ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

PartnerRe

Modeling Dependence in Catastrophe Risk Portfolio and Optimizing Capital Loading

A thesis submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY (ETH)
ZÜRICH

for the degree of
MASTER OF SCIENCE

presented by
NAN ZHAO

born on the 27 Febuary 1984
citizen of Bellevue - Switzerland

accepted on the recommendation of

PROF. DR. DIDIER SORNETTE

CHAIR OF ENTREPRENEURIAL RISKS

DEPARTEMENT OF MANAGEMENT, TECHNOLOGY AND
ECONOMICS

November 21, 2010

To My Parents,

for their support, love and care.

Acknowledgments

A number of people contributed directly or indirectly to the accomplishment of this thesis. Without their support, suggestions and continuous encouragement, this work could have not been possible.

First of all, I would like to thank my supervisor from ETH Zurich, Prof. Dr. Didier Sornette for his guidance and Dr. Gautier de Montmollin and Dr. Erik Ruettener from PartnerRe for giving me the opportunity to write this thesis in their Natural Catastrophe Business Unit, where I could learn from the top specialists in various areas. I would also like to thank the whole team at PartnerRe for their inspiring working environment which made the time spent within PartnerRe an unforgettable experience.

During my research, I have also received the help from many former ETH graduates who all volunteered to share their expertise in this field. I sincerely thank them for the valuable discussions. I owe special thanks to Marjan Beheshty and Davide Canestraro for their previous work and enlightenment.

Finally last but not least, I would like to thank my friends and my parents for their infinite support. I am forever indebted to them for their understanding, endless patience and encouragement.

Abstract

The disasters of hurricane Andrew (1992) and the Northridge earthquake (1994) will remain in the memories of many insurers or reinsurers. Owing to a total damage worth 45 billion in 1997 US dollars [Office, 2002, Froot, 1997], hurricane Andrew alone led to the bankruptcy of eleven insurance companies [Auffret, 2003], which all were holding highly concentrated risk positions. For those companies that survived, these tragedies made them reconsider their risk positions and raised their awareness on the importance of reliable risk management processes.

Since then insurers and reinsurers have endeavored to provide more accurate loss predictions of natural hazards for society, but also worked on a better diversification of their risk portfolio. Despite numerous measures undertaken, catastrophic losses have nevertheless been growing on a year-to-year basis, to the extent that it doubles every fourteen years. Hurricane Katrina in 2005, for example, led to one of the biggest natural catastrophes in the history of the USA. It caused a total economic damage estimated at 150 billion US dollars [Grossi and Kunreuther, 2006], leaving behind the destruction to approximately 80% of the industry infrastructure and private property in New Orleans.

With the increasing occurence of extreme natural hazards and its exploding costs, catastrophe risk has entered a new era. Risk management within a reinsurance company become increasingly a field of interdisciplinary work, requiring the expertise of highly skilled professionals in many core competencies, such as natural science in windstorms, probability mathematics, business management and the most up-to-date international and Swiss solvency regulations. This interdisciplinary thesis, following ETHZ MTEC department's objective to "develop the holistic thinking in order to plan and implement interdisciplinary projects with professional experts", will use this aforementioned approach to bridge these traditional core competencies into the analysis of the risk of a reinsurance portfolio, applying the most recent research from the academic world to solve concrete problems in the business world.

In the framework of this thesis, the relevance of this study based on the needs

of various stakeholders will be firstly analyzed, then the research moves to the general foundations, by defining the concept of risk and coherent risk measures, upon which the thesis is built. This being done, the data used for this thesis, originating from catastrophe models are thoroughly illustrated and explained before heading to the core part of assessing the dependencies within a portfolio or risks. In addition, various tools to investigate their underlying dependencies are provided, then using one key property of copulas, that they remain invariant under monotonous transformations to model their dependence structure.

Among the different models simulated, it appears that a Gaussian copula using the empirical distribution of the marginals provides the best goodness-to-fit to describe the underlying dependence structure of windstorm catastrophe losses in the North-Atlantic region. Furthermore, only very little tail dependence has been observed in this data. This result is quite surprising, as it was expected that extreme catastrophic events impact in several regions at a time. A discussion and the reasoning why this may occur are addressed in the last chapter.

During the research, the ties within the dataset have caused quite some troubles. For the reason that ties are the indication that the marginal distributions are non-continuous; this leads to non-unique copulas (Sklar's theorem). Furthermore most theorems in the field of copulas assume the hypothesis of continuous marginals, which then can not apply for the case of non-continuous marginals. To remedy this situation, a solution to split the dataset in four regions on the copula unitary square and its extension to a three-dimensional cube has been attempted. In the first case, this proposed solution offers a good alternative to solve the problem with ties, but it has been discovered that the solution is difficultly scalable beyond two dimensions. Other solutions and further research paths have then been proposed.

Table of Contents

Chapter 1

Introduction: Managing the Unpredictable

1.1 Natural catastrophe risk management

1.1.1 Relevance and motivation

On 24th August 2005, a hurricane of unprecedented strength hit the coast of the Gulf of Mexico. From Florida to Texas, it led to one of the most devastating natural catastrophes in the history of the United States of America. Winds of up to 240 km/h followed by heavy rainfall, left behind the destruction of approximately 80% of the industry infrastructure and private property. This also came with severe economic consequences for the population and society in New Orleans; the total economic damage was estimated at 150 Billion USD for the whole region, with cascading effects from homeowners to insurers and reinsurers.

Reinsurance companies (reinsurers) have played long since a decisive role in coping financially with extraordinary large loss events. Reinsurance, by definition, is a contract purchased by an insurance company (insurer), to transfer portions of its risk portfolio [1] to a reinsurer. By doing so, insurers benefit from a reduction of its risk exposures to extraordinary large events, which can in unfortunate cases result in the payment of very large claims and the insolvency of the company in the worst case. Furthermore this transfer of risk gives the additional advantage to insurers to provide less volatile, therefore more predictable financial results, which is relevant in the eyes of its shareholders.

From the reinsurers' perspective, they are awarded with a premium for every risk transferred to them. Hence for the success of the reinsurance business, being able to anticipate the losses due to natural catastrophe, such as earthquakes, cyclones,

---

[1]The portfolio of risk is a grouping or aggregation of individual insurance policy contracts.

floodings combined with portfolio management of these aggregated low probability, but large loss risks are central activities and a core part in its risk management strategy.

The aggregation of this set of risks is not a simple operation, like summation or addition, but depends very much on the dependence between the risks. For instance, in the case of fire hazards, each event happening, which is independent of one another, is localized. The risk of the aggregation of identically and independently distributed risk is quantified by the product of their marginal distributions. $\Pr\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \Pr\left(A_i\right)$. This is a trivial example of risk aggregation, but not all risks are independently distributed.

There are examples of hurricanes, in which each event may also have an impact on landfalls, power failure or social disorder from housebreaking to public riots. This may cast a larger footprint; hence, one has to take into consideration all of these additional dependent parameters for computing the risk of the hazard. The occurrence of such an event is likely to affect more substantially a higher number of assets covered by a portfolio of risks. Thus, a key element in computing the risk of this portfolio is the estimation of the dependencies of the risk, which plays an essential role in determining the portfolio loss distribution and consequently the estimation of capital needs to support this. Modeling this distribution through describing the dependencies between the risks will be one of the central tasks in this thesis.

1.1.2 Scope of this thesis

For any organization, the need for a robust and fully integrated risk management system are fundamentals of a competitive and sustainable strategy of a company. This lies in the framework of Enterprise Risk Management (ERM) techniques, defined by the Casualty Actuarial Society "as the discipline by which an organization in any industry assesses, controls, exploits, finances, and monitors risks from all sources for the purpose of increasing the organization's short- and long-term value to its stakeholders". In order to continuously deliver its promises, the top management team needs to have a good understanding of not only the company's total risks, but also the interactions within its risk portfolio.

In the case of reinsurance companies, natural catastrophe models are used to provide the technical inputs for a better support to risk management decisions. Only then the quantification of the overall risk for all its treaties has been correctly assessed and can begin a wider financial planning process, such as capital allocation analysis or better return on capital strategies. Currently, insurers and reinsurers are the stakeholders with the most widespread interest and integrated use of catastrophe models for pricing insurance premiums and financial sustainability reasons. The

capital markets also have an eager interest in catastrophe models in order to price catastrophe bonds more accurately.

The scope of this thesis is mainly limited to the catastrophe modeling and the capital allocation parts. The Figure 1.1 illustrates these major components.
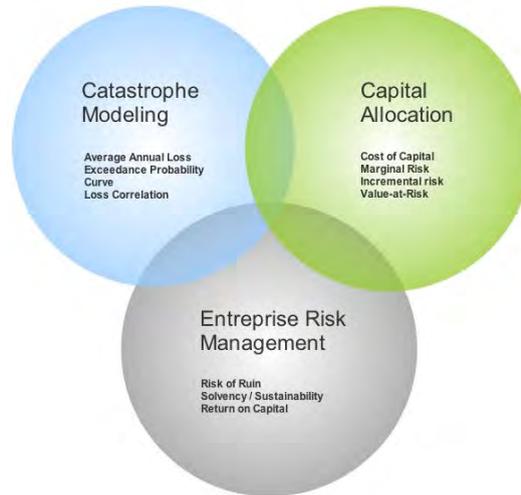
Figure 1.1: Role of catastrophe modeling in an insurance company's financial management.

Source: Grossi and Kunreuther, 2006 | Page 98.

1.1.3   Goal and thesis outline

The goal of this thesis is to assess, in the first step, the dependencies among elements within a reinsurance risk portfolio. These are then compared with the dependencies currently in use for simulating the aggregate risk in that region. In the second step, a copula model is proposed to fit the empirical data. And finally from the copula simulation and based on the European Solvency II and the Swiss Solvency Test standards, the adequate amount of capital for allocation is determined.

The structure of this thesis is as follows: Chapter 1 provides the general framework in which this thesis is built, its relevance and scope as well as the definitions of risk. Chapter 2 presents the origins of the data and explains the mechanisms of catastrophe risk models. We believe it is crucial to first have a good understanding how our data has been collected, before being able to construct a robust and further analysis. In Chapter 3 and 4, a recapitulation of all the theoretical foundations in dependence assessment and copula inference are illustrated with an overview of the most up-to date research and its corresponding literature and with the focus on the understanding of copulas, then complex mathematical details are summarized, applications and self-explanatory illustrations are shown to explain the theory. Therefore, this paper should be accessible to any user with general notions in

probability theory. After this theoretical overview, Chapter 5 provides the results of the implementation of the techniques covered. Then these results are discussed, alternative techniques are proposed before the most optimal copula model are selected to fit our dataset.

1.2 The virtues of diversification

Diversification in the context of this thesis refers to catastrophe risk diversification or the spreading of a variety of (re-)insurance contracts over a risk portfolio. Diversification is a vital tool for the sustainability of a reinsurance company, as it limits the probability of highly correlated risks happening at the same time. It is like when "putting all eggs in the same basket". If eggs were risks and the basket, a risk portfolio in case of an accident or a natural catastrophe, it would indeed be a risky scenario for insurers, which may face considerable losses by holding the majority of its signed contracts concentrating in a certain region. This can have very damaging effects on insurer's solvency, consequently it requires a higher amount of capital to cover this risk.

Historical events have shown the linkages between concentration of risks and default of insurers. Hurricane Andrew in 1992, for example, led to the insolvency of eleven insurers who all had highly concentrated risk profiles, either geographically, or in terms of asset mix [Forum, 2005]. Given the potential dangers of risk concentration, diversifying risk portfolios became a key element in the risk management of reinsurers. Intuitively, assuming X, Y two risk random variables and using the variance (Var) as a measure of risk, where it is known that the risk of the portfolio, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$. From this, it can be deduced that $\text{Var}(X, Y)$ is the smallest, when X and Y are independent giving $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. This is also known as diversification benefit. The higher the diversification, is the lower the capital protection is required against the risks of insolvency.

In a nutshell, diversifying strategies aim to minimize the total risk by exposing one's portfolio to different areas that would react differently to the same event and thereby mitigate the aggregate exposure from an individual source of risk. In this way, it allows reinsurances to insure positions which are too risky for insurances.

1.3 Risk measures and the new regulatory framework

1.3.1 Definition of risk

The concept of risk is understood intuitively by society as the possibility of some adverse consequences. It can appear to be easy to define and the Concise Oxford

English Dictionary defines risk as "hazard, a chance of bad consequences, loss or exposure to mischance". This broad definition may be satisfactory for a conversational usage. And for long, not all stakeholders in the risk profession (ie. economists, statisticians, insurance theorists etc.) have agreed on a common definition of risk, as a definition that is suitable for economists or statisticians may be worthless as an analytic tool for the insurance theorists [Vaughan and Vaughan, 1999]. A search for the definition of risk in the financial and insurance literature yield to many definitions, varying in the situational contexts and the specific applications. These definitions can sometimes be inconsistent, hence leading to an ambiguous usage of the word [Hubbard, 2009], and the construction based on it. Hence, a more explicit and precise definition is needed.

1.3.1.1 Risk, Uncertainty, Exposure, Peril, Hazard

There are several terminologies, which are often used in connection with the term risk and their subtle differences might not be clear at the first sight for the reader. As they are commonly used in the insurance vocabulary and later in this thesis, it seems appropriate to briefly clarify the meaning of risk in each terminology

Uncertainty refers to as a notion of indeterminate outcome which is characterized by doubt. In a case of uncertainty, there are always two possible outcomes and a positive one is possible. As nicely said in Vaughan and Vaughan [1999], Kaplan and Garrick [1981], the existence of risk - a condition or combination of circumstances in which there is a possibility of loss - creates uncertainty on the part of individuals when that risk is recognized. Some philosophers distinguish between "subjective" and "objective" risk [Holton, 2004] and uncertainty is seen with subjective risk, which is the person's perception of risk.

Exposure is a measure of the sensitivity of the value of a financial item (asset, liability or treaty) to changes in the relevant risk factor while risk is a measure of variability of the value of the item attributable to the risk factor.

Risk is also distinguished from *peril* and *hazards,* and it is not uncommon to see both words interchanged. Peril is a cause or source of a loss [Kaplan and Garrick, 1981], for example the peril of an earthquake, fire, is the loss accountable to the hazard. Hazard, is a condition that may create or increase the chance of a loss arising from a given peril. For example, natural hazards that increase the probability of loss from the perils of fire are the type of construction, the location of the property and the occupancy of the building.

For **insurance risk**, the subject of this thesis, it is defined conceptually as,

> *Risk is the likelihood of an event happening, and the severity of the negative consequences leading to an undesired outcome.*

It is noted that the definition is the combination of two parts, the possibility of a loss and the undesired outcome of a happening event.

1.3.2 Coherent measures of risk

Due to the conflicting opinions in the attempt of defining the concept of risk, Artzner et al. [1999] established instead a series of properties that a "good" definition of risk should have, so that it can meet the needs of all stakeholders. It is presented below in Definition 1.1 as "coherent measures of risk" and serves as a reference for how to assess risks.

Definition 1.1. (Coherent risk measures). A risk measure $\rho : \mathcal{M} \to \mathbb{R}$ on a convex cone $\mathcal{M}$, is called coherent if it satisfies the following four axioms. $\rho(L)$ can be interpreted as the amount of capital that would be needed to add to a position with a loss given by L.

Axiom 1.2. (Translation invariance). *For all $L \in \mathcal{M}$ and $l \in \mathbb{R}$, then $\rho(L+l) = \rho(L) + l$.*

Axiom 1.2 states simply that by adding or subtracting a value l to a position leading to the loss L, the capital requirements are altered exactly by that amount.

Axiom 1.3. (Subadditivity). *For all $L_1, L_2 \in M$, then $\rho(L_1 + L_2) \leq \rho(L_1) + \rho(L_2)$.*

The axiom of subadditivity states that with diversification benefits, the amount of capital needed to cover the accumulation of two risks is at the most the amount of the sum of capital of each of these risks. This happens in the case when both risk are fully dependent, in all other cases, the inequality holds.

Axiom 1.4. (Positive homogeneity). *For all $L \in \mathcal{M}$ and $\lambda \in \mathbb{N}^*$, then $\rho(\lambda L) = \rho(L + \ldots + L) \leq \lambda \rho(L)$.*

This axiom follows the two previous axioms.

Axiom 1.5. (Monotonicity). *For $L_1, L_2 \in M$ such that $L_1 \leq L_2$ almost surely we have $\rho(L_1) \leq \rho(L_2)$.*

This axiom is also obvious from an economic point of view.

Examples of applications using coherent measure of risk can be found in Malevergne and Sornette [2005] on page 10.

1.3.3 How to measure risk?

Just as the definition of risk is accepted, there should be also an agreement on the way in which risk should be measured. Intuitively, Markowitz suggested in the 50s, that "if the term *yield* were replaced by *expected yield* or *expected return*, and risk by *variance of return*, little change of the apparent meaning would result." This suggested that the variance of return might be a proxy for risk, but at the same time Markowitz distanced himself from this association. The reason is that despite this quotation giving a good understanding for the concept of risk, the application to a large extent limits to only normal distributed data.

In the following, this paper assesses the approaches which are the most commonly applied in practice, namely the notional-amount approach; factor sensitivity measures, risk measures based on loss distribution, or scenario-based risk measures and assesses them based on the definition of coherent measures of risk [Neslehova, 2009], before the best alternative is chosen for this thesis. The full text from these measured are taken can be found in McNeil et al. [2005].

1. The first is the oldest approach and simplest to apply. In the *notional-amount approach*, the risk of a portfolio is the sum of the values of the individual securities, weighted by a factor representing the riskiness of each. Suppose there is a portfolio consisting of d underlying risky positions with respective weights $w_1, \ldots, w_d$, the change in value of the portfolio over a given holding period can be written as $X = \sum_{i=1}^{d} w_i X_i$, where X_i denotes the change in value of the ith position. Measuring the risk of this portfolio essentially consists of determining its distribution function $F_X(x) = P(X \leq x)$. Although this method is still being used in the standardized approach of the Basel Committee, it nevertheless presents many flaws, namely the fact that it does not allow the illustration of diversification benefits. In a few steps, one can check that the subadditivity axiom does not hold, hence it is not a coherent measure of risk!

