Spring 2018

Models for high dimensional spatially correlated risks and application to thunderstorm loss data in Texas

Tobias Merk

Department:

MODELS FOR HIGH DIMENSIONAL SPATIALLY CORRELATED RISKS AND APPLICATION TO THUNDERSTORM LOSS DATA IN TEXAS

by

TOBIAS MERK

A THESIS

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

in

APPLIED MATHEMATICS

2018

Approved by

Dr. Akim Adekpedjou, Advisor
Dr. Gayla Olbricht
Dr. V. A. Samaranayake

ABSTRACT

Insurance claims caused by natural disasters exhibit spatial dependence with the strength of dependence being based on factors such as physical distance and population density, to name a few. Accounting for spatial dependence is therefore of crucial importance when modeling these types of claims. In this work, we present an approach to assess spatially dependent insurance risks using a combination of linear regression and factor copula models. Specifically, in loss modeling, observed dependence patterns are highly nonlinear, thus copula-based models seem appropriate since they can handle both linear and nonlinear dependence. The factor copula approach for estimating the spatial dependence reduces a complex dependence structure into a relatively easier task of estimating a spatial dependence parameter. Hence, we use a weighted sum of radial basis functions to model a spatial dependence parameter that determines the influence of each location. The methodology is illustrated using a thunderstorm wind loss dataset of Texas. Extensions to Matérn covariance functions and spatiotemporal models are briefly discussed.

ACKNOWLEDGMENTS

First of all I would like to thank my advisor Dr. Akim Adekpedjou. He always had time for me whenever I ran into problems and helped me understand specific topics. It was a pleasure working with you.

I also would like to express my gratitude to Prof. Hans-Joachim Zwiesler and Dr. Martin Bohner for giving me the opportunity to study at the Missouri University of Science and Technology. Thank you.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

Table Page

1. INTRODUCTION

Risk can be defined as *the possibility of loss, damage, injury, etc*. Everyone of us is exposed to some types of risks every day. Some of those risks are property loss related, some are health related and another type of risk may be losing his or her job due to an accident. To reduce the impact of those risks which ultimately will lead to financial losses, we buy insurance policies such as property insurance, life insurance and health insurance. The insurance transaction involves a policyholder (the person holding the insurance policy) and an insurer (the insurance company). One usually pays some amount of money (called *premium*) to the insurance company at a regularly scheduled interval in exchange for the insurer's promise to compensate the insured in the event of a loss. The loss may or may not be of financial nature, but it must be reducible to financial terms, e.g. a broken leg is not of financial nature but the cost for the operation is. An insurance risk is then represented by the uncertainty of how much money needs to be paid out to the insured. In mathematical terms, losses can be described by random variables, representing the amount of money that needs to be paid out to the policyholder.

Consider a portfolio with n policyholders. Assume that they are located in different regions such as zip codes, counties, cities and that they all have a property insurance. Of course those policyholders are independent. However, their location may have some similarities. The similarities between their locations can affect the event occurrences, thereby claim filing. Modeling the similarity between these locations enters the jurisdiction of spatial statistics because the similarity which occurs through space can be modeled with spatial correlation. This work pertains to the modeling of losses affected through space. Examples of space related losses are socio-economic factors or weather conditions. In the former case, assume the economy in a certain region is extremely bad, then there can be a

high correlation among unemployed people in nearby regions. This means, the portfolio of people holding unemployment insurances exhibits a high dependence for regions within a certain distance. In the latter case, a natural disaster in region A leading to policyholders filing claims can affect region B nearby. Consequently the same disaster will hit region B and claims at a similar magnitude will be filed as a result. So, accounting for spatial correlation in the modeling and analysis of claims is important. Ignoring the spatial dependence of such losses can result in wrong models and could lead to wrong estimates for reserves. The consequence of such an outcome would be insurance companies being unable to fulfill their promises of compensating the insured and can lead to ruin of the insurance company. It is therefore very important to factor the spatial aspect of the loss data into the analysis.

Investigating spatial data has always been a daunting task. Spatial dependence is expressed in terms of dependence between locations, so, it is a function of the distance. Techniques to investigate spatial data with dependence patterns include covariance structures based on distance between locations (Hua et al. [2017]). Covariance function and variogram (definition and details in Section 2) are two very popular approaches to do so. These are distance-based functions describing the degree of spatial dependence of an assumed underlying random field. Random fields are stochastic processes defined over a parameter space of dimensionality greater or equal to 2 taking values in an Euclidean space. Consider a family of random variables $\{X(t)\}_{t\in T}$, where $T \subseteq \mathbb{R}$ is some one-dimensional index set (usually time as a certain interval of the real line or a set of integers in the discrete case), then $X(t)$ is a random variable for each $t \in T$ and we call $\{X(t)\}_{t\in T}$ a stochastic process. In contrast, when $T \subseteq \mathbb{R}^p,\ p \geq 2$, $\{X(t)\}_{t\in T}$ is called random field. An example is given when $T = \{(Latitude, Longitude)\}$ is a set of locations, meaning that $X(t),\ t \in T$ is a random variable for each location. As a result, we obtain a "field" instead of a process.

Traditionally, these covariance-based methods are widely used in Geostatistics to describe spatial variability, e.g. to monitor groundwater quality, where nearby locations exhibit similar properties or in this case composition of water, i.e. high spatial dependence and for more distant locations we observe a much weaker (or even no) spatial dependence (Bárdossy [2006]). According to the "ArcGIS Pro" website (http://pro.arcgis.com/en/pro-app/help/analysis/geostatistical-analyst/what-is-geostatistics-.htm) "*Geostatistics* is a class of statistics used to analyze and predict the values associated with spatial or spatiotemporal phenomena. It incorporates the spatial (and in some cases temporal) coordinates of the data within the analyses".

Geostatistics can provide descriptive tools such as semivariograms to characterize the spatial pattern of continuous and categorical soil attributes (Goovaerts [1999]). In his paper, Goovaerts [1999] illustrated concepts such as sample semivariograms and the choice of an interpolation algorithm using multivariate soil data related to heavy metal contamination of an area of the Swiss Jura. "The growing interest of soil scientists in geostatistics arises because they increasingly realise that quantitative spatial prediction must incorporate the spatial correlation among observations" (Goovaerts [1999]).

Goovaerts et al. [2005] examined spatial variability of arsenic concentrations in groundwater in Michigan using a semivariogram approach. They showed a first step towards the assessment of the risk associated with exposure to low levels of arsenic in drinking water, specifically for the occurrence of bladder cancer. In addition, they used cross validation to assess the prediction performance. The disadvantage of this approach, however, is the fact that covariance-based structures can only account for linear dependence, not for nonlinear dependence. Moreover, the estimation of the variogram is a difficult task and empirical (based on observed data) variograms are sensitive to outliers (Bárdossy and

Kundzewicz [1990]). In a similar vein Yu et al. [2003] used variograms to assess concentrations of arsenic in water in Bangladesh to estimate arsenic-induced health effects.

Another application of (semi-) variograms is presented in Ly et al. [2011]. They developed different algorithms of spatial interpolation for daily rainfall data that was collected from 70 raingages within and surrounding the hilly landscape of the Ourthe and Ambleve catchments in Belgium over 30 years (1976–2005). Several semivariograms were fitted to daily rainfall data which were used to compare the interpolation performance of these algorithms based on validation raingages and cross validation. For each day, several variogram models were generated for all different raingages. One result of their study was that the semi-variance increased with larger separation distance, which implies that nearby rainfall data exhibits more similarity than those that are farther apart. This confirms the assumption of spatial dependence. While some applications consider hourly time steps to spatially interpolate with multivariate geostatistical method (Haberlandt [2007], Verworn and Haberlandt [2011]), others considered only monthly or annual time steps for spatial interpolation of precipitation (Goovaerts [2000]).

Bárdossy [2006] also dealt with a groundwater quality topic. He based his work on data collected from a large-scale groundwater quality measurement network in Baden-Württemberg (Germany). In contrast to covariance-based functions such as (semi-) variograms, he employed copulas to investigate the spatial dependence between four groundwater quality parameters, chloride, sulfate, pH, and nitrate. In his book, Nelsen [2006] describes copulas as functions that join multivariate distribution functions to their one-dimensional marginal distribution functions. "Copula" is a latin word that roughly translates to "link", "tie" or "bond". Alternatively, copulas can be seen as multivariate distribution functions whose one-dimensional margins are uniform on the interval $(0, 1)$.

Furthermore, Bárdossy [2006] calculated empirical copulas for the four mentioned groundwater parameter. For simulation and interpolation purposes, empirical copulas need to be fitted by theoretical ones (Bárdossy [2006]), which is why he compared the above computed empirical copulas to theoretical copulas and discovered that a non-Gaussian copula suites the data better. Bivariate empirical copulas are attractive alternatives to covariance-based functions, as corresponding rank correlations depict the strength of dependence independently of the marginals. Abe Sklar (1959) was the first to use the word copula in a mathematical or statistical sense (Nelsen [2006]). Copulas are specifically of interest because they can account for nonlinear dependence and dependence of heavy-tail random variables.

A very common familiy of copulas is the *Gaussian copula*. It is constructed from a multivariate normal distribution with a given correlation matrix. Since there is no closed formula for the *cumulative distribution function (cdf)* of the Normal distribution, the Gaussian copula has no analytical formula either. Another important class of copulas is the so called family of *Archimedean copulas*. Because of their simple form, the ease with which they can be constructed, and their many nice proerties, Archimedean copulas frequently appear in discussions of multivariate distributions (Nelsen [2006]). They also allow modeling dependence in arbitrary high dimensions, with only one parameter. Copulas have not been used in spatial contexts very often. They primarily attract interest in the financial sector (Embrechts et al. [2001]), where dependence between extremes can often be observed. They describe the dependence structure between random variables without information on the marginal distributions and are invariant to monotonic transformations of the marginals, which include logarithmic and/or Box-Cox transformations. This is a major advantage of copulas compared to the approaches using variograms, as those strongly depend on the marginal distribution (Bárdossy [2006]).

Other applications of copulas for spatial data include stochastic rainfall simulation that was presented by Michele and Salvadori [2003] and extreme value statistics discussed in Favre et al. [2004]. Based on the work of Joe [1996], Bedford and Cooke [2001], Bedford and Cooke [2002], and Kurowicka and Cooke [2006], Aas et al. [2009] used a pair-copula decomposition of a general multivariate distribution and propose a method for performing inference. Pair-copula decomposition is a procedure of decomposing a multivariate *probability density function (pdf)* of dependent random variables into a product of bivariate copulas. This decomposition is based on graphical models called vines and was introduced by Bedford and Cooke [2001] and Bedford and Cooke [2002]. The pair-copula approach in Aas et al. [2009] is applied to exhibit tail dependence in a financial dataset. Moreover, Aas et al. [2009] proposed a maximum pseudo-likelihood approach and corresponding algorithms for parameter estimation of the pair-copula decomposition.

Copulas offer an interesting opportunity to describe dependence structures for multivariate distributions (Bárdossy [2006]). Nevertheless, these copula-based approaches have their limits. "While the bivariate case is quite well understood, the estimation of higher dimensional copulas, however, is still an elaborate procedure" (Gräler and Pebesmaa [2011]). Some copulas can easily be extended to higher dimensions but many cannot. That is one reason, why the pair-copula decomposition is a powerful tool, because it is solely based on bivariate copulas which do not require higher dimensions. Built up on the pair-copula decomposition approach suggested by Aas et al. [2009], Gräler and Pebesmaa [2011] constructed bivariate copulas from a convex combination of copulas accounting for different distances by including the upper Fréchet-Hoeffding bound (which is a copula that describes perfect positive dependence) and the product-copula (which describes independence). Depending on the distance, these two famous copulas contribute to the strength of dependence that is assumed for the given distance. For smaller distancess the upper Fréchet-Hoeffding bound would play a major role due to assumed high correlation. Likewise, for large distances

the product-copula would contribute to the assumed independence. Another use of spatial data exploration using copulas is illustrated by Erhardt et al. [2015] who used vine copulas to model the dependence of temperature data between observation stations in Germany.

Hua et al. [2017] recently proposed a copula-based approach for assessing spatially dependent high dimensional risks using factor-copulas. Factor copula models for multivariate data are a recent development. Krupskii and Joe [2013] provided a detailed introduction to the topic and discussed several properties of factor copulas. They also talk about computational details and numerical issues for an implementation in the software R. Factor copula models are based on bivariate copulas that link observed data to latent variables. Spatially dependent data can exhibit very complex dependence structures. In contrast to commonly used geostatistical methods to model spatial dependence such as the semivariogram, copulas are especially able to capture nonlinear dependence. This is a major advantage of copulas. Moreover, copulas are very suitable for modeling non-normally distributed data such as insurance claims (McNeil et al. [2005]).

Insurance claims caused by natural disasters obviously exhibit high spatial dependence because nearby locations are affected in a similar vein and locations farther away may exhibit little to no effect of the same natural disaster. The strength of dependence may not only be determined by physical distance but also by population density since densely populated areas are more likely to exhibit more losses than sparsely populated areas. Hua et al. [2017] claim, their approach facilitates the challenge of modeling a complex spatial dependence structure into estimating a continuous function with spatial coordinates being the arguments. A good approach of estimating such a continuous function is using a weighted sum of radial basis functions that assigns a values to each location describing the influence

or effect of each single location compared to the others. We may call such a function spatial dependence parameter.

In their paper, they present two models, a spatial heterogeneity and a spatial dependence model. The former model is constructed via linear regression using date and population density as covariates. In addition, it also contains a spatial dependence parameter that is constructed via the discussed weighted sum of radial basis functions to explain the effect of each location. Adding a spatial dependence parameter to the regression model is necessary since linear models assume independent observations (which is obviously not the case here). The latter model is a spatial heterogeneity model that is based on the factor-copula approach. The proposed models are used to analyze a thunderstorm wind loss dataset consisting of insurance claims caused by thunderstorm winds in Texas (United States) in the years from 1996-2011. In addition, they briefly present extensions to spatiotemporal models and models for discrete data. In their main work, the value of the spatial dependence parameter at each location is only determined by the location. However, the spatial dependence parameter can be extended to have a spatial and a temporal component. This implies that every location would be assigned a value for the dependence parameter every time a loss occurred. A problem that may arise from this approach is the computational efficiency since it is already computational expensive without the temporal component. Another possible extension are so called "hurdle models" which are based on counting processes. The original approach ignores loss amounts of zero, however, one could use a Bernoulli random variable and a counting process to determine if and how many losses occur at given locations. More details about factor copula models are provided in Section 3.

The thesis is organized as follows. Section 2 introduces the mathematical concepts of copulas, spatially dependent data along with the theory of covariance-based estimators such as variograms, and other useful theory to facilitate the follow up of the thesis. The subsequent Section presents information on the thunderstorm dataset and a detailed analysis of the paper on which the work is based on. Concluding, a new approach based on Matérn covariance function will be briefly presented as a possible extension to investigate spatial data. Further extensions are given in Section 4, followed by a summary in Section 5.

2. METHODOLOGY

This section pertains to an introduction of some mathematical concepts as well as basic actuarial science concepts needed to facilitate the reading of this thesis. In the first section, we discuss terms such as risk, insurance policy, and insurance claim.

2.1. BASIC ACTUARIAL SCIENCE CONCEPTS

A *risk* is anything that has potential to lead to an unexpected adverse event or loss. So, a loss or severity is a random variable.

Definition 2.1.1 *(Loss)*
Let $(\Omega, \mathcal{F}, P)$ *be a probability space. A loss is a random variable* $X : \Omega \to \mathbb{R}$ *which assigns a scenario* $\omega \in \Omega$ *to a real value* $X(\omega) \in \mathbb{R}$.

Entities such as people or corporations buy insurance policies to safeguard against all or part of the financial losses that result from the occurrences of unexpected adverse events. Such unexpected events include fire, traffic accident, major illness, or natural disasters, e.g. hurricanes or tornados.

Definition 2.1.2 *(Insurance policy)*
An insurance policy is a contract between the insurer and the insured, known as the policyholder, which determines the claims which the insurer is legally required to pay in exchange for periodic payment, known as premium or insurance tariff.

