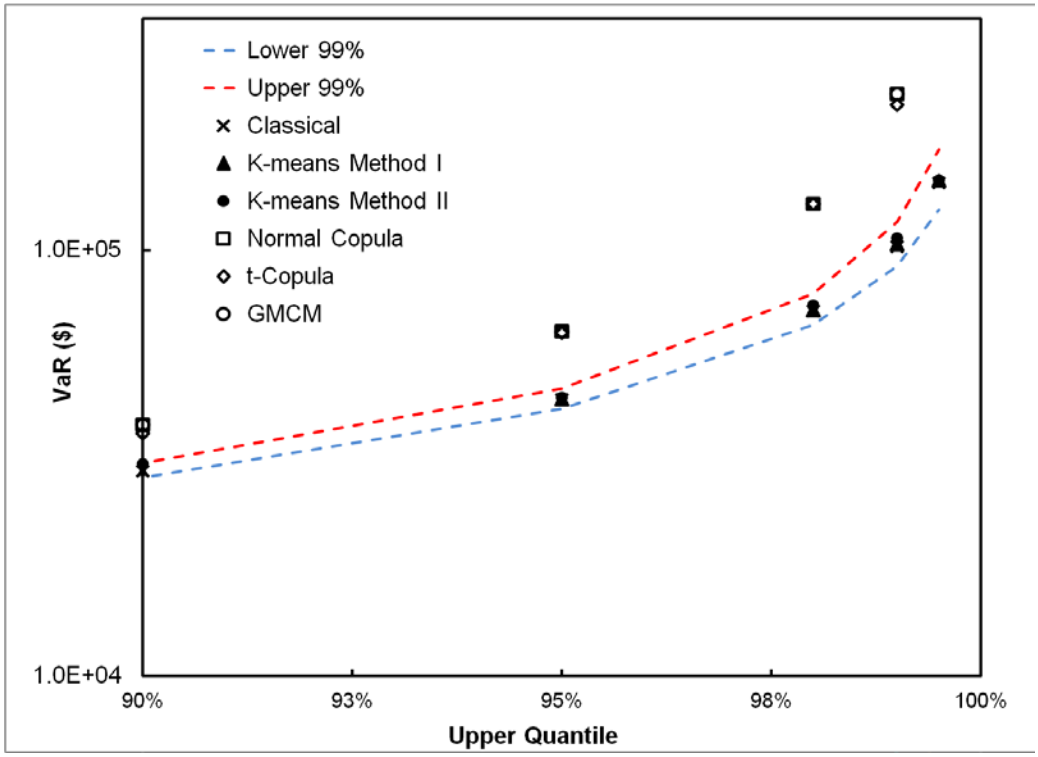


Figure 86 VaR Results for DPFS, DBP & Classical: Scenario (III)

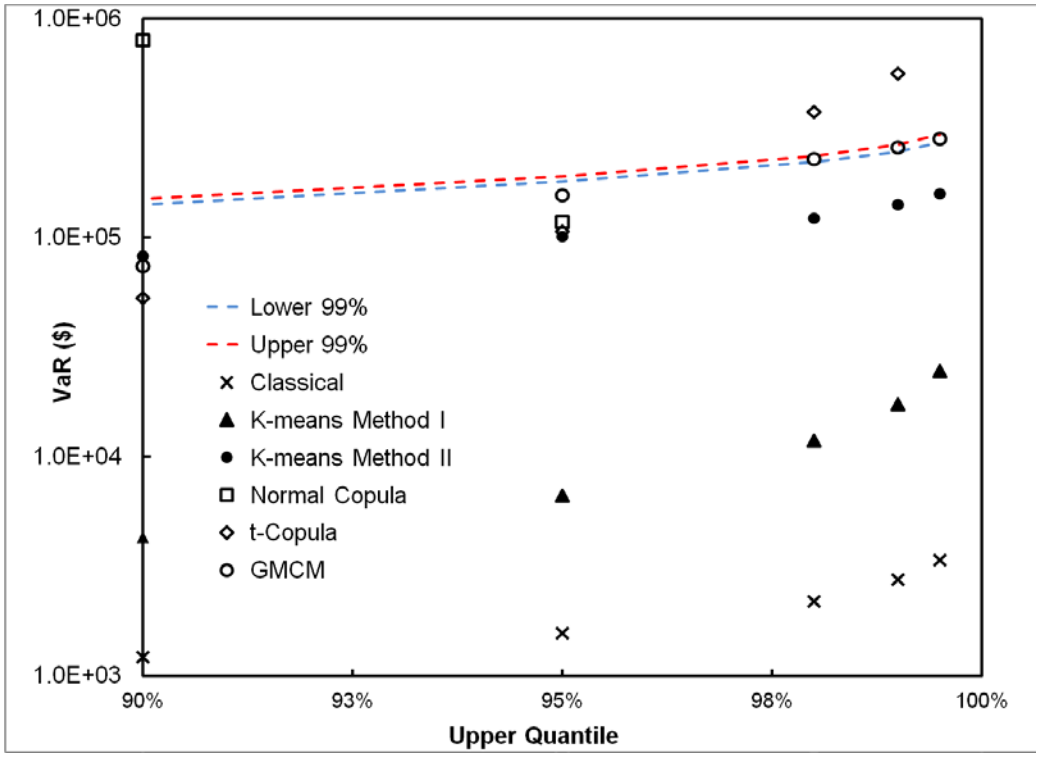


Figure 87 VaR Results for DPFS, DBP & Classical: Scenario (IV)

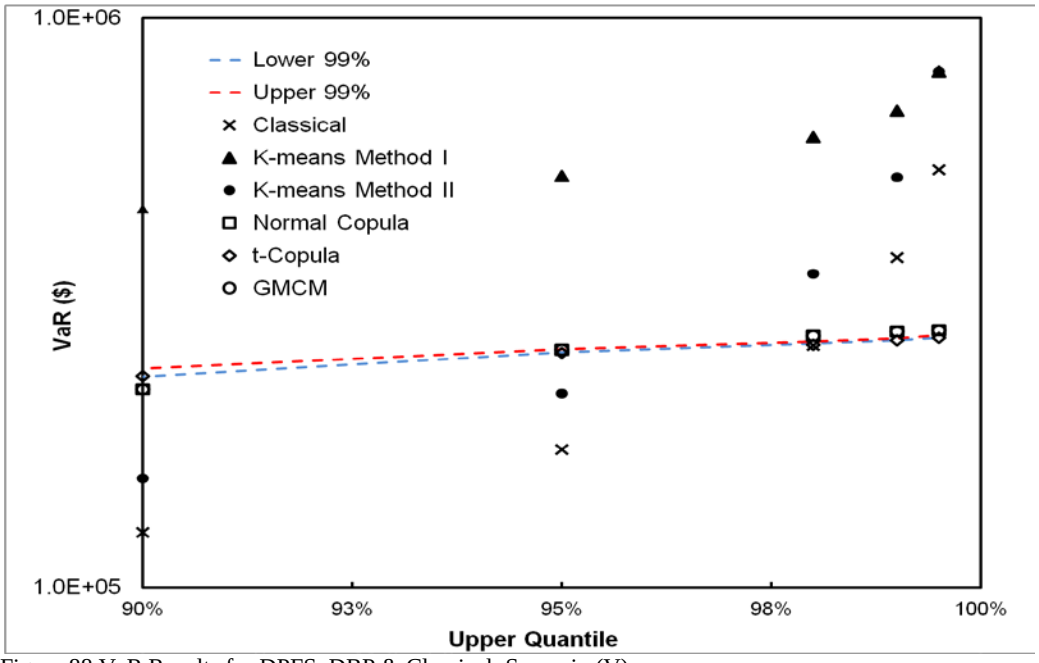


Figure 88 VaR Results for DPFS, DBP & Classical: Scenario (V)