2. *Factor sensitivity measures* provide the change in the portfolio value for a given predetermined change in one of the underlying risk factors [McNeil et al., 2005]. While this method provides useful information about the robustness of the overall portfolio value, it can however not measure the riskiness of an individual position. Moreover factor-sensitivity measures can create problems in the aggregation of portfolios. For instance risk measures based on this concept can not be aggregated across markets, to create a picture of the overall riskiness of a financial institution.

3. As to *scenario-based risk measures,* they consider a number of possible future risk-factor changes (scenarios) when evaluating the risk of a portfolio. The

risk of the portfolio is then measured as the maximum loss of the portfolio under all scenarios. In practice, it is used at the Chicago Mercantile Exchange (CME) to compute the initial margin for a simple portfolio. However, if this method is intended for a global application, the main problem is the difficulty to determine an appropriate set of scenarios and weighting factors, as not all type of scenarios can be anticipated.

4. *Risk measures based on loss distributions* derive from statistical quantities describing the loss distribution of the portfolio. It is the foundation of most modern risk measures. Mathematically, the Russian mathematician A. N. Kolmogorov (1933) presents in his work an axiomatic definition of randomness and probability as well as the lingua franca for discourses on risk and uncertainty. In Kolmogorov's language a probabilistic model is described by a triplet (Ω, F, P). An element ω of Ω represents a realization of an experiment. P denotes the probability measure and A is an element of F, the set of all events. The statement "the probability of a risk that an event A occurs" is denoted as $P(A)$. To model a situation where the insurance holds today a risky position with an uncertain future value, a mathematician would now define it as X being a random variable on the probability space (Ω, F, P); most of the modeling of a risky position X concerns its distribution function $F_X(x) = P(X \leq x)$. Several risky positions would then be denoted by a random vector $(X_1, \ldots, X_d)$, also written in bold as $\mathbf{X}$. The advantage of loss distributions is that it provides an accurate picture of the risk in a portfolio consisting of single instrument to the overall position of a financial institution. Furthermore, the loss distribution reflects diversification effects and can be compared across portfolios. It is however not a perfect measure, mainly due to crude statistical models for the loss distributions [McNeil et al., 2005]. These authors argue that this can not be an argument against using loss distributions, they would rather call for improvement in the way in which loss distribution is estimated. Examples of risk measures based on loss distributions are such as the ***Value-at-Risk*** or the ***Expected Shortfall*** and are used in this thesis as risk measures which are presented below.

Definition 1.6. (Value-at-Risk or VaR). Given some confidence level $\alpha \in (0,1)$. The VaR of the portfolio at confidence level α is given by the smallest number l so that the probability that the loss L exceeds l is no larger than $(1 - \alpha)$. Formally,

$$VaR_\alpha(L) = \inf\{l \in \mathbb{R} : P(L > l) \leq 1 - \alpha\} = \inf\{l \in \mathbb{R} : F_L(l) \geq \alpha\} \qquad (1.1)$$

In other words, VaR is thus simply the αth quantile of the loss distribution.

This paper intends to draw the reader's attention to the fact that Value-at-Risk is not a coherent measure of risk, as it does not satisfy subadditivity axiom as stated in Definition 1.1. In order to overcome this problem, another measure of risk, known as expected shortfall, has been defined on the basis of this:

Definition 1.7. (Expected shortfall or ES). For a loss L with $E(|L|) < \infty$ and distribution function F_L the expected shortfall at confidence level $\alpha \in (0,1)$ is defined as $ES_\alpha(L) = \frac{1}{1-\alpha} \int_0^1 q_u(F_L) \, du$, where $q_u(F_L) = F_L^{-1}(u)$ is the quantile function of F_L.

Per definition, the expected shortfall can also be written in a relationship with the Value-at-Risk:

$$ES_\alpha(L) = \frac{1}{1-\alpha} \int_0^1 q_u(F_L) \, du = \frac{1}{1-\alpha} \int_0^1 VaR_u(L) \, du = E[L \mid L \geq VaR_\alpha(L)] \quad (1.2)$$

and contrary to the VaR, the expected shortfall takes into account the shape of the tail. Instead of fixing a particular confidence level α, for all quantile levels $u > \alpha$, the ES takes into consideration the average of the tail of the loss distribution and thus "look further into the tail". This expected shortfall is also known as expected tail loss, XLoss, conditional VaR, TailVar or CVaR.

1.3.4 Solvency II and Swiss Solvency Test

Solvency II is a fundamental review of the capital adequacy regime for the European insurance industry, scheduled to come into effect by late 2012. It aims at establishing a revised set of EU-wide capital requirements and risk management standards to replace and strengthen the current solvency requirements. In a nutshell, Solvency II in analogy with the Basel II framework in the banking industry, also consists of three pillars with the aim of an increasing protection of the policy holder. Without going too much into the specifics of this resolution, the reader is referred to the website of the European Commission for the Solvency II project for more information [Commission, 2010]. The first pillar determines the minimal capital level calculated with the old Solvency I rules. However, the target capital however will be estimated using the expected shortfall, rather than Value-at Risk, using a level of 99.5% for calculations. This takes into consideration the previous remarks about VaR and ES as the distribution functions in the insurance industry are often skewed and heavy-tailed.

The Swiss Solvency Test (SST) is mandatory for all insurance companies which are domiciled in Switzerland together with their branches. Compatibility between

both systems have been assured, so that Swiss companies will not be at a competitive disadvantage to insurers which are domiciled in EU (and EEA) member countries.

In order to be consistent with the new regulations Solvency II, as well as the Swiss Solvency Test, this thesis is based upon these two measures of risk and their thresholds to assess the exposure risk.

Chapter 2

Catastrophe Risk Models

2.1 Relevance of natural catastrophe models

Climate change due to global warming, economic growth combined with an increasing density of urban population are indisputable trends which the world is facing today. In the outset of this new environment, natural hazards are not only becoming more frequent but also more damageable to the exposed population in the catastrophic zones. The series of natural disasters of hurricane Andrew (1992) and Northridge (1994) earthquake which led to unprecedented losses and the bankruptcy of eleven insurers are good examples. Since then, concerns have grown in the insurance industry to find additional financial resources to contain the exploding catastrophic losses and to forecast more accurately their catastrophe risk losses. More recently and despite the measures undertaken, hurricane Katrina (2005) in the USA or hurricane Lothar, closer to us have caused important damage to the infrastructure and severe losses to insurance companies. Considering the scope of its impacts, Katrina was one of the most devastating natural disasters in United States history. The devastating storm prompted its own endless flood of questions about how and why such a disaster had occurred. Many wondered to what extent human activities and global warming have to do with what seemed to be one more in a run of increasingly powerful and destructive Atlantic hurricanes? Indeed, since 1995, tropical storm and hurricane activity in the Atlantic has been well above normal. An average of 7.7 hurricanes and 3.6 major hurricanes developed each year between 1992 and 2005, compared to an average of five hurricanes and 1.5 major hurricanes in each of the previous 25 years National Hurricane Center [2009]. How can it be possible to anticipate the intensity and damage caused by a hurricane, so the government and population can be better prepared in the future? This issue is the main subject and will be addressed in the following paragraph.

In most fields of insurance, historical claims data is commonly used to forecast and assess the risk of an insured object. However in the area of catastrophe risks, only little historical or empirical data is available, as natural hazards occur only once every 10, 100 or 1000 years. This is especially the case when considering the additional geographical dimension of the windstorm tracks for example, when data is collected and some regions may have a maximum of one data point. This is hence largely insufficient to conduct statistical analysis making catastrophe losses difficult to forecast. Due to the lack of data, many companies have turned to catastrophe risk modeling, which remains as the only solution to assess the risk of natural catastrophe risk portfolio.

This is also where the data used in this work takes its origin and this chapter provides the fundamentals of the mechanisms of catastrophic risk models, which also guide the reinsurers' underwriting strategy to price the reinsurance treaties. Thus, in this first paragraph, the relevance of catastrophic models as an imperative tool for the industry has been explained. Then this paper illustrates its history, the general framework from its origin to its recent developments and the fundamentals in order to understand its present architecture and construction. By going through the different elements of this "black box", the inputs and outputs of these models will then be described to enable the reader to understand the form of the data which this thesis uses.

Figure 2.1: Tropical cyclones in the Atlantic

2.2   History and developments of catastrophe models

Natural catastrophe models have been developed through linking scientific studies in measuring natural hazards with advances in information technology and geographic

information systems (GIS). This combination of two separate field of studies - measuring hazard and mapping risk (as shown in Figure 2.2) came together in the early 1990s to simulate the study of the impact of natural catastrophe on property and industrial zones.
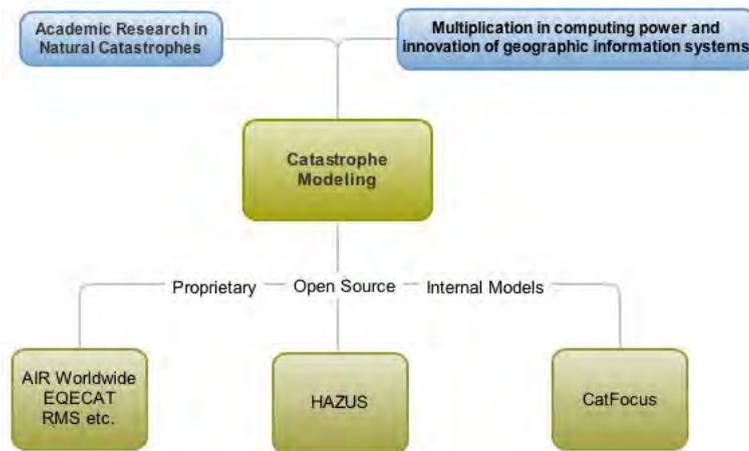
Figure 2.2: Developments of natural catastrophe modeling

1. *Risk mapping* combines information such as the current demography, state of the building (age, type and usage), scientific and financial data to determine the potential cost of catastrophes for a specified geographic area. In other words, an inventory of all the data characterizing all insured items are collected and stored. This is nothing new but lies in the earliest days of property insurance coverage. In the 1800's, residential insurers managed their risk by mapping the structures that they covered, using tacks on a wall-hung map to indicate their concentration of exposure. This rough technique was still in use until the 1960s until the first computer-based geographic information systems (GIS) software arrived. Since then, it became a better tool for conducting easier and more cost-effective studies on hazard and loss.

2. On the other side, *academic research of natural hazards* was the work of physicists at the very beginning. The first measurements of hurricane intensity and earthquake magnitude started in the 1800s after the modernization of the anemometer and the invention of the first modern seismograph. In the beginning of the twentieth century, as the understanding of the impact of natural hazards increased rapidly together with the improvements in scientific measuring techniques, the first data compiling hazard and loss studies to estimate the impact of natural disasters became available.

The combination of these two separate developments (as shown in Figure 2.2) lead to the advent of catastrophe modeling softwares. There are mainly three types,

proprietary models[1], open source model developed with a combination of public and private resources such as *Hazus* and the so-called in-house developed models, such as *CatFocus®* in the case of PartnerRe. The combination between in-house together with commercial models are used to provide a more independent and trustworthy estimation of risk.

2.3   How do they work?

Whatever natural hazard or type of catastrophe model is treated, the general idea behind remains the same. According to Grossi and Kunreuther [2006], there are four basic components in a catastrophe model, which are explained individually.

Figure 2.3: Structure of catastrophe models
Source: Grossi and Kunreuther, 2006 | Page 26.

The *hazard* component of catastrophe models comprises the simplified presentation of the complex properties of a natural hazard. Based on meteorological, physical and geophysical criteria, i.e. hazard source and attenuation models, it summarizes physical laws as well as historic and scientific hazard information. Its role consists of simulating these catastrophe events, statistically in coherence with the real events. For the case of a tropical cyclone, the risk based on historical events is characterized by its projected path and wind speed, along with other relevant parameters. The hazard component specific to tropical cyclones are discussed in more detail in the next section.

The *inventory* encompasses all the necessary information on the insured property. Each element is assigned to its geographic coordinates such as its latitude and longitude based on its street address, ZIP code, or another location descriptor.

---

[1]Today, the three main proprietary catastrophe modeling firms are: AIR Worldwide, Risk Management Solutions (RMS) and EQECAT (also known as EQE). Insurers, re-insurers, rating agencies, risk managers and major insurance brokers use licensed models from these firms.

Other factors such as the building height, age and occupancy types are also considered. It is very important to take all these risk factors into consideration and not to assume losses on average. As displayed in Figure 2.4, a large number of buildings are completely affected by a severe storm, while some stay intact. Some crucial risk factors, such as building maintenance and building structure can also make a big difference.

Figure 2.4: Different levels of damage for similar risks contributed heterogeneity in terms of losses.

Source: PartnerRe │ Tropical windstorm publication

Putting the hazard and inventory modules together enable the calculation of the *vulnerability* or susceptibility to damage of the structures at risk. In essence, this step quantifies the physical impact of the natural hazard phenomenon on the property at risk by computing the damage ratio. The vulnerability function calculates the loss for all risks by using the parameters and events in the hazard module. In the case of tropical cyclones, the losses for each simulated event, which are determined by its projected paths and wind speeds are computed. To practitioners, this module is also known as the engineering model, as it encompasses engineering techniques in the decision making of the vulnerability function (or damage quantification). Many studies on vulnerability have been undertaken, including a wide range of experiments and post-catastrophe on-site observations. It however differs from model to model, as the vulnerability function varies enormously between insurance lines (natural catastrophe, property, automobile etc.). This function can either be developed internally or sold as licensed software. In the case of the catastrophe model RMS for example, the justifications and statistical methods integrated in their models are not disclosed to the public.

From this measure of vulnerability, the loss to the inventory is evaluated. The *loss* module, also known as the actuarial module converts the calculated losses from all events into a risk premium, reflecting all relevant insurance and reinsurance

conditions, such as policy deductibles, by coverage (ie. site-specific or blanket deductibles), coverage limits and sub-limits, loss triggers, attachment points and limits for single or multiple location polices. The output is in all cases a loss file, containing the calculated losses for a portfolio of risks.

2.4   Tropical cyclone models

2.4.1   Introduction

Tropical cyclones are the most destructive of all atmospheric perils faced by the insurance industry. Of the ten costliest weather disasters in the history of the United States, six were the result of hurricanes[2]. Insuring against such risks requires the use of scientifically robust quantitative methods in order to assess these risks as the most accurately as possible. In the following section, the example of CatFocus® is used to illustrate the mechanisms of tropical cyclone models. CatFocus is the in-house catastrophe model currently in use in PartnerRe to provide a more detailed view on how this model functions by applying it to one specific natural hazard. For more detailed information about their soon–to-be-published study concerning tropical cyclones modeling, the reader is referred to the website of PartnerRe at www.partnerre.com.

Figure 2.5: Tracks of Atlantic tropical cyclones (1851—2005)
Source: National Hurricane Center (US) | Hurricane History

The term *tropical cyclone*[3] refers to a circulation of air that develops over the warm waters of the tropical latitudes between 20°N and 20°S. Their effects include powerful winds, heavy precipitation combined with huge waves and can be an important threat to coastal populations. Their strength progresses through regular

---

[2]Data from publications of US national climatic data center

[3]Cyclone is the generic name for any type of low pressure center that spins counterclockwise. Hurricane and typhoon are two names for tropical cyclones with winds of 65 knots (75 m.p.h.) or more.

stages of development from disturbance to a mature hurricane and its lifetime varies from two to five days per event, according to their intensity.

The origin of the cyclone can also be substantially influenced by its development. Through statistical research, it has been shown that cyclones in the Atlantic can be divided into three major tropical cyclone regions, based on its origin: Mid Atlantic regions, North Atlantic regions and Gulf region regions.

- *Mid Atlantic storms* have generally sufficient time to develop before making landfall, giving them the potential to gain considerable strength. They can gain the highest intensities among all three storms and show greater consistency in their movement over time.

- *North Atlantic storms* tend to move northwards, quickly reaching cooler waters that reduce their potential intensity. At these higher latitudes, north Atlantic storms enter the strong west wind drift where they rapidly degrade or undergo extra-tropical transition, gaining more extra-tropical characteristics (they lose their warm core and develop a cold core and frontal systems).

- In the *Gulf region of the Atlantic Ocean*, a large proportion of storm events move in different directions due to the absence of strong steering winds. In this region, more rapidly developing but short-lived storms are observed compared to the Mid Atlantic or North Atlantic regions owing to high sea surface temperatures. Gulf storms, despite tending to be of shorter duration, can reach wind intensities that are as high as those observed in the Mid Atlantic region.

There exist a multitude of differing methodologies used to evaluate tropical cyclone risk, all requiring a clear definition of the four previously defined components in Section 2.3. With all four components previously explained, an accent will put on the hazard component ie. on the generation of a statistically reliable set of tropical cyclone events using historical data.

2.4.2   Hazard component in tropical cyclones models

One key element in modeling tropical cyclones is the quality of meteorological data of past events. In order to simulate future events, several parameters are recorded by meteorological agencies at a six-hourly intervals. These quantities are the position of the latitude center of the cyclone eye, in order to model its path; the radius of maximum wind (Rmax), which is located near the eye wall and the radius of gale force (Rgale). Together they can give a good indication of the wind profile describing its size, shape of full wind field. This last piece of information, the wind field, is crucial for estimating the expected damage to each event passage on property. Using this data and a simulated maximum sustained wind speed (Wmax) based on Rmax

and Rgale, the wind field can be modeled using the Rankine vortex equation, which describes the rate at which wind speed reduces as a function of radial distance from the center of the cyclone, the effects of eye-wall cycles and how rotating winds are modified by the forward translation speed of a cyclone. All these characteristics (ie. Rgale, Rmax, Wmax) are to be taken into consideration within the simulation of our stochastic hazard model.
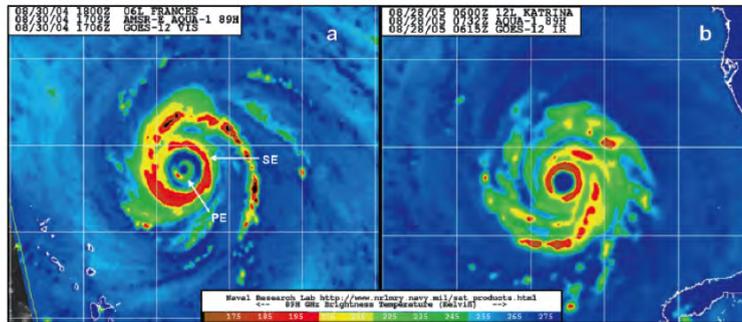
Figure 2.6: Satellite images of (a) hurricane Frances on September 30, 2004, and (b) hurricane Katrina August 28, 2005. An eye-wall replacement is underway at the time of the image in (a); convection in the primary eye-wall (marked PE) is weakening while convection in the secondary eye-wall (marked SE) strengthens. The warm (blue) ring between the primary and secondary eye-walls identifies the moat, an area associated with warm and dry, sinking air. In this event, the secondary eye-wall continued to contract and ultimately replaced the primary eye-wall. For comparison, hurricane Katrina in the image on the right (b) exhibits a single (primary) eye wall at the time of the image.