Insurance is a way of redistributing the society's assets, which (in case when the party that suffered loss has an insurance policy) will help the suffered party, covering their loss on the credit of those policyholders who did not suffer the loss. Insurance policy provides a guarantee of full or partial compensation for specified losses, illness, damage, or

sudden death in return for a periodic payment commonly known as *premium*, as stated in the terms of the contract. Insurance policies insuring properties are called *non-life insurance* policies, whereas those insuring human beings for their health or sudden death are called *life insurance* policies. In this work, we will only focus on non-life insurance. In non-life insurance, people commonly insure their cars, homes, or business, as well as other types of properties.

Definition 2.1.3 *(Insurance claim & loss)*

- *An insurance claim is a random variable X such that* $P(X \geq 0) = 1$. *It is either filed by the policyholder addressing the insurer or filed by the insurance company (as the insured) addressing the reinsurance company.*
- *The term loss is used to denote the payment that the insurer makes to the policyholder for the damage covered under the policy. Thus, whenever we say that there was a "loss" under a policy, we mean that the policyholder received a payment from the insurer.*

The financial operations of an insurer can be viewed in terms of a series of cash inflow and outflow. The inflow components are added to the reservoir of assets, while the reservoir is depleted by the outflow components. On the one hand, main inflow components for an insurance company are *premiums* paid by the policyholders. On the other hand, the main outflow components are *insurance claims*, *reinsurance premiums* and other operating costs. A very basic surplus model of the insurer at time t is given by

$$U(t) = u + \pi(t) - S(t), \quad t \geq 0,$$

where,

- $u = U(0)$ is the starting capital
- $\pi(t)$ is the aggregate premium income up to time t
- $S(t) = X_1 + \cdots + X_{N(t)}$ is the aggregate claim, where $N(t)$ is the number of claims observed in $[0, t]$, and X_i is the random variable representing the claim amount of the i^{th} claim.

Definition 2.1.4 *(Ruin)*

Ruin occurs if the surplus is negative, i.e. for this specific model if $U(t) < 0$.

To avoid ruin, actuaries, who work for insurance companies develop insurance models for the likelihood of occurrence of events and statistical models for fair premiums needed to fulfill their commitment towards the policyholder when the underlying event in the insurance contract occurs. Establishing fair premiums begins with risk classification, which involves the grouping of risks into various classes that share a homogeneous set of characteristics allowing the actuary to reasonably obtain fair pricing for each category (Antonio and Valdez [2010]). Everything that is insured must be classified. Insurance companies do this because they want to be as accurate as possible when setting up a premium. If the premium is too high, policyholders that have few risky characteristics may drop their policy and the company would only be insuring risky policyholders. If it is too low, the company may not be able to pay out claims, when unexpected events arise. Risk classification may be based on age, gender, type of car, zip code, previous driving record in the car insurance for example to name a few. Common approaches to estimate money needed to sustain adverse outcome such as ruin is the simple and double chain ladder method.

In this thesis, copulas are used to assess association between losses in different regions. Therefore, we discuss various copula results in the next section.

2.2. COPULA

The concept of copulas in a mathematical or statistical sense was first employed by Sklar [1959] (Nelsen [2006]). A copula is defined as a function that "joins together" one-dimensional distribution functions to form multivariate distribution functions (see Theorem 2.2.3). This theory was further developed by many authors such as Dall'Aglio et al. [1991] and Schweizer [1991]. Following Nelsen [2006], a bivariate copula can be defined as follows.

Definition 2.2.1 *(Copula)*
Let $I = [0, 1]$ *denote the unit interval. A bivariate copula* $C : I^2 \to I$ *is a function that fulfills the following properties:*

1. *For every* $u, v \in I$,

$$C(u, 0) = 0 = C(0, v)$$

and

$$C(u, 1) = u \quad \textit{and} \quad C(1, v) = v$$

2. *For every* $u_1, u_2, v_1, v_2 \in I$ *with* $u_1 \leq u_2$ *and* $v_1 \leq v_2$,

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0.$$

Since copulas map their values to the unit interval I, they are obviously defined on $[0, 1]$. The French mathematician Maurice Fréchet and Finnish-born statistician Wassily Hoeffding independently from each other obtained the basic best-possible bounds inequality for these functions. This result is given in the next theorem just below.

Theorem 2.2.2 *(Fréchet-Hoeffding bounds)*

Let $C : I^2 \to I$ *be a Copula and let* $(u, v) \in I^2$. *Define* $W(u, v) = max(u + v - 1, 0)$ *and* $M(u, v) = min(u, v)$. *Then,*

$$W(u, v) \leq C(u, v) \leq M(u, v). \tag{2.1}$$

Proof: Nelsen [2006], theorem 2.2.3.

The bounds in (2.1) are copulas themselves and are called *Fréchet-Hoeffding bounds*, named after the above mentioned people. Another important copula that is frequently encountered is the *product copula* $\Pi(u, v) = uv$. A simple way of presenting the graph of a copula is with a *contour diagram*, i.e. with graphs of its level sets $\{(u, v) \in I^2 \mid C(u, v) = t\}$ for selected constants $t \in I$. Note that the points $(t, 1)$ and $(1, t)$ are each members of the level set. Thus, there is no need to label the level sets in the diagram, as $C(t, 1) = t = C(1, t)$ already provide the constant for each level set (Nelsen [2006]). Figure 2.1 shows the contour diagram of the copulas $M(\cdot, \cdot)$, $\Pi(\cdot, \cdot)$ and $W(\cdot, \cdot)$.

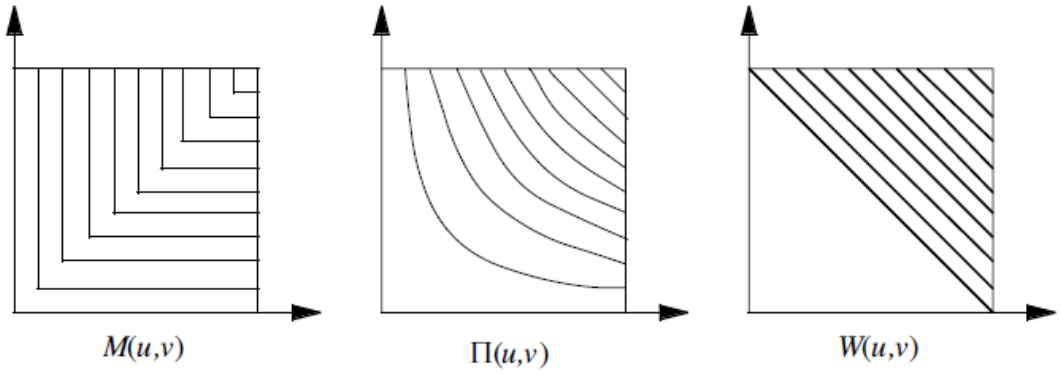

Figure 2.1. Contour diagrams of the copulas $M(\cdot, \cdot), \Pi(\cdot, \cdot), W(\cdot, \cdot)$

One may wonder how any of the above is related to probability theory, since everything so far is just deterministic. However, these results can be adopted to a statistical framework using random variables and distribution functions. In the statistical sense, a

random variable is a variable quantity whose values depend on chance and for which there exists a distribution function (Gnedenko [1962]). In what follows, random variables will be denoted by capital letters.

Consider a pair of random variables, X and Y, with cdfs $F_X(x)$ and $F_Y(y)$, respectively, and joint distribution function $H(x, y)$. To each pair of real numbers (x, y), we can associate three quantities: $F_X(x)$, $F_Y(y)$, and $H(x, y)$. In other words, each pair (x, y) leads to a point $(F_X(x), F_Y(y))$ in the unit square I^2, which corresponds to a number $H(x, y)$ in the unit interval I. This correspondence, which assigns the value of the joint distribution function to each ordered pair of values of the individual distribution functions, is indeed a function. Such functions are copulas (Nelsen [2006]). More useful than the formal definition of copulas is the link that can be established between copulas and distribution functions. This result is given in *Sklar's theorem* (Sklar [1959]).

Theorem 2.2.3 *(Sklar, 1959)*
Let $H(\cdot, \cdot)$ be a joint distribution function with margins $F(\cdot)$ and $G(\cdot)$. Then there exists a copula $C(\cdot, \cdot)$ such that for all $x, y \in \mathbb{R}$,

$$H(x, y) = C(F(x), G(y)). \tag{2.2}$$

If $F(\cdot)$ and $G(\cdot)$ are continuous, then $C(\cdot, \cdot)$ is unique: otherwise, $C(\cdot, \cdot)$ is uniquely determined on the space spanned by the range of $F(\cdot)$ and the range of $G(\cdot)$, i.e. $Range(F) \times Range(G)$.

Proof: Nelsen [2006], theorem 2.3.3.

From now on, assume $F(\cdot)$ and $G(\cdot)$ to be continuous. Equation (2.2) gives an expression for a joint distribution function in terms of a bivariate copula and two univariate distribution functions. Inverting this equation yields an expression of a copula in terms of

a joint distribution function and the inverses of its corresponding univariate margins. This means, a bivariate copula can be viewed as a joint distribution function of two random variables.

Corollary 2.2.4

Let $H(\cdot,\cdot)$, $F(\cdot)$, $G(\cdot)$ and $C(\cdot,\cdot)$ be as in the preceding theorem (Sklar) and let $F^{-1}(\cdot)$ and $G^{-1}(\cdot)$ be the inverse functions of $F(\cdot)$ and $G(\cdot)$, respectively. Then for any $(u,v) \in I^2$,

$$C(u,v) = H(F^{-1}(u), G^{-1}(v)). \tag{2.3}$$

Instead of working with the copula itself, one might rather consider its density. The copula itself does not allow an easy visualisation of the dependence , yet its density reveals the characteristics of the dependence (Bárdossy [2006]). The corresponding (bivariate) copula density is given by

$$c(u,v) = \frac{\partial C(u,v)}{\partial u \partial v},$$

or in terms of their corresponding *probability density functions (pdf)* and cdf

$$c(u,v) = \frac{h(F^{-1}(u), G^{-1}(v))}{f(F^{-1}(u))g(G^{-1}(v))},$$

where $h(\cdot,\cdot)$ is the joint density of X and Y, and $f(\cdot)$, $g(\cdot)$ are the univariate marginal densities of X and Y, respectively. Using the results from Sklar and Corollary 2.2.4, we show in an example how an expression of a bivariate copula of two random variable X and Y can be obtained, given the joint cdf $H(\cdot,\cdot)$ of X and Y.

Example 2.2.5 *(Gumbel's bivariate exponential distribution, (Gumbel [1960a]))*

Let H_θ be the joint distribution function given by

$$H_\theta(x,y) = \begin{cases} 1 - e^{-x} - e^{-y} + e^{-(x+y+\theta xy)} & , x \geq 0, y \geq 0, \\ 0 & , otherwise; \end{cases}$$

where $\theta \in I$ is a parameter. Then, for the marginal distribution functions we have

$F(x) = H_\theta(x, \infty) = 1 - e^{-x}$ *and analogously*

$G(y) = H_\theta(\infty, y) = 1 - e^{-y}$.

Therefore the margins are exponential with inverse functions $F^{-1}(u) = -\log(1-u)$ and $G^{-1}(v) = -\log(1-v)$ for $u, v \in I$. Hence, the corresponding copula is given by

$$C_\theta(u, v) = H_\theta(F^{-1}(u), G^{-1}(v)) = u + v - 1 + (1-u)(1-v)e^{-\theta \log(1-u)\log(1-v)}.$$

In the following theorem, we show that independent continuous random variables are characterized by the product copula $\Pi(u, v) = uv$. Its proof follows from equation (2.3) and the fact that X and Y are independent if and only if $H(x, y) = F(x)G(y)$.

Theorem 2.2.6 *(Independence)*

Let X and Y be continuous random variables. Then X and Y are independent if and only if their copula is identical to the product copula, i.e. $C_{XY}(\cdot, \cdot) = \Pi(\cdot, \cdot)$.

Consider again the preceding Example. Using the independence theorem, we can easily see that X and Y are not independent because their copula differs from the product copula $\Pi(\cdot, \cdot)$. In Section 1, we mentioned a major advantage of copulas compared to covariance-based methods, that is the invariance under strict monotonic transformation of the random variables. The following theorem describes this property in more details.

Theorem 2.2.7 *(Monotonic transformation)*

Let X and Y be continuous random variables with copula $C_{XY}(\cdot, \cdot)$. If α and β are strictly increasing on $Range(X)$ and $Range(Y)$, respectively, then $C_{\alpha(X)\beta(Y)}(\cdot, \cdot) = C_{XY}(\cdot, \cdot)$. Thus, $C_{XY}(\cdot, \cdot)$ is invariant under strictly increasing transformations of X and Y.

Proof: Nelsen [2006]

2.3. IMPORTANT CLASSES OF COPULAS

In this section, we present two very important classes of copulas: the Archimedean copulas and the Gaussian copulas. The former class contains a great variety of easily constructable copulas with many nice properties that we are going to see in the subsequent subsection, whereas the latter class consists of copulas that are rather complicated to handle because there exists no analytic closed form expression of these copulas. However, Gaussian copulas are important because they are constructed from the widely used multivariate Normal distribution and therefore describe the dependence structure of the multivariate Normal distribution.

2.3.1. Archimedean Copulas. Archimedean copulas are an important class of copulas that has a wide range of applications in finance and insurance due to a number of reasons: (1) They can easily be constructed and (2) the class subsumes many families of copulas, such as the Clayton, Frank, and Gumbel families, and (3), they possess nice properties (Nelsen [2006]). The class of Archimedean copulas allows for a great variety of different dependence structures including tail dependence, which is common in finance and insurance. They are also used in the field of Survival Analysis, where survival copulas in proportional hazard models are Archimedean (Segers [2013]) and in assessing portfolio credit risk (McNeil et al. [2005]). Furthermore, all commonly encountered Archimedean copulas have closed form expressions. For the definition of such a copula, we first need the concept of *pseudo-inverse*:

Definition 2.3.1 *(Pseudo-inverse)*
Let $\varphi : I \to [0, \infty)$ *be a continuous, strictly decreasing function such that* $\varphi(1) = 0$. *The* pseudo-inverse *of* φ *is the function* $\varphi^{[-1]} : [0, \infty] \to I$ *given by*

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t) & , 0 \leq t \leq \varphi(0), \\ 0 & , \varphi(0) \leq t \leq \infty. \end{cases} \tag{2.4}$$

Note that $\varphi^{[-1]}$ *is continuous and nonincreasing on* $[0, \infty]$*, and strictly decreasing on* $[0, \varphi(0)]$*. Moreover,* $\varphi^{[-1]}(\varphi(u)) = u$ *on* I *and*

$$\varphi(\varphi^{[-1]}(t)) = \begin{cases} t & , 0 \leq t \leq \varphi(0), \\ \varphi(0) & , \varphi(0) \leq t \leq \infty, \end{cases}$$

$$= min(t, \varphi(0)).$$

Finally, if $\varphi(0) = \infty$, then $\varphi^{[-1]}(\cdot) = \varphi^{-1}(\cdot)$.

Lemma 2.3.2 *(Archimedean copula)*

Let $\varphi : I \rightarrow [0, \infty]$ *be a continuous, strictly decreasing function such that* $\varphi(1) = 0$ *and let* $\varphi^{[-1]}$ *be the pseudo-inverse of* φ *defined by (2.4). Let* $C : I^2 \rightarrow I$ *be given by*

$$C(u, v) = \varphi^{[-1]}(\varphi(u) + \varphi(v)). \tag{2.5}$$

Then, C is a copula if and only if φ *is convex.*

Proof: Nelsen [2006], lemma 4.1.2 and theorem 4.1.4.

Copulas of the form (2.5) are called *Archimedean copula* with *generator* φ. The importance of this result is the characterization of the copula only through its generator. The behaviour of Archimedean copulas is completely determined through its generator. Hence, many properties and results of these copulas are only formulated in terms of the generator, rather than in terms of the copula itself. This often simplifies calculations. In the

following example, we show how to obtain an Archimedean copula from its generator φ. We only need to find the pseudo-inverse of the given generator followed by an application of equation (2.5).