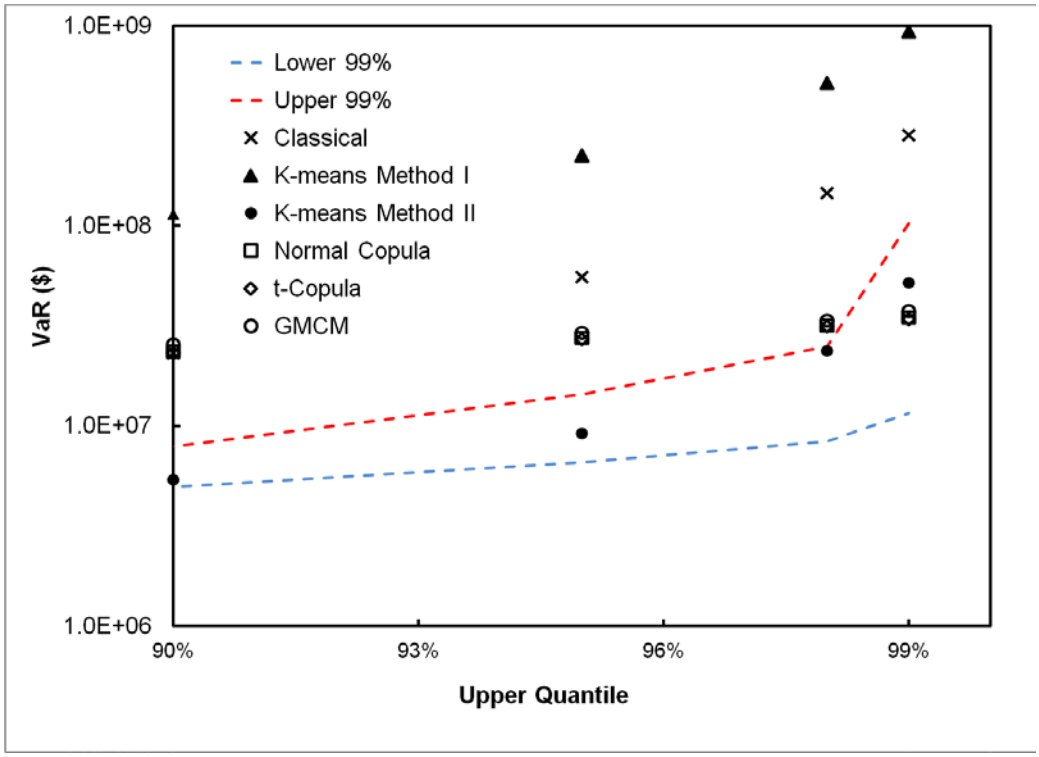


Figure 89 VaR Results for DPFS, DBP & Classical: Chemical Spills

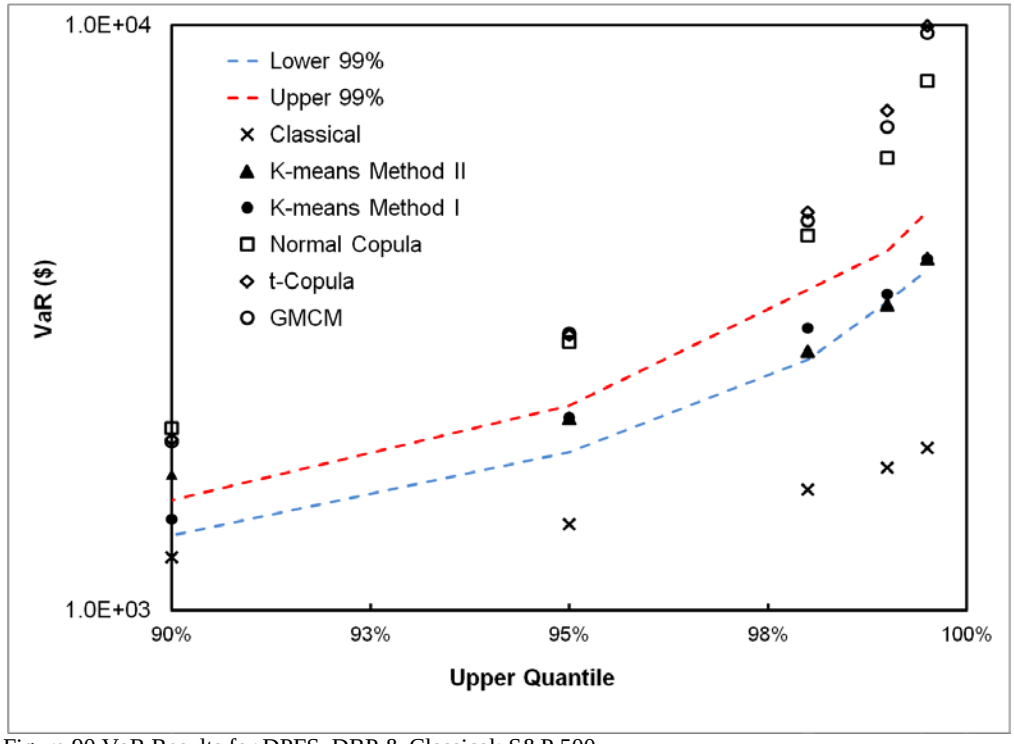


Figure 90 VaR Results for DPFS, DBP & Classical: S&P 500

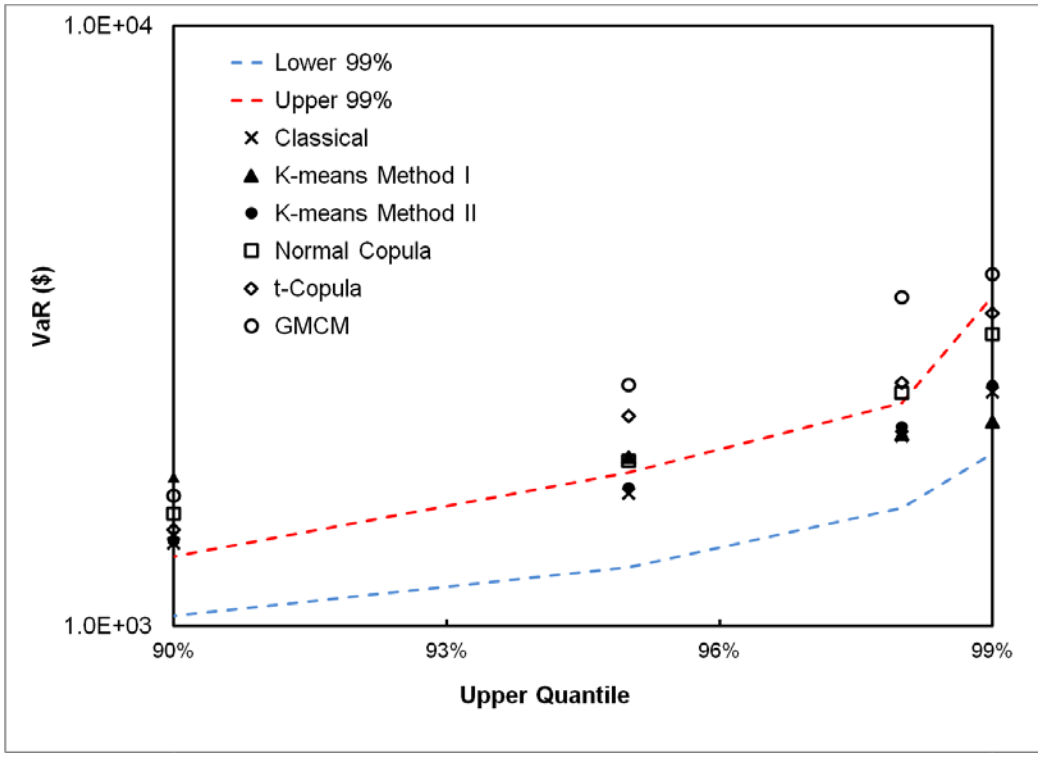


Figure 91 VaR Results for DPFS, DBP & Classical: DJIA

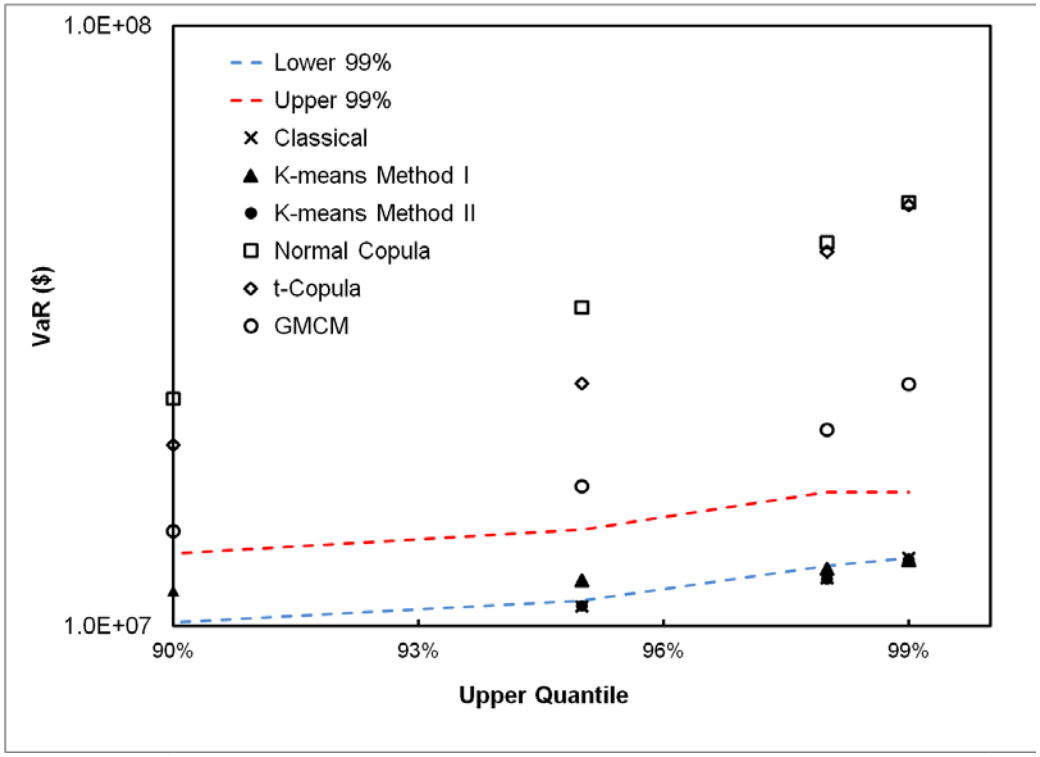


Figure 92 VaR Results for DPFS, DBP & Classical: Auto Accidents

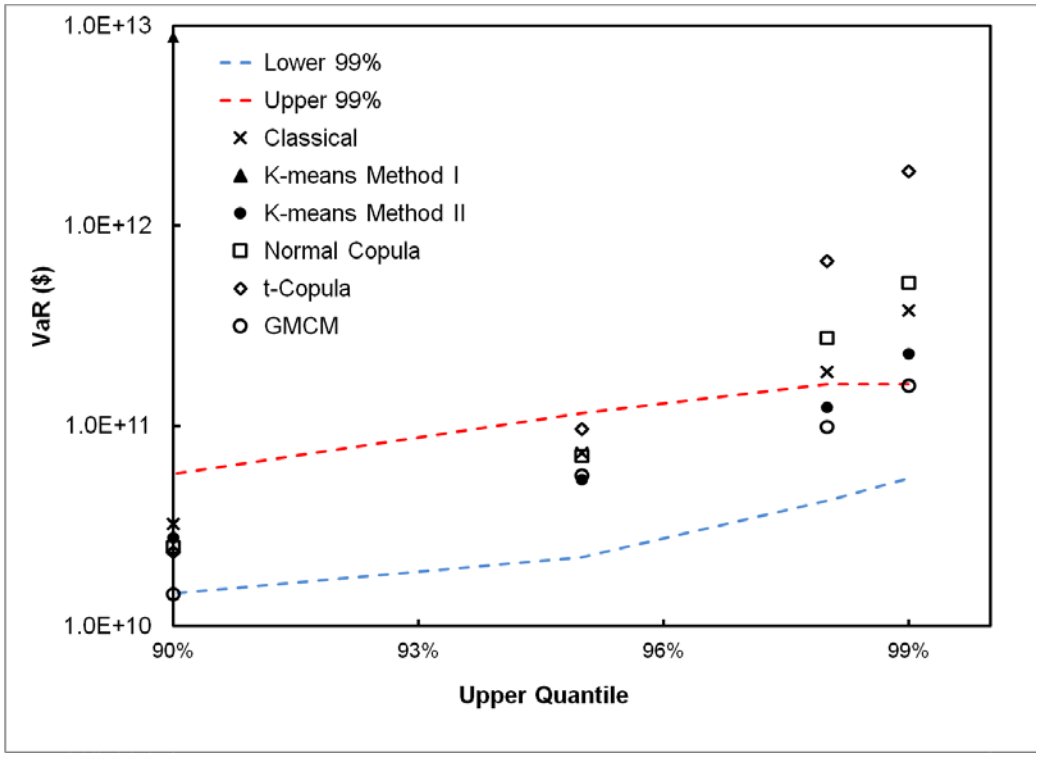


Figure 93 VaR Results for DPFS, DBP & Classical: US Hurricanes

4.2.4 Discussions and Implications

In this dissertation, comparison is done among the new methodologies with the classical methodology using five distinct scenarios of simulated data. The first two scenarios (i.e. Scenario (I) & (II)) correspond to partitioning the data based on severity and frequency. The DPFS K-means VaR computation performs well for this type of data as expected. The reason is that K-means approach is well adapted to separate the data when clear and a finite number of partitions exist. For these two scenarios, the copula based methodology over-estimates the VaR, while the classical methodology grossly underestimates the VaR. In the risk-management context, it is usually considered more unfavorably to underestimate the VaR for in that situation the company/institution could be severely in jeopardy of insolvency. On the other hand, if VaR is overestimated, the

company/institution loses opportunity cost of holding the extra EC for VaR which they could have invested for further gains. Thus, through the verification & validation procedure via large scale MCS, this dissertation has demonstrated that in cases like Scenarios (I) and (II), the DPFS K-means VaR estimation is more robust than the classical methodology.

Next, the results obtained from Scenario (III) are described. This case signifies the case of complete independence of severity and frequency. It is therefore expected that the classical methodology should work well for this type of data. The results indicate that while the classical methodology does perform well, the DPFS K-means VaR approach matches its performance. The K-means is robust enough to detect the case of independence in the data. The copula based methodology overestimates the VaR in this case. This is due to the fact that the copula is not a good fit and the empirical GoF test showed a non-statistically significant p-value.

Next, the results from Scenarios (IV) and (V) are presented. These two scenarios represent ideal cases for the DBP copula based parametric methodology, since there is a high degree of correlation between frequency and severity. Specifically, for Scenario (IV), the data is partitioned as a mixture, and thus should naturally follow a mixture model. In Scenario (V), the data is generated with a perfect *linear* correlation between frequency and severity, and thus the Gaussian/*t* copula should perform well. The results show that for Scenario (IV), the GMCM model calculates the most accurate VaR while the K-means methodology also works reasonably well. The classical methodology completely fails in this dataset. For Scenario (V), the *t*-Copula works the best while the

Gaussian copula is a close second. The GMCM is identical to the Normal copula since the mixture model found K=1 as the ideal mixture. It is interesting to note that the K-means algorithm and the classical methodology did not work well in this case. The reason is that a high number of splits are necessary for this type of data, since there is a perfect correlation between severity and frequency. The data mining literature argues that if splitting with a very high number, say K = 200, one may overfit the data. Thus, in this work there is a restriction of the number of possible splits to a small bound. Also, it is important to note that the classical methodology fails in this scenario (order of magnitude).

From the verification procedure, test instances are shown where each of the new methodologies work well. It is important to now validate the new methodologies from results using the real-world data. Since the truth is never known for these situations, the benchmark that the dissertation uses is the 99% bootstrap confidence bound for VaR obtained from historical data. To begin, the results from table (37) are analyzed which shows the chemical spills scenario. The DPFS K-means VaR (based on Method II) performs the best, while the *t*-Copula does reasonably well too. It is important to note that the classical methodology overestimates the VaR. This can be explained due to the fact that this is one of the cases similar to Case (II) from the theoretical calculations in the methodology section. The data is multi-modal and the classical overestimates primarily due to the fact that the true VaR involves a compound Poisson process where the frequency dominates.

Next, the results for the S&P 500 are examined. Note this data includes the Great Depression of 1929, along with the expansions of the 1950s, and the crashes in the last several decades. In this case, it is observed that the classical methodology underestimates the VaR. The reason for this is that there is a positive correlation between frequency and severity. The classical method underestimates the VaR because the severity portion dominates the frequency for this data. The DBP Gaussian copula VaR performs the next best after the K-means methodology, but overestimates the VaR.

Afterwards, the results for the DJIA are examined. This data starts from 1950 and goes till 2015. So it differs from the S&P 500 in that it does not include the Great Depression of the 1920's and it slow recovery from the 1930s. In this case, it is observed that the classical methodology performs well to estimate the VaR. The K-means robustly adjust for this type of data, and performs as well as the classical methodology. The normal copula also does well for this data for the tail regions of the VaR.

Next, the results for the automobile crashes data are examined. This data again starts from 1989 and goes till 1999. Thus as a small dataset, there are approximately only 120 months of data for the monthly VaR estimate. For this specific size dataset, the classical methodology performs well to estimate the VaR along with the K-means methods. Both of these methods do underestimate the VaR slightly, while the GMCM overestimates the VaR.

Finally, the results from the US Hurricanes are examined. This data is available annually from 1900 - 2005, and thus an annual VaR is estimated. This is one of the smallest datasets analyzed with approximately 200 total data points. The GMCM does the

best for this small dataset along with the K-means methodology. The classical methodology performs very poorly for this data as there is a large discrepancy between frequency and severity. A summary of the results are shown below in table (42).