Source: PartnerRe, Kossin and Sikowsky (2009)

2.4.3   Model output and description of the data

2.4.3.1   Output from the model

The output of the catastrophe models can either be illustrated by a GIS map of the potential loss or an event loss table based on the information of the losses incurred by each simulated catastrophe event, using the exceedance probability (EP) curve. In contrast to a GIS map of loss, which presents loss in a spatial manner, an EP curve portrays loss in a temporal manner.

1. GIS Map indicates the losses per region, due to a specific catastrophe. The more important the loss is, the more intense the color is on the map.

2. An EP curve is particularly valuable for insurers and reinsurers to manage the portfolio risks and optimize the insurance / reinsurance portfolio. For a given portfolio of structures at risk, an EP curve is a density probability distribution that a certain level of loss will be surpassed in a given time period.

By random permutation of the geophysical parameters, the historic event sets can be artificially enlarged, resulting in the so-called stochastic event sets. By choosing the scenarios, which affect a certain insurance portfolio, many hundred or thousand years of potential future losses can be simulated. In this way, important information about the individual claims distribution and the aggregate loss distribution for this portfolio can be obtained.
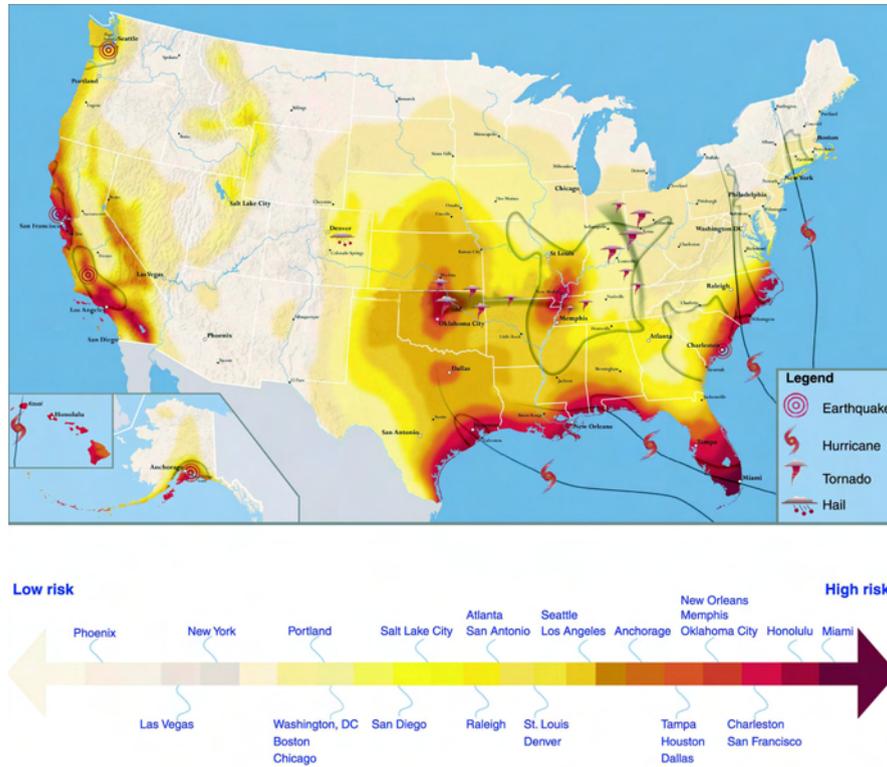
Figure 2.7: Example of GIS map
Source: Office [2002]

2.4.3.2 Description of the data

The focus of this thesis is on the South, South-East and North-East regions of the United States of America. Due to the aforementioned problem of scarcity of historical data, the data used below originates from the catastrophe models simulated by the Zurich modeling team of PartnerRe on the whole Business Unit Cat risk portfolio using the catastrophe model AIR CatRader version 11.0.1. The data provided by the catastrophe model is composed of the aggregated losses data of regions in the USA. Additionally, the marginal losses of each region are also provided. The dataset is composed of 8'005 data points.

	North-East	South-East	South
States	Connecticut, District of Columbia, Delaware, Massachusetts, Maryland, Maine, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Virginia, Vermont, West, Virginia	Alabama, Florida, Georgia, North Carolina, South Carolina	Arkansas, Kansas, Louisiana, Missis-sippi,Oklahoma, Texas

Table 2.1: States to region allocation

2.5   Discussion of catastrophe models

Historical data combined with stochastically simulated models are used in the prediction process. Hence, this requires high quality and reliable meteorological data on past events. However these data represent only up to maximum 150 years of historical measurements and are used to simulate very low probabilities events such as 1 in 250 or 1 in 500 years. Considering the intrinsic evolution of climate in the past million years, combined with external factors such as global warming, it is a pertinent question to ask whether these probabilities calculations are relevant for these extreme catastrophes. Despite the fact that catastrophe models can provide a good idea of the covered risk, the lack of data of extreme events outside of these 150 years of measurement, like hurricane Katrina or the changes of climate in recent decades, due to global warming which leads to increasing El Nino phenomenon may create a bias in the statistical inference and making less accurate predictions.

Chapter 3

Dependence Analysis

3.1 Introduction to copulas

In statistics, a copula (or dependence structure) describes the notion of dependence between random variables. It proceeds by linking them together to form a multi-variate distribution, which describes their joint behavior.

During the last ten years, copula modeling has kept the world of finance, insurance and numerous researchers busy. Despite copula being the subject of theoretical mathematical research already in the 1970s, it was only in the past decade that the industry fully realized the potential of this tool. Since then research in this field and applications to the finance industry has been subject of intense development. This whole adventure initially started off with this one simple actuarial question that a Swiss leading reinsurance company asked ETH Zurich mathematicians: "Given two marginal distributions X_1, X_2 with log-normal distribution functions $F_1 = LN(0, 1)$, $F_2 = LN(0, 16)$, how can one simulate from such a model if X_1 and X_2 have linear correlation $\rho = 0.5$ say" [Embrechts, 2009]. In the past, it was unknown to practitioners how to aggregate both risks. And at the same time, this question is crucial for the firm's overall risk assessment, as overestimating the dependence of risk positions can lead to higher capital requirements, which may handicap the competitiveness of the company due to the high capital costs. Or on the other side, underestimating the risk can put the company in the danger of insolvability in case of extraordinary catastrophe events, which be incur large losses for the reinsurance company.

More concretely, copulas have the advantage to showcase a complete and unique description of the dependence structure. In this sense, it provides **a way of isolating the description of the dependence structure** independently of the underlying marginal distribution. For example, the joint distribution of a random vector of

risk factors implicitly contains the information of both a description of the marginal behavior of individual risk factors, but as well a description of their dependencies. Imagine that the copula describing the underlying dependence of this random vector can be modeled by a Gaussian copula, then this dependence structure will remain the same whether the marginal distribution has a normal, student t or log-normal distribution.

Mathematically, the whole construction of copulas is expressed by the famous Sklar's theorem (Theorem 3.4), which allows the mapping of the individual marginal distribution functions $F_1, \ldots, F_n$ to the joint distribution function F via the copula C. In order to compare the random using the same scale, Sklar's theorem describes a copula based on a unitary scale of $[0, 1]$, Proposition 3.3 provides the tool to transform the marginal variable to an uniform distribution [0;1]. The copula can then be expressed as a multivariate distribution on the obtained uniforms, which is precisely a copula.

From a practical point of view, applications in copula inference have made a lot of progress since its initiation. Empirical applications have been successfully implemented in actuarial science, financial risk assessment or hydrology (See Malevergne and Sornette, 2003, Genest and Favre, 2007, which are also the two essential papers for reference in this thesis.). However it is in finance that it has been so extensively applied that it becomes almost the "newest fashion". With the growing popularity of copula applications and error-prone usage of this tool, it is believed that the application of the Gaussian copula to credit derivatives has been one of the reasons behind the global financial crisis of 2008–2009 [Donnelly and Embrechts, 2010].

In the context of this thesis, copulas will be applied to reinsurance data. For starters, the fundamental theory of dependence assessment and copulas used in this thesis are reviewed. This section however provides only the necessary tools for conducting copula inference but not the exhaustive theoretical background. For a deeper discussion about copulas, the reader is kindly referred to the Chapter 3 of Malevergne and Sornette [2005] or / and Chapter 5 of McNeil et al. [2005].

A basic definition of a copula is the following:

Definition 3.1. (Copula). A n-dimensional copula is a multivariate joint distribution function defined on the n-dimensional unit cube $[0, 1]^n$ with marginal uniform distributions on the interval [0, 1].

Specifically, $C : [0, 1]^n \to [0, 1]$, where $C(\mathbf{u}) = C(u_1, \ldots, u_n)$ is an n-dimensional copula, if:

1. $C(u_1, \ldots, u_n)$ is increasing in each of its component u_i. This is due to the fact that C is a multivariate distribution function.

2. The standard marginal distributions are uniform, which is equivalent to say that:

$$C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i \quad \forall i \in \{1, \ldots, n\}, \; u_i \in [0, 1]. \qquad (3.1)$$

3. C satisfies the rectangle inequality: for all $(a_1, \ldots, a_n)$, $(b_1, \ldots, b_n) \in [0, 1]^n$ with $a_i \leq b_i$ we have

$$\sum_{i_1=1}^{2} \cdots \sum_{i_n=1}^{2} (-1)^{i_1 + \ldots + i_n} \, C(u_{1i_1}, \ldots, u_{ni_n}) \geq 0, \qquad (3.2)$$

where $u_{j1} = a_j$ and $u_{j2} = b_j \quad \forall j \in \{1, \ldots, n\}$.

Proposition 3.2. *These two transformations hold:*

- **Quantile transformation**: Let $U \sim U(0, 1)$, F be a distribution function with generalized inverse F^{-1}. Then $P\left(F^{-1}(U) \leq x\right) = F(x)$.

- **Probability transformation**: Let X be a random variable with continuous distribution function F. Then $F(X) \sim U(0, 1)$.

Proof. The proof of this proposition can be found in page 186 of McNeil et al. [2005]. $\qquad \square$

Proposition 3.3. (Invariance under monotone transformations). *Let* $\mathbf{X} = (X_1, \ldots, X_n)'$ *be a random vector with copula C and let $T_1, \ldots, T_n$ be strictly increasing functions. Then* $(T_1(X_1), \ldots, T_n(X_n))'$ *also has copula C.*

Proof. A proof of this proposition can be found in page 188 of McNeil et al. [2005]. $\qquad \square$

Proposition 3.2 and 3.3 are fundamental properties of copula modeling. Because of the invariance under strictly increasing transformations property, the copula remains the same, regardless of the transformations. These properties are useful to transform the data to an uniform distribution.

Theorem 3.4. (Sklar, 1959). *Let F be a joint distribution function with margins $F_1, \ldots, F_n$. Then there exists a copula $C : [0, 1] \to [0, 1]$ such that, for all $x_1, \ldots, x_n$ in $R \in [-\infty, \infty]$,*

$$F(x_1, \ldots, x_n) = C_\theta(F_1(x_1), \ldots, F_n(x_n)). \qquad (3.3)$$

If the margins are continuous, then C is unique; otherwise C is uniquely determined on $Ran(F_1) \times Ran(F_2) \times \ldots \times Ran(F_n)$, where $Ran(F_1)$ denotes the range of F_i.

Conversely, if C is a copula and $F_1, \ldots, F_n$ are uni-variate distribution functions, then the function F defined above is a joint distribution function with margins $F_1, \ldots, F_n$.

One can also evaluate Theorem 3.4 for $x_i = F_i^{-1}(u_i)$ and obtain an explicit expression for C:

$$C_\theta(u_1, \ldots, u_n) = F\left(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)\right). \tag{3.4}$$

Proof. A proof of this theorem can be found in Schweizer and Sklar [1983]. □

For example, consider two random variables X and Y, with continuous cumulative distribution functions F_X and F_Y. The probability transform is applied separately in order to obtain two uniform random variables $U = F_X(X)$ and $V = F_Y(Y)$ with the same dependence structure as X and Y. Based on these two uniform marginal distributions, various copulas C_θ are constructed to fit the data. It will be examined in the next chapter, how C_θ is chosen and fitted to model the dependence structure of the data.

Theorem 3.5. (Fréchet-Hoeffding bounds). *Let X be a random vector with marginal distribution functions $F_1, \ldots, F_d$ and joint distribution function F. Let $u_i = F_i(x_i)$, then, for every copula $C_\theta(u_1, \ldots, u_n)$, the bounds:*

$$W = \max\left\{\sum_{i=1}^{d} u_i + 1 - d, 0\right\} \leq C(u) \leq \min(u_1, \ldots, u_d) = M. \tag{3.5}$$

The lower and upper bounds constitute the Fréchet-Hoedffding bounds and are the strongest form of dependence that random variables can exhibit. The upper bound is the so-called comonotonicity copula, representing the perfect positive dependence of the joint distribution of the random vector $(U, ..., U)$. The lower bound is the countermonotonicity copula and represents perfect negative dependence. This is however valid only in the two-dimensional case and is then the joint distribution of $(U, 1-U)$.

Proof. The proof of this theorem can be found in page 189 of McNeil et al. [2005]. □

3.2 Dependence assessment

3.2.1 Dependence measures

In mathematics, dependence refers to a statistical measure of the relationship between two data sets. Dependence measures indicate the strength and the direction

of a relationship between two random variables and yield a scalar value between $[-1 : 1]$, namely the correlation coefficient. In this section, three kinds of dependence measures are discussed: the Pearson linear correlation, rank correlations and the coefficients of tail dependence.

3.2.1.1 Linear correlation

Dependence can be described by Pearson's correlation coefficient. As it is easy to implement, it is one of the most popular measure of dependence and is obtained by dividing the covariance of two variables by the product of their standard deviations. Formally it is defined as the following:

Definition 3.6. (Pearson's linear correlation). The correlation coefficient $\rho_{X,Y}$ between two random variables X and Y is

$$\rho_{X,Y} = \text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \qquad (3.6)$$

where σ_X and σ_Y express their respective standard deviations and μ_X and μ_Y the expected values.

Pearson's correlation expresses the extent to which random variables are linearly "proportional" to each other. It assesses the dependence by looking at the percent of variation of one variable, based on the unitary change of the second variable.

From a practical point of view, Pearson's linear correlation is a widely used tool to assess dependence. However, it should not be taken as a default measure of dependence despite its popularity, as it leads to an improper estimation of dependence. This is because Pearson's correlation coefficient measures only the linear dependency between two random variables and fails to capture other kinds of dependencies as illustrated in Figure 3.1. For this reason, it is only suitable for normal or more generally elliptical distributions. In the following paragraph, some further shortcomings that users may encounter when using the linear correlation are presented in a more extensive manner [Embrechts et al., 2002]:

Fallacies surrounding the linear correlation:

1. If X and Y are independent, then the correlation is zero. But the converse is false: if $\rho = 0$, the random variables are not obviously independent. For example in the last row of Figure 3.1, the correlation is zero, but clearly they are not independent.

2. The necessary invariance property under strictly increasing linear transformations of copula models, can not be applied to Pearson's linear correlation. As
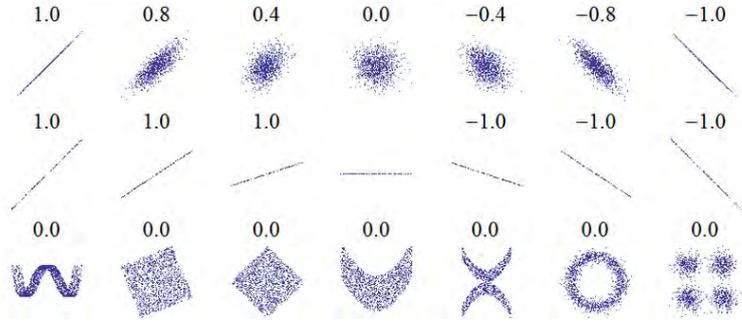
Figure 3.1: Linear correlation examples
Source: Wikipedia; linear correlation

seen in the formula above, Pearson's correlation coefficient depends on both the marginal distribution and the dependence structure. In case of elliptical distributions, the invariance property remains valid. But beyond this category of distributions, two real-valued random variables, where $T : \mathbb{R} \to \mathbb{R}$ a increasing transformation, then $\rho_{X,Y}(T(X), T(Y)) \neq \rho(X, Y)$.

3. Another pitfall of Pearson's linear correlation is that according to the definition, the variances of X and Y have to be finite. This restriction to finite-variance models is not ideal especially for estimating heavy tailed distributions. For example, actuaries who model losses in different business lines with infinite-variance distributions may not describe the dependence of their risks by using linear correlation.

3.2.1.2 Rank correlations

In order to overcome the aforementioned problems related to Pearson's linear correlation, rank correlations have been introduced to provide a better alternative for measuring dependence in the case of non-elliptical distributions.

Rank correlations are also scalar measures of dependence, but they only depend on the dependence structure (ie. copula) of a bivariate distribution and not on its marginal distributions, unlike linear correlation, which depends on both [McNeil, Frey, and Embrechts, 2005]. This is indeed the case, as by looking at the ranks alone, the dependence does not vary under strictly increasing transformation, hence the copula remains the same.