Example 2.3.3

Let $\varphi(t) = -\log t,\ t \in I$. *Because* $\varphi(0) = \infty$, $\varphi^{[-1]}(t) = \varphi^{-1}(t) = \exp(-t)$. *Using notation (2.5) yields*

$$C(u, v) = \exp\Big(- \big[(-\log u) + (-\log v)\big]\Big) = uv = \Pi(u, v).$$

Thus, the product copula $\Pi(\cdot, \cdot)$ *is an Archimedean copula.*

In the same way, one can show that the lower Fréchet-Hoeffding bound $W(u, v)$ is an Archimedean copula as well. The corresponding generator is given by $\varphi(t) = 1 - t,\ t \in I$. As we can see, constructing Archimedean copulas is straightforward. We only need to find suitable generator functions fulfilling all the necessary properties. Theoretically any function $\varphi(\cdot)$ that fulfills the generator properties (such as continuity, strictly decreasing, etc. as given in the definition of Archimedean copula) can be used as a generator. However, not all of them may be usefull. We refer to Nelsen [2006] for a list of important Archimedean copulas including their generator functions. Many generator functions involve a parameter which gives them some flexibility. Examples of important one-parameter families of Archimedean copulas are (Nelsen [2006]):

- Clayton family:
  $C_\theta(u, v) = \big[max\big(u^{-\theta} + v^{-\theta} - 1, 0\big)\big]^{-\frac{1}{\theta}}, \quad \varphi_\theta(t) = \frac{1}{\theta}\big(t^{-\theta} - 1\big), \quad \theta \in [-1, \infty)\backslash\{0\}.$
  The Clayton copula was first introduced by Clayton [1978]. It is mostly used to study correlated risks because of their ability to capture lower tail dependence. Tail dependence is a dependence measure that looks at the concordance (see section 2.4) between extreme values (tail of the joint distribution) of the random variables X and Y. Special cases are $C_{-1}(\cdot, \cdot) = W(\cdot, \cdot)$, $C_0(\cdot, \cdot) = \Pi(\cdot, \cdot)$ and $C_\infty(\cdot, \cdot) = M(\cdot, \cdot)$.

- Gumbel family:

  $C_\theta(u,v) = \exp\Big(-\big[(-\log u)^\theta + (-\log v)^\theta\big]^{\frac{1}{\theta}}\Big), \quad \varphi_\theta(t) = (-\log t)^\theta, \quad \theta \in [1,\infty]$

  The Gumbel copula (Gumbel [1960b]) is used to model asymmetric dependence in the data. If outcomes are expected to be strongly correlated at high values but less correlated at low values, then the Gumbel copula is an appropriate choice. The Gumbel family is often used in stock market analysis (Mahfoud [2012]). Special cases are $C_1(\cdot,\cdot) = \Pi(\cdot,\cdot)$ and $C_\infty(\cdot,\cdot) = M(\cdot,\cdot)$.

- Frank family:

  $C_\theta(u,v) = -\frac{1}{\theta}\log\Big(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\Big), \quad \varphi_\theta(t) = -\log\Big(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\Big), \quad \theta \in \mathbb{R}\backslash\{0\}.$

  The Frank copula is suitable for modeling data characterized by weak tail dependence. This family can be used to fit bivariate loss distributions (Bouyé et al. [2000]). Special cases are $C_{-\infty}(\cdot,\cdot) = W(\cdot,\cdot)$, $C_0(\cdot,\cdot) = \Pi(\cdot,\cdot)$ and $C_\infty(\cdot,\cdot) = M(\cdot,\cdot)$.

Unlike Archimedean copulas, the Gaussian copulas do not have a closed form expression because it involves the inverse of the cdf of the standard Normal distribution. The next subsection gives more details on Gaussian copulas.

2.3.2. Gaussian Copulas.

Definition 2.3.4 *(Gaussian copula)*
Let $u = (u_1,\ldots,u_d) \in I^d$. The copula of the d-variate normal distribution is given by

$$C(u) = \Phi_d(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_d)),$$

where Φ^{-1} is the inverse of the univariate cdf of the standard normal distribution and

$$\Phi_d(x_1,\ldots,x_d) = \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_d}\frac{1}{\sqrt{(2\pi)^d\det\Sigma}}\exp\Big(-\frac{1}{2}(y-\mu)^\top\Sigma^{-1}(y-\mu)\Big)dy,$$

with $\det\Sigma$ *being the determinant of* Σ*, denotes the joint distribution function of the general d-variate standard normal distribution with* $y = (y_1, \ldots, y_d)$*, expectation vector* $\mu = (\mu_1, \ldots, \mu_d)$ *and covariance matrix* Σ*. In the bivariate case the copula expression can be written as*

$$\begin{aligned} C(u, v; \rho) &= \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \rho) \\ &= \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(- \frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \right) dxdy, \end{aligned}$$

where ρ *is the usual linear correlation coefficient of the corresponding bivariate normal distribution.*

2.4. DEPENDENCE CONCEPTS

Copulas provide a natural way to assess dependence between 2 random variables X and Y. Linear correlation (or Pearson's correlation) is most frequently used in practice to quantify dependence. However, since linear correlation is not a copula-based measure of dependence, it can often be misleading and should not be taken as the canonical quantity (Embrechts et al. [2001]). In the following, we are going to explore some copula-based measures of the degree of monotonic dependence between X and Y, also known as "measure of association" such as Kendall's tau and Spearman's rho. Both quantities are defined in terms of concordance and discordance. An explanation of concordance is given in the next definition.

Definition 2.4.1 *(Concordance)*

A pair of random variables (X, Y) *is said to be* concordant *if large values of X tend to be paired with large values of Y and small values of X to be paired with small values of Y. To be*

more precise, let (x_i, y_i) *and* (x_j, y_j) *be two observations from a pair of continuous random variables* (X, Y)*. Then, these two pairs of observations are* concordant *if* $(x_i - x_j)(y_i - y_j) > 0$ *and* discordant *if* $(x_i - x_j)(y_i - y_j) < 0$

The measure of association known as *Kendall's tau* is named after the British statistician *Maurice Kendall*, who developed it in 1938. As we will see in the definition below, it can be interpreted as the probability of concordance minus the probability of discordance.

Definition 2.4.2 *(Kendall's tau)*
Let $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ *be a random sample of n observations from a pair of continuous random variables* (X, Y)*. There are* $\binom{n}{2}$ *distinct pairs* (x_i, y_i) *and* (x_j, y_j)*,* $i \neq j$*, of observations in the sample, and each pair is either concordant or discordant. The sample version of Kendall's tau is defined as*

$$t = \frac{N_c - N_d}{N_c + N_d} = \frac{N_c - N_d}{\binom{n}{2}},$$

where N_c *denotes the number of concordant pairs and* N_d *the number of discordant pairs. The corresponding population version for pairs of continuous independent and identically distributed (*iid*) random variables* (X_1, Y_1) *and* (X_2, Y_2) *is defined as the probability of concordance minus the probability of discordance:*

$$\tau_{X,Y} = P\big[(X_1 - X_2)(Y_1 - Y_2) > 0\big] - P\big[(X_1 - X_2)(Y_1 - Y_2) < 0\big], \tag{2.6}$$

where obviously $-1 \leq \tau_{X,Y} \leq 1$.

For $\tau_{X,Y} = 1$, we have perfect positive monotonic correlation, i.e. $Y = f(X)$ for some monotonic increasing function f, and for $\tau_{X,Y} = -1$ we have perfect negative monotonic correlation, i.e. $Y = f(X)$, where f is some monotonic decreasing function. In order to demonstrate the role that copulas play in concordance and measures of association,

we first define a "concordance function" Q, which is the difference of the probabilities of concordance and discordance between two vectors (X_1, Y_1) and (X_2, Y_2) of continuous random variables with possibly different joint distributions H_1 and H_2, but with common margins F and G. We then show that this functions depends on the distribution of (X_1, Y_1) and (X_2, Y_2) only through their copulas (Nelsen [2006]).

Theorem 2.4.3 *(Concordance function)*
Let (X_1, Y_1) and (X_2, Y_2) be independent vectors of continuous random variables with joint distribution functions H_1 and H_2, respectively, with common margins F (for X) and G (for Y). Let C_1 and C_2 denote the copulas of (X_1, Y_1) and (X_2, Y_2), respectively, so that $H_1(x, y) = C_1(F(x), G(y))$ and $H_2(x, y) = C_2(F(x), G(y))$. Let Q denote the difference between the probabilities of concordance and discordance of (X_1, Y_1) and (X_2, Y_2), i.e.,

$$Q = P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - P\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right]$$

Then

$$Q = Q(C_1, C_2) = 4 \iint_{I^2} C_2(u, v) dC_1(u, v) - 1 \tag{2.7}$$

Proof: Nelsen [2006], theorem 5.1.1.

Using the preceding theorem, we can now formulate Kendall's tau in terms of copulas. Because (X_1, Y_1) and (X_2, Y_2) in (2.6) are identically distributed, both random vectors have the same copula C. The following theorem gives an expression of Kendall's tau for a pair of continuous random variables (X, Y) in terms of their copula.

Theorem 2.4.4 *(Copula expression of Kendall's tau)*
Let X and Y be continuous random variables with copula C. Then, the population version of Kendall's tau for X and Y (denoted by either $\tau_{X,Y}$ or τ_C) is given by

$$\tau_{X,Y} = \tau_C = Q(C, C) = 4 \iint_{I^2} C(u, v) dC(u, v) - 1. \tag{2.8}$$

Note that the integral in (2.8) can be expressed as the expected value of the function $C(U, V)$, where $U, V \sim U(0, 1)$ with joint distribution function C, i.e.

$$\tau_{X,Y} = \tau_C = 4\,\mathbb{E}(C(U, V)) - 1.$$

Proof: Nelsen [2006], theorem 5.1.1.

In general, evaluating Kendall's tau in (2.8) requires the evaluation of the double integral. For Archimedean copulas, however, a close form expression is provided in the next corollary:

Corollary 2.4.5

Let X and Y be random variables with an Archimedean copula C generated by φ. Then, the population version τ_C of Kendall's tau for X and Y is given by

$$\tau_C = 1 + 4 \int_0^1 \frac{\varphi(t)}{\varphi\prime(t)} dt. \tag{2.9}$$

One of the reasons, Archimedean copulas are easy to work with is that expressions are often formulated in terms of the generator rather than in terms of the copula (Nelsen [2006]). The following example shows how to compute Kendall's tau for the Clayton family.

Example 2.4.6

Let C_θ be a member of the Clayton family of Archimedean copulas with generator $\varphi_\theta(t) = \frac{1}{\theta}\left(t^{-\theta} - 1\right)$. Then for $\theta \geq -1$,

$$\frac{\varphi_\theta(t)}{\varphi\prime_\theta(t)} = \frac{t^{\theta+1} - t}{\theta} \text{ when } \theta \neq 0, \text{ and } \frac{\varphi_0(t)}{\varphi\prime_0(t)} = t \log t;$$

so that (using (2.9))

$$\tau_\theta = \frac{\theta}{\theta + 2}.$$

Spearman's rho, named after the English psychologist *Charles Spearman* is a non-parametric measure of rank correlation (statistical dependence between the rankings of two variables). Just like Kendall's tau, it assesses how well the relationship between two variables can be described using a monotonic function. As with Kendall's tau, the measure of association known as *Spearman's rho* is also based on concordance and discordance (Nelsen [2006]).

Definition 2.4.7 *(Spearman's rho)*
Let (X_1, Y_1)*,* (X_2, Y_2) *and* (X_3, Y_3) *be three independent random vectors with a common joint distribution function* H *(whose margins are* F *and* G*) and copula* C*. The population version* $\rho_{X,Y}$ *of Spearman's rho is defined to be proportional to the probability of concordance minus the probability of discordance for the two vectors* (X_1, Y_1) *and* (X_2, Y_3) *(the pair* (X_3, Y_2) *could also be used equally as well), i.e. a pair of vectors with the same margins, but one vector has distribution function* H*, while the components of the other are independent:*

$$\rho_{X,Y} = 3\Big(P\big[(X_1 - X_2)(Y_1 - Y_3) > 0\big] - P\big[(X_1 - X_2)(Y_1 - Y_3) < 0\big]\Big)$$

Moreover, $-1 \leq \rho_{X,Y} \leq 1$.

Note that while the joint distribution function of (X_1, Y_1) is $H(x, y)$, the joint distribution function of (X_2, Y_3) is $F(x)G(x)$ (because X_2 and Y_3 are independent). Thus, the copula of X_2 and Y_3 is the product copula $\Pi(\cdot, \cdot)$. Similar to Kendall's tau, using (2.7), this measure of association can also be expressed in terms of the corresponding copula (Nelsen [2006]).

Theorem 2.4.8

Let X and Y be continuous random variables with copula C. Then, the population version of Spearman's rho for X and Y is given by

$$\rho_{X,Y} = \rho_C = 3Q(C, \Pi) = 12 \iint_{I^2} uv dC(u, v) - 3 \tag{2.10}$$

$$= 12 \iint_{I^2} C(u, v) du dv - 3. \tag{2.11}$$

As with Kendall's tau, Spearman's rho can also be expressed in terms of Expectation. Furthermore, it can be related to Pearson's correlation coefficient.

Corollary 2.4.9

Let $X \sim F$ and $Y \sim G$, and let $U = F(X)$ and $V = G(Y)$. Obviously $U, V \sim U(0, 1)$. Then, (2.10) can be written as

$$\rho_{X,Y} = \rho_C = 12\,\mathbb{E}(UV) - 3 \; = \frac{\mathbb{E}(UV) - \frac{1}{4}}{\frac{1}{12}} \; = \frac{\mathbb{E}(UV) - \mathbb{E}(U)\,\mathbb{E}(V)}{\sqrt{var(U)}\sqrt{var(V)}}$$

$$= \frac{cov(U, V)}{\sqrt{var(U)}\sqrt{var(V)}} \; = \rho(F(X), G(Y)),$$

where ρ denotes Pearson's correlation coefficient. As a consequence, Spearman's rho for a pair of continuous random variables coincides with Pearson's correlation coefficient for the random variables $U = F(X)$ and $V = G(Y)$.

In fact, (2.11) can be rewritten as

$$\rho_C = 12 \iint_{I^2} [C(u, v) - uv] du dv, \tag{2.12}$$

which means ρ_C in (2.12) is proportional to the signed volume between the graphs of the copula $C(\cdot,\cdot)$ and the product copula $\Pi(\cdot,\cdot)$. Thus, ρ_C is a measure of "average distance" between the distribution of X and Y (represented by $C(\cdot,\cdot)$) and independence (represented by $\Pi(\cdot,\cdot)$) (Nelsen [2006]). One last useful theorem in this context talks about perfect correlation in terms of Kendall's tau or Spearman's rho.

Theorem 2.4.10

Let X and Y be continuous random variables with copula C, and let κ denote Kendall's tau or Spearman's rho. Then the following are true:

1. $\kappa(X,Y) = 1 \Leftrightarrow C(\cdot,\cdot) = M(\cdot,\cdot)$.
2. $\kappa(X,Y) = -1 \Leftrightarrow C(\cdot,\cdot) = W(\cdot,\cdot)$.

Proof: Embrechts et al. [1999].

For more details and information on copulas, we refer to Nelsen [2006].

2.5. SPATIAL DEPENDENCE

Before talking about spatial dependence, we first need to clarify what spatial (or spatio-temporal) data are. Spatio-temporal data are data provided with a geographical and a time component. The geographical component is usually given in form of coordinates (Latitude and Longitude) and the time as a date. However, spatial datasets that do not have a temporal dimension can occur in many areas of science, e.g. the spatial data may be from a "snapshot" in time (e.g., liver-cancer rates in U.S. counties in 2009, groundwater quality in Michigan in October 2005, etc.), or they may be taken from a process that is not evolving in time. One fundamental scientific problem that arises is understanding the evolution of processes over time, particularly in environmental studies (e.g., the evolution of sea-ice coverage in the Arctic and/or changes in sea level) (Cressie and Wikle [2011]).

Many models in various fields are built after analysing past data. For instance, insurance risks for extreme weather conditions such as floods or hurricanes can only be estimated after analysing data from the previous years. Moreover, insurance policies from the same region can be highly dependent, because such weather conditions in one region affect many people in the same way.