Table 42 VaR Results Summary for all Data; X = Optimal Method Δ = Method Tied

		VaR Estimation		
Data Type		DPFS K-means	DBP Copula	Classical
Simulated Scenarios Data	Scenario I: Hi/Med/Low Severity → Hi/Med/Low Frequency (One-to-Many)	X		
	Scenario II: Hi/Med/Low Severity → Hi/Med/Low Frequency (1-to-1)	X		
	Scenario III: Frequency Independent Severity	Δ		Δ
	Scenario IV: Hi/Low Severity Mixture → Hi/Low Frequency		X	
	Scenario V: Perfect Correlation between Frequency & Severity		X	
Real-World Data	S&P 500	X		
	DJIA	Δ		Δ
	Chemical Spills		X	
	Automobile Accidents	Δ		Δ
	US Hurricanes		X	

Overall, in the real-world data it is observed that for two cases, the classical methodology does reasonably well; (1) DJIA; and (2) Automobile accidents. However, in both of these, they still underestimate the VaR slightly. The DPFS K-means VaR methodology performs well in all of the real-data sets, while the DBP GMCM and Gaussian copula VaR perform well in all but the S&P 500 and Chemical spills data. It is

interesting to note that the copula based methods *never* underestimate the VaR. As previously mentioned, underestimating the VaR is usually considered worse than overestimation from a risk management perspective. Now the next question that arises next is the following: Which method should one use in what situation (parametric versus non-parametric)? Overall, I elaborate this process through a flow-chart shown in figure (94).

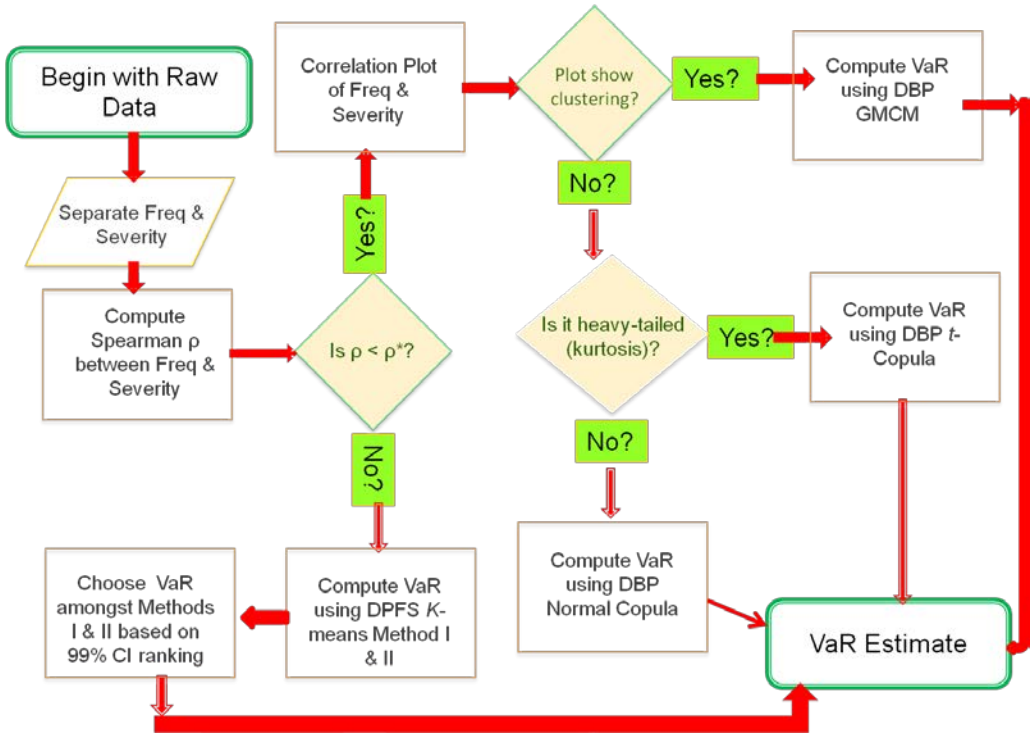


Figure 94 Flow-Chart for Choosing Amongst VaR Estimation Methodologies

CHAPTER 5: CONCLUSIONS

This dissertation addresses several methodological and empirical foundations within quantitative risk management, specifically within the context of modern ORM. Two fundamental questions are addressed in the methodological portion:

(1) Can a flexible severity distribution be found to model loss severity?

(2) For calculating VaR, can more robust methodologies be developed which can handle both situations of dependence and independence between frequency and severity?

This dissertation investigates these questions and provides an important contribution to the field by identifying flexible severity distributions which can robustly model Aggregate Loss distributions without requiring the current trial-and-error GoF tests approach. Secondly, this dissertation also establishes a more robust VaR estimation procedure which does not have any inherent assumptions and can work in both situations where the current best practices are suitable and where the current methodology fails. Two distinct methodologies are developed: (1) Non-parametric approach: DPFS K-means VaR; and (2) parametric approach: DBP copula VaR. The strengths and weakness of the aforementioned two methodologies are validated using real-world data (across multiple diverse domains), and verified using scenario simulated data, and through a rigorous mathematical argument. Similar to data mining techniques, there is no universal method

which has been found to robustly estimate VaR in all situations. It is concluded that future practitioners can use both methodologies depending on where each methodology works better. I have shown in the previous section, a flow-chart diagram which provides a "recipe" for determining which methodology to use based on the situation at hand (figure (94)).

One of the key strengths in the empirical analysis portion of this dissertation is that unlike in most modern risk management papers, the data that is analyzed here is available for validation to anyone with internet access. In addition, data coming from the following different domains are studied:

(1) Financial loss data;

(2) Government loss data from Chemical Spills;

(3) Insurance losses; and

(4) Natural calamities.

The LNG is a found to be robust and a flexible distribution which can accurately fit different types of loss severity data. The Burr distribution is also good for this purpose; however, there is no natural statistical interpretability of the parameters for the Burr as there is for the LNG. The problem with both of these distributions is that for low variability in the parameter estimates, one needs large datasets.

In addition, this dissertation develops five simulated data scenarios to verify the two new methodologies for robust VaR estimation. In both simulated scenarios and in real-world data, the new methodologies perform at least as well and in most cases significantly better than the current best practices.

5.1 Key Findings and the Implications of the Study

The key conclusions derived from this dissertation are the following:

(1) Lognormal-Gamma distribution is ideal to fit severity data. However, the caveat here is that LNG is a three parameter distribution, and the MLE works best with larger data for lower variance in the parameter estimates.

(2) Burr distribution does overall the best for severity fitting. However, unlike LNG, there are no closed form MLE expressions for the parameter estimates, and there are no natural interpretations for the parameter values.

(3) The classical methodology works well on "average" assuming that there is clear independence among the frequency and severity. The limits of this assumption are tested through large scale MCS for simulated data. In 4 out of 5 distinct scenarios it is found that the classical methodology grossly under/over estimates the VaR. In addition, a mathematical justification is provided that shows instances where the current method underestimates the true VaR and a case where it overestimates the true VaR.

(4) It has been found that the DPFS K-means VaR is adaptable at handling most types of real-life and scenario based data for loss VaR estimation. In the real-world data, in all cases the K-means is either the best methodology or a close second.

(5) It is observed that the DBP copula VaR never underestimates the VaR. It works the best in very specific scenarios where a very strong correlation (e.g. $\rho > 0.40$) is found between frequency and severity.

(6) From this work, it is recommended that practitioners' should use both DPFS K-means VaR and DBP copula VaR and choose the appropriate one based on correlation characteristics in their dataset. Both of these methodologies are not computationally

expensive (run within minutes in modern PCs), and practitioners should use both. Also unlike GoF tests, there is very little specific background mathematical knowledge necessary for practitioners to apply these two methodologies.

(7) Insurance companies, government agencies, and financial institutions interested in more accurately estimating the appropriate holding amount of Economic Capital would directly benefit from this work. More robust VaR estimation helps risk managers' better handle the overall budget/cash flow in their organization. In addition, it provides them insight into what KRIs they should be focusing on in the new fiscal year to potentially lower the VaR in the future.

5.2 Limitations due to Sample Size, Computational Requirements and Parametric Approach Assumptions

The current MLE approach requires a large data set size to fit the severity parameters with low variance estimates for LNG and Burr distributions. Based on my practical industry experience, a minimum of 50 data points is needed for LNG and approximately 500 data points are required for Burr for robust parameter estimation. This MLE requirement can cause problems in analyzing small sample data such as limited publicly available datasets like US hurricanes and automobile accidents. This limitation may not be pertinent if sufficiently large datasets are publicly available from the insurance and natural calamities domain.

For the DPFS K-means VaR the inherent limitation is that there is a high cost in trying to obtain the ideal split number, i.e. K. The silhouette technique is computationally intensive, and requires trial-and-error to find the K value. Therefore, in this study the value of K is bounded to be a maximum of three. This is based on the methodology from

figure (2) [14], which argues that risk can be broken down into the high/medium/low categorical descriptions. In cases where there are much finer gradations, the K-means will require intense computational time to obtain the ideal split value.

For DBP copula VaR methodology, the inherence limitation is the correct choice of the severity and frequency statistical distributions. For this study, Poisson is chosen as the frequency and lognormal is chosen as the severity. The reason is that Poisson is known to be robust for frequency distributions. Lognormal distribution is a heavy-tailed distribution and tends to model financial systems (based on Black-Scholes theory for example). Ideally speaking, LNG or Burr should model generic severity data better. However some of the datasets, namely hurricanes and automobile accidents, are quite small in size. Thus the parameter estimates from MLE would have a high variance which would add another layer of error to the VaR estimate.

5.3 Future Work

This dissertation contributes by providing a good starting ground for researchers to expand on more computationally efficient and robust methods for estimating VaR without assuming independence of frequency and severity. For the flexible severity distribution portion, an opportunity for future improvement is to move away from MLE to estimate the parameters and instead use the Minimum Hellinger Distance Estimator (MHDE) [95-103]. This methodology has a non-parametric foundation of fitting kernel density to the data and then minimizing a distance to the parametric density of interest. Some preliminary work in this area does show promise in fitting datasets of small size.

Regarding the robust estimation of DBP copula VaR, one area of immediate future work is developing a *t*-Mixture Copula Model. The Gaussian copula does not have heavy tail capability while the Student *t* distribution in general does model heavy tails well. Extending the work of [90], it should be possible to develop a *t*-Mixture Copula Model (tMCM) which can fit heavy-tail data much better than the standard Gaussian copula. This phenomenon can perhaps explain why for the real-world data the *t*-Copula tends to fit the data marginally better than the Gaussian copula. Bayesian approaches [104] to this area can also be looked into.

In addition, future research work can be done in using more sophisticated unsupervised learning techniques instead of the K-means algorithm. For example, there are cases where K-medoids tends to be more robust to noise and outliers as compared to K-means. This is because it minimizes the absolute distance between the points (L_1 norm) instead of a sum of squared Euclidean distances (L_2 norm). However, time-wise there is a computational burden for this methodology. There are new research areas such as fuzzy clustering which perhaps can better partition the loss severity and loss frequency data. It would be interesting to look at this possibility.

Empirically speaking, it would be interesting to find more actuarial based loss datasets. For example, earthquake and fire losses would be interesting to further investigate in the natural calamities domain. It would be good if further insurance based large datasets could be obtained. Finally, I have a direct longer term goal to analyze US individual tax return identity theft losses for the US Treasury from a modern ORM framework. The permission to use and anonymize this type of dataset requires longer

time, but it is possible to obtain from the US Treasury. It is a definite goal in the near future to analyze this data using the new methodologies developed in this dissertation. This will directly aid Congress in better allocating the US Treasury's budget in terms of how much they can expect to lose due to identity theft of US tax returns.

Finally, one of the theoretical areas of future work comes from modeling Aggregate Loss distribution. The current best practices use FFT based models. There are some recent thoughts on using the Discrete Wavelet Transform [105-115] to compute an analytical estimate for the aggregate loss, but this only works for gamma severity and Poisson frequency. An interesting area to pursue would be to test if theoretically one can expand this to include the Burr and LNG distributions for the severity component of modern ORM.