Practically, rank correlations use non-parametric measures such as Spearman's ρ and Kendall's τ, by looking at the probability of concordance and discordance of ranks of the data. The specifics are explained in the next section.

Definition 3.7. (Ranks). Given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from a pair of continuous random variables (X, Y) with a bi-variate distribution $H(x, y)$ which

characterizes their joint behavior. The rank of each observation R_i and S_i are defined as

$$R_i = \sum_{k=1}^{n} 1\left\{X_k \leq X_i\right\} \text{ and } S_i = \sum_{k=1}^{n} 1\left\{Y_k \leq Y_i\right\}. \tag{3.7}$$

Under the assumption that X and Y are continuous random variables, the probability of ties is zero and the ranks defined explicitly.

- KENDALL'S TAU (τ) [KENDALL, 1938]

  Definition 3.8. (Kendall's tau). Let X_1 , $\widetilde{X}_1$ be iid random variables with continuous marginal distribution F_1 and X_2 , $\tilde{X}_2$ be iid random variables with continuous marginal distribution F_2. Then the Kendall's tau is defined by:

$$\rho_\tau(X_1, X_2) = \left(P((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0)\right) - \left(P((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) < 0\right) \tag{3.8}$$

$$= E(sign((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2))). \tag{3.9}$$

  Given two points in $\mathbb{R}^2$, denoted by (x_1, x_2) and $(\tilde{x}1, \tilde{x}2)$, are stated to be *concordant* if $(x_1 - \tilde{x}1)(x_2 - \tilde{x}2) > 0$ and to be *discordant* if $(x_1 - \tilde{x}1)(x_2 - \tilde{x}2) < 0$. Kendall's tau represents the measure of probability of concordance between (X_1, X_2) and $(\tilde{X}_1, \tilde{X}_2)$ minus the probability of discordance. For example, if X_2 tends to increase with X_1, then the probability of concordance is expected to be high in comparison with the probability of disconcordance [Malevergne and Sornette, 2005].

  Definition 3.9. The empirical estimator of Kendall's tau is defined as :

$$\tau_n = \frac{P_n - Q_n}{\binom{n}{2}} = \frac{4}{n(n-1)} P_n - 1, \tag{3.10}$$

  where P_n and Q_n are the number of concordant and discordant pars in the sample, and τ_n an asymptotically unbiased estimator of ρ_τ_[Genest and Favre, 2007].

  Beyond two dimensions, the Kendall's tau can be represented by a matrix of a random vector $\rho_\tau(\mathbf{X}) = cov(sign((\mathbf{X} - \tilde{\mathbf{X}}))$, where $\tilde{X}$ is an independent copy of X. For practical reasons in copula modeling (as seen in Step 3 of Algorithm 4.3.1), the matrix of Kendall's tau needs to be semi-definite positive. Since the last expression can be illustrated as a covariance matrix, $\rho_\tau(\mathbf{X})$ is obviously positive semi-definite [McNeil et al., 2005].

- SPEARMAN'S RHO (ρ) [SCARSINI, 1984]

  Definition 3.10. (Spearman's rho). Let X_1 and X_2 be two random variables with continuous marginal distribution functions F_1 and F_2. Then Spearman's rho is given by:

  $$\rho_s(X_1, X_2) = \text{corr}\left(F_1(X_1), F_2(X_2)\right). \tag{3.11}$$

  In other words, Spearman's ρ_s is simply the linear correlation of the probability transformed random variables, which for continuous random variables is the linear correlation of their unique copula.

  The empirical estimator of ρ_s is ρ_n, which is the correlation between the pairs of the ranks (R_i, S_i) and is given by:

  $$\rho_n = \frac{\sum\limits_{i=1}^{n} \left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right)}{\sum\limits_{i=1}^{n} \left(R_i - \bar{R}\right)^2 \sum\limits_{i=1}^{n} \left(S_i - \bar{S}\right)^2} \in [-1, 1] \tag{3.12}$$

  where

  $$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{n+1}{2} = \frac{1}{n}\sum_{i=1}^{n} S_i = \bar{S}. \tag{3.13}$$

  The Spearman's ρ_s matrix for the general multivariate random vector $\mathbf{X}$ is given by $\rho(\mathbf{X}) = \text{corr}\left(F(X_1), F(X_2), \ldots, F(X_d)\right)$ and must be again be positive semi-definite [Joe, 1997].

3.2.2 Tail dependencies

Due to the losses in relationship with the likelihood of extreme weather events occurring in more than one region, tail dependency provides a measure to estimate pair wise *extremal dependence*. In other words it measures the strengths of dependencies in the tails of a bi-variate distribution. This piece of information that the tails of data sets are asymptotically dependent or independent is especially important when fitting copulas to empirical data, as some models (eg. Gaussian copula) are asymptotically independent in the tail, while others have tail dependence. Such knowledge can influence significantly the estimation of the necessary capital for the firm.

Definition 3.11. (Asymptotic dependence measure λ). Let X and Y be random variables with continuous distribution functions F_X and F_Y. The coefficient

of upper tail dependence of X and Y are defined as:

$$\lambda_u(X,Y) = \lim_{u \to 1^-} P\left[X > F_X^{-1}(u) \mid Y > F_Y^{-1}(u)\right] \tag{3.14}$$

$$= \lim_{u \to 1^-} \frac{P\left[X > F_X^{-1}(u) \cap Y > F_Y^{-1}(u)\right]}{P\left[Y > F_Y^{-1}(u)\right]}. \tag{3.15}$$

For example, if X and Y represent the losses of two peril zones of probability u, their coefficient of tail dependence λ_u looks at the probability that X exceeds its q-quantile, given that Y exceeds its q-quantile, and then consider the limit as q goes to 1. The roles of X and Y are interchangeable.

The upper tail dependence looks at the limit of the probability that Y is greater than its q-quantile, given that X exceeds its q-quantile. If $\lambda_u > 0$, then X and Y present tail dependence and large events tend to occur simultaneously with probability λ_u. If $\lambda_u = 0$, they show signs of asymptotically independence in the upper tail.

Definition 3.12. Analogously, the coefficient of lower tail dependence is given by:

$$\lambda_l(X,Y) = \lim_{u \to 0^+} P\left[X \le F_X^{-1}(u) \mid Y \le F_Y^{-1}(u)\right] \tag{3.16}$$

$$= \lim_{u \to 0^+} \frac{P\left[X \le F_X^{-1}(u) \cap Y \le F_Y^{-1}(u)\right]}{P\left[Y \le F_Y^{-1}(u)\right]} \tag{3.17}$$

provided a limit λ_u and $\lambda_l \in [0,1]$ exists. If $\lambda_l > 0$, then X and Y are stated to have lower tail dependence.

One can also interpret the last expression in terms of a Value-at-Risk. In fact, the quantiles $F_X^{-1}(u)$ and $F_Y^{-1}(u)$ are nothing but the Values-at-Risk of X and Y at the confidence level u. Thus, the coefficient λ_u simply provides the probability that X exceeds the VaR at level u, assuming that Y has exceeded the VaR at the same probability level u, when this level goes to one [Malevergne and Sornette, 2005].

Tail dependence can also be written by using copulas:

Theorem 3.13. (Copula representation of the asymptotic dependence λ). Taking the case of upper tail dependence λ_u and let C be the copula of the variables X and Y, then

$$\lambda_u(X,Y) = \lim_{u \to 1^-} \frac{1 - 2u - C\left(F(X), G(Y)\right)}{1 - u}. \tag{3.18}$$

Proof. A proof of this theorem can be found in McNeil et al. [2005]. $\qquad\square$

This thesis draws the readers' attention to the fact that $\lambda_u = 0$ alone is not a sufficient condition to assert the asymptotic independence of the copula and a complementary tail dependence measure $\bar{\lambda}$ was developed by Coles et al. [1999] to confirm the H_0 hypothesis of tail independence:

Definition 3.14. (Asymptotic dependence measure $\bar{\lambda}$). Let X and Y be random variables with continuous distribution functions F_X and F_Y. The asymptotic dependence measure $\bar{\lambda}$ of upper tail dependence of X and Y is defined as:

$$\bar{\lambda}_u = \lim_{u \to 1} \frac{2 \log\left(\Pr\left(X > F_X^{-1}(u)\right)\right)}{\log\left(\Pr\left(X > F_X^{-1}(u), Y > F_Y^{-1}(u)\right)\right)} - 1, \tag{3.19}$$

where $-1 \leq \bar{\lambda}_u \leq 1$.

$\bar{\lambda}$ represents the rate at which $P\left[X \leq F_X^{-1}(u) \mid Y \leq F_Y^{-1}(u)\right]$ approaches zero. When $\bar{\lambda}_u = 1$, the variables are asymptotically dependent and if $-1 \leq \bar{\lambda}_u < 1$, the variables are asymptotically independent. Hence, the pair $(\lambda, \bar{\lambda})$ together can provide information about the extremal dependence at the tails. To sum up, if $0 < \lambda_u \leq 1$ and $\bar{\lambda}_u = 1$, it is a sufficient to ascertain the tail dependence between two random variables. Alternatively in order to ascertain tail independence the condition $\lambda_u = 0$ must be fulfilled, but this also requires as a necessary condition that $-1 \leq \bar{\lambda}_u < 1$.

3.2.3 Graphical tools for dependence assessment

Graphical tools [Genest and Favre, 2007] are used to visualize the underlying dependence in our data set. Alternatively, they can also be used to check the adequacy of the fitted copula model by comparing it against the empirical copula. The graphical tools for dependence assessment are mainly all rank based, so that only the information of the dependence structure between random variables is illustrated. Other tools plotting the complete multivariate distribution, such as scatter plots for example are not ideal for this reason as it also illustrates the marginal behavior of X and Y. Three rank based plots have been implemented in the framework of this thesis and are explained in the following.

3.2.3.1 Rank plot

Rank plots aim to visualize the copula by plotting the respective ranks of two random variables (X, Y) onto their respective (empirical) marginal distributions (F, G). In other words, for each observation (X_i, Y_i), a graph $(F_i(X), G_i(Y))$ is plotted for all

observations $i = 1, \ldots, n$.

$$U_i = F_n(X_i) = \frac{1}{n} \sum_{k=1}^{n} 1 \{X_k \leq X_i\} = \frac{R_i}{n} \qquad (3.20)$$

$$V_i = G_n(Y_i) = \frac{1}{n} \sum_{k=1}^{n} 1 \{Y_k \leq Y_i\} = \frac{S_i}{n}. \qquad (3.21)$$

3.2.3.2 Chi plot

Chi plots are intended to reveal more detailed and explicit information regarding the nature of association of two random variables (X, Y). The elements of the Chi plot illustrate the pair (λ_i, χ_i), for $\lambda_i, \chi_i \in [0, 1]$, are formally defined as:

$$\lambda_i = 4 sign \left(\tilde{F}_i, \tilde{G}_i \right) \max \left(\tilde{F}_i^2, \tilde{G}_i^2 \right) \qquad \forall i \in (1, \ldots, n) \qquad (3.22)$$

$$\chi_i = \frac{H_i - F_i G_i}{\sqrt{F_i (1 - F_i) G_i (1 - G_i)}} \qquad (3.23)$$

where

$$H_i = \frac{1}{n-1} \# \{j \neq i : X_j \leq X_i, Y_j \leq Y_i\} \qquad (3.24)$$

$$F_i = \frac{1}{n-1} \# \{j \neq i : X_j \leq X_i\}, \quad G_i = \frac{1}{n-1} \# \{j \neq i : Y_j \leq Y_i\}$$

$$\tilde{F}_i = F_i - \frac{1}{2}, \quad \tilde{G}_i = G_i - \frac{1}{2}.$$

Graphically, this method is designed so that the plot (λ_i, χ_i) is approximately horizontal under the null hypothesis of independence, $H_0 : C = uv$, one would expect $H_i \approx F_i \times G_i$ for all i. Under other forms of association, it produces characteristic patterns as illustrated in Figure 3.2. Intuitively χ_i indicates the departure from the independence condition of two marginal distributions (X, Y) and measures the failure of the bi-variate distribution function to factorize into a product of marginal distribution functions. λ measures the distance from (X_i, Y_i) to the bivariate median. For a full description of how to construct a chi plot and the theory behind it, the reader is referred to Fisher and Switzer [1985, 2001].

3.2.3.3 Kendall's plot

Kendall's plot (or K-plots) are initially inspired from the QQ-plot which looks at the order statistics of the quantiles of marginal distributions. This procedure retains the Chi plot's property of invariance to monotone transformations of the marginal distribution and combines it with the advantage of being easier to interpret as the curvature of the graph displays in a definite way the copula. It has furthermore
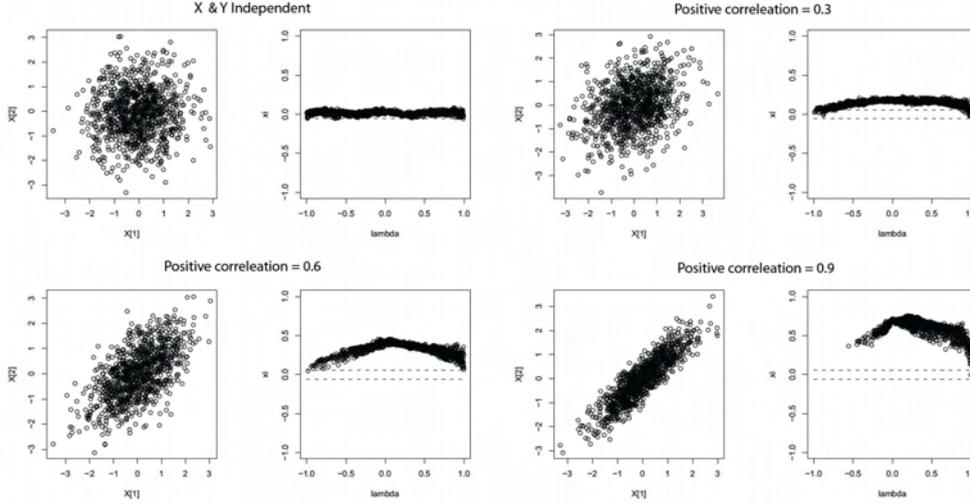
Figure 3.2: Scatter plot and Chi-plot examples with normal distribution samples
Source: Canestaro [2010]

the benefit of being readily extensible in a multivariate context [Genest and Boies, 2003].

The empirical joint distribution, H (as defined in Equation 3.24) is shown on the x-axis and it is compared to the independence condition of the ith statistics of the same copula on the y-axis. If the independence hypothesis holds, the points of the K-plot will lie along the main diagonal $y = x$, as shown in the first graph of Figure 3.3. A deviation from the main diagonal indicates dependence in the data. The pair $(W_{i:n}, H_{(i)})$ is plotted for $i \in (1, \ldots, n)$, where $H_{(1)} < H_{(2)} < \ldots < H_{(n)}$ are the order statistics related to the quantities $H_1, H_2, \ldots, H_n$. $W_{i:n}$ is defined as the expected value of the ith order statistic from a random sample of size n from the random variable $W = C(U, V) = H(X, Y)$. When $H_0 : H(X, Y) = F(X)G(Y)$, $W_{i:n}$ is given by:

$$W_{i:n} = n \binom{n-1}{i-1} \int_0^1 \omega k_0\left(\omega\right) \left\{K_0\left(\omega\right)\right\}^{i-1} \left\{1 - K_0\left(\omega\right)\right\}^{n-1} d\omega \qquad (3.25)$$

where $K_0\left(\omega\right)$ is the distribution function under H_0 and is defined as

$$K_0\left(\omega\right) = \omega - \omega \log\left(\omega\right). \qquad (3.26)$$

For more details of this method, the reader is recommended to have a deeper look at Genest and Boies [2003] on page 3-4.
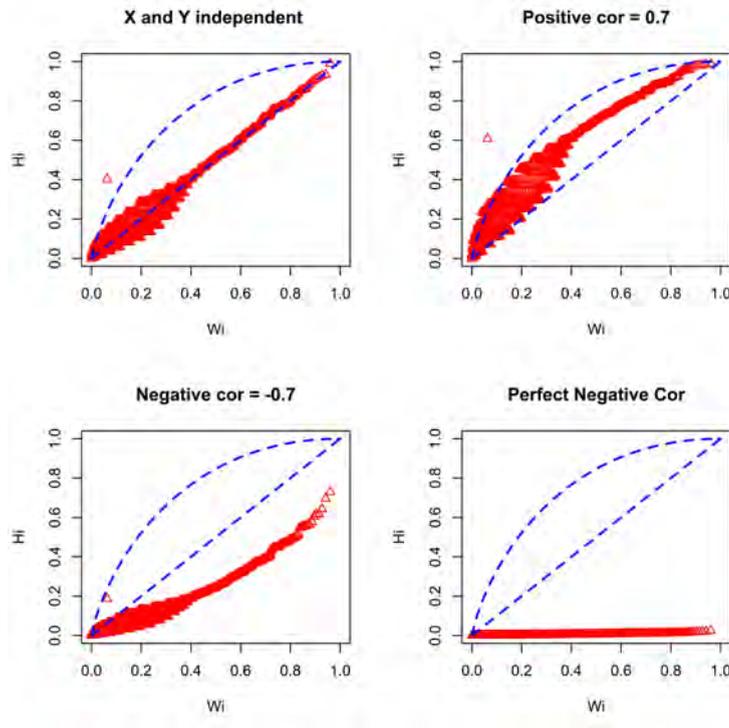
Figure 3.3: Kendall's plot examples with normal distribution samples

3.2.4 Tests of independence

Testing for independence between two continuous random variables X and Y is an important problem that has been the object of much attention in the past century. Despite a great amount of literature available on the subject, classical tests of independence between two continuous random variables based on Pearson's correlation, such as Student's test and Fisher's Z-transform continue to be the most commonly used in practice. These tests only measure the degree of linear association for the normal paradigm and beyond as stated in limitations of the linear correlation coefficient in Section 3.2.1.1, their effectiveness is questionable. Therefore the test of independence should be based on invariant statistics, such as the copula based quantities like Spearman's ρ or Kendall's τ, which are not affected by the shape of the margins.