In the environmental sciences, proximity in space and time is a particularly relevant factor. "Nearby" is a relative notion, relative to the spatial and temporal scales of the phenomenon under study. For example, in the spatial case, a toxic-waste-disposal site may directly affect a neighbourhood of a few square kilometers; a coal-burning power plant may directly affect a heavily populated region of many tens of square kilometers, and an increase in greenhouse gases will affect the whole planet (Cressie and Wikle [2011]). To paraphrase a famous geographer named Waldo Tobler, while everything is related to everything else, things that are close together tend to be more related than things that are far apart. Terrain elevations, soil types, and surface air temperatures, for instance, are more likely to be similar at points two meters apart than at points two kilometers apart ("https://www.e-education.psu.edu/natureofgeoinfo/c1_p18.html"). This is called *spatial dependence*.

In the following we present a few functions that are commonly used to assess spatial dependence. First, we present a model-based measure of the spatial statistical dependence in a geostatistical process, called *variogram*.

Definition 2.5.1 *(Variogram)*
Let $h = (h_1, \ldots, h_d) \in \mathbb{R}^d$ *and let* $\left\{Y(s) \mid s \in D \subset \mathbb{R}^d\right\}$ *be a real-valued spatial process defined on a domain D of the d-dimensional Euclidean space* $\mathbb{R}^d$*, and suppose that differences of variables that are h units apart, vary in a way that depends only on h. Specifically,*

suppose that

$$2\gamma_Y(h) = var\Big(Y(s+h) - Y(s)\Big), \quad for\ all\ s, s+h \in D.$$

The qunatity $2\gamma_Y(h)$*, which is a function only of the difference h between the spatial locations s and* $(s+h)$*, is called stationary* variogram.

If $2\gamma_Y(h)$ can be written as a function of the euclidean norm of h (denoted as $\|h\|_2$), the variogram is said to be *isotropic*. Isotropic variograms are functions purely of the distance between two spatial locations, regardless of the direction, i.e. it does not matter if one location is east, west, etc. of the other. The only aspect that matters is the difference between the two locations. A *semivariogram* (i.e. one half the variogram) of $Y(\cdot)$ is given by $\gamma_Y(h)$. Another common function in this context is the so called *covariance function*.

Definition 2.5.2 *(Covariance function)*
Under the same assumptions as in the preceding definition, we can define the covariance function $C_Y(\cdot)$ *as*

$$C_Y(h) = cov\Big(Y(s+h), Y(s)\Big), \quad for\ all\ s, s+h \in D, \tag{2.13}$$

and specify the mean function to be constant, i.e.

$$\mathbb{E}\Big(Y(s)\Big) = \mu, \quad for\ all\ s, s+h \in D. \tag{2.14}$$

The restrictions (2.13) and (2.14) define the class of second-order stationary *processes in D, with stationary* covariance function $C_Y(\cdot)$.

Instead of the covariance function $C_Y(\cdot)$, researchers often prefer working with the *correlation function* $\rho_Y(\cdot)$ given by

$$\rho_Y(\cdot) = \frac{C_Y(\cdot)}{C_Y(0)}.$$

There exists a relationship between the semivariogram and the covariance function. Assuming the existence of the stationary covariance function given in (2.13), the seminvariogram of $Y(\cdot)$ exists and is given by

$$\begin{aligned} 2\gamma_Y(h) &= var\Big(Y(s+h) - Y(s)\Big) \\ &= var\Big(Y(s+h)\Big) + var\Big(Y(s)\Big) - 2cov\Big(Y(s+h), Y(s)\Big) \\ &= 2var\Big(Y(s)\Big) - 2cov\Big(Y(s+h), Y(s)\Big) \\ &= 2C_Y(0) - 2C_Y(h) \\ \Leftrightarrow \gamma_Y(h) &= C_Y(0) - C_Y(h), \quad h \in \mathbb{R}^d. \end{aligned}$$

This implies, that theoretically $\gamma_Y(0) = 0$. However, at an infinitesimally small separation distance, the semivariogram often exhibits a value greater than 0, that is $\gamma(h) \to c_0 > 0$ as $h \to 0$. c_0 has been called *nugget effect* by Matheron [1962]. It is believed by geostatisticians that this discontinuity can be made up of both measurement error and spatial dependence at scales smaller than the available distances between observations (Cressie and Wikle [2011]). In practice nothing can be said about the variogram at distances smaller than $\min\Big(\big\|s_i - s_j\big\|\Big)$ for $1 \leq i < j \leq n$. For more information on spatial statistics, we refer to Cressie and Wikle [2011].

3. REGRESSION MODELS FOR SPATIAL DATA

In Section 2, we introduced copulas, discussed some nice properties of these functions and discussed spatial dependence that comes along with spatial data. Moreover, we introduced spatial data as data that involves a geographical location usually in terms of Latitude and Longitude. Insurance risks such as thunderstorm winds contain a high level of spatial dependence as population density and geographical distance significantly affect the insurance losses (Hua et al. [2017]). One would assume that more insurance claims are filed at densely populated locations than at locations with lower populations densities. In addition, we can assume that if an insurance claim due to thunderstorm winds is filed at a given location s, it is very likely that other claims within a "small" distance to the location s are filed as well. In this context, "small" is a relative notion. Generally speaking, locations closer to s exhibit a higher probability of also filing a claim than locations further away from s. This implies that proximity in space yields a higher correlation and we therefore experience spatial dependence in the thunderstorm wind dataset that we are going to analyze. Assuming independent insurance claims for spatial data would not correctly reflect the reality and will therefore not be a good assumption.

In this Section we will discuss two approaches of modeling spatial data (beside the covariance based methods discussed in Section 2), a spatial heterogeneity and a spatial dependence model. The former model is based on a linear regression model and therefore models the expected loss at each location, while the latter uses a factor-copula approach to account for the dependence among losses at different geographic locations. The advantage of a copula-based model over a covariance-based model, is the accountability for nonlinear dependence (Hua et al. [2017]). The main challenge for modeling the spatial dependence among random losses at different locations is to construct feasible dependence structures.

Assume $Y(s),\ s \in D$, where D is the domain of locations s, is a random variable describing the loss at location s. The dependence structure between losses $Y(s_i)$ at different locations can be very complex. We will model a spatial dependence parameter based on radial basis functions (see section 3.2) to account for the spatial effects. The radial basis function approach assigns a higher importance to closer locations and therefore models the spatial dependence because random variables at locations close to each other should interact. Conclusively, the spatial dependence parameter describes the influence of a location compared to all other locations. This spatial dependence parameter can be estimated directly from the data.

Since linear regression models assume independence among different observations (which is obviously not given in our data), we may include a spatial dependence parameter into the regression model to account for this dependence. Assume s is a geographical location and $\theta(s)$ is a spatial dependence parameter describing the influence of location s. In this context, a general regression model is given by

$$\mu = \mathbb{E}[Y|s,t,x] = \theta(s) + \beta^\top X(t,x),$$

where Y is a random variable denoting the losses, s = (Latitude, Longitude) is a geographical location, t is a time index (e.g. date) and x denotes the population density. The design matrix $X(t,x)$ contains the covariates "time" and "population density" and $\beta = (\beta_0, \ldots, \beta_p)^\top$ is the parameter vector to be estimated.

In the spatial dependence model we use latent factors to account for the complex spatial dependence among locations. Latent factors are variables that are not directly observed, e.g. quality of life or happiness in economics. Assume the random variable V to be a latent factor that is connected with each $Y(s)$, then the dependence between the

two random variables $Y(s)$ and V can be modeled through a bivariate copula $C(u, v; \theta(s))$, with $\theta(s)$ being the spatial dependence parameter for location s. Then, we can assume that random variables at different geographic locations $Y(s)$ to be independent conditional on V, which implies that the dependence structure among $Y(s)$ can be obtained by integrating over the support of V. As a result, the factor copula approach does not model complex connections directly, but models each specific location separately and the latent factor V takes care of the interdependence among the locations (Hua et al. [2017]). After introducing both models, we are going to present a maximum likelihood estimation based on the joint likelihood function of the spatial heterogeneity and spatial dependence model in order to obtain optimal parameters.

In order to obtain a better fit for the data, we have to account for inflation and remove any potential outliers that may significantly affect the model.

3.1. THE DATA

The thunderstorm wind loss dataset we consider contains property damage losses due to thunderstorm winds in Texas, United States, from 1996 to 2013. These data are obtained from the National Climatic Data Centre (NCDC) of the National Oceanic and Atmospheric Administration (NOAA). According to the *NOAA National Severe Storms Laboratory (NSSL)* (a federal research laboratory under NOAA's Office of Oceanic and Atmospheric Research), a thunderstorm is a rain shower including thunder and lightning. There are several kinds of property damages associated with thunderstorm winds (*NSSL*):

- Flash flooding as a result of rainfall
- Fires caused by lightnings
- Car and window damages due to hail

- Strong winds (up to more than 120 mph) responsible for knocked down trees, power lines and mobile homes

The loss amounts (in US Dollars (USD)) are adjusted by the consumer price index (cpi) to the 2013 level to minimize the inflation effect. More specifically, denote the losses of each year by $X_1, \ldots, X_{18}$, where $X_1 = (X_{11}, X_{12}, \ldots, X_{1n_1})$ corresponds to the n_1 losses in 1996, $X_2 = (X_{21}, X_{22}, \ldots, X_{2n_2})$ corresponds to the n_2 losses in 1997, etc., and let $\alpha = (\alpha_1, \ldots, \alpha_{18})$ be the consumer price index to the 2013 level (e.g. $\alpha_1 = 1.485$ for an inflation increase of 48.5% from 1996 to 2013). Then, we adjust the losses as $Y_i = \alpha_i X_i$. This transformation does not change the underlying distribution of the losses since it is linear. In general, let $c > 0$ and let $F_X(x)$ be the cdf of the random variable X. Furthermore, assume a new random variable Y is created by multiplying X by a constant c, i.e. by $Y = cX$. Then,

$$F_Y(y) = P(Y \leq y) = P(cX \leq y) = P(X \leq y/c) = F_X(y/c)$$

and hence,

$$f_Y(y) = \frac{1}{c} f_X(y/c).$$

Therefore, multiplying random variables by a constant does not change the underlying distribution as we stay in the same family of distributions. However, the parameters may change.

Loss severities of zero are excluded from the dataset, i.e. we only consider nonzero losses. We also excluded observations from February 10, 1998 as Texas experienced a series of rare winter storms. These high amounts of property damage losses are very untypical for February in the remaining data (Hua et al. [2017]). The first few observations of the cpi adjusted data can be seen in Table 3.1 with property damages being displayed in thousand USD. If not otherwise mentioned, we always refer to property damage in thousand USD. The data consists of information on the County and City, where the property damage occurred, as well as the exact geographic location given as coordinates (Latitude and Longitude). For

coding convenience, the date is given in the *yyyy/mm/dd* format, where *yyyy* is the year, e.g. $yyyy = 1996$, *mm* corresponds to the month, e.g. $mm = 01$ for January and *dd* represents the day of the month, e.g. $dd = 17$ for the 17^{th} day of the corresponding month. For instance, $1996/01/17$ in Table 3.1 refers to January 17, 1996. This is not a common date format in daily life, but since we are rather interested in the year and month of the insurance claim instead of the day, it is more convenient to have the date available in this format.

Table 3.1. Thunderstorm Wind Loss Dataset

County	City	Date	Property damage	Latitude	Longitude
Hood Co.	Cresson	1996/01/17	7.425	32.53	-97.63
Hill Co.	Lake Whitney	1996/01/17	37.125	31.90	-97.38
Johnson Co.	Burleson	1996/01/17	111.375	32.53	-97.32
Tarrant Co.	Crowley	1996/01/17	7.425	32.58	-97.32
McLennan Co.	Waco	1996/01/17	37.125	31.55	-97.15
McLennan Co.	Waco	1996/01/17	2.970	31.55	-97.15
Hill Co.	Hillsboro	1996/01/17	22.275	32.00	-97.13
Tarrant Co.	Arlington	1996/01/17	22.275	32.73	-97.12

Table 3.2 shows a summary of the cpi adjusted property losses (without the already excluded rare February winter storms from 1998). We notice that the 3^{rd} quantile is relatively small and therefore at least 75% of the data is concentrated close to zero, yet the maximum is 400000. This implies that we are dealing with highly right skewed data, which are typical for insurance claims (McNeil et al. [2005]), especially with thunderstorm data that result in high claims.

Table 3.2. Summary of Loss Amounts in Thousands

Sample size	Min.	1^{st} quantile	Median	Std. Dev	Mean	3^{rd} quantile	Maximum
7554	0.01	4.29	10.82	4709.27	144.31	32.04	400000

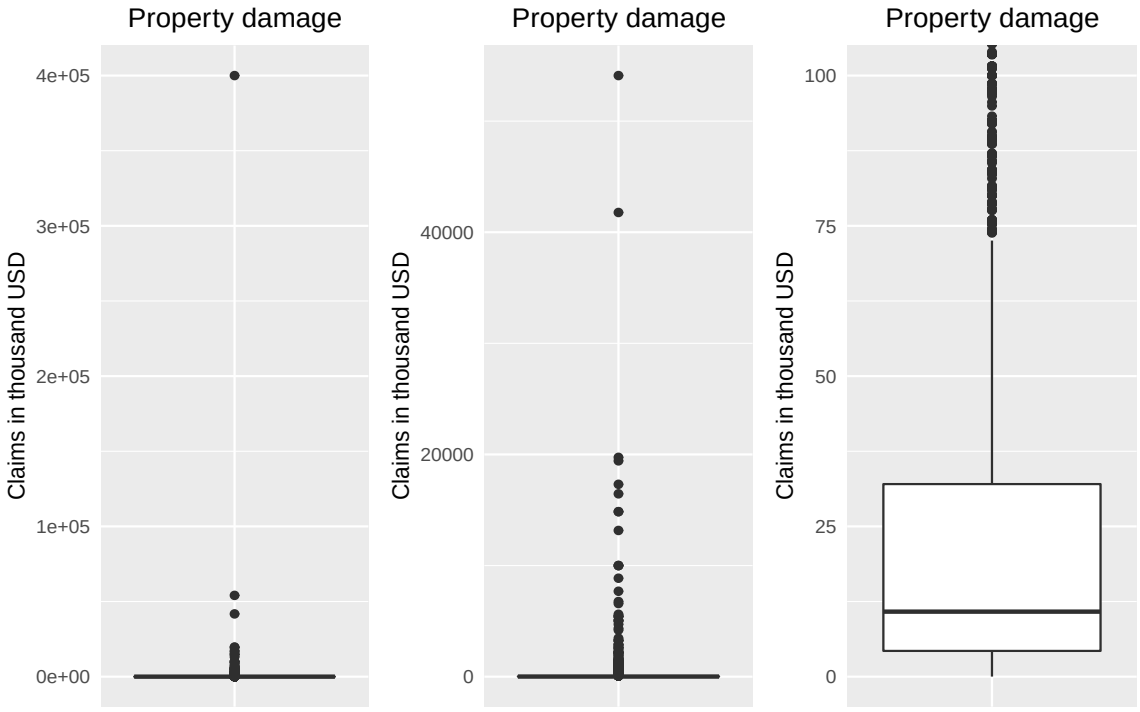

Figure 3.1. Summary of Loss Amounts in Thousands. Left: complete dataset. Centre: excluding outliers. Right: zoomed in to show majority of data.

We suspect the data contains some outliers and decided to exclude them to facilitate modeling. In order to clean the data, we generate the boxplot given in Figure 3.1 to visualize the distribution of the property damages. The boxplot on the left shows the complete dataset (without the winter storms from 1998), i.e. the same as the summary given in Table 3.2.

We notice that the maximum observation is a lot larger than all other observations and can be considered an outlier. Thus, we may want to exclude this specific observation to allow for a good model fit. Excluding this single observation results in a distribution of claims displayed in the boxplot in the middle in Figure 3.1. It is still impossible to even see the box that contains 50% of the data. As already mentioned, these data are highly skewed to the right and most observations are located somewhere close to zero (≤ 70).

After having excluded the large observation, we can zoom in to obtain the boxplot on the right to illustrate where the majority of the data can be found. As it is common for boxplots, the lower bound of the box corresponds to the 1^{st} sample quartile (Q_{25}), the horizontal line inside the box displays the sample median, whereas the 3^{rd} sample quartile (Q_{75}) is shown as the upper bound of the box. The black dots display observations that fall outside the interval $[Q_{25} - 1.5IQR,\ Q_{75} + 1.5IQR]$, where $IQR = Q_{75} - Q_{25}$ is the *interquartile range*. In this case, values not belonging to the aforementioned interval only occur in the right tail of the distribution, which is common for insurance data [*McNeil et al.*, 2005]. From now we will work with the dataset obtained after having excluded outliers. The newly obtained data is now summarized in Table 3.3. Excluding this single largest observation obviously significantly affects the mean and standard deviation but has little to no effect on the quartiles and the median.