APPENDIX A: SAMPLE SET OF R CODES FOR ANALYSIS

Main Scripts

```
############Copula Analysis

mydata.ScenarioII <- read.xlsx(xlsxFile="C:/Users/sabyguharay/Documents/GMU/Fourth
Year/CopulaWork/ScenarioII_Data.xlsx", sheet="Input_Data_R", colNames=TRUE);
good.data.ScenarioII <- data.frame(mydata.ScenarioII)
attach(good.data.ScenarioII);
head(good.data.ScenarioII);  #check that stuff makes sense
dim(good.data.ScenarioII);
#good.data.ScenarioII <- good.data.ScenarioII[,-4];  #Removes Column 4 if needed
names(good.data.ScenarioII)
#good.data.ScenarioII$NA. <- NULL;   in case I have to remove a column

save.image("~/.RData");

#########################################################Comparison of ########################

library(openxlsx);
help(read.xlsx)

classical.severity.ScenarioII <- read.xlsx(xlsxFile="C:/Users/sabyguharay/Documents/GMU/Fourth
Year/CopulaWork/ScenarioII_Data.xlsx", sheet="Severity", colNames=TRUE);
head(classical.severity.ScenarioII);


classical.severity.ScenarioII <- read.csv("C:/Users/sabyguharay/Documents/GMU/Fourth
Year/CopulaWork/Scenario_II_Severity.csv", head=TRUE);


classical.severity.ScenarioII <- data.frame(classical.severity.ScenarioII)
attach(classical.severity.ScenarioII);
head(classical.severity.ScenarioII);  #check that stuff makes sense
names(classical.severity.ScenarioII)

hist(log(classical.severity.ScenarioII$Raw.Severity), nclass=500)    #Classical Severity
summary(good.data.ScenarioII$Lambda_hat)            #Classical Frequency
summary(good.data.ScenarioII$Mu_hat)
summary(good.data.ScenarioII$Sig_hat)
#####################################################End of ########################



dev.off();
par(mfrow=c(2,2));
hist(good.data.ScenarioII$Mu_hat, nclass=200, xaxt='n', col="yellow", main=expression(paste("Histogram of
Estimated ", mu)), xlab=expression(hat(mu)), ylab="Density", cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5);
#abline(v=1, col="gray60", lwd=2)
```

```
#abline(v=10, col="lightgray", lwd=2)
axis(side=1, at=seq(0,4,0.5), labels=seq(0,4,0.5))

hist(good.data.ScenarioII$Sig_hat, nclass=50, prob=TRUE, xaxt='n', col="green", main=expression(paste("Histogram
of Estimated ", sigma)), xlab=expression(hat(sigma)), ylab="Density",  cex.lab=1.5, cex.axis=1.5, cex.main=1.5,
cex.sub=1.5);
#abline(v=0.5, col="black", lwd=2)
axis(side=1, at=seq(0.75,4.0,0.25), labels=seq(0.75,4.0,0.25))

hist(good.data.ScenarioII$Lambda_hat, nclass=50, prob=TRUE, col="red", xaxt='n',
main=expression(paste("Histogram of Estimated ", lambda)), xlab=expression(hat(lambda)), ylab="Mass Function",
cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5);
#abline(v=5, col="blue", lwd=2)
#abline(v=50, col="magenta", lwd=2)
axis(side=1, at=seq(20,70,5), labels=seq(20,70,5))

hist(log(classical.severity.ScenarioII$Raw.Severity), nclass=50, prob=TRUE, col="black", xaxt='n', main="Histogram
of Severity Losses: Scenario II", xlab="Log(Losses)", ylab="Density",  cex.lab=1.5, cex.axis=1.5, cex.main=1.5,
cex.sub=1.5);
#abline(v=5, col="blue", lwd=2)
#abline(v=50, col="magenta", lwd=2)
axis(side=1, at=seq(-5,15,1), labels=seq(-5,15,1))

#Create Data Object
data.ScenarioII<- data.frame(lambdaHat = good.data.ScenarioII$Lambda_hat, muHat=good.data.ScenarioII$Mu_hat,
sigmaHat=good.data.ScenarioII$Sig_hat)
#Order is Lambda, Mu, Sigma


#First check the Spearman's Correlation
cor(data.ScenarioII,method='kendall')
cor(data.ScenarioII,method='spearman')
pairs.panels(data.ScenarioII, method="spearman")   #Default is Pearson's Correlation
pairs.panels(data.ScenarioII)   #Default is Pearson's Correlation

#Then do F(data); i.e. transform to U[0,1] scale
#Pseudo observations are the observations in the [0,1] interval.
uDat.ScenarioII<- pobs(as.matrix(data.ScenarioII))
pairs.panels(uDat.ScenarioII)


#Compute the Gaussian/t/Gumbel Copula
normal.cop <-normalCopula(c(0.9,-0.9, 0.9),dim=3,dispstr="un")  #Starting point is c(0.9,-0.9,0.9)
fg.ScenarioII <- fitCopula(copula=normal.cop, data=uDat.ScenarioII, optim.method="Nelder-Mead", method='ml')
rhos.ScenarioII<- coef(fg.ScenarioII)
rhos.ScenarioII

#My Stuff  (Need to Understand t-Copulas a little better)
myCop.t <- tCopula(rep(0, 3), dim = 3, dispstr = "un", df=4, df.fixed=TRUE)
#myCop.t <- tCopula(rep(0, 3), dim = 3, dispstr = "un")
myCop.t <- ellipCopula(family = "t", dim = 3, dispstr = "un",
                       param = c(-0.9, 0.5, 0.05), df = 4)
fg.t.ScenarioII <- fitCopula(copula=myCop.t, data=uDat.ScenarioII, optim.method="Nelder-Mead", method='mpl')
rhos.t.ScenarioII<- coef(fg.t.ScenarioII)
rhos.t.ScenarioII
fg.t.ScenarioII@estimate[4] <- round(fg.t.ScenarioII@estimate[4]);
fg.t.ScenarioII@estimate = fg.t.ScenarioII@estimate[-4]

#Trying other copulas
```

```
#gum.cop.4 = archmCopula(family="gumbel", dim=3, param=3)
#myCop.clayton <- archmCopula(family = "clayton", dim = 3, param = 3)
#fg.gumbel <- fitCopula(copula=gum.cop.4, data=uDat.ScenarioII, method="mpl");
#rhos.gumbel <- coef(fg.gumbel)
#rhos.gumbel

#fg.clayton <- fitCopula(copula=myCop.clayton, data=uDat.ScenarioII, method="mpl");
#rhos.clayton <- coef(fg.clayton)
#rhos.clayton




#Create the Multivariate Distribution from Copula Object
mvdc.normal.ScenarioII <- mvdc(copula=normalCopula(fg.ScenarioII@estimate,dim=3,dispstr="un"),
                        margins=c("emp","emp", "emp"),paramMargins=
                          list(list(obs=data.ScenarioII$lambdaHat),
                              list(obs=data.ScenarioII$muHat),
                              list(obs=data.ScenarioII$sigmaHat)));

mvdc.t.ScenarioII <- mvdc(copula=tCopula(fg.t.ScenarioII@estimate,dim=3,dispstr="un"),
                     margins=c("emp","emp", "emp"),paramMargins=
                       list(list(obs=data.ScenarioII$lambdaHat),
                          list(obs=data.ScenarioII$muHat),
                          list(obs=data.ScenarioII$sigmaHat)));


#mvdc.gumbel <- mvdc(copula=gumbelCopula(fg.gumbel@estimate,dim=3),
#                    margins=c("emp","emp", "emp"),paramMargins=
#                      list(list(obs=data.ScenarioII$lambdaHat),
#                          list(obs=data.ScenarioII$muHat),
#                          list(obs=data.ScenarioII$sigmaHat)));



#Randomly Sample from MVDC
size.data.ScenarioII = dim(good.data.ScenarioII)[1];

random.mvdc.normal.ScenarioII <- rMvdc(n=size.data.ScenarioII, mvdc.normal.ScenarioII)
head(random.mvdc.normal.ScenarioII)
colnames(random.mvdc.normal.ScenarioII) <- c("lambda","mu", "sigma")

random.mvdc.t.ScenarioII <- rMvdc(n=size.data.ScenarioII, mvdc.t.ScenarioII)
head(random.mvdc.t.ScenarioII)
colnames(random.mvdc.t.ScenarioII) <- c("lambda","mu", "sigma")


random.mvdc.normal.ScenarioII.xlsx <- rMvdc(n=10000, mvdc.normal.ScenarioII)
random.mvdc.t.ScenarioII.xlsx  <- rMvdc(n=10000, mvdc.t.ScenarioII)

#random.mvdc.gumbel <- rMvdc(n=1000, mvdc.gumbel)
#head(random.mvdc.gumbel)

L_Component.ScenarioII = random.mvdc.normal.ScenarioII[,1];
Mu_Component.ScenarioII = random.mvdc.normal.ScenarioII[,2];
Sig_Component.ScenarioII= random.mvdc.normal.ScenarioII[,3];
```

```
x.comp.ScenarioII = c(L_Component.ScenarioII,good.data.ScenarioII$Lambda_hat);
y.comp.ScenarioII = c(Mu_Component.ScenarioII,good.data.ScenarioII$Mu_hat);
z.comp.ScenarioII = c(Sig_Component.ScenarioII,good.data.ScenarioII$Sig_hat);
df.ScenarioII = data.frame(cbind(x.comp.ScenarioII, y.comp.ScenarioII, z.comp.ScenarioII));
df.ScenarioII$fac <- factor(rep(LETTERS[1:2], each = length(good.data.ScenarioII$Mu_hat)))

#plot3d(x.comp,y.comp,z.comp,pch=20,col='blue', xlab="Lambda Hat", ylab="Mu Hat", zlab="Sigma Hat");
plot3d(df.ScenarioII$x.comp.ScenarioII, df.ScenarioII$y.comp.ScenarioII, df.ScenarioII$z.comp.ScenarioII,
col=as.numeric(df.ScenarioII$fac), xlab="Lambda Hat", ylab="Mu Hat", zlab="Sigma Hat");
legend3d("topleft", legend = paste('Type:', c('Copula Surface', 'Empirical Surface')), pch = 16, col = c("red", "black"),
cex=1.0, inset=c(0.01))
#scatterplot3d(df$x.comp, df$y.comp, df$z.comp, color=as.numeric(df$fac), xlab=expression(hat(lambda)),
ylab=expression(hat(mu)), zlab=expression(hat(sigma)));



L_Component.t.ScenarioII = random.mvdc.t.ScenarioII[,1];
Mu_Component.t.ScenarioII = random.mvdc.t.ScenarioII[,2];
Sig_Component.t.ScenarioII= random.mvdc.t.ScenarioII[,3];

x.comp.t.ScenarioII = c(L_Component.t.ScenarioII,good.data.ScenarioII$Lambda_hat);
y.comp.t.ScenarioII = c(Mu_Component.t.ScenarioII,good.data.ScenarioII$Mu_hat);
z.comp.t.ScenarioII = c(Sig_Component.t.ScenarioII,good.data.ScenarioII$Sig_hat);
df.t.ScenarioII = data.frame(cbind(x.comp.t.ScenarioII, y.comp.t.ScenarioII, z.comp.t.ScenarioII));
df.t.ScenarioII$fac <- factor(rep(LETTERS[1:2], each = length(good.data.ScenarioII$Mu_hat)))

#plot3d(x.comp,y.comp,z.comp,pch=20,col='blue', xlab="Lambda Hat", ylab="Mu Hat", zlab="Sigma Hat");
plot3d(df.t.ScenarioII$x.comp.t.ScenarioII, df.t.ScenarioII$y.comp.t.ScenarioII, df.t.ScenarioII$z.comp.t.ScenarioII,
col=as.numeric(df.t.ScenarioII$fac), xlab="Lambda Hat", ylab="Mu Hat", zlab="Sigma Hat");
legend3d("topleft", legend = paste('Type:', c('Empirical Surface', 'Copula Surface')), pch = 16, col = c("red", "black"),
cex=1, inset=c(0.02))
#scatterplot3d(df$x.comp, df$y.comp.3, df$z.comp, color=as.numeric(df$fac), xlab=expression(hat(lambda)),
ylab=expression(hat(mu)), zlab=expression(hat(sigma)));

#Export table to XLSX so Monte Carlo can be done in MATLAB
#Must Close the XLSX File for the writing to be done

options(java.parameters = "-Xmx8000m");

write.table(x=random.mvdc.normal.ScenarioII.xlsx, file="C:/Users/sabyguharay/Documents/GMU/Fourth
Year/CopulaWork/copula_normal_ScenarioII.csv", quote=FALSE, sep=",",
          col.names=TRUE);

write.table(x=random.mvdc.t.ScenarioII.xlsx, file="C:/Users/sabyguharay/Documents/GMU/Fourth
Year/CopulaWork/copula_t_ScenarioII.csv", quote=FALSE, sep=",",
          col.names=TRUE);


#write.xlsx(x=random.mvdc.normal, file="C:/Users/sabyguharay/Documents/GMU/Fourth
Year/CopulaWork/Copula_SimData_III.xlsx", sheetName="Copula_Surface_Normal_big",
#          col.names=TRUE, append=TRUE);

#write.xlsx(x=random.mvdc.t, file="C:/Users/sabyguharay/Documents/GMU/Fourth
Year/CopulaWork/Copula_SimData_III.xlsx", sheetName="Copula_Surface_t_big",
#          col.names=TRUE, append=TRUE);
```

```
x.comp.ScenarioIv = c(L_Component.ScenarioIV,good.data.ScenarioIv$Lambda_hat);
y.comp.ScenarioIv = c(Mu_Component.ScenarioIV,good.data.ScenarioIv$Mu_hat);
z.comp.ScenarioIv= c(Sig_Component.ScenarioIV,good.data.ScenarioIv$Sig_hat);
df.ScenarioIV = data.frame(cbind(x.comp.ScenarioIV, y.comp.ScenarioIv, z.comp.ScenarioIV));
df.ScenarioIV$fac <- factor(rep(LETTERS[1:2], each = length(good.data.ScenarioIV$Mu_hat)))

#plot3d(x.comp,y.comp,z.comp,pch=20,col='blue', xlab="Lambda Hat", ylab="Mu Hat", zlab="Sigma Hat");
plot3d(df.ScenarioIV$x.comp.ScenarioIV, df.ScenarioIV$y.comp.ScenarioIV, df.ScenarioIV$z.comp.ScenarioIV,
col=as.numeric(df.ScenarioIV$fac), xlab="Lambda Hat", ylab="Mu Hat", zlab="Sigma Hat");
legend3d("topleft", legend = paste('Type:', c('Copula Surface', 'Empirical Surface')), pch = 16, col = c("red", "black"),
cex=1.0, inset=c(0.01))
#scatterplot3d(df$x.comp, df$y.comp, df$z.comp, color=as.numeric(df$fac), xlab=expression(hat(lambda)),
ylab=expression(hat(mu)), zlab=expression(hat(sigma)));



L_Component.t.ScenarioIv = random.mvdc.t.ScenarioIv [,1];
Mu_Component.t.ScenarioIV = random.mvdc.t.ScenarioIv [,2];
Sig_Component.t.ScenarioIV= random.mvdc.t.ScenarioIv [,3];

x.comp.t.ScenarioIV = c(L_Component.t.ScenarioIV,good.data.ScenarioIV$Lambda_hat);
y.comp.t.ScenarioIV = c(Mu_Component.t.ScenarioIV,good.data.ScenarioIV$Mu_hat);
z.comp.t.ScenarioIV = c(Sig_Component.t.ScenarioIV,good.data.ScenarioIV$Sig_hat);
df.t.ScenarioIV = data.frame(cbind(x.comp.t.ScenarioIV, y.comp.t.ScenarioIV, z.comp.t.ScenarioIV));
df.t.ScenarioIV$fac <- factor(rep(LETTERS[1:2], each = length(good.data.ScenarioII$Mu_hat)))

#plot3d(x.comp,y.comp,z.comp,pch=20,col='blue', xlab="Lambda Hat", ylab="Mu Hat", zlab="Sigma Hat");
plot3d(df.t.ScenarioII$x.comp.t.ScenarioII, df.t.ScenarioII$y.comp.t.ScenarioII, df.t.ScenarioII$z.comp.t.ScenarioII,
col=as.numeric(df.t.ScenarioII$fac), xlab="Lambda Hat", ylab="Mu Hat", zlab="Sigma Hat");
legend3d("topleft", legend = paste('Type:', c('Empirical Surface', 'Copula Surface')), pch = 16, col = c("red", "black"),
cex=1, inset=c(0.02))
#scatterplot3d(df$x.comp, df$y.comp.3, df$z.comp, color=as.numeric(df$fac), xlab=expression(hat(lambda)),
ylab=expression(hat(mu)), zlab=expression(hat(sigma)));

#Export table to XLSX so Monte Carlo can be done in MATLAB
#Must Close the XLSX File for the writing to be done

options(java.parameters = "-Xmx8000m");

write.table(x=random.mvdc.normal.ScenarioII.xlsx, file="C:/Users/sabyguharay/Documents/GMU/Fourth
Year/CopulaWork/copula_normal_ScenarioII.csv", quote=FALSE, sep=",",
        col.names=TRUE);

write.table(x=random.mvdc.t.ScenarioII.xlsx, file="C:/Users/sabyguharay/Documents/GMU/Fourth
Year/CopulaWork/copula_t_ScenarioII.csv", quote=FALSE, sep=",",
        col.names=TRUE);
```