3.2.4.1 Procedure for testing independence

As described in Genest and Favre [2007], a test of independence based on Spearman's ρ and Kendall's τ is performed.

For Spearman's ρ   Under the null hypothesis of the independence between X and Y, $H_0 : C = \Pi$, the distribution ρ_n defined as the asymptotically unbiased estimator

of ρ_s is close to normal with zero mean and variance $\frac{1}{n-1}$. Where n is the sample size. So for an approximate level of $\alpha = 0.05$, one may reject H_0 if

$$\sqrt{n-1} * |\rho_S| > z_{\alpha/2} = 1.96. \tag{3.27}$$

For Kendall's τ An alternative test of independence can be based on τ_n since under H_0, this statistics is close to normal with zero mean, variance equal to $\frac{2(2n+5)}{9n(n-1)}$ and n is the sample size. H_0 is rejected at 5% level if

$$\sqrt{\frac{9n(n-1)}{2(2n+5)}} * |\rho_\tau| > z_{\alpha/2} = 1.96. \tag{3.28}$$

Chapter 4

Dependence Modeling and Copulas

4.1   Introduction to copula models

A joint distribution function of a random vector of risk factors contains implicitly the information of both a description of the marginal behavior of individual risk factors and their dependence. The copula approach provides a way of isolating the description of their *dependence structure* and helps in a better understanding of dependence relationship between random variables. Just as there are unlimited types of dependencies, there exist also as many copulas models which all differ in the types of dependence they represent. A family of copulas describes a range of dependence structures, which have several parameters, all relate to the strength and form of the dependence. As to the marginal behavior, a copula model possesses the flexibility to combine a variety of possible marginal distributions. This is in fact one important advantage of copula models, as they don't make any assumptions on the marginal distributions. Then how to model the data dependence structure in the form of a copula? This is one of the most commonly asked question for practitioners, which will be answered in this thesis.

In the present chapter, some of the most important copula families, subdivided into the following three categories are presented. *Fundamental copulas* encompass the most elementary dependence structures such as full dependence and independence; *implicit copulas* are extracted from well-known multivariate distributions such as Gaussian or Student-t and do not have any closed form expression; explicit copulas are represented through a close form expression. When conducting copula inference for a multidimensional dataset, the usage of one single parameter θ leaves

limited flexibility to model the multidimensional character of the dependence structure Nelson [2005]. For this purpose multi-parameter *Archimedean copulas*, such as nested copulas as described in the book of Joe [1997] (pp. 150-153) has been developed, but they are complex to implement in practice.

After presenting the aforementioned copulas families, methodologies on how to fit them to the data are explained. There is unfortunately no hat trick or a universal rule to choose one copula and each fitting is different for each individual case. It would also be much easier if our marginals are normally distributed. However is not the case with frequently heavy tails present in the used catastrophe data. Then after, this is to be followed by algorithms to simulate copulas and finally the some goodness-to-fit testing will be discussed.

4.1.1   Fundamental copulas

The family of fundamental copulas represents the most elementary copulas at the lower or higher boundaries for all copula families.

The *independence copula* exhibits the case, where all the components of the marginal random variables are perfectly independent from each other. By the law of probability, the joint distribution of independent random variables can be decomposed into the product of the marginal probabilities and therefore given by:

$$C_{independent}(u_1, \ldots, u_d) = \prod (u_1, \ldots, u_d) = \prod_{i=1}^{d} u_i. \qquad (4.1)$$

The *comonotonicity copula* is the Fréchet upper bound copula (M) from Theorem 3.5 and is given by:

$$C_{comonotonic}(u_1, \ldots, u_d) = \min\{u_1, \ldots, u_d\}. \qquad (4.2)$$

It relates to perfectly dependent continuous margins, in the sense that they are almost surely strictly increasing or decreasing transformations of each other, so that $u_i = T_i(u_1)$. It represents the distribution function of the vector $\mathbf{U} = U(U, \ldots, U)$.

Finally the *countermonotonicity copula* is the two-dimensional Fréchet lower bound (W) copula from Theorem 3.5 binding continuous marginals who are perfectly negatively dependent. It is given by:

$$C_{countermonotonic}(u_1, u_2) = \max\{u_1 + u_2 - 1, 0\}. \qquad (4.3)$$

This means that if $u_1, u_2 \sim U(0, 1)$, the copula corresponds exactly to a bivariate distribution functions of random vectors that are either U or $1 - U$. It is not possible to extend the concept of countermonotonicity to more than two dimensions.
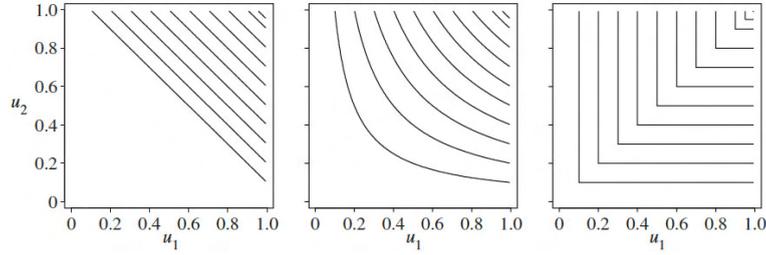
Figure 4.1: Contour plots of the three fundamental copulas. From left to right: countermonotonicity, independence and comonotonicity copulas.

Source: Alexander McNeil et al. $\mid$ Quantitative Risk Management $\mid$ Page 190.

4.1.2 Implicit copulas

4.1.2.1 The Gaussian copula

The Gaussian copula is constructed according to Sklar's theorem by using the multivariate normal distribution as the name suggests. It is defined formally as:

$$C_P^{Ga}(u) = \Phi_P^d(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_d)), \tag{4.4}$$

where Φ_P^d denotes the joint d-variate standard normal distribution function with the correlation matrix parameter P, which is estimated using the dependence measures cited in Section 3.2.1. Φ^{-1} denotes the inverse of the distribution function of univariate standard normal distributions.

There are two interesting cases for P. One is when $P = I_d$, where I_d is the identity matrix of d dimensions, then the independence copula is obtained. And if $P = J_d$, a $d \times d$ matrix consisting entirely of ones, the comonotonicity copula or a perfectly positively dependent copula is obtained.

Comments about this copula: In practice, the Gaussian copula has the advantage of being easy to construct and to simulate, but it limits the application only to tail independent data sets. When one looks at the tail of the distribution, the coefficients of upper and lower tail dependence for the Gaussian copula are constantly zero, regardless how highly the underlying dependence of the data is in reality (a proof can be found in McNeil et al. [2005] on page 211.). In the framework of this thesis and in presence of sufficient amount of data, the marginal distributions have been modeled by using their empirical distributions. For this specific case, the latter will be named hereafter as the "**Meta-Gaussian copula**".

4.1.2.2 The Student t copula

A famous copula model is the Student t copula. Let a d-dimension vector $\mathbf{X}$ have the following stochastic representation:

$$\mathbf{X} = \mu + \frac{\sqrt{\nu}}{\sqrt{S}}\mathbf{Z}$$

where $\mu \in \mathbb{R}^d$, $S \equiv \chi_\nu^2$ and $\mathbf{Z} \equiv \mathcal{N}_d(0, \Sigma)$ are independent. Then $\mathbf{X}$ has a d-dimensional t_ν distribution with mean μ and covariance matrix $\frac{\nu}{\nu-2}\Sigma$. The copula of $\mathbf{X}$ is defined as:

$$C_{\nu,P}^t(\mathbf{u}) = \mathbf{t}_{\nu,P}(t_\nu^{-1}(u_1), ..., t_\nu^{-1}(u_n)),$$

where P is a correlation matrix and ν the number of degrees of freedom. t_ν is the distribution function of a standard uni-variate t distribution.

Comments about this copula: As in the case of the Gaussian copula, the comonotic copula is obtained with $P = J_d$, a $d \times d$ matrix consisting entirely of ones. However if P is diagonal or is the identity matrix I_d, this does not mean that the components are independent, but are just uncorrelated (see Section 3.2.1.1 for fallacies of the linear correlation). Student-t copulas have however the advantages of capturing tail dependencies, as illustrated in Figure 4.2c.

4.1.3 Explicit copulas

The concept of Archimedean copulas has been approached for the first time in the paper of Schweizer and Sklar [1983], but the term of "Archimedean copulas" appeared only after the two papers of Genest and Mackay [2003] were published. Since then, Archimedean copulas have became a widely used class of copulas as they allow for a greater variety of dependence structures, which cannot be modeled with implicit copulas. This is particularly interesting for asymmetric dependence structures, for example in the stock market where the dependence for big losses is much greater than for big gains. However as this paper analyses a catastrophe risk portfolio, where the data represents the losses due to natural hazards, the interest is mainly in the upper tail of large losses. The lower tail representing smaller losses are less relevant for capital allocations in the reinsurance industry. For this reason, modeling two tails are not be covered within the framework of this study.

Definition 4.1. (Archimedean copula). Let $\phi : [0,1] \to [0,\infty]$ be a continuous, strictly decreasing and convex function so that $\phi(1) = 0$ and $\phi(0) = \infty$. Let $\phi^{-1} : [0,\infty] \to [0,1]$ be the inverse of ϕ, then the function $C : [0,1]^d \to [0,1]$ defined by

$$C\left(u_1, \ldots, u_d\right) = \phi^{-1}\left(\phi\left(u_1\right) + \ldots + \phi\left(u_d\right)\right) \qquad (4.5)$$

is called the Archimedean copulas if and only if ϕ^{-1} is completely monotonic on $[0, \infty)$ ie.

$$\left(-1\right)^d \frac{\partial^k}{\partial u^k} \phi^{-1}\left(u\right) \geq 0 \quad \forall k \geq 1 \qquad (4.6)$$

The function ϕ is called the generator of the copula. $\phi\left(0\right) = \infty$ then the generator is stated to be strict.

4.1.3.1 Archimedean copulas

In order to model the dependence of the data, the most commonly used copulas have been examined in further detail. For a more extended study in this field, a collection of twenty-two one-parameter families of Archimedean copulas can be found in Table 4.1 of Nelson [2006].

The *Gumbel copula,* with the generating function

$$\phi\left(u\right) = \left(-\log\left(u\right)\right)^{\theta}, \qquad \theta \geq 1 \qquad (4.7)$$

interpolates between independence and perfect dependence by having θ to represent the strength of the dependence. In particular, if $\theta = 1$, the independence copula is obtained, while the limit as $\theta \to \infty$ is the comonotonicity copula.

The *Clayton copula,* with the generating function

$$\phi\left(u\right) = \frac{u^{-\theta} - 1}{\theta}, \qquad \theta > 0 \qquad (4.8)$$

is an asymmetric copula allowing a heavy tail dependence. Furthermore, the Clayton copulas is known as a comprehensive copula as it can articulate all three fundamental copulas, specifically:

- If $\theta \to 0$, C_{θ}^{Cl} approaches the independence copulas, Π.

- If $\theta \to \infty$, C_{θ}^{Cl} approaches the comonotonicity copula, M.

- If $\theta \to 1$, C_{θ}^{Cl} approaches the countermonotonic copulas, W.

The *Frank copula,* with the generating function

$$\phi\left(u\right) = -\log\left(\frac{\exp\left(-\theta u\right) - 1}{\exp\left(-\theta\right) - 1}\right), \qquad \theta \neq 0 \qquad (4.9)$$

Copula family	Copula C_θ	Parameter	C_U	C_I	C_L
Normal	$\Phi_\theta^n(\Phi^{-1}(u_1),...,\Phi^{-1}(u_n))$,	$-1 \leq \theta \leq 1$	$\theta = 1$	$\theta = 0$	$\theta = -1$
t-copula	$\mathbf{t}_{\nu,\theta}(t_\nu^{-1}(u_1),...,t_\nu^{-1}(u_n))$	$-1 \leq \theta \leq 1$	$\theta = 1$	NA	NA
Clayton	$max\left[(u^{-\theta} + v - \theta - 1, 0)\right]^{\frac{1}{\theta}}$	$-1 \leq \theta \leq \infty$	$\theta \to \infty$	$\theta \to 0$	$\theta \to -1$
Frank	$\frac{1}{\theta} * \ln\left[1 + \frac{e^{\theta u}-1}{e^{\theta v}-1}e^{-1} - 1\right]$	$-\infty \leq \theta \leq \infty$	$\theta \to \infty$	$\theta \to 0$	$\theta \to -\infty$
Gumbel	$\exp\left[-\left((-\ln u)^\theta + (-\ln v)^\theta\right)^{\frac{1}{\theta}}\right]$	$1 \leq \theta \leq \infty$	$\theta \to \infty$	$\theta \to 1$	NA

Table 4.1: Definition of 5 copula families with their form, parameter spaces and respective comonotonicity copula C_U, independence copula C_I and countermonotonicity copula C_L.

(a) Clayton copula (b) Gumbel copula (c) Student t copula with $\nu = 1$ (d) Gaussian copula
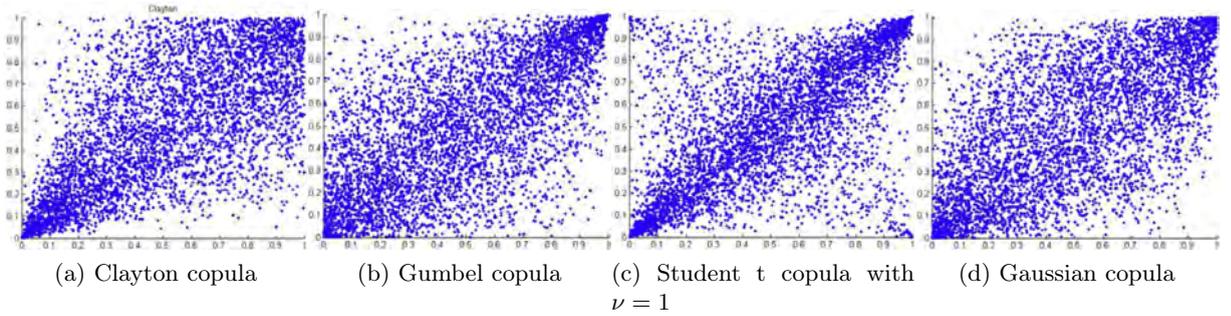
Figure 4.2: Scatter plots of the simulation of selected copula models with $\rho_\tau = 0.5$

4.2 Methodology for fitting copula to data

There are mainly three methodologies, when estimating the parameters (ρ, θ) of a copula family $\{C_{\rho_1,...,\rho_n,\theta,}\}$ to model the dependence between two random variables X and Y: firstly by using fully parametric method, known also as the "inference functions for margins" (IFM). The second is the semi-parametric and the third the moments estimators method. These methods are explained respectively in Shih and Louis [1995], Genest et al. [1995] and Genest and Favre [2007], and it is also reviewed from a practical point of view in the following section.

4.2.1 Parametric approach

The inference function for margins (IFM) method approach consists of estimating uni-variate parameters from separately maximizing uni-variate likelihoods, and

then estimating dependence parameters from separate bi-variate likelihoods or from a multivariate likelihood. Given F, G two marginal distributions with their respective densities f, g, parametrized with α_1, α_2 and copula C_θ, the bi-variate density function is given by $h_{\alpha_1, \alpha_2, \theta} = f_{\alpha_1}(x) g_{\alpha_2}(y) c_\theta(F_{\alpha_1}(x), G_{\alpha_2}(y))$. From the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, the likelihood to be maximised is:

$$
\begin{aligned}
L(\alpha_1, \alpha_2, \theta) &= \sum_{i=1}^{n} \log\{h_{\alpha_1, \alpha_2, \theta}(X_1, Y_1)\} \\
&= \sum_{i=1}^{n} \log\{f_{\alpha_1}(X_i)\} + \sum_{i=1}^{n} \log\{g_{\alpha_2}(X_i)\} + \\
&\quad \sum_{i=1}^{n} \log\{c_\theta(F_{\alpha_1}(X_i), G_{\alpha_2}(Y_i))\}.
\end{aligned}
\tag{4.10}
$$

Once the parameters of $\hat{\alpha}_1, \hat{\alpha}_2$ are calculated through the maximisation of $\sum_{i=1}^{n} \log\{f_{\alpha_1}(X_i)\}$ and $\sum_{i=1}^{n} \log\{g_{\alpha_2}(X_i)\}$, then $\hat{\theta}$ can be obtained by maximising the rest of the Equation 4.11.

One major advantage of the IFM method is that it makes inference for many multivariate models computably feasible as long as the margins are chosen correctly. Otherwise, this presents some important disadvantages which may cause the estimate $\hat{\theta}$ to be biased and a wrong estimation of the dependence parameter.

4.2.2   Semi-parametric approach

Similar to the method presented above, this approach overcomes the problem in the estimation of the marginals, but uses the empirical distribution function instead. Hence, this method maximises the rank based log-likelihood:

$$
l_\theta = \sum_{i=1}^{n} \log\{c_\theta(F_{\alpha_1}(X_i), G_{\alpha_2}(Y_i))\}
\tag{4.11}
$$

is also called the pseudo-maximum likelihood by Genest and Rivest (1993). It can be shown in Genest et al. [1995] that $\hat{\theta}$ is a consistent estimator of θ, as $n \to \infty$ and asymptotically normal distributed with mean θ and variance $\frac{\sigma^2}{n}$:

$$
\theta_n \approx \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right)
\tag{4.12}
$$

where σ^2 depends entirely on the true underlying copula C_θ (Proposition 2.1 of Genest *et al.*, 1995). Due to its advantages of being margin free and thus invariant with respect to increasing transformations, this approach is adopted for the work.

4.2.3   Method-of-Moments approach

Due to the limitations of computer simulation in its early days, statistical methods based on rank correlations are been developed in Iman and Conover (1983) to estimate the copula parameter θ. This technique is particularly useful when the multivariate data is limited and formal estimation of a full multivariate model is unrealistic. This general method explores a theoretical relationship between one of the rank correlations and the parameters of the copula and substitute empirical values of the rank correlation into this relationship to get estimates of some or all of the copula parameters [McNeil et al., 2005].