Table 3.3. Summary of Loss Amounts in Thousands (after adjustments)

Sample size	Min.	1st quantile	Median	Std. Dev	Mean	3rd quantile	Maximum
7553	0.01	4.29	10.82	1003.08	91.37	32.04	54100

In the following section we will introduce the spatial heterogeneity model that is based on linear regression and thus, models the first moments of losses at different locations (Hua et al. [2017]).

3.2. THE SPATIAL HETEROGENEITY MODEL

Before getting into details of the spatial heterogeneity model, let us take a look at the distribution of the claims throughout the year illustrated in Figure 3.2. We observe that thunderstorm wind damages are much higher during the spring and summer months, compared to the fall and winter months, i.e. there are clearly seasonal patterns in the loss amounts. According to the *(NSSL)*, thunderstorms are most likely to occur in the spring and

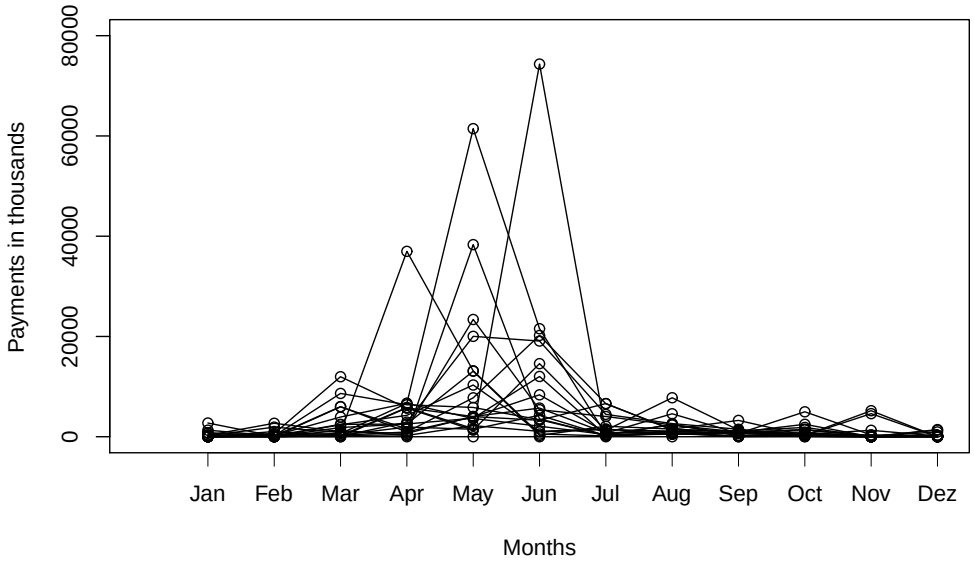

Figure 3.2. Seasonality of Loss Amounts (Thousands) from Year 1996 to 2013

summer months. Consequently, to model this seasonality, we use trigonometric functions such as $\sin(\cdot)$ and $\cos(\cdot)$ in the spatial heterogeneity model to account for these seasonal pattern.

As already explained, population density may significantly affect the losses at different locations and from Figure 3.2, we observe that "time of the year" has an effect on the losses as well. Recall that the spatial heterogeneity model is based on linear regression and it seems reasonable to include county-level population density and "time of the year" as covariates. "Time of the year" refers to the month in which the claim occurred (see Figure 3.2). Population densities are obtained from the U.S. Census Bureau (http://www.census.gov/) and are calculated as the population in the corresponding county divided by the area of the county. Linear regression assumes the residuals to be normally distributed. Modeling the natural logarithm of the losses supports the normal distribution assumption for the residuals (Hua et al. [2017]). To that end, let Y_i, $i = 1, \ldots, n$, with sample size $n = 7553$, be the

natural logarithm of the loss amounts adjusted by the inflation indexes. In addition, let D be the set of all possible coordinates. The model is given by

$$\mu = \mathbb{E}[Y|s,t,x] = \theta(s;w,K,\gamma) + \beta_0 + \beta_1 t + \beta_2 sin(\omega t) + \beta_3 cos(\omega t) + \beta_4 x, \tag{3.1}$$

where $s \in D$ is a two dimensional vector containing the coordinates of the location; $\theta(s;w,K,\gamma)$ is a function that accounts for the geographical effects at s, i.e. its influence on other locations (see below); $t = 1,\ldots,216$ is the month indicator from 1996 to 2013 (18 years is equivalent to 216 months), for example $t = 13$ corresponds to January 1997; $\omega = \frac{2\pi}{12}$ such that the period of the trigonometric functions becomes 12 (months), to account for the seasonal patterns that repeat every year. Finally, x denotes the population density of the county where s is located. The regression parameters to be estimated are $(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4)$.

To account for spatial dependence we will use radial basis functions to approximate the function $\theta(s;w,K,\gamma)$. Radial basis functions are usually applied to approximate functions or data which are only known at a finite number of points or too difficult to evaluate otherwise (Powell [1981], Cheney [1966]). A radial basis function $\phi(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ is a real-valued function whose value at a given argument x depends only on the distance from some fixed point c, called a centre, i.e. $\phi(x;c) = \phi(\|x-c\|)$. $\|\cdot\|$ is typically the Euclidean norm. Radial basis functions originated in the context of neural networks in the work by Broomhead and Lowe [1988]. The family of radial basis functions consists of many useful kernel smoothing functions, such as *Gaussian kernels*, *inverse multiquadratic kernels* and *thin plate spline kernels*. Following Hua et al. [2017], a real-valued continuous function $\theta : \mathbb{R}^2 \to \mathbb{R}$ can be approximated by

$$\theta(s;w,K,\gamma) \approx \sum_{k=1}^{K} w_k \phi\big(\,\|s-e_k\|_2\,\big), \tag{3.2}$$

where $w = (w_1, \ldots, w_K) \in \mathbb{R}^K$ are the weights that can be estimated by solving a system of linear equations because the approximation function is linear in the weights, $\phi : \mathbb{R}_+ \to \mathbb{R}$ is a radial basis function (see (3.3)), γ is the shape parameter for the basis function ϕ, and $e_k,\ k = 1, \ldots, K$ are some "reasonable" K prespecified centres. With "reasonable" we mean, those K centres should be chosen in a sense that they represent the data as good as possible, e.g. more centres should be found where many observations are located. It would not make sense to choose many centres far away from the majority of observations. The norm $\|\cdot\|_2$ is the usual Euclidean norm, describing the distance between the locations s and e_k. Other norms are also allowed, but we retain the Euclidean norm. Some commonly used radial basis functions are

$$\begin{aligned} \text{Gaussian: } \phi(x) &= \exp\big(-\gamma x^2\big) \\ \text{Inverse multi-quadratic: } \phi(x) &= \frac{1}{\sqrt{x^2 + \gamma^2}} \\ \text{Thin plate spline: } \phi(x) &= x^2 \log x \end{aligned} \tag{3.3}$$

Let $y_i,\ i = 1, \ldots, n$ be the logarithm of the observed losses. One way to estimate the weights $w = (w_1, \ldots, w_K)$ in (3.2) is to construct a system of linear equations

$$\sum_{k=1}^{K} w_k \phi\big(\|s_i - e_k\|_2\big) = y_i, \quad i = 1, \ldots, n, \tag{3.4}$$

for each distinct observation (s_i, y_i). Equation (3.4) can be rewritten as

$$\Big(\phi\big(\|s_i - e_1\|_2\big), \ldots, \phi\big(\|s_i - e_K\|_2\big)\Big) \begin{pmatrix} w_1 \\ \vdots \\ w_K \end{pmatrix} = y_i. \tag{3.5}$$

From (3.5) we obtain the following system of linear equations

$$\underbrace{\begin{pmatrix} \phi(\,\|s_1 - e_1\|_2\,) & \dots & \phi(\,\|s_1 - e_K\|_2\,) \\ \vdots & & \vdots \\ \phi(\,\|s_n - e_1\|_2\,) & \dots & \phi(\,\|s_n - e_K\|_2 \end{pmatrix}}_{=A\in\mathbb{R}^{n\times K}} \underbrace{\begin{pmatrix} w_1 \\ \vdots \\ w_K \end{pmatrix}}_{=w\in\mathbb{R}^{K\times 1}} = \underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{=y\in\mathbb{R}^{n\times 1}} . \tag{3.6}$$

Obtaining the solution w to the problem $Aw = y$ may require the use of a generalized inverse of the matrix A. For solving such a system in R, we refer to the package *limSolve*, which uses the Moore-Penrose generalized inverse of the matrix A to solve the corresponding system of linear equations. Another possible way to find the weights in (3.4) is to estimate the weights as the least square estimates of the regression coefficients from linear regression (which leads to the exact same result).

We use a smooth function $\theta(s; w, K, \gamma)$ as given in equation (3.2) to explain the effect of a specific location s. In order to get a good approximation for $\theta(s; w, K, \gamma)$ in (3.2) we use a clustering algorithm to partition the data into K clusters. The K centres of the clusters are then used as the K prespecified points e_k given in the approximation of $\theta(s; w, K, \gamma)$.

The number of clusters $K \leq n$ can be any arbitrary number (one approach to choose a good number is discussed further below). However, from the statistical viewpoint, large values of K may lead to overfitting and poor prediction (Hua et al. [2017]). A possible option is to select the K centres and group the data according to a certain clustering algorithm, such as the *K-means* clustering. Note that the K-means method is just one way of partitioning the data. We just need a partition over which to approximate the function $\theta(s; w, K, \gamma)$ such that the data is more or less evenly clustered based on the geographical information only. We will use the K-means algorithm to partition the data based on geographical locations, so that a geographical area can be divided into K smaller areas that are "representative"

of the whole area (explanation of representative further below). The function $\theta(s; w, K, \gamma)$ can then be approximated at those K centres of the clusters. Given the number of clusters K, the K-means method assigns those K centers so that the within-cluster sum of squares is minimized. The clustering aims to partition n observations into K clusters in which each observation belongs to the cluster with the nearest mean (Hua et al. [2017]). More specifically, let S_K be the partition, then

$$S_K = \arg \min_{S; S=\cup S_k} \sum_{k=1}^{K} \sum_{s \in S_k} \|s - mean(s, s \in S_k)\|_2^2,$$

where $\cup$ is the union operation of mutually exclusive subsets of S (Hua et al. [2017]).

Having partitioned the locations into K groups, we can now estimate $\theta(s; w, K, \gamma)$ as given in equation (3.2) based on claim observations $y_i = y(s_i),\ i = 1, \ldots, n$, where s_i is the coordinate of the i^{th}-location. We choose the number of clusters with the K-means algorithm based on how much variability is explained by those clusters. This means that one should choose a number of clusters such that adding another cluster does not "significantly improve" the within-group sum of squares. A vague approach is given by the elbow method that looks at the percentage of variance explained as a function of the clusters. Statistical techniques for obtaining an optimal number of clusters can be based on information criteria such as AIC or BIC.

Figure 3.3 shows the relationship between the number of clusters and the within-group sum of squares. We observe that when there are about 20 or more clusters, the total within-group variability is significantly decreased and adding more clusters does not improve the variability much.

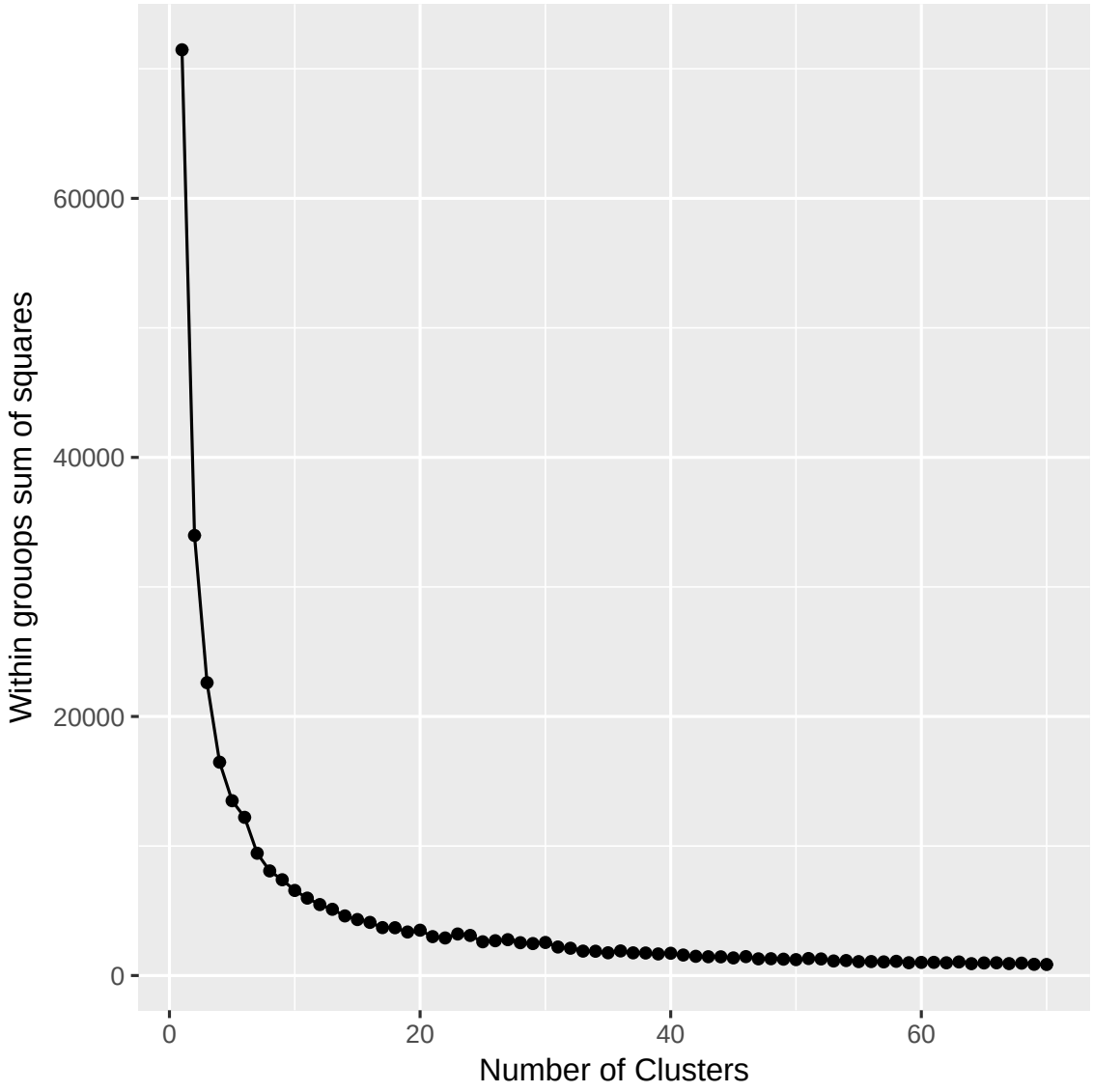

Figure 3.3. Withing-Group Sum of Squares vs. Number of Clusters

According to Figure 3.3, we believe it is appropriate to choose 20 clusters to represent the data. Figure 3.4 displays a result of the K-means algorithm for $K = 20$, where different colors stand for different clusters and the black stars in each cluster shows the centre of each cluster according to the K-means algorithm. We observe when the insurance claims are densely distributed (e.g. east Texas), then the clustering algorithm assigns many smaller clusters for this region. This implies that the centres of the clusters are also within closer

distances to each other. In contrast, when the insurance claims are sparsely distributed (e.g. west Texas), we obtain larger clusters and conclusively the centres are further apart from each other. Using the centres of the clusters as the K prespecified locations $e_1, \ldots, e_K$ in equation (3.2) represents the data in a sense that areas with densely filed insurance claims are assigned more clusters (and respectively more centres) than areas with sparsely distributed insurance claims. This implies that areas with a higher claim density are assigned more weight compared to areas with a smaller claim density.