APPENDIX B: SAMPLE MATLAB PROGRAMS

Main scripts

```
clc;
clear all;


%
%%%%%%%%%%%%%%%Generate Simulated Data
Numb_Sim_Months = 10000;
Numb_loss_Month = zeros(Numb_Sim_Months*1, 1);
%Month_region = zeros(Numb_Sim_Months, 1);
Mean_Severity_Month = zeros(Numb_Sim_Months*1, 1);
Median_Severity_Month = zeros(Numb_Sim_Months*1, 1);
Sum_Severity_Month = zeros(Numb_Sim_Months*1, 1);
Min_Severity_Month = zeros(Numb_Sim_Months*1, 1);
Max_Severity_Month = zeros(Numb_Sim_Months*1, 1);
Month_counter = 1:Numb_Sim_Months;
Month_counter = Month_counter';

%

lambda_unique = 14;

%%%%%%End of Frequency Parameters


%

mu_true = 5;
sig_true = 2;


%%%%%%%%%

Freq_Sev_Matrix = [];


All_severity = zeros(Numb_Sim_Months*100,1);

Threshold = 0.01;
 counter=0;
% Simulation for
tic;
for i = 1:Numb_Sim_Months
   %Low Frequency
      Numb_loss_Month(i) = poissrnd(lambda_unique);
      %Generate
```

```
      Y_ln = logsamp(mu_true,sig_true,Threshold,Numb_loss_Month(i));
      counter = counter+Numb_loss_Month(i);
      Mean_Severity_Month(i) = mean(Y_ln);
      Median_Severity_Month(i) = median(Y_ln);
      Sum_Severity_Month(i) = sum(Y_ln);
   if (i == 1)
      All_severity(1:counter) = Y_ln;
   else
      All_severity(counter-numel(Y_ln)+1:counter) = Y_ln;
   end
   if mod(i, 100000) == 0
      disp(i);
   end
end
Clean_All_severity=All_severity(1:find(~All_severity,1)-1);
t=toc;
t;
disp(t);      %%%%%%%%%%%%%%%%%%%%%%%280 Seconds


%hist(Sum_Severity_Month);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%Store Matrix
for Losses

tic;
Month_counter_severity = [];
Frequency_counter_severity = [];
Median_Sev_Freq_matrix = [];
Mean_Sev_Freq_matrix = [];

Clean_Median_Severity_Month = Median_Severity_Month;
Clean_Mean_Severity_Month = Mean_Severity_Month;
Clean_Median_Severity_Month(isnan(Clean_Median_Severity_Month)) = 0;
Clean_Mean_Severity_Month(isnan(Clean_Mean_Severity_Month)) = 0;


for i=1:Numb_Sim_Months
   Month_counter_severity = [Month_counter_severity; repmat(Month_counter(i), Numb_loss_Month(i),1)];
   Frequency_counter_severity = [Frequency_counter_severity;
repmat(Numb_loss_Month(i),Numb_loss_Month(i),1)];
end


Severity_Month_Freq_matrix = cat(2, Frequency_counter_severity, Clean_All_severity);

Median_Sev_Freq_matrix = cat(2, Numb_loss_Month, Clean_Median_Severity_Month);
Mean_Sev_Freq_matrix = cat(2, Numb_loss_Month, Clean_Mean_Severity_Month);

Month_Counter_Frequency = [];


for i = 1:Numb_Sim_Months
   test = repmat(Month_counter(i), Numb_loss_Month(i), 1);
   Month_Counter_Frequency = cat(1,Month_Counter_Frequency,test);
end
```

```
Month_Severity_Matrix = cat(2,Month_Counter_Frequency, Clean_All_severity);
Freq_Severity_Matrix = cat(2,Month_counter, Numb_loss_Month, Mean_Severity_Month, Median_Severity_Month,
Sum_Severity_Month);

   %%%%%%%%%%%%%%Export Data to XLSX
    full_path = 'C:\Users\sabyguharay\Documents\GMU\Fourth Year\Theoretical_RawData_Scenario.xlsx';
    Utility_header = {'Month Number','No. of Losses', 'Mean Loss', 'Median Loss', 'Aggregate Loss Sum'};
    xlRange = 'A1';
    xlswrite(full_path, Utility_header, 'Freq_Severity_Matrix',xlRange);
    xlRange_2 = 'A2';
    xlswrite(full_path, Freq_Severity_Matrix, 'Freq_Severity_Matrix', xlRange_2);

    Utility_header_2 = {'Month Number','Raw Severity'};
    xlswrite(full_path, Utility_header_2, 'Month_Severity_Matrix',xlRange);
    xlswrite(full_path, Month_Severity_Matrix, 'Month_Severity_Matrix', xlRange_2);
t=toc;
disp(t);  % takes 60 seconds

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%END OF
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%Empirical Analysis

T_Log_0p4 = prctile(HighFreq_severity, 0);
T_Log_25p4 = prctile(HighFreq_severity, 25);
T_Log_33p4 = prctile(HighFreq_severity, 33);
T_Log_50p4 = prctile(HighFreq_severity, 50);
T_Log_67p4 = prctile(HighFreq_severity, 67);
T_Log_75p4 = prctile(HighFreq_severity, 75);
T_Log_90p4 = prctile(HighFreq_severity, 90);
T_Log_95p4 = prctile(HighFreq_severity, 95);
T_Log_98p4 = prctile(HighFreq_severity, 98);
T_Log_99p4 = prctile(HighFreq_severity, 99);
T_Log_995p4 = prctile(HighFreq_severity, 99.5);
T_Log_999p4 = prctile(HighFreq_severity, 99.9);
T_Log_9995p4 = prctile(HighFreq_severity, 99.95);
T_Log_9999p4 = prctile(HighFreq_severity, 99.99);
T_Log_mean = mean(HighFreq_severity);

new_arr = [T_Log_mean T_Log_0p4 T_Log_25p4 T_Log_33p4 T_Log_50p4 T_Log_67p4 T_Log_75p4 T_Log_90p4
T_Log_95p4 T_Log_98p4 T_Log_99p4 T_Log_995p4 T_Log_999p4 T_Log_9995p4 T_Log_9999p4]';
%LNG_VaR_2 = flipud(new_arr);
Empirical_HF_Quantiles= new_arr;
Empirical_HF_Quantiles



T_Log_0p4 = prctile(LowFreq_severity, 0);
T_Log_25p4 = prctile(LowFreq_severity, 25);
T_Log_33p4 = prctile(LowFreq_severity, 33);
T_Log_50p4 = prctile(LowFreq_severity, 50);
T_Log_67p4 = prctile(LowFreq_severity, 67);
T_Log_75p4 = prctile(LowFreq_severity, 75);
T_Log_90p4 = prctile(LowFreq_severity, 90);
T_Log_95p4 = prctile(LowFreq_severity, 95);
```

```
T_Log_98p4 = prctile(LowFreq_severity, 98);
T_Log_99p4 = prctile(LowFreq_severity, 99);
T_Log_995p4 = prctile(LowFreq_severity, 99.5);
T_Log_999p4 = prctile(LowFreq_severity, 99.9);
T_Log_9995p4 = prctile(LowFreq_severity, 99.95);
T_Log_9999p4 = prctile(LowFreq_severity, 99.99);
T_Log_mean = mean(LowFreq_severity);

new_arr = ['T_Log_mean T_Log_0p4 T_Log_25p4 T_Log_33p4 T_Log_50p4 T_Log_67p4 T_Log_75p4 T_Log_90p4
T_Log_95p4 T_Log_98p4 T_Log_99p4 T_Log_995p4 T_Log_999p4 T_Log_9995p4 T_Log_9999p4]';
%LNG_VaR_2 = flipud(new_arr);
Empirical_LF_Quantiles = new_arr;
Empirical_LF_Quantiles


T_Log_0p4 = prctile(MedFreq_severity, 0);
T_Log_25p4 = prctile(MedFreq_severity, 25);
T_Log_33p4 = prctile(MedFreq_severity, 33);
T_Log_50p4 = prctile(MedFreq_severity, 50);
T_Log_67p4 = prctile(MedFreq_severity, 67);
T_Log_75p4 = prctile(MedFreq_severity, 75);
T_Log_90p4 = prctile(MedFreq_severity, 90);
T_Log_95p4 = prctile(MedFreq_severity, 95);
T_Log_98p4 = prctile(MedFreq_severity, 98);
T_Log_99p4 = prctile(MedFreq_severity, 99);
T_Log_995p4 = prctile(MedFreq_severity, 99.5);
T_Log_999p4 = prctile(MedFreq_severity, 99.9);
T_Log_9995p4 = prctile(MedFreq_severity, 99.95);
T_Log_9999p4 = prctile(MedFreq_severity, 99.99);
T_Log_mean = mean(MedFreq_severity);

new_arr = ['T_Log_mean T_Log_0p4 T_Log_25p4 T_Log_33p4 T_Log_50p4 T_Log_67p4 T_Log_75p4 T_Log_90p4
T_Log_95p4 T_Log_98p4 T_Log_99p4 T_Log_995p4 T_Log_999p4 T_Log_9995p4 T_Log_9999p4]';
Empirical_MF_Quantiles = new_arr;
Empirical_MF_Quantiles


Log_0p4 = prctile(Sum_Severity_Month, 0);
Log_25p4 = prctile(Sum_Severity_Month, 25);
Log_50p4 = prctile(Sum_Severity_Month, 50);
Log_90p4 = prctile(Sum_Severity_Month, 90);
Log_95p4 = prctile(Sum_Severity_Month, 95);
Log_98p4 = prctile(Sum_Severity_Month, 98);
Log_99p4 = prctile(Sum_Severity_Month, 99);
Log_995p4 = prctile(Sum_Severity_Month, 99.5);
Log_999p4 = prctile(Sum_Severity_Month, 99.9);
Log_9995p4 = prctile(Sum_Severity_Month, 99.95);
Log_9999p4 = prctile(Sum_Severity_Month, 99.99);
Log_99999p4 = prctile(Sum_Severity_Month, 99.999);
Log_999995p4 = prctile(Sum_Severity_Month, 99.9995);
Log_999999p4 = prctile(Sum_Severity_Month, 99.9999);


Log_mean = mean(Sum_Severity_Month);

new_arr = [Log_mean Log_0p4 Log_25p4 Log_50p4 Log_90p4 Log_95p4 Log_98p4 Log_99p4 Log_995p4
Log_999p4 Log_9995p4 Log_9999p4 Log_99999p4 Log_999995p4 Log_999999p4]';
%LNG_VaR_2 = flipud(new_arr);
VaR_Data_Historical = new_arr;
```

```
VaR_Data_Historical

%%%%%%%%%%%%%%Plot of the Severity Region
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%
subplot(2,1,1)      % add first plot in 2 x 2 grid
hist(log(Clean_All_severity));
xlabel({'Loss Severity (Log)'},'FontSize',16, 'FontName','Times New Roman');
ylabel({'Count'},'FontSize',16, 'FontName','Times New Roman');
title({'Simulated Data of the Daily Loss Severity for Scenario III'},...
   'FontSize',16, 'FontName','Times New Roman');

subplot(2,1,2)      % add first plot in 2 x 2 grid
hist(Numb_loss_Month);
xlabel({'# Losses per Month'},'FontSize',16, 'FontName','Times New Roman');
ylabel({'Count'},'FontSize',16, 'FontName','Times New Roman');
title({'Frequency Distribution for Scenario III'},...
   'FontSize',16, 'FontName','Times New Roman');

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

[~, ~, raw] = xlsread('C:\Users\sabyguharay\Documents\GMU\Fourth Year\Real Data\Files Ready for
Analysis\MATLAB_FEEDER.xlsx','Data_For_Pivot_Table','U8:U307');
raw(cellfun(@(x) ~isempty(x) && isnumeric(x) && isnan(x),raw)) = {''};

%% Replace non-numeric cells with NaN
R = cellfun(@(x) ~isnumeric(x) && ~islogical(x),raw); % Find non-numeric cells
raw(R) = {NaN}; % Replace non-numeric cells

%% Create output variable
data = reshape([raw{:}],size(raw));

%% Allocate imported array to column variable names
Empirical_Agg_Loss = data(:,1);
%Clean_Mean_Severity_Month = data(:,2);
Numb_Sim_boot = 10000;

%% Clear temporary variables
clearvars data raw cellVectors R;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%END OF IMPORTING THE


Sample_length=200;
Quantile_length = 7;
Bootstrap_n = Numb_Sim_boot;      %10000 months of Bootstrap sample
Bootstrap_matrix = zeros(Bootstrap_n*100, Quantile_length);
%Distribution
tic;
for j=1:(Bootstrap_n*100)
    New_Loss = datasample(Empirical_Agg_Loss,Sample_length,'Replace',true);
    Bootstrap_matrix(j,1) = prctile(New_Loss, 0);
    Bootstrap_matrix(j,2) = prctile(New_Loss, 25);
    Bootstrap_matrix(j,3) = prctile(New_Loss, 50);
    Bootstrap_matrix(j,4) = prctile(New_Loss, 90);
    Bootstrap_matrix(j,5) = prctile(New_Loss, 95);
    Bootstrap_matrix(j,6) =  prctile(New_Loss, 98);
    Bootstrap_matrix(j,7) =  prctile(New_Loss, 99);
```

```
        if mod(j, Bootstrap_n*10) == 0
            disp(j);
        end
end
%t=toc;
%t


Final_Quantiles_Matrix_Spills = zeros(2,7);
quantile_interest = 99;

for i=1:7
    Final_Quantiles_Matrix_Spills(1,i) = prctile(Bootstrap_matrix(:,i), 100-quantile_interest);
    Final_Quantiles_Matrix_Spills(2,i) = prctile(Bootstrap_matrix(:,i), quantile_interest);
end

Export_xls = Final_Quantiles_Matrix_Spills';
% t=toc;
% disp(t);  %1040 seconds

%%%%%%%%%%%%%%%%%%%%%%%%
    %%%%%%%%%%%%%%%Export Data to XLSX
     full_path = 'C:\Users\sabyguharay\Documents\GMU\Dissertation Final Defense\Print Results Thesis Level
Good.xlsx';
     Utility_header = {'Lower','Upper'};
     xlRange = 'M3';
     xlswrite(full_path, Utility_header, 'Real-World Data New',xlRange);
     xlRange_2 = 'M4';
     xlswrite(full_path, Export_xls, 'Real-World Data New', xlRange_2);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%End of Writing to XLSX
%t=toc;