Copula	$\rho_\tau =$	$\rho_S =$
Gaussian	$\frac{2}{\pi} \arcsin(\rho)$	$\frac{6}{\pi} \arcsin(\frac{\rho}{2}) \approx \rho$
Student t	$\frac{2}{\pi} \arcsin(\rho)$	-
Gumbel	$1 - \frac{1}{\theta}$	-
Clayton	$\frac{\theta}{\theta+2}$	-
Frank	$1 - 4\theta^{-1}\left(1 - D_1\left(\theta\right)\right)$	$1 - \frac{12}{\theta}\left(D_1\left(\theta\right) - D_2\left(\theta\right)\right)$
Note: $D_k\left(x\right) = \frac{k}{x^k} \int_0^x \frac{t_k}{e^t-1} dt$ namely the "Debye" function		

Table 4.2: Calibration using rank correlations for 5 families of copulas, where ρ_s and ρ_τ are the rank correlations determined respectively through Spearman's rho and Kendall's tau, ρ and θ the copula parameters.

4.3   Simulating copulas

Monte-Carlo simulations are used to build pseudo observations and different copula models are applied to calculate the capital needed for this portfolio. The following steps illustrate how to proceed:

4.3.1   Algorithm for simulating implicit copulas

1. Calculate the Spearman's Rho and Kendall's Tau rank correlations.

2. For the Student t copula, use the maximum likelihood test to determine the degrees of freedom

3. Find the upper triangular matrix (A) of the correlation matrix (P) by using the Cholesky decomposition.

4. Generate n independent random variates $u_1, \ldots, u_n$ from $U(0,1)$.

5. Set $x = A\mathbf{u}$.

6. Calculate $\hat{F}^{-1}(u_i)$, for $i = 1, \ldots, n$, where $\hat{F}$ is the estimated marginal distribution function.

- $\hat{F}$ is Φ in the standard normal distribution for the Gaussian copula. t_ν for the student t distribution with ν degrees of freedom for the t copula.

  - If a sufficiently large database is available, the empirical distribution function may be taken.

7. Sum up the potential losses over the different risk types.

8. Compute the empirical distribution function and implement the goodness-to-fit.

9. Calculate the E(S) and VaR.

4.3.2   Algorithm for simulating Archimedean copulas

1. Perform the Spearman's Rho and Kendall's Tau rank correlations.

2. Simulate two independent U (0,1) random variates s and t.

3. Set $\omega = K^{-1}(t)$, where K is the distribution function C (u,v) and defined as $K(t) = t - \frac{\phi(t)}{\phi'(t^+)}$.

4. Set $\omega = \phi^{[-1]}(s\phi(\omega))$ and $v = \phi^{[-1]}((1-s)\phi(\omega))$.

5. Continue starting from step 6 of the previous algorithm.

4.4   Goodness-to-fit assessment

In order to find the copula, which fits the data the best, various goodness-to-fit tests [Genest and Favre, 2007], either graphical tools or formal tests have been developed. However, in the framework of this thesis, the fitted copula models is simply compared with the empirical data models as the empirical distributions are available.

4.4.1   Graphical tools

This is a simple application of the methods described in Section 3.2.3. In order to choose the most adequate model between several different copulas, it can examine, by plotting on the same graph, the simulated data set using the methodology in 4.3.

A second option, which is related to the K-plots, consists of comparing the empirical distribution K_n of the variables $W_1, \ldots, W_n$ with the estimated K_{θ_n}. And it can be illustrated how well they fit.

4.4.2   Formal blanket tests of goodness-to-fit

Formal tests methodologies for testing goodness-to-fit of copula models are just at its beginnings. In Genest and Favre [2007], a critical review of six rank-based blanket procedures for goodness-of-t testing for any class of copulas are proposed. Two of the tests discussed in this paper are based on the empirical copula, two are based on Kendall's transform and two are based on Rosenblatt's transform. To compare the relative power of these procedures, a Monte Carlo study involving a large number of copula alternatives and dependence structures have been carried out. The results showed that there was unfortunately no single test which is preferable to all the others, but the one based on the empirical copula and one of the procedures based on Rosenblatt's transform yield the best goodness-of-fit tests for copula models. However, the disadvantage with the latter is that it relies on a non-unique (and hence somewhat arbitrary) Rosenblatt's transform.

4.4.3   Test based on the empirical copula

The empirical goodness-to-fit tests compare the distance between the empirical copula C_n and the copula model C_{θ_n} given by $\|C_n - C_{\theta_n}\|$ over the whole data set:

$$S = n \int_{[0,1]^d} \{C_{\theta_n}(u) - C_n(u)\}^2 \, dC_{\theta_n}(u) = \sum_{i=1}^{n} \{C_{\theta_n}(U_i) - C_n(U_i)\}^2. \qquad (4.13)$$

Large values of S lead to the rejection of the copula model.

There are mainly three main tests for goodness-to-fit tests. In Malevergne and Sornette [2003], the authors propose two additional measures more suitable for copula models. Instead of looking at the maximum distance between both models, they prefer to take the average distance on the whole distribution.

1. The Kolmogorov distance is the maximum local distance along the quantile which most often occurs in the bulk of the distribution. The average distance examines the distance for the whole distribution.

2. Anderson-Darling distance puts the emphasis on the tails of the two distributions. The average Anderson-Darling distance also looks at the distance for the whole distribution.

3. Cramer-von-mises criterion is an alternative to the average Kolmogorov test.

The reader is referred to the paper of Malevergne and Sornette [2003] for more details about the test procedures or the test methodologies.

Chapter 5

Application to Windstorm Losses

5.1 Introduction

The theory studied in the two previous chapters is applied to the complete risk portfolio of PartnerRe's natural catastrophe business unit in the North America region. The data originates from simulated windstorm losses by using AIR CatRader version 11.0.1 on the complete PartnerRe's risk portfolio in coastal regions in the US. This data is hereafter denoted as "event loss data".

In this following section, a summary of the data is firstly provided, then followed by a rigorous assessment of dependence of the data by using independence testing, graphic visualizations of the dependence of the data, including the rank plot, Chi plot, Kendall's plot and tail dependence plot are all computed. In a further step, all three dependence measures are computed. These first results are compared with the portfolio model of PartnerRe. Using the same copula than PartnerRe to simulate the dependence structure of its risk portefolio, differences between both models have been discovered. These differences are discussed, alternative methodologies are proposed, goodness-to-fit tested, before concluding on some recommendation for PartnerRe's future risk management operations.

5.2 Summary of the data

As described in Section 2.4.3.2, the dataset is composed of 8'005 data points, each representing a simulated North-Atlantic windstorm occuring in one of the three region. A sample of the dataset is illustrated in Table 5.1. The columns 4-6 represents the losses, the windstorm caused in that respective region, and a 0 means simply that there were no losses.

Event #	Event ID	Frequency	Event Loss occured in NE region	Event Loss occured in SE region	Event Loss occured in S region
1	02agulf_1019	0.000288095	44'782	0	845'382
2	02agulf_1033	0.000288095	31'378'898	0	0
3	02agulf_1870	0.000288095	735'660	48'672	27
⋮	⋮	⋮	⋮	⋮	⋮
8'005	gulf1931	0.000288095	3'136	776	2'970'128

Table 5.1: Sample of the dataset

From the summary of the data given in Table 5.2, one can notice the skewed and heavy tailed data, with mostly zero loss from the first quantile up to median and important losses starting from the third quantile. Looking at the mean and maximum losses, the South-East region appears to be the most exposed to hurricane losses, whereas the South region presents the most skewed distribution, which shows more damaging losses.

	North-East	South-East	South	Total
Min	0	0	0	0
1st Quantile	0	0	0	0
Median	0	0	0	8
Mean	333	476	187	997
3rd Quantile	0	3	3	236
Max	33'064	25'831	23'995	33'064
Standard Deviation	2'193	2'111	1'112	3'235
Skewness (in unitary scale)	9.35	6.57	11.6	5.01

Unit: In millions of USD

Table 5.2: Summary of the event loss data of PartnerRe's risk portfolio in coastal regions in the USA.

5.3 Preliminary dependence analysis

5.3.1 Test of independence

Using the methodology described in Section 3.2.4 and the rank correlations presented in Section 5.4 to perform the test of independence two-by-two for the data set composed of 8'005 events, it is found that all the computed values are superior to $z_{\alpha/2} = 1.96$. Therefore one can confirm the rejection of H_0 hypothesis and prove that there exists a dependence relationship between the variables.

5.3.2 Graphical tools

The rank plot, Chi plot and Kendall's plot have been implemented in order to compare the dependence of the coastal regions two-by-two. However, only the rank plots and Chi plots are displayed below as it is not possible to compute the Kendall's plot properly owing to the enormous size of the dataset. As described in Section 3.2.3.3, the Kendall's plot computes the pair (W_i, H_i), where the equation W_i is defined by: $W_{i:n} = n \binom{n-1}{i-1} \int_0^1 \omega k_0(\omega) \{K_0(\omega)\}^{i-1} \{1 - K_0(\omega)\}^{n-1} d\omega$. Due to the exponential effect of i or n and as $n > 1050$, the function gets so small that the computation leads to an output of NaN in the software R. The reason behind, as explained in the lecture notes in statistical computing with R by Robert Gray @ Harvard is, as soon as the amount below 2^(-1022) is computed, the computer considers it as zero because numbers smaller than this have larger relative errors than machine precision. For future implementations of the Kendall's plot, the reader is advised to have a data set below 1'000 items.

For this reason, only the rank plots and Chi plots are displayed and discussed in the six graphs in Figure 5.1.
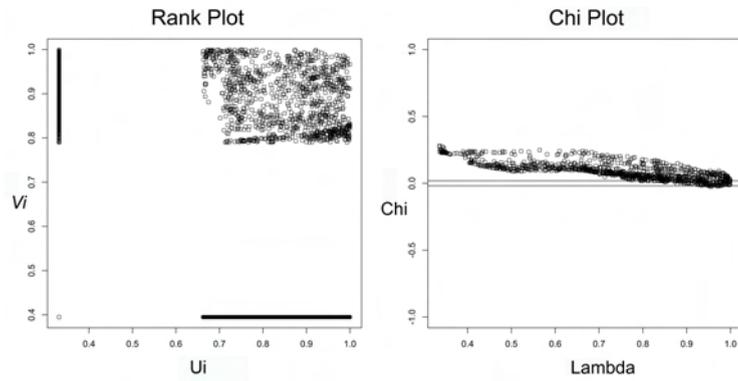
(a) South-East (U_i) and North-East (V_i) regions
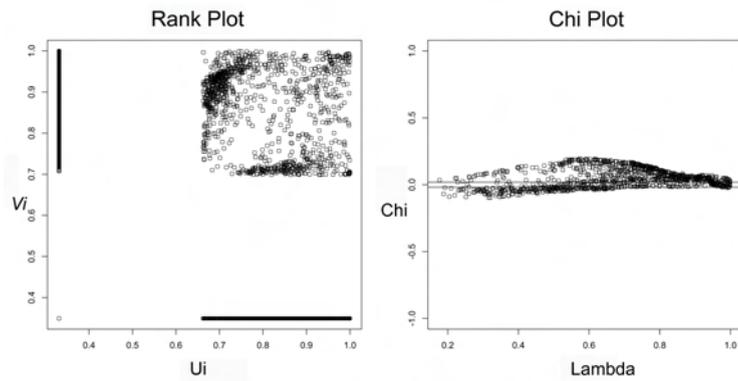
(b) South-East (U_i) and South (V_i) regions
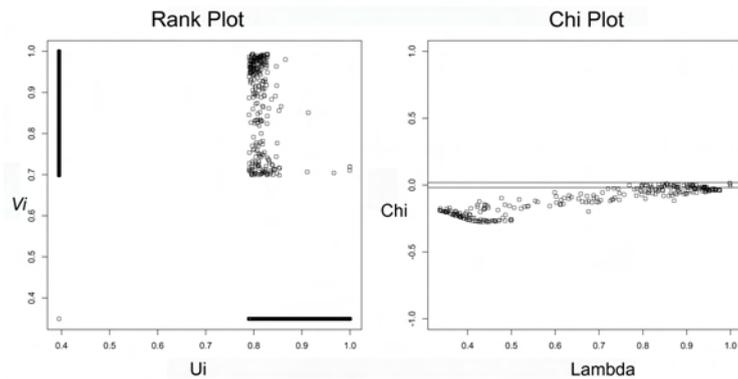
(c) South (U_i) and North-East (V_i) regions

Figure 5.1: Rank plot and Chi plot of event loss data

Discussion of the results

As a preliminary comment, one can observe a rather uncommon form of the rank plot which is only defined in a limited part of the space $[0, 1]^2$, furthermore the presence of two perpendicular lines on the axises may also draw the readers' attention. Despite the fact that the rank plot is defined on the whole space $[0, 1]^2$, the result illustrated above is due to the large number of zeros within the data. All the random variables

which both have zero losses are all assigned the same rank. If there are n ties, then the rank of all these elements will be n, and the number following n with be given the rank $n+1$. In the case of catastrophe model simulated data, there can be many observed ties with the value zero. The two perpendicular lines seen on the axises or the rank plot are a good example, they illustrate a loss for one of the random variables and zero for the other.

Analysis of graph 5.1a: South-East vs. North-East regions: The rank plot shows that there are three groups of data, which presents a higher data concentration. Two among which are near the axises x and y, which illustrate the hurricanes contributing to larger losses for one region will create smaller damage for the other. The third group can be found in the middle of the graph with similar scale of losses in both regions. This latter result suggest a good positive dependence between the two random variables. These first observations are confirmed by the Chi plot, which illustrates the positive dependence with the majority of the points falling above and outside of the confidence bands. One can however notice a tendency for the data to lean closer to the x axis as $\lambda \rightarrow 1$. This indicates a decrease in the dependence for very large losses and shows that the extreme events for one region tend not to incur catastrophic losses in both regions. This is rather unexpected, as such stronger hurricanes as Katrina tend to lead to more important damage in both regions. This remark will be developed in the course of the next chapters and discussed in the conclusion. The application of tail dependence to the data set is presented in depth in the Section 5.3.

Analysis of graph 5.1b: South-East vs. South regions: The similar three groups are also observed in the rank plot for this pair of data. Two groups among three are situated near the horizontal axis, which indicates that more damages in one region is occurring than in the other. In the third group, a concentration of data in the upper right corner can be observed. This illustrates significant amount of events causing important damages to both regions and shows the first signs of tail dependence. This will be discussed in further detail the next section. The Chi plot illustrates positive dependence compensated with a small portion of the data negatively correlated. But in overall, the random variables should show positive dependence in the losses.

Analysis of graph 5.1c: South vs. North-East regions: The rank plot shows two groups of data on the upper and lower parts of the graph, which originates from large losses in one region and very small losses in the other. Geographically, these are two very far apart regions, where the windstorms tend not to strike simultaneously unless for extreme events. From the Chi plot, it appears that these two

regions are negatively dependent. Finally, again, there doesn't seem to be signs of tail dependence for these two regions.

5.3.3   Tail Dependencies

In the previous paragraph, the dependencies of the whole distribution are plotted. This however can not be generalized to the behavior in the tail. Based on empirical assumptions of natural catastrophes, tail dependencies tend to increase the more destructive catastrophic events become and it has so far not been observed.

Ghoudi, Khoudraji and Rivest [Ghorbal et al., 2009] proposed a test for measuring the dependence at the tails, which determines if it can be characterized by an extreme-value copula. This latter can be considered to provide models for the dependence structure between rare events. Extreme-value copulas not only arise naturally in the domain of extreme-value theory, they can also be a convenient choice to model general positive dependence structures. For further details about this methodology or extreme-value copulas, the reader is referred to Gudendorf and Segers [2009]. The null hypothesis being tested is the dependence structure fitting an extreme value copula against the alternative that it is not. However after going through the paper, one can realize that this test can only be applied to distributions with continuous margins, which is unfortunately not the case in the present situation owing to the high amount of zeros within the data. Instead an empirical plot is constructed to visualize the dependencies in the tails.

5.3.3.1 λ_u Tail dependence coefficient

One can recall the definition of upper tail dependency:

$$
\begin{aligned}
\lambda_u(X,Y) &= \lim_{u \to 1^-} P\left[X > F_X^{-1}(u) \mid Y > F_Y^{-1}(u)\right] \\
&= \lim_{u \to 1^-} \frac{P\left[X > F_X^{-1}(u) \cap Y > F_Y^{-1}(u)\right]}{P\left[Y > F_Y^{-1}(u)\right]}.
\end{aligned}
$$

Given X and Y, two random variables with distribution F_X and F_Y. Using the second part of the equation, one plot for each quantile α, the empirical number of realizations of X which is greater than $F_X(\alpha)$ given that the corresponding values of Y are greater than $F_Y(\alpha)$, with a convergence towards the tail dependence when $\alpha \to 1$.

In the framework of this thesis, only the upper tail for tail dependency is analysed, as the lower tail, which represents the smallest losses is less relevant for capital allocation decisions.