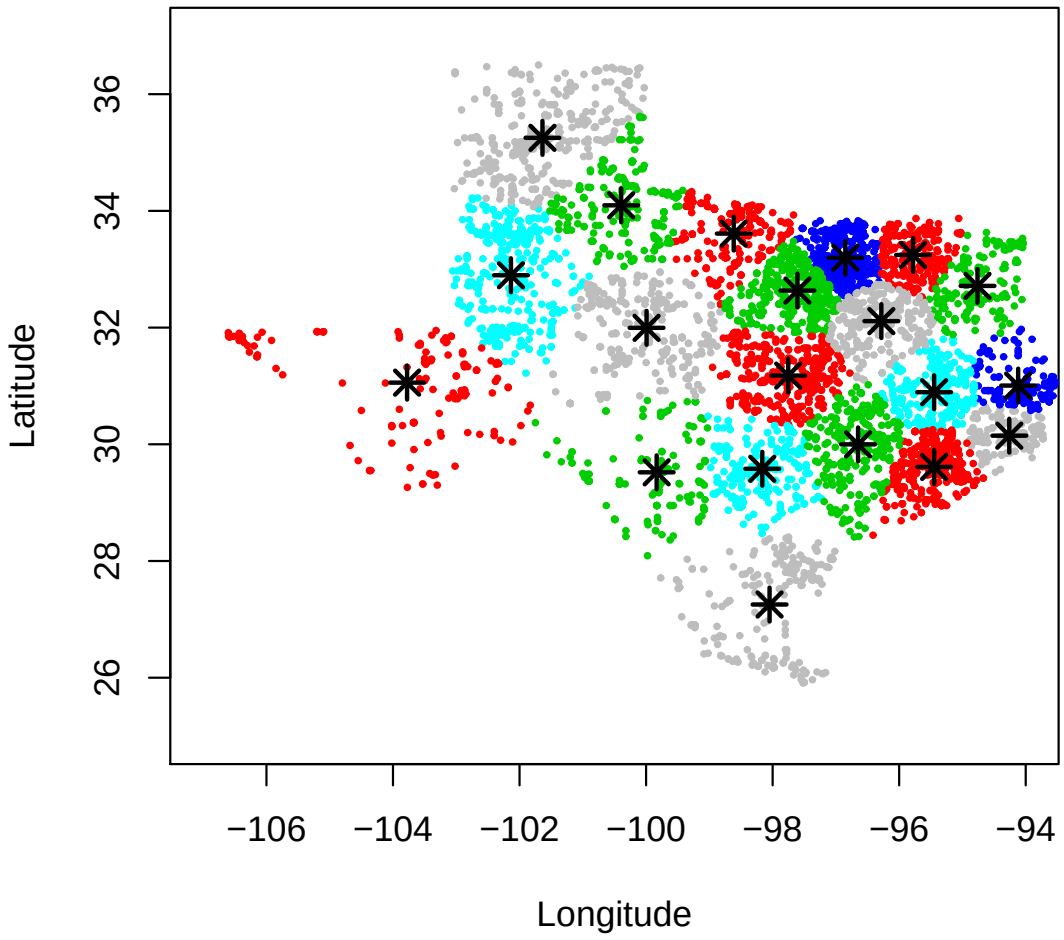

Figure 3.4. 20 Chosen Clusters

In order to get an estimate for $\theta(s; w, K, \gamma)$, we need to find the best shape parameter γ for the radial basis function (if one exists). This can be done by minimizing the *root mean square error (RMSE)*. That is

$$\gamma = \arg\min_{\gamma} \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i(\gamma) - y_i)^2}{n}}, \tag{3.7}$$

where y_i is the natural logarithm of the i^{th} observed property loss and $\hat{y}_i(\gamma)$ is the estimated logarithm of the i^{th} property loss in (3.4) with shape parameter γ.

Figure 3.5 illustrates how the value of gamma affects the RMSE for a Gaussian basis function. It seems appropriate to choose $\gamma = 0.04$ for the Gauss basis function as it minimizes the RMSE. The corresponding RMSE is 1.5823.

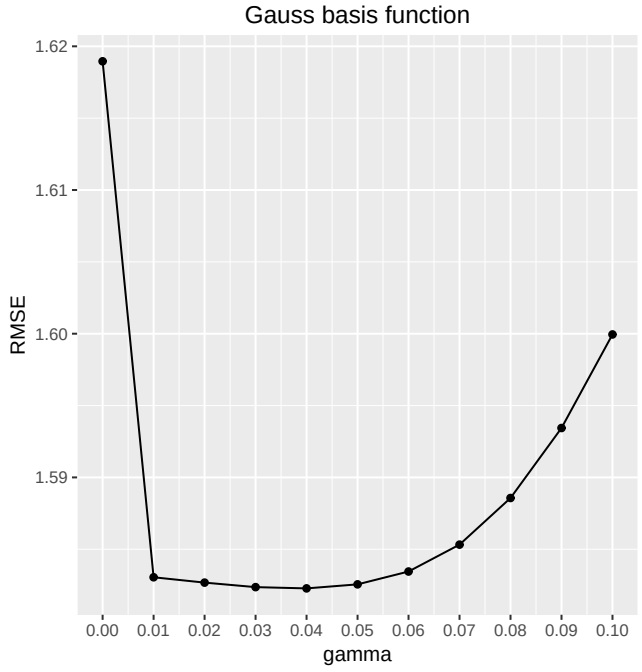

Figure 3.5. RMSE vs. Shape Parameter of Gaussian Basis Function

In the same vein, we can find an appropriate value for γ for the *inverse multi-quadratic* basis function. The result is displayed in Figure 3.6. Here it seems appropriate to choose $\gamma = 2$. The corresponding RMSE is 1.5734.

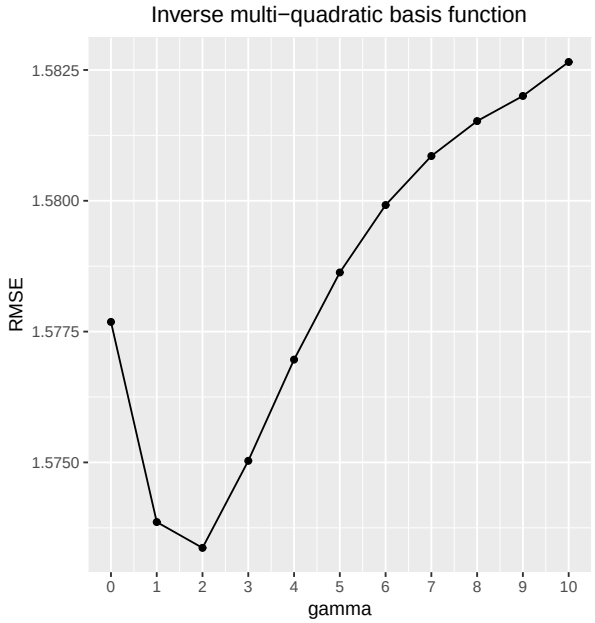

Figure 3.6. RMSE vs. Shape Parameter of Inverse Multi-Quadratic Basis Function

Recall that for approximating $\theta(s; w, K, \gamma)$, the 20 weights for the radial basis function are obtained as solution to the system of linear equations given in (3.6) or via linear regression (which yields the exact same result). Now, we fix the weights and the value for the shape parameter γ (which was chosen according to the RMSE) and approximate $\theta(s; w, K, \gamma)$ as in (3.2). After approximating $\theta(s; w, K, \gamma)$, we use the estimated $\hat{\theta}(s; w, K, \gamma)$ to normalize the response variable $Y(s) - \hat{\theta}(s; w, K, \gamma)$. Then, the rest of the parameters are estimated by the linear regression model given in (3.1):

$$\begin{aligned}\mu - \hat{\theta}(s; w, K, \gamma) &= \mathbb{E}[Y - \hat{\theta}(s; w, K, \gamma)|s, t, x] \\ &= \beta_0 + \beta_1 t + \beta_2 sin(\omega t) + \beta_3 cos(\omega t) + \beta_4 x.\end{aligned}$$

The type of radial basis function determines the speed of decay and can be chosen based on overall model-fitting performance, such as AIC. Among the three presented radial basis functions, the Gaussian basis function has the smallest AIC value of the regression model and therefore performs best.

Remark

When K is relatively large ($K > 10$), the AIC values become unstable (Hua et al. [2017]). The larger the value of K, the larger the number of clusters, and conclusively more weights need to be estimated which will lead to run time issues further below when maximizing the likelihood function. Therefore, one should consider a parameter calibration for the number of clusters K, and the shape parameter γ. For more details, we refer the reader to Hua et al. [2017].

For the same reason given in the remark, Hua et al. [2017] proposed the number of cluster to be $K = 4$. For this special case, we obtain $\gamma = 0.26$ as the optimal shape parameter for the Gaussian basis function. As the Gaussian basis function did a better job in the previous calculations, we retain this basis function. We estimated the weights of $\theta(s; w, K, \gamma)$ for $K = 4$ and $\gamma = 0.26$ in (3.6). Now, we fix the weights $w = (w_1, \ldots, w_4)$ and estimate $\hat{\theta}(s; w, K, \gamma)$ in (3.2) with the fixed weights. Then, as before, we normalize the response variable and the new regression model becomes

$$\begin{aligned} \mu - \hat{\theta}(s; w, K, \gamma) &= \mathbb{E}[Y - \hat{\theta}(s; w, K, \gamma)|s, t, x] \\ &= \beta_0 + \beta_1 t + \beta_2 sin(\omega t) + \beta_3 cos(\omega t) + \beta_4 x. \end{aligned}$$

The result for this regression model is given in Table 3.4.

Table 3.4. Regression summary with fixed weights

Coefficients	Estimate	S.E.
Intercept	$+9.576e+00$	$3.875e-02$
t	$-1.376e-03$	$2.934e-04$
$sin(\omega t)$	$+1.179e-01$	$2.793e-02$
$cos(\omega t)$	$+1.091e-01$	$3.229e-02$
Pop. density	$+2.035e-04$	$2.802e-05$
σ	$+1.601e+00$	$1.842e-02$
AIC	28553.69	

Since the estimated regression coefficient for the cumulated months t is negative ($\hat{\beta}_1 = -0.001376$), the model suggests that, over the years, the average property damage loss amount due to thunderstorm winds in Texas has been slightly decreasing, after adjusting for inflation. This reflects the trend in the data correctly. In a similar vein we observe that the estimated regression coefficient for population density ($\hat{\beta}_4 = 0.0002035 > 0$) is positive and therefore affects the loss amount in such a way that a higher population density leads to a higher average loss amount (which is what we would expect). One may argue that although $\hat{\beta}_4$ is positive, it is very small and therefore has little effect on the response. However, since we model the logarithm of the losses, a population density increase of, for example, 100 will lead to a multiplicative factor of $\exp(100\hat{\beta}_4) \approx 1.02$ of the estimated loss. This does not seem a lot, but assume we want to compare the estimated loss at two cities with population densities 500 and 3000, respectively, then we obtain a factor of $\exp(3000\hat{\beta}_4)/\exp(500\hat{\beta}_4) = \exp(2500\hat{\beta}_4) \approx 1.66$. This implies, if we fix all other covariates, the estimated losses in a city with population density 3000 is 1.66 times higher than in a city with a density of only 500. Therefore, we can conclude that $\hat{\beta}_4$ has indeed an important effect on the estimated loss amount.

The next section pertains to an introduction to *factor copula models*. As outlined in Section 2, copulas are ideal to describe (nonlinear) dependence structures. After the introduction of factor-copula models, we will construct an overall likelihood function that considers the dependence structure from the factor copula in order to estimate the best regression parameters for the spatial heterogeneity model.

3.3. THE SPATIAL DEPENDENCE MODEL

Following Krupskii and Joe [2013], let $X = (X_1, \ldots, X_n)$ be a random n-dimensional vector with joint cdf $F_X(x)$, $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and let $F_{X_i}(x_i)$ be the marginal cdf of X_i for $i = 1, \ldots, n$. By Sklar, the copula C_X corresponding to F_X, is a multivariate uniform cdf such that $F_X(x_1, \ldots, x_n) = C_X(F_{X_1}(x_1), \ldots, F_{X_n}(x_n))$. If F_X is continuous, the copula C_X is unique. Now, let $U_i = F_{X_i}(X_i)$ for $i = 1, \ldots, n$, then $U_i \sim U(0, 1)$. The joint cdf of the vector $U = (U_1, \ldots, U_n)$ is then given by $C(u_1, \ldots, u_n)$, where C is an n-dimensional copula. In the following, we assume all copulas are continuous and their densities exist.

Assume $U_1, \ldots, U_n$ to be conditionally independent given p latent variables $V_1, \ldots, V_p$. The random variables V_i, $i = 1 \ldots, p$ can be assumed to follow distributions such as the standard Normal distribution for the continuous case or a multinomial distribution in the discrete case for instance (McNeil et al. [2005]) but without loss of generality, we can assume $V_i \overset{iid}{\sim} U(0, 1)$, $i = 1, \ldots, p$. The assumption of uniform latent variables simplifies many calculations. Furthermore, let the conditional cdf of U_i given $V_1, \ldots, V_p$ be denoted by $F_{i|V_1,\ldots,V_p}$. Then, the *p-factor copula model* (Krupskii and Joe [2013]) is given by

$$C(u_1, \ldots, u_n) = \int_{[0,1]^p} \prod_{i=1}^{n} F_{i|V_1,\ldots,V_p}(u_i|v_1, \ldots, v_p) dv_1 \cdots dv_p. \tag{3.8}$$

In this model, $F_{i|V_1,\ldots,V_p}$ needs to be appropriately expressed in terms of a sequence of bivariate copulas that link the observed variable U_i to the latent variable V_k.

3.3.1. One-Factor Copula Model. We first study the case of $p = 1$ latent variable in (3.8). For $i = 1, \ldots, n$, denote the joint cdf and density of (U_i, V_1) by C_{i,V_1} and c_{i,V_1} respectively. Since $U_1, V_1 \sim U(0,1)$, then $F_{i|V_1}$ is just a partial derivative of the copula C_{i,V_1} with respect to the second argument (Krupskii and Joe [2013]). That is,

$$F_{i|V_1}(u_i|v_1) = C_{i|V_1}(u_i|v_1) = \frac{\partial C_{i,V_1}(u_i, v_1)}{\partial v_1}.$$

With $p = 1$, equation (3.8) becomes

$$C(u_1, \ldots, u_n) = \int_0^1 \prod_{i=1}^{n} F_{i|V_1}(u_i|v_1)dv_1 = \int_0^1 \prod_{i=1}^{n} C_{i|V_1}(u_i|v_1)dv_1. \tag{3.9}$$

Note that $\frac{\partial}{\partial u}C_{i|V_1}(u|v_1) = \frac{\partial^2}{\partial u \partial v_1}C_{i,V_1}(u, v_1) = c_{i,V_1}(u, v_1)$. Then (3.9) implies by differentiation that the density of the 1-factor copula is

$$c(u_1, \ldots, u_n) = \frac{\partial^n C(u_1, \ldots, u_n)}{\partial u_1, \ldots, \partial u_n} = \int_0^1 \prod_{i=1}^{n} c_{i,V_1}(u_i, v_1)dv_1.$$

In the model, dependence is defined by n bivariate linking copulas $C_{1,V_1}, \ldots, C_{n,V_1}$. There are no constraints among these bivariate copulas. Note that any conditional independence model for absolutely continuous random variables, conditioned on one latent variable, can be written in this form. The main advantage of the model is that it allows for different types of tail dependence structure. If all bivariate linking copulas are lower (upper) tail dependent, then all bivariate margins of U are also lower (upper) tail dependent respectively. Thus, with appropriately chosen linking copulas, asymmetric dependence structure as well as tail dependence can be easily modeled (Krupskii and Joe [2013]).