%disp(t);   %28 seconds

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%End of Chemical
Spills
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%%%%Begin
% [~, ~, raw] = xlsread('C:\Users\sabyguharay\Documents\GMU\Fourth Year\Excel Frequency Analysis\SP500
WRDS 1925 - Current.xlsx','FinalResults','M7:M1074');
% raw(cellfun(@(x) ~isempty(x) && isnumeric(x) && isnan(x),raw)) = {''};
%
% %% Replace non-numeric cells with NaN
% R = cellfun(@(x) ~isnumeric(x) && ~islogical(x),raw); % Find non-numeric cells
% raw(R) = {NaN}; % Replace non-numeric cells
%
% %% Create output variable
% data = reshape([raw{:}],size(raw));
%
% %% Allocate imported array to column variable names
% Empirical_Agg_Loss_SP500 = data(:,1);
% %Clean_Mean_Severity_Month = data(:,2);
%
% %% Clear temporary variables
% clearvars data raw cellVectors R;
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%END OF IMPORTING SP500 Historical VaR

Sample_length_SP500=1000;
%tic;
Bootstrap_matrix_SP500 = zeros(Bootstrap_n*100, Quantile_length+1);
%Distribution
for j=1:(Bootstrap_n*100)
        New_Loss = datasample(Empirical_Agg_Loss_SP500,Sample_length_SP500,'Replace',true);
        Bootstrap_matrix_SP500(j,1) = prctile(New_Loss, 0);
        Bootstrap_matrix_SP500(j,2) = prctile(New_Loss, 25);
        Bootstrap_matrix_SP500(j,3) = prctile(New_Loss, 50);
        Bootstrap_matrix_SP500(j,4) = prctile(New_Loss, 90);
        Bootstrap_matrix_SP500(j,5) = prctile(New_Loss, 95);
        Bootstrap_matrix_SP500(j,6) =  prctile(New_Loss, 98);
        Bootstrap_matrix_SP500(j,7) =  prctile(New_Loss, 99);
        Bootstrap_matrix_SP500(j,8) =  prctile(New_Loss, 99.5);
        if mod(j, Bootstrap_n*10) == 0
            disp(j);
        end
end


Final_Quantiles_Matrix_SP500 = zeros(2,Quantile_length+1);
%quantile_interest = 95;

for i=1:(Quantile_length+1)
    Final_Quantiles_Matrix_SP500(1,i) = prctile(Bootstrap_matrix_SP500(:,i), 100-quantile_interest);
    Final_Quantiles_Matrix_SP500(2,i) = prctile(Bootstrap_matrix_SP500(:,i), quantile_interest);
end
% t=toc;
% disp(t);      %16 Seconds



Export_xls = Final_Quantiles_Matrix_SP500';
disp('I am on the SP500');
%%%%%%%%%%%%%%%%%%%%%%%%%%Export Results to Excel
    %%%%%%%%%%%%%%%Export Data to XLSX
     full_path = 'C:\Users\sabyguharay\Documents\GMU\Dissertation Final Defense\Print Results Thesis Level
Good.xlsx';
     Utility_header = {'Lower','Upper'};
     xlRange = 'M22';
     xlswrite(full_path, Utility_header, 'Real-World Data New',xlRange);
     xlRange_2 = 'M23';
     xlswrite(full_path, Export_xls, 'Real-World Data New', xlRange_2);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%End of Writing to XLSX

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%END OF
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%Begin
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%Begin DJIA Analysis
% [~, ~, raw] = xlsread('C:\Users\sabyguharay\Documents\GMU\Fourth Year\Real Data\Latest Data from
Yahoo.xlsx','Input_DJIA_MATLAB','E2:E798');
% raw(cellfun(@(x) ~isempty(x) && isnumeric(x) && isnan(x),raw)) = {''};
%
% %% Replace non-numeric cells with NaN
% R = cellfun(@(x) ~isnumeric(x) && ~islogical(x),raw); % Find non-numeric cells
% raw(R) = {NaN}; % Replace non-numeric cells
%
% %% Create output variable
% data = reshape([raw{:}],size(raw));
%
% %% Allocate imported array to column variable names
% Empirical_Agg_Loss_DJIA = data(:,1);
% %Clean_Mean_Severity_Month = data(:,2);
%
% %% Clear temporary variables
% clearvars data raw cellVectors R;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%END OF IMPORTING


% tic;
Sample_length_DJIA=500;
Bootstrap_matrix_DJIA = zeros(Bootstrap_n*100, Quantile_length);
%Distribution
for j=1:(Bootstrap_n*100)
    New_Loss = datasample(Empirical_Agg_Loss_DJIA,Sample_length_DJIA,'Replace',true);
    Bootstrap_matrix_DJIA(j,1) = prctile(New_Loss, 0);
    Bootstrap_matrix_DJIA(j,2) = prctile(New_Loss, 25);
    Bootstrap_matrix_DJIA(j,3) = prctile(New_Loss, 50);
    Bootstrap_matrix_DJIA(j,4) = prctile(New_Loss, 90);
    Bootstrap_matrix_DJIA(j,5) = prctile(New_Loss, 95);
    Bootstrap_matrix_DJIA(j,6) =  prctile(New_Loss, 98);
    Bootstrap_matrix_DJIA(j,7) =  prctile(New_Loss, 99);
    if mod(j, Bootstrap_n*10) == 0
      disp(j);
    end
end


Final_Quantiles_Matrix_DJIA = zeros(2,7);
%quantile_interest = 95;

for i=1:7
  Final_Quantiles_Matrix_DJIA(1,i) = prctile(Bootstrap_matrix_DJIA(:,i), 100-quantile_interest);
  Final_Quantiles_Matrix_DJIA(2,i) = prctile(Bootstrap_matrix_DJIA(:,i), quantile_interest);
end
% t=toc;
% disp(t);    %11 Seconds

Export_xls = Final_Quantiles_Matrix_DJIA';
disp('I am on the DJIA');
%%%%%%%%%%%%%%%%%%%%%%%%Export Results to Excel
  %%%%%%%%%%%%%%Export Data to XLSX
   full_path = 'C:\Users\sabyguharay\Documents\GMU\Dissertation Final Defense\Print Results Thesis Level
Good.xlsx';
```

```
    Utility_header = {'Lower','Upper'};
    xlRange = 'M59';
    xlswrite(full_path, Utility_header, 'Real-World Data New',xlRange);
    xlRange_2 = 'M60';
    xlswrite(full_path, Export_xls, 'Real-World Data New', xlRange_2);


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%END
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%



%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%Begin
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%%%%%Begin Auto Analysis
% [~, ~, raw] = xlsread('C:\Users\sabyguharay\Documents\GMU\Fourth Year\Real
Data\automobile.xlsx','PivotTable','N5:N119');
% raw(cellfun(@(x) ~isempty(x) && isnumeric(x) && isnan(x),raw)) = {''};
%
% %% Replace non-numeric cells with NaN
% R = cellfun(@(x) ~isnumeric(x) && ~islogical(x),raw); % Find non-numeric cells
% raw(R) = {NaN}; % Replace non-numeric cells
%
% %% Create output variable
% data = reshape([raw{:}],size(raw));
%
% %% Allocate imported array to column variable names
% Empirical_Agg_Loss_Auto = data(:,1);
% %Clean_Mean_Severity_Month = data(:,2);
%
% %% Clear temporary variables
% clearvars data raw cellVectors R;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%END OF IMPORTING Auto Historical VaR

Sample_length_Auto = 100;
%tic;
Bootstrap_matrix_Auto = zeros(Bootstrap_n*100, Quantile_length);
%Distribution
for j=1:(Bootstrap_n*100)
      New_Loss = datasample(Empirical_Agg_Loss_Auto,Sample_length_Auto,'Replace',true);
      Bootstrap_matrix_Auto(j,1) = prctile(New_Loss, 0);
      Bootstrap_matrix_Auto(j,2) = prctile(New_Loss, 25);
      Bootstrap_matrix_Auto(j,3) = prctile(New_Loss, 50);
      Bootstrap_matrix_Auto(j,4) = prctile(New_Loss, 90);
      Bootstrap_matrix_Auto(j,5) = prctile(New_Loss, 95);
      Bootstrap_matrix_Auto(j,6) =  prctile(New_Loss, 98);
      Bootstrap_matrix_Auto(j,7) =  prctile(New_Loss, 99);
      if mod(j, Bootstrap_n*10) == 0
         disp(j);
      end
end
```

```
Final_Quantiles_Matrix_Auto = zeros(2,7);
%quantile_interest = 95;

for i=1:7
    Final_Quantiles_Matrix_Auto(1,i) = prctile(Bootstrap_matrix_Auto(:,i), 100-quantile_interest);
    Final_Quantiles_Matrix_Auto(2,i) = prctile(Bootstrap_matrix_Auto(:,i), quantile_interest);
end


Export_xls = Final_Quantiles_Matrix_Auto';
disp('I am on the Autos');
%%%%%%%%%%%%%%%%%%%%%%%%%%Export Results to Excel
    %%%%%%%%%%%%%%%Export Data to XLSX
     full_path = 'C:\Users\sabyguharay\Documents\GMU\Dissertation Final Defense\Print Results Thesis Level
Good.xlsx';
     Utility_header = {'Lower','Upper'};
     xlRange = 'M40';
     xlswrite(full_path, Utility_header, 'Real-World Data New',xlRange);
     xlRange_2 = 'M41';
     xlswrite(full_path, Export_xls, 'Real-World Data New', xlRange_2);
  %%%%%%%%%%%%%%%End of Export Data to XLSX
% t=toc;
% disp(t);      %30 Seconds
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%END Auto Analysis
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%Begin Hurricane Analysis
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%%%%Begin Hurricane Analysis
% [~, ~, raw] = xlsread('C:\Users\sabyguharay\Documents\GMU\Fourth Year\Real Data\US
Hurricanes.xlsx','Input_Data_MATLAB','E2:E107');
% raw(cellfun(@(x) ~isempty(x) && isnumeric(x) && isnan(x),raw)) = {''};
%
% %% Replace non-numeric cells with NaN
% R = cellfun(@(x) ~isnumeric(x) && ~islogical(x),raw); % Find non-numeric cells
% raw(R) = {NaN}; % Replace non-numeric cells
%
% %% Create output variable
% data = reshape([raw{:}],size(raw));
%
% %% Allocate imported array to column variable names
% Empirical_Agg_Loss_Hurricane = data(:,1);
% %Clean_Mean_Severity_Month = data(:,2);
%
% %% Clear temporary variables
% clearvars data raw cellVectors R;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%END OF IMPORTING Hurricane Historical VaR

tic;
Sample_length_Hurricane = 105;
```

```
Bootstrap_matrix_Hurricane = zeros(Bootstrap_n*100, Quantile_length);
%Distribution
for j=1:(Bootstrap_n*100)
        New_Loss = datasample(Empirical_Agg_Loss_Hurricane,Sample_length_Hurricane,'Replace',true);
        Bootstrap_matrix_Hurricane(j,1) = prctile(New_Loss, 0);
        Bootstrap_matrix_Hurricane(j,2) = prctile(New_Loss, 25);
        Bootstrap_matrix_Hurricane(j,3) = prctile(New_Loss, 50);
        Bootstrap_matrix_Hurricane(j,4) = prctile(New_Loss, 90);
        Bootstrap_matrix_Hurricane(j,5) = prctile(New_Loss, 95);
        Bootstrap_matrix_Hurricane(j,6) =  prctile(New_Loss, 98);
        Bootstrap_matrix_Hurricane(j,7) =  prctile(New_Loss, 99);
        if mod(j, Bootstrap_n*10) == 0
            disp(j);
        end
end



Final_Quantiles_Matrix_Hurricane = zeros(2,7);
%quantile_interest = 95;


for i=1:7
    Final_Quantiles_Matrix_Hurricane(1,i) = prctile(Bootstrap_matrix_Hurricane(:,i), 100-quantile_interest);
    Final_Quantiles_Matrix_Hurricane(2,i) = prctile(Bootstrap_matrix_Hurricane(:,i), quantile_interest);
end


Export_xls = Final_Quantiles_Matrix_Hurricane';
disp('I am on the Last One: Hurricanes');



%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%Begin Analysis
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%%%Begin Analysis
% [~, ~, raw] = xlsread('C:\Users\sabyguharay\Documents\GMU\Fourth Year\Real
Data\US.xlsx','Input_Data_MATLAB','E2:E107');
% raw(cellfun(@(x) ~isempty(x) && isnumeric(x) && isnan(x),raw)) = {''};
%
% %% Replace non-numeric cells with NaN
% R = cellfun(@(x) ~isnumeric(x) && ~islogical(x),raw); % Find non-numeric cells
% raw(R) = {NaN}; % Replace non-numeric cells
%
% %% Create output variable
% data = reshape([raw{:}],size(raw));
%
% %% Allocate imported array to column variable names
% Empirical_Agg_Loss_Hurricane = data(:,1);
% %Clean_Mean_Severity_Month = data(:,2);
```