The graphics displayed in Figure 5.2 represent the tail dependencies for quantile $\alpha > 0.5$ of the dataset. One can notice that the dependency coefficient λ_u decreases significantly, as $\alpha \to 1$. This indicates the absence of tail dependence and that an important catastrophe in one region will not create losses of the similar scale in another region.
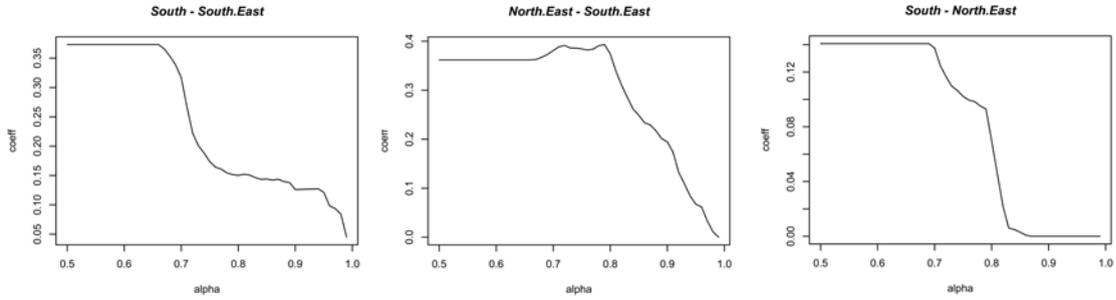
(a) λ_u Tail dependence coefficient
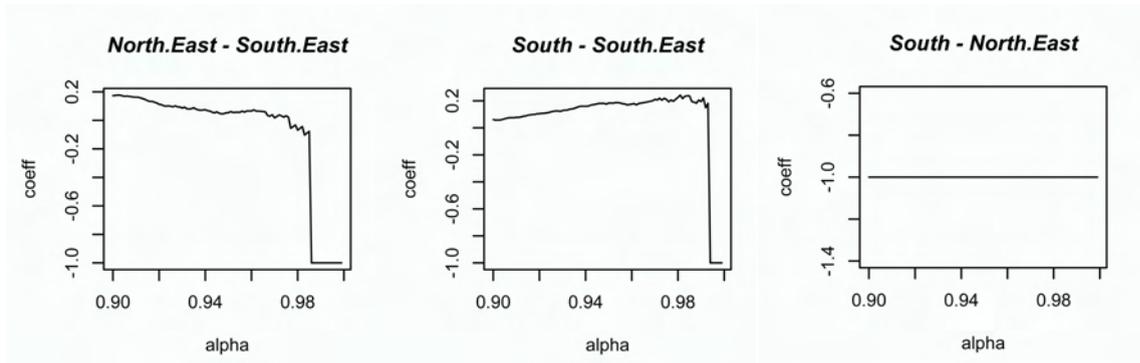
(b) $\bar{\lambda}_u$ Tail dependence coefficient

Figure 5.2: Tail dependencies of event loss data

5.3.3.2 $\bar{\lambda}_u$ Tail dependence coefficient

In order to confirm the absence of tail dependence assessed previously, the $\bar{\lambda}_u$ dependence coefficient is computed and with $\bar{\lambda}_u < 1$, the hypothesis of an absence of dependence in the tails of the marginal distributions in these regions is confirmed.

5.4   Dependence measurement

In this section, the dependency measurements such as the linear correlation, then rank correlations between the random variables have to be computed. Both dependence measurements, as illustrated in Table 5.4 will be used later on to fit the

adequate copula to the data. Using the methods of moments approach as described in Section 4.2.3, one can fit the dependencies measured for different copula families.

	ρ		
	S-E	N-E	S
South-East (S-E)	1	-0.01	0.038
North-East (N-E)	-0.01	1	-0.026
South (S)	0.038	-0.026	1

(a) Linear correlation

	ρ_s			ρ_τ		
	S-E	N-E	S	S-E	N-E	S
South-East (S-E)	1	0.27	0.08	1	0.23	0.07
North-East (N-E)	0.27	1	-0.19	0.23	1	-0.17
South (S)	0.08	-0.19	1	0.07	-0.17	1

(b) Spearman's rho (ρ_s) and Kendall's tau (ρ_τ)

Table 5.3: Dependencies measured for the event loss data

As an example, using the dependence measured from Pearson's correlation coefficient and the rank correlations to form a Gaussian copula fitting the event loss data, it can be observed that despite different correlation matrices, both models fit the data rather well as shown in Figure 5.3. Despite this good result, one should be careful when using the Pearson's linear correlation as it can lead to a wrong estimation of the dependence for non-normal distributed data, as mentioned in the flaws regarding linear correlation (Section 3.2.1.1).
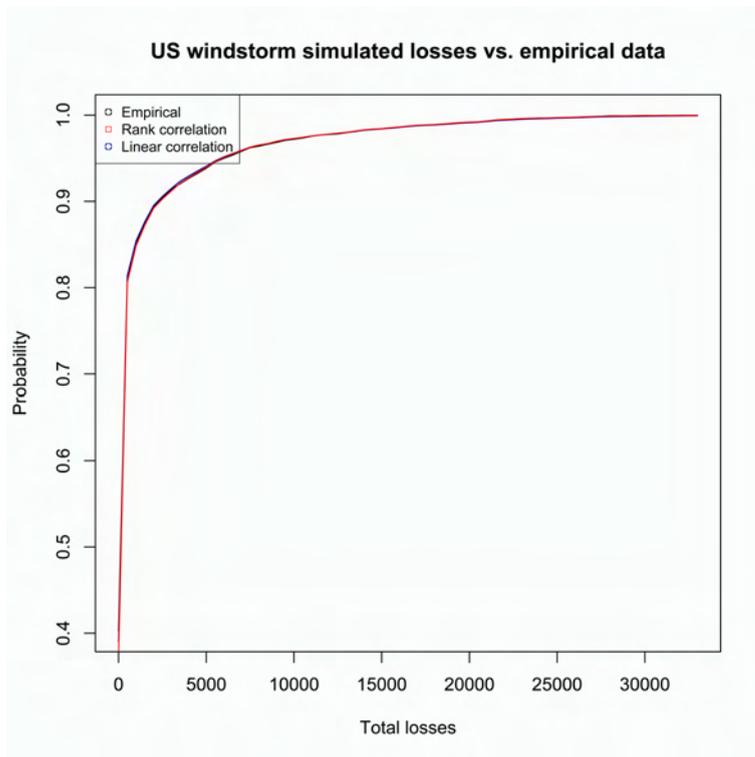
(a) For the whole distribution
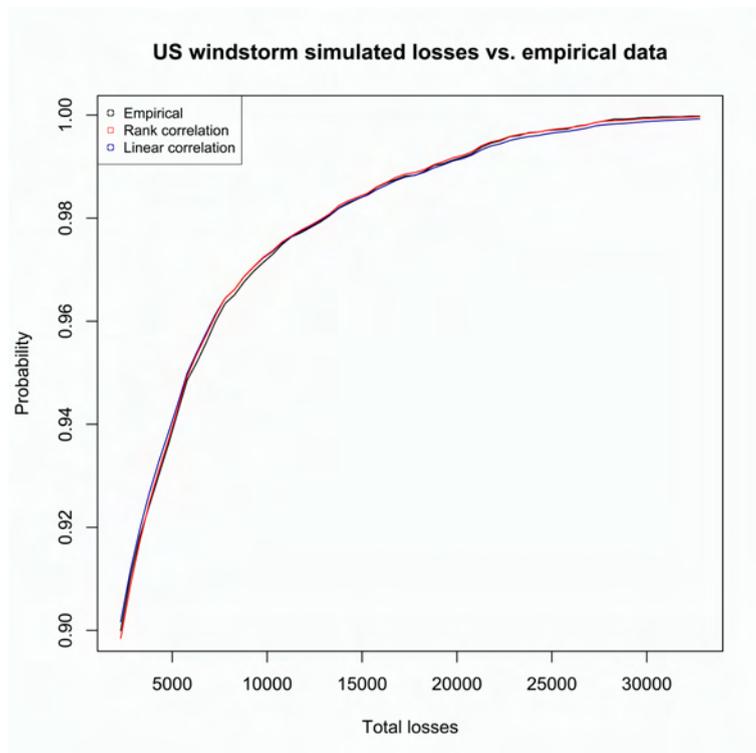
(b) Starting from the 0.9th quantile

Figure 5.3: Fit for a Meta-Gaussian copula using linear correlation and rank correlations compared to the simulated aggregated sum across regions of the event loss data

5.5   Comparison with PartnerRe's internal portfolio model

PartnerRe's portfolio model assesses the company's overall natural catastrophe risk exposure based on a Meta-Gaussian copula with a correlation matrix as shown in Table 5.4. Before performing any copula inference, a comparison has been performed to compare the simulated Meta-Gaussian copula using PartnerRe's portfolio model with a Meta-Gaussian copula using the underlying dependencies calculated previously for the event loss data. It is expected to find similar distributions. Let X be the aggregated sum of event loss of the South-East, North-East and South regions, the Value-at-Risk and Expected Shortfall are computed by using the European solvency standards at 99.5% to compare the thresholds of each model. As it can be seen in Table 5.5, the disparities start to become rather consequent above the 90% quantile. For $ES_{99.5}(X)$, an overestimation of the capital allocation of 9.6 billion USD can be observed. Due to the disparity with the portfolio model, the research begins by trying to explain these differences.

	ρ		
	S-E	N-E	S
South-East (S-E)	1	0.36	0.15
North-East (N-E)	0.36	1	0.02
South (S)	0.15	0.02	1

Table 5.4: Dependencies of PartnerRe's portfolio model using a Meta-Gaussian copula

5.5.1   Discussion

Ties in the data

One initial interrogation has been raised about the impact of the high amount of zeros within the data and the way to treat them. One can recall that copulas do not make any assumptions about the form of marginal distributions, but about the form of the relationships between marginals. This relationship is calculated by using the dependence measures discussed in Section 3.2.1 using Sklar's theorem (Theorem

	Empirical aggregation of three zones	Simulation using PartnerRe's correlation matrix.	Difference
E(X)	984	1'007	-23
$VaR_{90\%}(X)$	2'285	2'176	109
$VaR_{95\%}(X)$	6'000	5'819	181
$VaR_{99\%}(X)$	18'713	19'445	-732
$VaR_{99.5\%}(X)$	22'240	23'495	-1'255
$VaR_{99.9\%}(X)$	27'792	33'062	-5'270
$ES_{99.5}(X)$	**25'913**	**29'731**	**-3'818**
$ES_{99.9}(X)$	30'274	39'872	-9'598
Skewness	5.02	5.73	

Unit: In millions of USD

Table 5.5: Comparison between the sum across regions of the event loss data and the sum across regions of a simulated Meta-Gaussian copula using the dependencies from PartnerRe's portfolio model.

3.4), $F(x_1, ..., x_d) = C_\theta(F_1(x_1), ..., F_d(x_d))$, which links the marginal distributions to the multivariate distribution. However, this link or copula is **unique** only if the marginals are continuous random variables. In the situation of a continuous random variable, the probability of each point is zero and having ties in the sample would mean that the distribution is not monotone, hence not continuous. Without this assumption, it seems that the copula is unique, which seems to be the case with the data used in this thesis. Hence the rank correlation computed does not represent the unique underlying dependence of the data.

Attempts to overcome the problem

This problem of ties in computing rank correlations is the object of some recent papers Denuit and Lambert [2005], Neslehova [2007] and a tour in the most advanced research in the field of non continuous rank correlations provided only a limited solution to this problem. Denuit and Lambert [2005], Neslehova [2007] proposed generalizations to compute Spearman's rho and Kendall's tau for arbitrary non-continuous random variables. Their solution presents the following two difficulties, on one hand, these newly created measures do not reach the bounds ±1 for counter-monotonic and comonotonic copulas. The second difficulty lies in the convergence of their method to the real rank correlation. Back testing shows that even with a large n, the convergence is slow, sometimes not constant and limited to the body of the data. In fact, this method is applicable for the tails of the distributions as the quantile $q \to 1$ or $q \to 0$, the measure $\rho_\tau \to 0$.

Without a generalised solution possible, there have been various attempts, specific to the present data set, to overcome the problem related to ties within the

data. The first idea consisted simply of deleting them, as suggested in many literature articles. If the ties appear in a small amount in proportion with the rest of the data, this would indeed have a negligible effect on the dependence structure. However in the case here, almost half of the database consists of zeros either for one region or for both regions and deleting them would lead to a biased estimation of the dependence structure. For the data with zero losses in both regions, one can argue their relevance in assessing dependence and only "dilute the data". But as the three analyzed regions only represent partly of the event data set and as a simulation could have produced a non-zero damage in other regions considered by the model. Thus, throwing away these events would also induce a further bias in the inference results.

After quite some time spent in cogitating and discussing with colleague researchers, no solution seemed to appear until one discussion with a PhD student working on a similar set of data. His solution is based on a two-dimensional space and is presented in the next paragraph. For the ones interested in the original work, it can be found at Canestaro [2010].

Solution proposition for the problem

To overcome this difficulty with the following ties $(0,0)$, $(0,y)$, $(x,0)$, the complete data set is split in four parts, as illustrated below in Figure 5.4. Using the transformation of their marginal distributions (Proposition 3.3), the data is transformed onto the square unit; Part A is be the set of data with no damages $(0,0)$ simulated for both regions. Part B and C are the marginal data of losses in only one region and no damage in the other. And finally part D collects the data with losses in both regions. For the simulation of part A, B, C, it can be easily computed by using their marginal empirical distributions. As no ties remain in the remaining part D data, dependency modeling through copula theory can be applied.

For the simulation of the complete data set, using conditional probability, the probability of occurrence of each part are taken into account the probability of having a Part A, B, C or D data. As the simulation of part A, B and C are trivial, only the results of copula inference in part D data is illustrated in depth in the following.
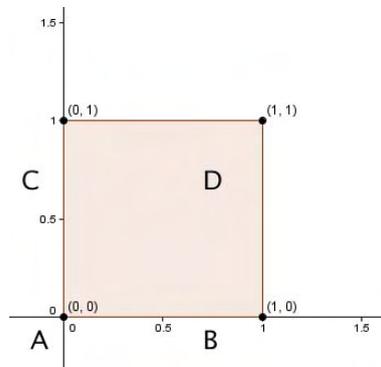
Figure 5.4: The data set split into 4 parts

5.6   Applying new methodology

Taking for starters the previously presented methodology applied to a bi-variate example of the North-East and South-East regions, the data is first split into these four categories. In the first case, the part of data containing (0,0) is simulated in proportion with the original data. For the two other categories (x,0) and (0,y), their empirical marginal distribution is then used to simulate the data from an uniform distributed random variable. For the case of (x,y), as no ties still remain - it is now a continuously distributed data, hence satisfying the conditions of Sklar's theorem. This allows to use the tools available in copula inference in order to fit the data to the most appropriate copula.

	All data	Part A: Data of the form (0,0)	Part B: Data of the form (x,0)	Part C: Data of the form (0,y)	Part D: Data of the form (x,y)
Size	8'005	5'103	1'888	783	1'018

Table 5.6: Amount of data for each part

5.6.1   Rank correlation for part D data

Considering now the part D of the data exclusively, the absence of ties within this part of the data set guarantees finding an unique copula describing the dependence structure of the data. Then by computing the pair wise rank correlations, the **Spearman's Rho** (ρ_s) and **Kendall's Tau** (ρ_τ) are computed between the South-East vs. North-East regions. Tables 5.6 illustrates these findings and shows a negative dependence between these two peril zones. The paper draws the readers'

attention that this does not mean that the catastrophe events happening in these two regions are negatively correlated. But instead it is the modeled losses on the whole PartnerRe's risk portfolio, based on a set of catastrophe events for treaties in this region, which are negatively dependent. This is rather a good sign of portfolio optimization, as uncorrelated risks are taken in the portfolio.

	ρ_s		ρ_τ	
	S-E	N-E	S-E	N-E
South-East (S-E)	1	-0.34	1	-0.23
North-East (N-E)	-0.34	1	-0.23	1

Table 5.7: Rank Correlation applied to for part D of event loss data for the South-East and North-East regions

5.6.2   Fitting copula to part D data

Using one million simulation points and the algorithm presented in the previous chapter, various copulas models are fit to the empirical data. It appears that the Meta-Gaussian copula fits the data the best both in the body and in the tail of the distribution. From the theory, the Meta-Gaussian copula is well known to be tail independent, which confirms once more our assumptions about the tail independence of data. This also explains why the tail dependent copulas such as the Student-t, Clayton and Gumbel copulas are not suitable for modeling our dependence structure as they largely overestimate for the tail region. Table 5.7 presents the simulated copula summed across regions and compared with the original data set or "event loss data".

	Gaussian Copula with Gaussian Marginals	Meta-Gaussian copula	Student-t copula ($\nu = 264$)	Clayton copula ($\theta = -0.37$)	Frank copula ($\theta = -2.17$)	Gumbel copula ($\theta = 1$)	Aggregated sum across regions of event loss data
Min	-18'460	2.284e-05	1.253e-05	0.202	4.670e-06	5.031e-06	0.03
1st Quantile	3'180	178	151	152	191	103	163
Median	7'473	1'112	1062	980	1'109	890	1'105
E(X)	7'464	4'022	4'066	3'934	4'029	4'017	4'046
3rd Quantile	11'780	5'096	5'108	4'755	5'062	4'898	5'086
$VaR_{90\%}(X)$	15'581	12'918	13'273	12'804	12'916	13'173	13'123
$VaR_{95\%}(X)$	17'903	20'153	20'404	20'634	20'208	20'712	20'128
$VaR_{99\%}(X)$	22'304	27'345	27'992	27'594	27'700	29'726	27'386
$VaR_{99.5\%}(X)$	23'911	29'804	30'565	30'565	30'448	32'994	29'310
$VaR_{99.9\%}(X)$	27'016	33'062	34'684	38'220	34'561	42'147	33'038
$ES_{99.5}(X)$	**25'838**	**32'271**	**33'773**	**35'774**	**33'707**	**37'819**	**31'277**
$ES_{99.9}(X)$	28'850	35'189	39'579	43'876	39'610	46'066	33'061

Unit: In millions of USD

Table 5.8: Fitting copula to part D of event loss data

5.6.3 Goodness-to-fit test for part D data

Using the Cramer-von Mises two-sample criterion, the goodness to fit test looks at the differences between the aggregated sum across regions of simulated loss data and the fitted copula models. As shown in Table 5.8, the Meta-Gaussian appears to be the copula with the best fit to the data. The Student-t copula model is ranked second in the goodness-to-fit test. Recall that for the Student-t copula, the higher the degree of freedom, the more independent the tails are. Using the maximum likelihood function, the best fitting parameter ν for the Student-t copula is 264, whose behavior approaches the Meta-Gaussian copula.