For the special case of bivariate Normal linking copulas, let $C_{i,V_1}(\cdot,\cdot)$ be the bivariate Normal copula with correlation $\alpha_{i1},\ i = 1,\ldots,n$. Let Φ, ϕ denote the standard Normal cdf and density respectively and let $\Phi_2(\cdot,\cdot;\rho)$ be the bivariate normal cdf with correlation ρ. Then $C_{i,V_1}(u_i, v_1) = \Phi_2(\Phi^{-1}(u_i), \Phi^{-1}(v_1); \alpha_{i1})$ and

$$F_{i|V_1}(u_i, v_1) = C_{i|V_1}(u_i|v_1) = \Phi\left(\frac{\Phi^{-1}(u) - \alpha_{i1}\Phi^{-1}(v_1)}{\sqrt{1-\alpha_{i1}^2}}\right).$$

3.3.2. Two-Factor Copula Model. Consider the case for $p = 2$. Let C_{i,V_1} be the copula of (U_i, V_1) as before and let $C_{i,V_2;V_1}$ be the copula for $F_{i|V_1} = F_{U_i|V_1}$ and $F_{V_2|V_1}$. Note that $F_{V_2|V_1}$ is the $U(0,1)$ cdf since we assume V_2 and V_1 to be independent. Then, equation (3.8) becomes (Krupskii and Joe [2013])

$$\begin{aligned} C(u_1,\ldots,u_n) &= \int_0^1\int_0^1 \prod_{i=1}^n F_{i|V_2,V_1}(u_i|v_1,v_2)dv_1dv_2 \\ &= \int_0^1\int_0^1 \prod_{i=1}^n C_{i|V_2;V_1}(C_{i|V_1}(u_i|v_1)|v_2)dv_1dv_2, \end{aligned} \tag{3.10}$$

where $C_{i|V_2;V_1}(C_{i|V_1}(u|v_1)|v_2) = \frac{\partial}{\partial v_2}C_{i,V_2;V_1}(C_{i|V_1}(u|v_1), v_2)$.

A copula density reveals more about the dependence than its cdf. Therefore, differentiating (3.10) with respect to $u_1,\ldots,u_n$, implies that the 2-factor copula density is

$$c(u_1,\ldots,u_n) = \int_0^1\int_0^1 \prod_{i=1}^n \left[c_{i,V_2;V_1}(C_{i|V_1}(u_i|v_1), v_2)\cdot c_{i,V_1}(u_i, v_1)\right]dv_1dv_2. \tag{3.11}$$

For general factor copula models, the bivariate link copulas do not need to belong to the same parametric family. However, for our spatial dependence model, we require that the link copulas are of the same family. For the special case of bivariate Normal linking copulas, let C_{i,V_1} and $C_{i,V_2;V_1}$ be the bivariate Normal copula with correlations α_{i1} and $\gamma_i = \alpha_{i2}/(1-\alpha_{i1}^2)^{1/2}$ respectively, $i = 1,\ldots,n$. Here, α_{i2} is the correlation of

$Z_i = \Phi(U_i)$ and $W_2 = \Phi(V_2)$, so that the independence of V_1 and V_2 implies that γ_i is the partial correlation of Z_i and W_2 given $W_1 = \Phi(V_1)$. In general the correlation is given as $\rho_{Z,W_2:W_1} = [\rho_{Z,W_2} - \rho_{Z,W_1}\rho_{W_2,W_1}]/[(1-\rho^2_{Z,W_1})(1-\rho^2_{W_2,W_1})]^{1/2}$. In our case, $\rho_{W_1,W_2} = 0$ due to the independence between V_1 and V_2. Then $\rho_{Z,W_2:W_1} = \rho_{Z,W_2}/(1-\rho^2_{Z,W_1})^{1/2}$ or in terms of α, $\gamma_i = \alpha_{i2}/(1-\alpha^2_{i1})^{1/2}$. Then, we have (Krupskii and Joe [2013])

$$\begin{aligned} C_{i|V_2;V_1}(C_{i|V_1}(u_i)|v_1)|v_2) &= \Phi\left(\left[\frac{\Phi^{-1}(u) - \alpha_{i1}\Phi^{-1}(v_1)}{\sqrt{1-\alpha^2_{i1}}} - \gamma_i\Phi^{-1}(v_2)\right] \Big/ \sqrt{1-\gamma_i^2}\right) \\ &= \Phi\left(\frac{\Phi^{-1}(u) - \alpha_{i1}\Phi^{-1}(v_1) - \gamma_i\sqrt{1-\alpha^2_{i1}}\Phi^{-1}(v_2)}{\sqrt{(1-\alpha^2_{i1})(1-\gamma_i^2)}}\right). \end{aligned}$$

In this work, we will use a 2-factor copula model. In the likelihood function that will be discussed further below, we will evaluate an integral of the form given in (3.11). In order to solve such a double integral, one needs to use numerical integration. For instance, one can use *Gauss-Hermite quadrature* or *Gauss-Legendre quadrature*. Gauss-quadrature is a numerical approximation of the definite integral of a function, usually stated as a weighted sum of function values at specified points within the domain of integration (for more on quadrature, see for example Press et al. [2007] (Section 4.6. "Gaussian Quadratures and Orthogonal Polynomials"). It can be shown that the evaluation points x_{k_i} are not equidistant but the roots of a polynomial belonging to a class of orthogonal polynomials (see Press et al. [2007] or Stoer and Bulirsch [2002]).

Assuming the parameters are θ_{i1} for C_{i,V_1} and θ_{i2} for $C_{i,V_2;V_1}$, we obtain

$$c(u_1,\ldots,u_n;\theta) \approx \sum_{k_1=1}^{nq}\sum_{k_2=1}^{nq} w_{k_1}w_{k_2}\prod_{i=1}^{n}\left[c_{i,V_2;V_1}(C_{i|V_1}(u_i|x_{k_1};\theta_{i1}), x_{k_2};\theta_{i2})) \cdot c_{i,V_1}(u_i, x_{k_1};\theta_{i1})\right], \tag{3.12}$$

where w_{k_i} are the quadrature weights, x_{k_i} are the roots of an orthogonal polynomial function, and nq is the number of quadrature points (Krupskii and Joe [2013]). When the nodes x_{k_i} are chosen to be the zeros of the *Legendre polynomials*, the method is called *Gauss-Legendre quadrature*. This method is attributed to the German mathematician Johann Carl Friedrich Gauß (1777 – 1855) and the French mathematician Adrien-Marie Legendre (1752 – 1833).

In what follows, we present a Maximum-likelihood approach that is based on both the spatial heterogeneity and spatial dependence model to find the optimal parameter estimates for the spatial heterogeneity model using the dependence structure from the spatial dependence model.

Joint likelihood. In order to obtain the best result, we can combine the spatial heterogeneity model and the spatial dependence model. Therefore, we use a Maximum-likelihood estimation based on the joint likelihood function containing both the marginal spatial heterogeneity model and the spatial dependence model, which is given by

$$L(\beta, \theta(s; w, K, \gamma)|y) = c(F_1(y_1), \ldots, F_n(y_n); \theta(s; w, K, \gamma)) \prod_{i=1}^{n} f_i(y_i; \beta, \sigma), \tag{3.13}$$

where $c(\cdot)$ is the n-dimensional copula density (that can be obtained from the factor copula approach), $f_i(y_i; \beta, \sigma)$ is the normal density with mean $\beta_0 + \beta_1 t_i + \beta_2 sin(\omega t_i) + \beta_3 cos(\omega t_i) + \beta_4 x_i$ and variance σ^2 (from the spatial heterogeneity model). $\theta(s; w, K, \gamma)$ is the dependence parameter of the copula and the value of $\theta(s; w, K, \gamma)$ is determined by the geographical location s through the radial basis function given in equation (3.2). In contrast to the spatial heterogeneity model, the spatial dependence parameter $\theta(s; w, K, \gamma)$ is now embedded in the copula density and cannot be estimated with the previous methods. $\theta(s; w, K, \gamma)$ needs to be estimated in the likelihood function with the other parameters. Thus, we need to

maximize the likelihood function in 10 different parameters which is a computationally difficult problem.

To approximate the copula density as given in (3.12), we use the Gauss-Legendre quadrature with the number of quadrature points being 35, i.e. $nq = 35$ (see R-package *CopulaModel* for explanation for number of quadrature points). The computations were conducted in the software R with numerical optimization. Optimizing the full likelihood as given above is numerically difficult since we need to estimate many parameters, the evaluation of the copula density is itself a numerical procedure and the number of observations we have is large ($n = 7553$). Therefore, instead of maximizing the multivariate likelihood function given in (3.13), we use a *pairwise (or composite) likelihood* approach, i.e. for $i, j = 1, \ldots, n$

$$
\begin{aligned}
(\hat{\beta}, \hat{\sigma}, \hat{w}) &= \arg\max_{(\beta,\sigma,w)} \sum_{i>j} L_{i,j}(\beta, \theta(s; w, K, \gamma)|y) \qquad (3.14)\\
&= \arg\max_{(\beta,\sigma,w)} \sum_{i>j} \left[c(F_i(y_i), F_j(y_j); \theta(s; w, K, \gamma)) \cdot f_i(y_i; \beta, \sigma) \cdot f_j(y_j; \beta, \sigma)\right],
\end{aligned}
$$

where $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_4)$, $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_4)$ and $\theta(s; w, K, \gamma)$ as given in (3.2). For more information on composite likelihood, we refer to Heagerty and Lele [1998] or Curriero and Lele [1999].

The optimization of the likelihood in (3.13) is a difficult problem regarding run time or speed of calculations. As initial values for optimizing the objective function we used the values from Table 3.4, which seem to be a reasonable choices to ensure that the iterations converge. We refer to Krupskii and Joe [2013] for detailed discussions about numeric issues on implementing factor copulas and the R package *CopulaModel* associated with the book *Dependence Modeling with Copulas*, Joe [2014] for implementations of factor copulas. However, this package cannot be installed via the regular "install.packages()" command in

R. We refer to *http://copula.stat.ubc.ca/* for details on installation of the R package. Useful functions in this context are provided by *dfact2cop()* which evaluates a bivariate copula density of a 2-factor copula and *pfact2cop()* which can help evaluate the bivariate cdf of a 2-factor copula model. These functions are only provided for the bivariate case, which is why we chose to compute the MLE with the pairwise likelihood approach given in (3.14) instead of as given in (3.13). The MLEs are reported in Table 3.5.

We can assume that $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ in equation (3.1) approximately follows a multivariate Normal distribution with the mean being the MLEs given in Table 3.5 and variance-covariance matrix being approximated by the Fisher information matrix $\hat{\Sigma} = \{\sigma_{ij}^2\}_{i,j=0,\ldots,5}$, which can be obtained as the inverse of the estimated negative Hessian matrix. The estimated Hessian matrix is obtained from the optimization procedure. An approximate 95% CI for β_i is then given by $[\hat{\beta}_i + z_{\frac{\alpha}{2}}\sigma_{ii}, \hat{\beta}_i - z_{\frac{\alpha}{2}}\sigma_{ii}]$, where $z_{\frac{\alpha}{2}} = \Phi^{-1}(\frac{\alpha}{2})$ is the $\frac{\alpha}{2}$ - quantile of the standard Normal distribution.

Table 3.5. MLEs of Model with 95% CI

Coefficients	Estimate	Lower CI	Upper CI
$\hat{\beta}_0$: Intercept	$+9.590e+00$	$+9.514e+00$	$+9.667e+00$
$\hat{\beta}_1$: t	$-1.820e-03$	$-2.398e-03$	$-1.241e-03$
$\hat{\beta}_2$: $sin(\omega t)$	$+6.887e-02$	$+1.380e-02$	$+1.239e-01$
$\hat{\beta}_3$: $cos(\omega t)$	$+9.906e-02$	$+3.538e-02$	$+1.627e-01$
$\hat{\beta}_4$: Pop. density	$+2.330e-04$	$+1.327e-04$	$+1.269e-03$
$\hat{\sigma}$	$+1.59e+00$		
$\hat{w}_1$	$-2.867e-01$		
$\hat{w}_2$	$-7.763e-01$		
$\hat{w}_3$	$-2.107e-01$		
$\hat{w}_4$	$+8.568e-02$		

According to Table 3.5, we propose the following model to describe the thunderstorm wind loss data from Texas:

$$\mu = \mathbb{E}[Y|s,t,x] = \hat{\theta}(s,w,K,\gamma) + \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 sin(\omega t) + \hat{\beta}_3 cos(\omega t) + \hat{\beta}_4 x, \tag{3.15}$$

where $\hat{\theta}(s,w,K,\gamma) = \sum_{k=1}^{4} \hat{w}_k \phi\big(\|s - e_k\|_2\big)$ and $\phi(x) = \exp(-0.26x^2)$ being the Gaussian basis function with shape parameter $\gamma = 0.26$. Figure 3.7 visualizes the estimated spatial dependence parameter $\hat{\theta}(s,w,K,\gamma)$. The large blue colored area is located around Houston, which is the largest city in Texas by population. Beside the blue colored area, $\hat{\theta}(s,w,K,\gamma)$ is mostly close to zero and therefore has little effect on the response.

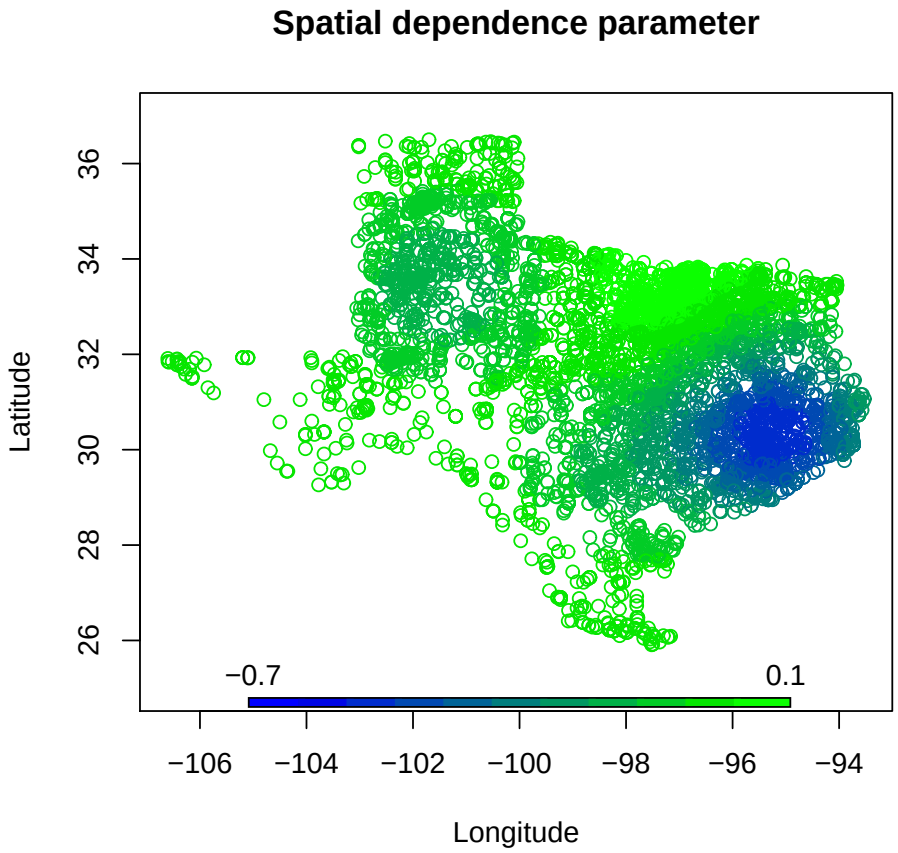

Figure 3.7. Estimated spatial dependence parameter

Prediction. Using equation (3.1) and the corresponding MLEs given in Table 3.5, we can conduct risk assessment for a given set of values of covariates such as location, population density and month. Let $x = (1, t, \sin(\omega t), \cos(\omega t), \text{population density})^\top$ be a given vector of covariates. Then, we can calculate the standard error of the predicted loss $\mu = \mathbb{E}[Y|s,t,x]$ as $\hat{\sigma}_\mu^2 = \hat{var}(\mathbb{E}[Y|s,t,x]) = x^\top \hat{\Sigma} x$. As already explained, the risks $\exp(Y)$ are highly right skewed which makes the median a better measure of risk severity than the mean. Thus, we can assess the risk severity as $\exp(\hat{\mu})$ with μ defined in (3.1). Because $\hat{\mu}$ is an estimator of the mean μ and $\exp(Y)$ is highly right skewed, we can think of $\exp(\hat{\mu})$ as an approximation of the median. A $(1-\alpha)100\%$ confidence interval of the risk $\exp(\mu)$ is given as $[\exp(\hat{\mu} + z_{\frac{\alpha}{2}}\hat{\sigma}_\mu),\ \exp(\hat{\mu} - z_{\frac{\alpha}{2}}\hat{\sigma}_\mu)]$, where $z_{\frac{\alpha}{2}} = \Phi^{-1}(\frac{\alpha}{2})$ is the $\frac{\alpha}{2}$ quantile of the standard Normal distribution. Table 3.6 shows some example of risk assessment for some cities for 2014 including a 95% confidence interval (CI) for the median risk. The values are rounded to the next closest integers. To check the accuracy of these predictions we compare the result to actually observed data in the database.