REFERENCES

[1] Kahneman, Daniel, and Amos Tversky. "Prospect theory: An analysis of decision under risk." *Econometrica: Journal of the Econometric Society* (1979): 263-291.
[2] Taleb, Nassim Nicholas. *The black swan: The impact of the highly improbable fragility*. Vol. 2. Random House, 2010.
[3] Beck, Ulrich. *World at risk*. Polity, 2009.
[4] Posner, Richard A. *Catastrophe: risk and response*. Oxford University Press, 2004.
[5] Mercier, Hugo, and Dan Sperber. "Why do humans reason? Arguments for an argumentative theory." *Behavioral and brain sciences* 34.02 (2011): 57-74.
[6] Taylor, John B., and John C. Williams. *A black swan in the money market*. No. w13943. National Bureau of Economic Research, 2008.
[7] Gabaix, Xavier. *Power laws in economics and finance*. No. w14299. National Bureau of Economic Research, 2008.
[8] Borwein, Jonathan M., David H. Bailey, and David Bailey. *Mathematics by experiment: Plausible reasoning in the 21st century*. Natick, MA: AK Peters, 2004.
[9] Resnick, Sidney I. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.
[10] Guharay, Sabyasachi and KC Chang. "An Application of Data Fusion Techniques in Quantitative Operational Risk Management." In *Information Fusion (Fusion), 2015 18th International Conference on*, pp. 1914-1921, IEEE Press, 2015.
[11] Accords, Basel. "Operational Risk - Supervisory Guidelines for the Advanced Measurement Approaches" (2011), http://www.bis.org/publ/bcbs196.pdf.
[12] "Basel II: Revised international capital framework", Basel Committee on Banking Supervision document, www.bis.org/publ/bcbsca.htm.
[13] Jöhnemark, Alexander. "Modeling Operational Risk." (2012), Master of Science Thesis.
[14] Advisory, Op Risk, and Towers Perrin. "A New Approach for Managing Operational Risk: Addressing the Issues Underlying the 2008 Global Financial Crisis." *Sponsored by Joint Risk Management, Section Society of Actuaries, Canadian Institute of Actuaries & Casualty Actuarial Society* (2010).
[15] Staudt, Andy. "Tail risk, systemic risk and copulas." *Casualty Actuarial Society E-Forum, Fall 2010-Volume 2*. 2010.
[16] A. Chernobai, and S. Rachev, "Applying robust methods to operational risk modeling" *Journal of Operational Risk*, vol. 1, no. 1, pp. 27-41, 2006.
[17] Frachot, Antoine, Olivier Moudoulaud, and Thierry Roncalli. "Loss distribution approach in practice." *The Basel Handbook: A Guide for Financial Practitioners* (2004): 527-554.

[18] G. Mignola and R. Ugoccioni, "Sources of Uncertainty in modeling Operational Risk Losses", *The Journal of Operational Risk*, vol. 1, no. 2, pp. 35, 2006.
[19] Masdemont, Josep J., and Luis Ortiz-Gracia. "Haar wavelets-based approach for quantifying credit portfolio losses." *Quantitative Finance* 14, no. 9 (2014): 1587-1595.
[20] Casella, George, and Roger L. Berger. *Statistical inference*. Vol. 2. Pacific Grove, CA: Duxbury, 2002.
[21] De Fontnouvelle, Patrick, De Jesus-Rueff, John S. Jordan, and Eric S. Rosengren. "Using loss data to quantify operational risk." *Available at SSRN 395083* (2003).
[22] Rachev, Svetlozar T., Anna Chernobai, and Christian Menn. "Empirical examination of operational loss distributions." In *Perspectives on Operations Research*, pp. 379-401. DUV, 2006.
[23] Dutta, Kabir, and Jason Perry. "A tale of tails: an empirical analysis of loss distribution models for estimating operational risk capital." (2006).
[24] B. Ergashev, "Estimating the lognormal-gamma model of operational risk using the Markov Chain Monte Carlo method", *The Journal of Operational Risk*, vol. 4, no. 1, pp. 35, 2009.
[25] Guillen, Montserrat, Faustino Prieto, and José María Sarabia. "Modeling losses and locating the tail with the Pareto Positive Stable distribution." Insurance: Mathematics and Economics 49.3 (2011): 454-461.
[26] De Fontnouvelle, Patrick, Eric Rosengren, and John Jordan. "Implications of alternative operational risk modeling techniques." *The Risks of Financial Institutions*. University of Chicago Press, 2007. 475-512.
[27] Dahen, Hela, and Georges Dionne. "Scaling models for the severity and frequency of external operational loss data." *Journal of Banking & Finance* 34.7 (2010): 1484-1496.
[28] Gomes, Erika, and Henryk Gzyl. "Disentangling frequency models." *The Journal of Operational Risk* 9.2 (2014): 3.
[29] Badescu, Andrei L., Gong Lan, X. Sheldon Lin, and Dameng Tang. "Modeling correlated frequencies with application in operational risk management." *The Journal of Operational Risk* 10.1 (2015): 1.
[30] Panjer, Harry H. "Recursive evaluation of a family of compound distributions." *Astin Bulletin* 12.01 (1981): 22-26.
[31] Böcker, Klaus, and Claudia Kluppelberg. "Operational VaR: a closed-form approximation." *RISK-LONDON-RISK MAGAZINE LIMITED*- 18.12 (2005): 90.
[32] Böcker, Klaus, and Jacob Sprittulla. "Operational VAR: meaningful means." *Risk Magazine* 12 (2006): 96-98.
[33] Jin, Tao, and Jiandong Ren. "Recursions and fast Fourier transforms for certain bivariate compound distributions." *Journal of Operational Risk* 4 (2010): 19.
[34] Opdyke John D., and Alexander Cavallo. "Estimating operational risk capital: the challenges of truncation, the hazards of maximum likelihood estimation, and the promise of robust statistics." *The Journal of Operational Risk* 7.3 (2012): 3.
[35] Opdyke, J. D. "Estimating Operational Risk Capital with Greater Accuracy, Precision, and Robustness." *Precision, and Robustness (October 1, 2013)* (2013).

[36] Ruppert, David. *Statistics and data analysis for financial engineering*. New York, NY: Springer, 2011.
[37] Frachot, Antoine, Thierry Roncalli, and Eric Salomon. "The correlation problem in operational risk." *Operational Risk Risk's Newsletter* (2004).
[38] Di Clemente, Annalisa, and Claudio Romano. "A copula-extreme value theory approach for modeling operational risk." *Operational Risk Modeling and Analysis* (2004): 189-208.
[39] Böcker, Klaus, and Claudia Klüppelberg. "Modeling and measuring multivariate operational risk with Lévy copulas." *Journal of Operational Risk* 3.2 (2008): 3-27.
[40] Fantazzini, Dean, Luciana Dalla Valle, and Paolo Giudici. "Copulae and operational risks." *International Journal of Risk Assessment and Management* 9.3 (2008): 238-257.
[41] Abbate, Donato, Elise Gourier, and Walter Farkas. "Operational risk quantification using extreme value theory and copulas: from theory to practice." *Journal of Operational Risk* 3 (2009).
[42] Dalla Valle, Luciana. "Bayesian copulae distributions, with application to operational risk management." *Methodology and Computing in Applied Probability* 11.1 (2009): 95-115.
[43] Böcker, Klaus, and Claudia Klüppelberg. "Multivariate models for operational risk." *Quantitative Finance* 10.8 (2010): 855-869.
[44] Brechmann, Eike Christian, Claudia Czado, and Sandra Paterlini. "Modeling dependence of operational loss frequencies." *Available at SSRN 2345342* (2013).
[45] Darbellay, G. A. *An adaptive histogram estimator for the mutual information (Extended Version)*. No. 1936. Research Report, 1998.
[46] Moddemeijer, Rudy. "A statistic to estimate the variance of the histogram-based mutual information estimator based on dependent pairs of observations." *Signal Processing* 75.1 (1999): 51-63.
[47] Dionisio, Andreia, Rui Menezes, and Diana A. Mendes. "Mutual information: a measure of dependency for nonlinear time series." *Physica A: Statistical Mechanics and its Applications* 344.1 (2004): 326-329.
[48] Dionísio, Andreia, Rui Menezes, and Diana A. Mendes. "Entropy-based independence test." *Nonlinear Dynamics* 44.1-4 (2006): 351-357.
[49] Kraskov, Alexander, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information." *Physical review E* 69.6 (2004): 066138.
[50] Li, Jianping, Xiaoqian Zhu, Yongjia Xie, Jianming Chen, Lijun Gao, Jichuang Feng, and Wujiang Shi. "The mutual-information-based variance-covariance approach: an application to operational risk aggregation in Chinese banking." *The Journal of Operational Risk* 9, no. 3 (2014): 3.
[51] Aue, Falko, and Michael Kalkbrener. "LDA at work: Deutsche Bank's approach to quantifying operational risk." *Journal of Operational Risk* 1.4 (2006): 49-93.
[52] Chapelle, Ariane, Yves Crama, Georges Hübner, and Jean-Philippe Peters. "Practical methods for measuring and managing operational risk in the financial sector: A clinical study." *Journal of Banking & Finance* 32, no. 6 (2008): 1049-1061.