	Gaussian Copula with Gaussian Marginals	Meta-Gaussian	Student-t $(\nu = 264)$	Clayton $(\theta = -0.3)$	Frank $(\theta = -2.17)$	Gumbel $(\theta = 1)$
Cramer-von Mises criterion	231	0.356	0.36	0.918	0.708	11.273

Table 5.9: Cramer-von Mises criterion for goodness-to fit of various copula models

In order to illustrate the goodness-to-fit, the graph illustrated in Figure 5.5 shows the cumulative distribution function of the aggregated losses of the South-East and North-East regions of the US, simulated through various copula models compared with the empirical aggregated losses (or raw data) from the catastrophe models. It is plotted from the 0.9th quantile and it can be observed per quantile that the Meta-Gaussian has the best goodness-to-fit compared to the empirical aggregated losses. The Student-t copula comes second. But with such a high degree of freedom ($\nu = 264$), one can argue that its behavior tend to resemble the Meta-Gaussian copula. All other simulated copula models show less successful results, especially the Gaussian copula with standard normal marginal distribution. From this example it can be seen the importance of modelling the marginal distribution correctly.
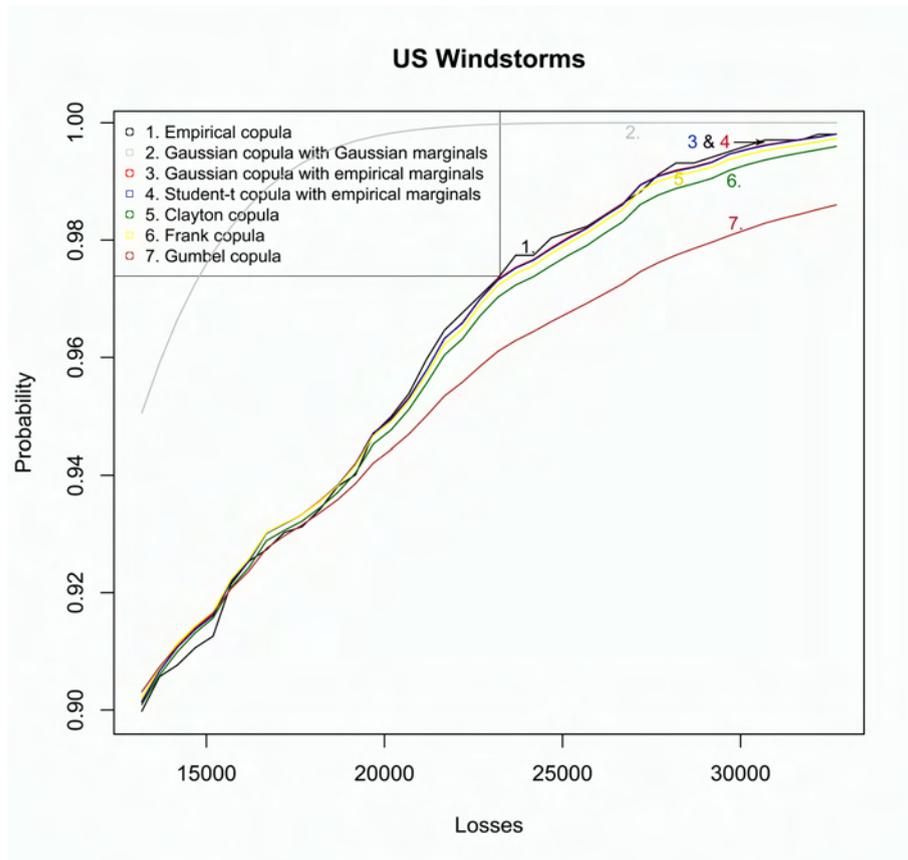
(a) From 0.9th quantile and onwards
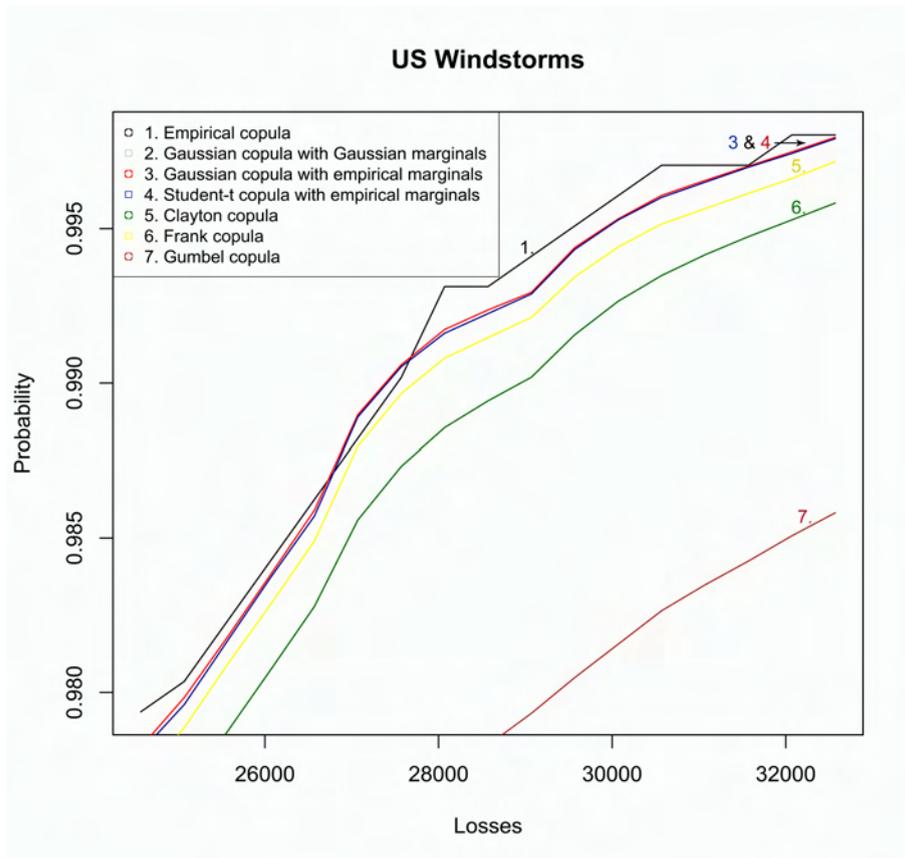
(b) From the 0.98th quantile and onwards

Figure 5.5: Comparison between the fit of copula models on the aggregated empirical windstorm losses in the USA at the tails

5.6.4   Simulation of all data parts A, B, C, D

After the statistical inference conducted for part D data, the data of part A, B, and C have also been simulated according to their probability in the whole data set. Table 5.10 provides the summary of the simulated losses across regions and all parts versus the aggregated across regions of the event loss data. As observed, the overall fit from the body to the tail of both distributions are very good. Using the European Solvency II standard, the expected shortfall at 99.5% gives only an error of 0.05% in comparison with the empirical data.

	Aggregated sum across regions of the simulation of all parts	Aggregated sum across regions of event loss data	Difference
Min	0	0	0
1st Quantile	0	0	0
Median	0	0	0
E(X)	796	780	16
3rd Quantile	41	40	1
$VaR_{90\%}(X)$	1'483	1'478	5
$VaR_{95\%}(X)$	5'011	4'993	18
$VaR_{99\%}(X)$	17'434	17'808	-374
$VaR_{99.5\%}(X)$	21'463	21'367	96
$VaR_{99.9\%}(X)$	27'494	27'664	-170
$ES_{99.5}(X)$	**25'543**	**25'558**	-15
$ES_{99.9}(X)$	30'604	30'274	330

Unit: In millions of USD

Table 5.10: Comparison between aggregated sum of event loss data and the simulated data for all parts A, B, C, D

5.7   From two dimensions to three dimensions

Previously, an example applied to a two-dimensional data set is illustrated. In a next step, this methodology will be extended to the complete portfolio of all three regions. In a similar way, the three dimensional data set will be split in eight different parts, namely (0,0,0), (x,0,0), (0,y,0), (0,0,z), (x,y,0), (x,0,z), (0,y,z), (x,y,z):

	All data	Form (0,0,0)	Form (x,0,0)	Form (0,y,0)	Form (0,0,z)	Form (x,y,0)	Form (x,0,z)	Form (0,y,z)	Form (x,y,z)
Size	8'005	3'544	1'058	783	1'559	815	830	0	216

- For the first fours parts containing maximum one univariate distribution ((0,0,0), (x,0,0), (0,y,0), (0,0,z)) these parts can be simulated by using their empirical distribution in proportion with their contribution to the whole dataset.

- For the bi-variate distributions ((x,y,0), (x,0,z), (0,y,z)), these three parts are simulated by using the aforementioned methodology described using the example between North-East and South-East regions, in Section 5.6. A bi-dimensional copula inference has been conducted together with the goodness-to-fit test.

- For the form (x,y,z), the search for the best copula model leads to the result of a Meta-Gaussian copula once again.

5.7.1 Results and discussion

After fitting and simulating for each part, it is found that the Meta-Gaussian copula provides the best goodness-to-fit to the empirical data. Aggregating all data parts together, Table 5.11 illustrates the aggregated sum of the event loss data compared with the aggregated sun for the event loss data. Significant and inconstant disappearances can be observed in the result of the simulation, which raises interrogations about the scalability of this methodology.

On one hand, the whole data set is split into many smaller parts, which can lead to less precise copula and statistical inference, on the other hand, the multiple copula estimations, which approximates each time the empirical data by using copula models, summed together may lead to the observed differences in Table 5.11.

Risk based capital measure	Aggregated sum across regions of the simulation of all parts	Aggregated sum across regions of event loss data	Difference
Min	0	0	0
1st Quantile	0	0	0
Median	0	7	-7
E(X)	710	984	-274
3rd Quantile	10	223	-213
$VaR_{90\%}(X)$	1'043	2'286	-1243
$VaR_{95\%}(X)$	4'444	6'001	-1557
$VaR_{99\%}(X)$	16'409	18'713	-2304
$VaR_{99.5\%}(X)$	21'382	22'241	-859
$VaR_{99.9\%}(X)$	27'730	27'792	-62
$ES_{99.5}(X)$	**25'774**	**25'914**	**-140**
$ES_{99.9}(X)$	33'365	30'274	3091

Unit: In millions of USD

Table 5.11: Comparison between the simulation and the empirical data for a tri-dimensional data set

Chapter 6

Conclusion

Within the framework of this thesis, the whole methodology on how to manage catastrophe risks within a reinsurance company is illustrated. A key element of this is the dependence analysis between different regions, which allow the reinsurance to better mitigate risk through optimizing their risk portfolio diversification. By conducting statistical inference to model the portfolio dependence structure, the data originating from catastrophe models is used instead of the one from historical events. The reason is mainly due to the lack of data of the latter, which makes any statistical work very difficult.

From the data of catastrophe models, it is observed that the Gaussian copula using the empirical marginal distributions presents the best fit for the data used. This confirms the initial concerns in observing no tail dependence among the random variables, putting this in context of the current climate situation showing extreme weather in many parts of the world. Extreme events appear to be happening much more often and more consistently than ever before. This summer, the heat waves baked the eastern part of the United States, parts of Africa and eastern Asia, above all Russia, which lost millions of acres of wheat and thousands of lives in a drought worse than any other in the historical record. Yet the current statistical and mathematical tools for modeling extreme events might not be the most adequate to deal with such extreme events, as these outliers are very difficult to predict based on the data of the past two hundred years only. Hence there is thus an urgent need to improve current models for statistical extremes, and to develop new methodology, in order to better understand, predict and manage these risks. Currently extremal modeling for single stationary time series is now well-established, but methods for multiple time series and spatial data are less well-developed, and a major effort is needed to develop approaches for dealing with realistic problems, in applications where spatial-temporal variation and non-stationarity are key elements [Davison,

2009].

Another recommendation for catastrophe modelers is about their internal construction. Most of the geophysical models only use linear correlation [Dong, 2001] to describe dependencies. This might explain the results why the Gaussian copula fits the data the best. However as mentioned in Section 3.2.1.1, linear correlation presents many flaws in assessing dependence and proves to be inappropriate in many cases. For a better estimation of dependence, rank correlations are more suitable for measuring non-linear dependence.

Thirdly comparing the simulated results by using the correlation matrix from PartnerRe's portfolio model and the data originating from AIR CatRader, a similar behavior in the body can be noticed. But significant disappearances can be found especially in the tail of the distribution, as illustrated in Table 5.5. The portfolio model presents heavy tail behavior, which may reflect the experience of portfolio managers when modeling catastrophe loss data. In the framework of this thesis, using our computed rank correlations, significant differences have been found with the one from the portfolio model (Table 5.3). However, its validity can not be guaranteed due to the ties within the data. In any case, it is important to note that these dependencies vary in time according to the new treaties signed in the region as well regional and climate variations. Hence a regular adaptation of these parameters is necessary.

Finally the methodology proposed to simulate catastrophe risk data is applicable for a two-dimensional data set. However it seems to be difficultly scalable beyond, which unfortunately limits its implementation in practice for multi-dimensional data sets.

Proposals for future research

One of the biggest difficulties in this research is to find the appropriate way to treat the ties at zero. As currently there is not an available tool for non-continuous copula inference or semi-continuous copula inference, a suggestion would be to first estimate how much continuous data and semi-continuous data may differ having the same dependence structure. This can give us a good estimate of our relative error in using traditional methods such as Kendall's Tau and Spearman's Rho and signal the necessity to conduct alternative inference methodologies. The methodology applied within the framework of this thesis has been applied to a two and three dimension data set. The multi-dimensional criteria are crucial, considering that tropical cyclones can impact many countries at the same time. However this procedure may be limited to lower dimensions in case of not sufficient amount of data.

A separate direction to take is working on the generalization of a Kendall's Tau or Spearman's Rho for semi-continuous data only. Looking at the difficulties for finding a solution for non-continuous data in general Neslehova [2007], it would be an interesting approach to constrain the research to semi-continuous data with ties only at the data point 0. This would however provide a very limited solution but nevertheless very useful for the reinsurance industry.

List of Tables

List of Figures

Bibliography

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Math. Finance*, 9(3):203–228, 1999.

Davide Canestaro. PhD dissertation, UniversitaÕ Degli Studi Di Firenze, 2010. This PhD Thesis has not yet been published.

S. Coles, J. Heffernan, and J. Tawn. Dependence measures for extreme value analyses. *Extremes*, 2:339–365, 169, 255, 1999.

European Commission. Background to the solvency 2 project, May 2010.

Anthony Davison. Risk, rare events and extremes. 2009. Website.

Michel Denuit and Philippe Lambert. Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, 93:40–57, 2005.

W. Dong. *Modern Portfolio Theory with Application to Catastrophe Insurance*, chapter Building a More Profitable Portfolio. Reactions Publishing Group, London, 2001.

Catherine Donnelly and Paul Embrechts. The devil is in the tails: actuarial mathematics and the subprime mortgage crisis. *Astin Bulletin*, 40(1):1–33, 2010.

P. Embrechts, A. McNeil, and D. Straumann. *Risk Management: Value at Risk and Beyond*. Correlation and dependence in risk management: properties and pitfalls, 2002.

Paul Embrechts. Copulas: A personal view. *Journal of Risk and Insurance*, 76: 639–650, 2009.

N.I. Fisher and P. Switzer. Chi-plots for assessing dependence. *Biometrika*, 72(2): 253–265, 1985.

N.I. Fisher and P. Switzer. Graphical assessment of dependence: Is a picture worth a 100 tests? *Am. Stat*, 55(3):233–239, 2001.

Chief Risk Officiers Forum. A framework for incorporating diversification in the solvency assessment of insurers. *The Chief Risk Officer Forum*, pages 203–228, 2005.

C. Genest and J.-C. Boies. Detecting dependence with kendall plots. *Am. Stat*, 57 (4):275–284, 2003.

C. Genest and J. Mackay. The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician*, 40:280–283, 2003.

C. Genest, K. Ghoudi, and L.-P Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distribution. *Biometrika*, 82: 543–552, 1995.

Christian Genest and Anne-Catherine Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12:347–368, 2007.

Noomen Ben Ghorbal, Christian Genest, and Johanna Neslehova. On the ghoudi, khoudraji and rivest test for extreme-value dependence. *The Canadian Journal of Statistics*, 37:534–552, 2009.

Patrici Grossi and Howard Kunreuther. *Catastrophe Modeling: A New Approach to managing risk*. Huebner International Series on Risk, Insurance and Economic Security. Springer, 2006.

Gordon Gudendorf and Johan Segers. Extreme-value copulas. *arXiv.org*, 2009.

Glyn A. Holton. Defining risk. *Financial Analysts Journal*, 60(6):19–25, 2004.

Douglas Hubbard. *The Failure of Risk Management: Why It's Broken and How to Fix It*, chapter 5. John Wiley and Sons, New Jersey, 2009.

H. Joe. *Multivariate models and dependence concepts*. Chapman and Hall, London, 1997.

Stanley Kaplan and B. John Garrick. On the quantitative definition of risk. *Risk Analysis*, 1(1), 1981.

M. G. Kendall. A new measure of rank correlation. *Biomedica*, 30:81–89, 1938.

Y. Malevergne and D. Sornette. Testing the gaussian copula hypothesis for financial assets dependences. *Quantitative Finance*, 3:231–250, 2003.

Y. Malevergne and D. Sornette. *Extreme Financial Risks*, pages 99–246. Springer, 2005.

Alexander J. McNeil, Ruediger Frey, and Paul Embrechts. *Quantitative Risk management*, chapter 1,2,5,6, pages 184–234. Princeton University Press, New Jersey, 2005.

USA National Hurricane Center. Hurricane history, 2009.

Roger B. Nelson. Dependence modeling with archimedean copulas. *Proceedings of the Second Brazilian Conference on Statistical Modeling in Insurance and Finance*, pages 45–54, 2005.

Roger B. Nelson. *An introduction to copulas*. Springer, New York, 2006.

Johanna Neslehova. On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis*, 98:544–567, 2007.

Johanna Neslehova. *An Introduction to Copulas*. ETHZ D-Math, 2009. Lecture Notes.

United States General Accounting Office. Catastrophe insurance risks: The role of risk-linked securities and factors affecting their use. Report to the chairman, committee on financial services, house of representatives, Harvard Business School and National Bureau of Economic Research, 2002.

M. Scarsini. On measures of concordance. *Stochastica*, 8:201–218, 1984.

B. Schweizer and A. Sklar. *Probabilistic Metric Spaces*. Elsevier,, New York, 1983.

J.H. Shih and T.A. Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51:1384Đ1399, 1995.

Emmett J. Vaughan and Therese Vaughan. *Fundamentals of Risk and Insurance*, chapter 1, pages 1–10. John Wiley and Sons, Inc., New Jersey, 1999.