Table 3.6. Risk assessment for 2014

	San Antonio, July	Colorado City, June	Midland, June	Dallas, May
Median Risk	9409	8815	11037	14854
Lower 95% CI	8475	8079	9915	13061
Upper 95% CI	10444	9618	12286	16893
Observed Median	9843	8366	11318	49213

We notice that the prediction works well for many cases, however, in May 2014 Dallas experienced severe thunderstorm winds that lead to large losses and our prediction does not match the observed value.

3.4. MATÉRN CORRELATION FUNCTION AS MEASURE OF DEPENDENCE

An alternative to estimating the dependence parameter between locations with the radial basis functions is provided by the Matérn covariance function as a measure of dependence. The Matérn covariance (named after the Swedish statistician Bertil Matérn) is a covariance function used in spatial statistics, geostatistics, machine learning, and other statistical applications. It is commonly used to define the statistical covariance between measurements made at two points that are h units apart from each other. Since the covariance only depends on distances between points, it is stationary. If the distance is the Euclidean distance, the Matérn covariance is also isotropic (Minasny and McBratney [2005]). The Matérn isotropic covariance function is given by (Cressie and Wikle [2011]; Handcock and Stein [1993]; Stein [1999])

$$C(h) = \frac{2^{1-v}}{\Gamma(v)}\left(\frac{h}{r}\right)^{v} K_v\left(\frac{h}{r}\right), \tag{3.16}$$

where h is the separation distance, $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$ is the Gamma-function, $K_v(\cdot)$ is a modified Bessel function of the second kind of order v (Minasny and McBratney [2005]; Abramowitz and Stegun [1972]). The modified Bessel functions of the second kind are sometimes also called the Basset functions or MacDonald functions (Spanier and Oldham [1987], p.499). Eventually $r > 0$ is the distance parameter which measures how quickly the correlations decay with distance, and $v > 0$ is the smoothness parameter. The model was first introduced by *Matérn* in 1960, but was deduced earlier by *Whittle* in 1954 (constrained to $v = 1$). An alternative parameterization of equation (3.16) has been suggested by Handcock and Wallis [1994]:

$$C(h) = \frac{2^{1-v}}{\Gamma(v)}\left(\frac{2v^{1/2}h}{r}\right)^{v} K_v\left(\frac{2v^{1/2}h}{r}\right),$$

which allows r to be less dependent on v (Stein [1999]).

The Matérn model has great flexibility for modelling the spatial covariance compared with the standard models because of its smoothness parameter v. The parameter v, which controls the smoothness of the spatial process, should be determined from the spatial data. When v is small ($v \to 0$) it implies that the spatial process is rough, and when it is large ($v \to \infty$) that the process is smooth. If v is of the form $(n + 1/2)$, $n \in \mathbb{N}_0$, then $C(h)$ is the product of a polynomial of degree m in (h/r) and $\exp(-h/r)$ (Minasny and McBratney [2005]):

$$
\begin{aligned}
v = 1/2 &\Rightarrow C(h) = \exp(-h/r) \\
v = 3/2 &\Rightarrow C(h) = \left[(h/r) + 1\right] \exp(-h/r) \\
v = 5/2 &\Rightarrow C(h) = \left[(h/r)^2 + 3(h/r) + 3\right] \exp(-h/r)
\end{aligned}
$$

Instead of using radial basis functions to estimate the spatial dependence parameter, we can use the Matérn covariance function as a measure of spatial dependence.

Recall: The spatial dependence parameter $\theta(s; w, K, \gamma)$ can be approximated by equation (3.2)

$$\theta(s; w, K, \gamma) \approx \sum_{k=1}^{K} w_k \phi\big(\|s - e_k\|_2 \big),$$

where $\phi(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ is a radial basis function. In this approximation we can now replace $\phi(\cdot)$ with the Matérn covariance function $C(\cdot)$

$$\theta(s; w, K, \gamma) \approx \sum_{k=1}^{K} w_k C\big(\|s - e_k\|_2 \big)$$

and conduct the same computations as mentioned in the previous subsections.

Theoretically, the parameters v and r need to be estimated from the data. However, this is a cumbersome procedure and not the aim of this work. We choose v according to what can be found in other literature. A good choice may be $v = 3/2$ (Gneiting et al. [2010]). We also assume $r = 1$. Running the same procedure as explained above, yields the predictions for 2014 as given in Table 3.7.

Table 3.7. Risk assessment for 2014 using Matérn covariance function

	San Antonio, July	Colorado City, June	Midland, June	Dallas, May
Median Risk	9944	8301	9689	11776
Lower 95% CI	8958	7579	8704	10354
Upper 95% CI	11038	9092	10786	13392
Observed Median	9843	8366	11318	49213

This result is a little different to the prediction we obtained before. One may consider choosing other values for the parameters to optimize this procedure, i.e. by estimating the parameters from the data.

4. EXTENSIONS

SPATIO-TEMPORAL MODEL

The proposed model can be extended to a spatio-temporal model by letting $\theta(s; w, K, \gamma)$ not only depend on the geographical location but also on the time, i.e. s = (Latitude, Longitude, time). In this case we not only need to partition the data geographically using clusters but also a partition along the time line $[0, t_{max}]$ is needed , where t_{max} is the maximum time t considered. This yields a modification of K as $K = K_l \times K_t$, where K_l is the number of clusters and K_t is the number of partitions along time (Hua et al. [2017]).

LOSS FREQUENCIES

Our original data contained many loss amounts of zero. We can extend the proposed model to modeling spatial dependence for loss frequencies. For that purpose we let $M(s) \in \{0, 1, 2, \dots\}$ be the number of losses at location s. $M(s)$ can be written as $M(s) = I(s)J(s)$ with $I(s) \sim Ber(p)$, where p is the probability of success, and $J(s)$ is a count variable, i.e. a discrete random variable. Furthermore, we assume $I(s)$ and $J(s)$ to be independent. Then

$$P[M(s) = m] = \begin{cases} P[I(s) = 0] & , m = 0 \\ P[I(s) = 1]P[J(s) = m] & , m = 1, 2, 3, \dots \end{cases}$$

An example of $J(s)$ can be a shifted Poisson distribution whose parameter $\lambda(s)$ depends on the geographical location s

$$P[J(s) = m] = \exp(-\lambda(s))\frac{(\lambda(s))^{m-1}}{(m-1)!}, \ m = 1, 2, \dots$$

We can apply the spatial dependence concepts to model dependence among $J(s)$ using a one-factor copula model. The corresponding overall likelihood function is given by

$$L(p, \alpha, \theta | m_1, \ldots, m_n) = \int_0^1 \prod_{i=1}^{n} (1-p)^{I(m_i=0)} \left[p(C_{i|v}(F_J(m_i)|v) - C_{i|v}(F_J(m_i-1)|v)) \right]^{I(m_i>0)} dv,$$

where F_J is the cdf of $J(s)$, p is the parameter for the Bernoulli random variable $I(s)$, and α are the regression coefficients for $\lambda(s)$. As before, θ are the dependence parameters that can be written as a function of s, that is $\theta(s; w, K, \gamma)$. This implies we can use the proposed approach for estimating $\theta(s; w, K, \gamma)$ (Hua et al. [2017]). For more information on factor copula models for discrete data, we refer to Nikoloulopoulos and Joe [2015].

5. CONCLUSIONS

In this work, we illustrated how linear models and factor copula models can be used in order to model spatially dependent data such as insurance losses that are caused by thunderstorm winds (or natural disasters in general). The main challenge of estimating complex spatial dependence structures is reduced to estimating a spatial dependence parameter $\theta(s; w, K, \gamma)$. The spatial dependence parameter that is estimated via a weighted sum of radial basis functions is embedded in the bivariate copulas in the likelihood function which makes it computationally extremely expensive to maximize the likelihood function. Note that we have used a bivariate Gaussian copulas in the methods but other families of bivariate copulas can also be considered. The R-package "CopulaModel" also provides functions for some other copula families such as the Frank or Gumbel family. Regarding run time, one clearly needs to consider the amount of cluster that are being used for estimating $\theta(s; w, K, \gamma)$ as it heavily influences the optimization procedure of the likelihood function. As demonstrated, the proposed approach allows us to make efficient loss predictions as it was shown for some cities in Texas. We have also discussed potential extensions of the proposed models that can serve as dissertation research problems.

REFERENCES

K. Aas, C. Czadob, A. Frigessic, and H. Bakkend. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44:182–198, 2009.

M. Abramowitz and I. E. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 10th Printing*. Department of Commerce, National Bureau of Standards, 1972.

K. Antonio and E. A. Valdez. Statistical concepts of a priori and a posteriori risk classification. *Amsterdam: Universiteit van Amsterdam*, 2010.

T. Bedford and R.M. Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32:245–268, 2001.

T. Bedford and R.M. Cooke. Vines–a new graphical model for dependent random variables. *The Annals of Statistics*, 30:1031–1068, 2002.

E. Bouyé, V. Durrleman, A. Nikeghbali, G. Riboulet, and T. Roncalli. Copulas for finance; a reading guide and some applications. 2000.

A. Bárdossy. Copula-based geostatistical models for groundwater quality parameters. *Water Resour. Res.*, 42, 2006.

A. Bárdossy and Z. W. Kundzewicz. Geostatistical methods for detection of outliers in groundwater quality spatial fields. *Journal of Hydrology*, 115:343–359, 1990.

D. S. Broomhead and D. Lowe. Multivar iable functional interpolation and adaptive networks. *Complex Systems Publications, Inc.*, 2:321–355, 1988.

E. W. Cheney. *Introduction to Approximation Theory*. McGraw-Hill, 1966.

D.G. Clayton. Model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65: 141–151, 1978.

N. Cressie and C.K. Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, Inc, 2011.

F. C. Curriero and S. Lele. A composite likelihood approach to semivariogram estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, 4(1):9–28, 1999.

G. Dall'Aglio, S. Kotz, and G. Salinetti eds. *Advances in Probability Distributions with Given Marginals*. Springer, 1991.

P. Embrechts, A. McNeil, and D. Straumann. Correlation and dependence in risk management: Properties and pitfalls. 86, 1999.

P. Embrechts, F. Lindskog, and A. McNeil. Modelling dependence with copulas and applications to risk management. *Handbook of Heavy Tailed Distributions in Finance*, pages Chapter 8, 329–384, 2001.

T.M. Erhardt, C. Czado, and U. Schepsmeier. R-vine models for spatial time series with an application to daily mean temperature. *Biometrics*, 71(2):323–32, 2015.

A.-C. Favre, S. El Adlouni, L. Perreault, N. Thiémonge, and B. Bobée. Multivariate hydrological frequency analysis using copulas. *Water Resour. Res*, 40, 2004.

B. V. Gnedenko. The theory of probability (translated from the russian kurs theorii veroyatnostei, ed. 2, by b. d. seckler). *Science*, 138:422–423, 1962.

T. Gneiting, W. Kleiber, and M. Schlather. Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491), 2010.

P. Goovaerts. Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma*, 89(1-2):1–45, 1999.

P. Goovaerts. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228:113–129, 2000.

P. Goovaerts, G. AvRuskin, J. Meliker, M. Slotnick, G. Jacquez, and J. Nriagu. Geostatistical modeling of the spatial variability of arsenic in groundwater of southeast michigan. *Water Resour. Res.*, 41, 2005.

B. Gräler and E. Pebesmaa. The pair-copula construction for spatial data: a new approach to model spatial dependency. *Procedia Environmental Sciences*, 7:206–211, 2011.

E.J. Gumbel. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55:698–707, 1960a.

E.J. Gumbel. Distributions des valeurs extrêmes en plusiers dimensions. *Publications de l'Institut Statistique de l'Université de Paris*, 9:171–173, 1960b.

U. Haberlandt. Geostatistical interpolation of hourly precipitation from rain gauges and radar for a large-scale extreme rainfall event. *Journal of Hydrology*, 332:144–157, 2007.

M. S. Handcock and M. L. Stein. A bayesian analysis of kriging. *Technometrics*, 35: 403–410, 1993.

M. S. Handcock and J.R. Wallis. An approach to statistical spatial–temporal modeling of meteorological fields (with discussion). *Journal of the American Statistical Association*, 89:368–390, 1994.

P. J. Heagerty and S. R. Lele. A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93(443):1099–1111, 1998.

L. Hua, M. Xia, and S. Basu. Factor copula approaches for assessing spatially dependent high-dimensional risks. *North American Actuarial Journal*, 21(1):147–160, 2017.

H. Joe. Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. *IMS Lecture Notes - Monograph Series*, 28, 1996.

H. Joe. *Dependence Modeling with Copulas*. Chapman and Hall, 2014.

P. Krupskii and H. Joe. Factor copula models for multivariate data. *Journal of Multivariate Analysis*, 120:85–101, 2013.

D. Kurowicka and R. Cooke. *Uncertainty Analysis with High Dimensional Dependence Modelling*. John Wiley & Sons Ltd, 2006.

S. Ly, C. Charles, and A. Degré. Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the ourthe and ambleve catchments, belgium. *Hydrol. Earth Syst. Sci.*, 15:2259–2274, 2011.

M. Mahfoud. Bivariate archimedean copulas: an application to two stock market indices. *BMI Paper*, 2012.

G. Matheron. *Traité de Geostatistique Apliquée, Tome I*. Memoires du Bureau de Recherches Geologiques et Minieres, 1962.

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management Concepts, Techniques and Tools*. Princeton University Press, 2005.

C. D. Michele and G. Salvadori. A generalized pareto intensity-duration model of storm rainfall exploiting 2-copulas. *J. Geophys. Res*, 108(D2), 2003.

B. Minasny and A. B. McBratney. The matérn function as a general model for soil variograms. *Geoderma*, 128:192–207, 2005.

R. B. Nelsen. *An Introduction to Copulas (Second Edition)*. Springer, 2006.

A. K. Nikoloulopoulos and H. Joe. Factor copula models for item response data. *Psychometrika*, 80(1):126–150, 2015.

M. J. D. Powell. *Approximation Theory and Methods*. Cambridge University Press, 1981.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing (3rd ed.)*. Cambridge University Press, 2007.

B. Schweizer. *Thirty years of copulas. In: Dall'Aglio G, Kotz S, Salinetti G (eds) Advances in Probability Distributions with Given Marginals.* Springer, 1991.

J. Segers. Copulas: An introduction part ii: Models; presentation at columbia university, new york city, 9–11 oct. *Université catholique de Louvain (BE), Institut de statistique, biostatistique et sciences actuarielles*, 2013.

A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231, 1959.

J. Spanier and K. B. Oldham. The basset kv(x). *An Atlas of Functions*, page 499– 507, 1987.

M. L. Stein. *Interpolation of Spatial Data; Some Theory for Kriging*. Springer, 1999.

J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis (3rd ed.)*. Springer, 2002.

A. Verworn and U. Haberlandt. Spatial interpolation of hourly rainfall effect of additional information, variogram inference and storm properties. *Hydrol. Earth Syst. Sci.*, 15: 569–584, 2011.

W. H. Yu, C. M. Harvey, and C. F. Harvey. Arsenic in groundwater in bangladesh: A geostatistical and epidemiological framework for evaluating health effects and potential remedies. *Water Resour. Res.*, 39(6):1146, 2003.

VITA

Tobias Merk was born in Ochsenhausen, Germany, on September 18, 1993. In February 2016, he graduated from Ulm University with a B.Sc. in Mathematics. In his Bachelor degree, he spent the fall semester 2014 at the University of Ottawa (Canada) and interned at Boehringer Ingelheim (Biberach an der Riß, Germany) during the summer of 2015. In 2016, he started his Master at Ulm University and completed his first two semesters there, followed by an internship at Daimler AG in Ulm. In Fall 2017, he enrolled at Missouri S&T as a Master student in Applied Mathematics with emphasis on statistics. He received his M.Sc. Degree in Applied Mathematics from Missouri University of Science and Technology in May 2018.