[53] Cope, Eric, and Gianluca Antonini. "Observed correlations and dependencies among operational losses in the ORX consortium database." *Journal of Operational Risk* 3.4 (2008): 47-74.
[54] Cope, Eric W., Giulio Mignola, Gianluca Antonini, and Roberto Ugoccioni. "Challenges and pitfalls in measuring operational risk from loss data." *Journal of Operational Risk* 4, no. 4 (2009): 3-27.
[55] Colombo, Andrea and Desando Stefano. "Developing and Implementing Scenario Analysis Models to Measure Operational Risk at Intesa San Paolo". *The MathWorks News & Notes*, 91606v00. Available at http://www.mathworks.com/tagteam/52614_91606v00_intesa_upd.pdf.
[56] Liu, Hsiang-Hsi and Mauricio Cortes. "An Assessment of the efficiency of operational risk management in Taiwan's banking industry: an application of the stochastic frontier approach". *Journal of Operational Risk* 10.1(2014):127-156.
[57] Meel, A., L. M. O'Neill, J. H. Levin, Warren D. Seider, U. Oktem, and N. Keren. "Operational risk assessment of chemical industries by exploiting accident databases." *Journal of Loss Prevention in the Process Industries* 20, no. 2 (2007): 113-127.
[58] Panjer, Harry H. *Operational risk: modeling analytics*. Vol. 620. John Wiley & Sons, 2006.
[59] Grammig, Joachim and Kai-Oliver Maurer. "Non-monotonic hazard functions and the autoregressive conditional duration model." *Econometrics Journal*, Vol. 3, 2000, pp. 16–38.
[60] Mehta, Neelesh B. and Andreas F. Molisch. "Approximating a sum of random variables with a Lognormal." *Wireless Communications, IEEE Transactions on* 6.7 (2007): 2690-2699.
[61] Martinez, Wendy L., and Angel R. Martinez. Computational statistics handbook with MATLAB. Vol. 22. CRC press, 2007.
[62] MacKay, David JC. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
[63] Mojena, Richard. "Hierarchical grouping methods and stopping rules: An evaluation." The Computer Journal 20.4 (1977): 359-363.
[64] Kaufman, Leonard, and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.
[65] Tibshirani, Robert, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63.2 (2001): 411-423.
[66] Caliński, Tadeusz, and Jerzy Harabasz. "A dendrite method for cluster analysis." *Communications in Statistics-theory and Methods* 3.1 (1974): 1-27.
[67] Guharay, Sabyasachi, KC Chang and J Xu. "Robust Estimation of Value-at-Risk Through Correlated Frequency and Severity Model." In *Information Fusion (Fusion), 2016 19th International Conference on*, pp. 995-1002. IEEE Press, 2016.
[68] Nelsen, Roger B. *An introduction to copulas*. Springer Science & Business Media, 2007.
[69] Sklar, M. Fonctions de répartition à n dimensions et leurs marges. Université Paris 8, 1959.

[70] Kojadinovic, Ivan, and Jun Yan. "Modeling multivariate distributions with continuous margins using the copula R package." *Journal of Statistical Software* 34.9 (2010): 1-20.
[71] Panchenko, Valentyn. "Goodness-of-fit test for copulas." *Physica A: Statistical Mechanics and its Applications* 355.1 (2005): 176-182.
[72] Genest, Christian, Bruno Rémillard, and David Beaudoin. "Goodness-of-fit tests for copulas: A review and a power study." *Insurance: Mathematics and economics* 44.2 (2009): 199-213.
[73] Chen, Xiaohong, and Yanqin Fan. "Estimation of copula-based semiparametric time series models." *Journal of Econometrics* 130.2 (2006): 307-335.
[74] Genest, Christian, and Bruno Rémillard. "Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models." *Annales de l'IHP Probabilités et statistiques*. Vol. 44. No. 6. 2008.
[75] Kole, Erik, Kees Koedijk, and Marno Verbeek. "Selecting copulas for risk management." *Journal of Banking & Finance* 31.8 (2007): 2405-2423.
[76] Rémillard, Bruno, and Olivier Scaillet. "Testing for equality between two copulas." *Journal of Multivariate Analysis* 100.3 (2009): 377-386.
[77] Dobrić, Jadran, and Friedrich Schmid. "A goodness of fit test for copulas based on Rosenblatt's transformation." *Computational Statistics & Data Analysis* 51.9 (2007): 4633-4642.
[78] Kolev, Nikolai, Ulisses dos Anjos, and Beatriz Vaz de M. Mendes. "Copulas: a review and recent developments." *Stochastic Models* 22.4 (2006): 617-660.
[79] Kojadinovic, Ivan, Jun Yan, and Mark Holmes. "Fast large-sample goodness-of-fit tests for copulas." *Statistica Sinica* (2011): 841-871.
[80] Dobrić, Jadran, and Friedrich Schmid. "Testing goodness of fit for parametric families of copulas—application to financial data." *Communications in Statistics—Simulation and Computation*® 34.4 (2005): 1053-1068.
[81] Scaillet, Olivier. "Kernel-based goodness-of-fit tests for copulas with fixed smoothing parameters." *Journal of Multivariate Analysis* 98.3 (2007): 533-543.
[82] Savu, Cornelia, and Mark Trede. "Goodness-of-fit tests for parametric families of Archimedean copulas." *Quantitative Finance* 8.2 (2008): 109-116.
[83] González-Manteiga, Wenceslao, and Rosa M. Crujeiras. "An updated review of Goodness-of-Fit tests for regression models." *Test* 22.3 (2013): 361-411.
[84] Genest, Christian, and Bruno Rémillard. "Discussion of "Copulas: tales and facts", by Thomas Mikosch." *Extremes* 9.1 (2006): 27-36.
[85] Kojadinovic, Ivan, and Jun Yan. "A goodness-of-fit test for multivariate multiparameter copulas based on multiplier central limit theorems." *Statistics and Computing* 21.1 (2011): 17-30.
[86] Brechmann, Eike Christain, and Claudia Czado. "Risk management with high-dimensional vine copulas: An analysis of the Euro Stoxx 50." *Statistics & Risk Modeling* 30.4 (2013): 307-342.
[87] Genest, Christian, Ivan Kojadinovic, Johanna Nešlehová, and Jun Yan. "A goodness-of-fit test for bivariate extreme-value copulas." *Bernoulli* 17.1 (2011): 253-275.

[88] Berg, Daniel. "Copula goodness-of-fit testing: an overview and power comparison." *The European Journal of Finance* 15.7-8 (2009): 675-701.
[89] Genest, Christian, Jean-François Quessy, and Bruno Rémillard. "Goodness-of-fit procedures for copula models based on the probability integral transformation." *Scandinavian Journal of Statistics* 33.2 (2006): 337-366.
[90] Tewari, Ashutosh, Michael J. Giering, and Arvind Raghunathan. "Parametric characterization of multimodal distributions with non-Gaussian modes." *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, 2011.
[91] Bilgrau, Anders Ellern, et al. "GMCM: Unsupervised clustering and meta-analysis using gaussian mixture copula models." *Journal of Statistical Software* 70.2 (2016): 1-23.
[92] Data available from http://www.finance.yahoo.com.
[93] Data available at National Response Center, http://www.nrc.uscg.mil/.
[94] Charpentier, Arthur, ed. *Computational Actuarial Science with R*. CRC Press, 2014.
[95] Vidyashankar, Anand N., and Jie Xu. "Stochastic optimization using hellinger distance." *Proceedings of the 2015 Winter Simulation Conference*, pp. 3702-3713, IEEE Press, 2015.
[96] Cheng, An-lin, and Anand N. Vidyashankar. "Minimum Hellinger distance estimation for randomized play the winner design." *Journal of statistical planning and inference* 136.6 (2006): 1875-1910.
[97] Hooker, Giles, and Anand N. Vidyashankar. "Bayesian model robustness via disparities." *Test* 23, no. 3 (2014): 556-584.
[98] Sriram, T. N., and A. N. Vidyashankar. "Minimum Hellinger distance estimation for supercritical Galton–Watson processes." *Statistics & probability letters* 50, no. 4 (2000): 331-342.
[99] Xie, Wei, Barry L. Nelson, and Russell R. Barton. "A Bayesian framework for quantifying uncertainty in stochastic simulation." *Operations Research* 62, no. 6 (2014): 1439-1452.
[100] Cheng, An-lin, and Anand N. Vidyashankar. "Minimum Hellinger distance estimation for randomized play the winner design." *Journal of statistical planning and inference* 136, no. 6 (2006): 1875-1910.
[101] Basu, Ayanendranath, Hiroyuki Shioya, and Chanseok Park. *Statistical inference: the minimum distance approach*. CRC Press, 2011.
[102] Prause, Annabel, Ansgar Steland, and Mohammed Abujarad. "Minimum Hellinger distance estimation for bivariate samples and time series with applications to nonlinear regression and copula-based models." *Metrika* 79.4 (2016): 425-455.
[103] Hooker, Giles. "Consistency, efficiency and robustness of conditional disparity methods." *Bernoulli* 22.2 (2016): 857-900.
[104] Guharay, Sabyasachi and KC Chang. "Application of Bayesian Simulation Framework in Quantitatively Measuring Presence of Competition in Living Species." In *Proceedings of the 2015 Winter Simulation Conference (WSC)*, pp. 4033-4044, IEEE Press, 2015.

[105] Ishitani, Kensuke, and Kenichi Sato. "An analytical evaluation method of the operational risk using fast wavelet expansion techniques." *Asia-Pacific Financial Markets* 20.3 (2013): 283-309.
[106] Pacelli, Vincenzo, and Michele Azzollini. "An artificial neural network approach for credit risk management." *Journal of Intelligent Learning Systems and Applications* 3.2 (2011): 103.
[107] Cifter, Atilla. "Value-at-risk estimation with wavelet-based extreme value theory: Evidence from emerging markets." *Physica A: Statistical Mechanics and its Applications* 390.12 (2011): 2356-2367.
[108] Kozaki, M., and A-H. Sato. "Application of the Beck model to stock markets: Value-at-Risk and portfolio risk assessment." *Physica A: Statistical Mechanics and its Applications* 387, no. 5 (2008): 1225-1246.
[109] He, Kaijian, Lijun Wang, Yingchao Zou, and Kin Keung Lai. "Value at risk estimation with entropy-based wavelet analysis in exchange markets." *Physica A: Statistical Mechanics and its Applications* 408 (2014): 62-71.
[110] He, Kaijian, Kin Keung Lai, and Guocheng Xiang. "Portfolio value at risk estimate for crude oil markets: a multivariate wavelet denoising approach." *Energies* 5.4 (2012): 1018-1043.
[111] Yang, Jianhui, and Peng Lin. "Dynamic risk measurement of futures based on wavelet theory." *Computational Intelligence and Security (CIS), 2011 Seventh International Conference on*. IEEE, 2011.
[112] Berger, Theo. "Forecasting Based on Decomposed Financial Return Series: A Wavelet Analysis." *Journal of Forecasting* (2015).
[113] He, Kaijian, Rui Zha, Yanhui Chen, and Kin Keung Lai. "Forecasting Energy Value at Risk Using Multiscale Dependence Based Methodology." *Entropy* 18, no. 5 (2016): 170.
[114] Wang, Hongqian, Kaijian He, and Kin Keung Lai. "Multivariate EMD-based Portfolio Value at Risk Estimate for Electricity Markets." In *Business Intelligence and Financial Engineering (BIFE), 2013 Sixth International Conference on*, pp. 211-215. IEEE, 2013.
[115] Wang, Hongqian, Kaijian He, and Yingchao Zou. "EMD Based Value at Risk Estimate Algorithm for Electricity Markets." In *Computational Sciences and Optimization (CSO), 2014 Seventh International Joint Conference on*, pp. 445-449. IEEE, 2014.

BIOGRAPHY

Sabyasachi Guharay received his Bachelor of Science in Engineering in Operations Research & Financial Engineering from Princeton University in 2003. He received his M.A. in Statistics from Wharton School, University of Pennsylvania in 2006. He has worked on quantification of Operational Risk as a risk management consultant at Towers Watson consulting from 2007-2009 and Stamford Risk Analytics in 2009. From late 2009 till present, he works as a Senior Operations Researcher for the U.S. Department of Treasury where he is directly responsible for quantification of US Tax Return Identity Theft. He published three journal papers in the past, and currently he is in the process of submitting two journal papers based on this dissertation work. In addition, he published three papers in peer-reviewed conference proceedings including 2015 Winter Simulation Conference and the 18th & 19th International Conference on Information Fusion.