



Factor Copula Approaches for Assessing Spatially Dependent High-Dimensional Risks

Lei Hua, Michelle Xia & Sanjib Basu

To cite this article: Lei Hua, Michelle Xia & Sanjib Basu (2017) Factor Copula Approaches for Assessing Spatially Dependent High-Dimensional Risks, North American Actuarial Journal, 21:1, 147-160, DOI: [10.1080/10920277.2016.1246251](https://doi.org/10.1080/10920277.2016.1246251)

To link to this article: <https://doi.org/10.1080/10920277.2016.1246251>



Published online: 01 Feb 2017.



Submit your article to this journal [↗](#)



Article views: 209



View related articles [↗](#)




View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

Factor Copula Approaches for Assessing Spatially Dependent High-Dimensional Risks

Lei Hua ¹, Michelle Xia,¹ and Sanjib Basu²

¹*Division of Statistics, Northern Illinois University, DeKalb, Illinois*

²*School of Public Health, University of Illinois at Chicago, Chicago, Illinois*

In this article, we propose an innovative approach for modeling spatial dependence among losses from various geographical locations. The proposed model converts the challenging task of modeling complex spatial dependence structures into a relatively easier task of estimating a continuous function, of which the arguments can be the coordinates of the locations. The approach is based on factor copula models, which can capture various linear and nonlinear dependence. We use radial basis functions as the kernel smoother for estimating the key function that models all the spatial dependence structures. A case study on a thunderstorm wind loss dataset demonstrates the analysis and the usefulness of the proposed approach. Extensions to spatiotemporal models and to models for discrete data are briefly introduced, with an example given for modeling loss frequency with excess zeros.

1. INTRODUCTION

Data that are spatially distributed are increasingly becoming available nowadays because of the development of modern technologies such as satellites, smart phones, and telematics installed in vehicles. Insurance risks such as natural disasters, contagious diseases, and traffic accidents often exhibit spatial dependence, with the strength of dependence depending on various factors such as physical distance and population density. For property-and-casualty insurance, losses from natural disasters possess spatial dependence. In health insurance, pandemic and epidemic diseases are inherently spatial. Traditional loss models treat the territory variable as a rating factor, which may not adequately account for the spatial dependence structure. This may lead to an underestimation of the overall risk, and the problem may be amplified when the loss distributions are heavy-tailed.

In spatial modeling, spatial heterogeneity models first moments of losses at different locations (e.g., by using traditional regression methods), while spatial dependence considers the dependence among random elements at different geographical locations. For spatial dependence, a commonly adopted assumption is that the dependence is relatively stronger for elements that are closer to each other. Commonly used techniques for spatial dependence include covariance structures based on distance, for example, by assuming spatial processes with locations being the indexes. For comprehensive monographs on spatial statistics, we refer the readers to Cressie (1993) and Cressie and Wikle (2011). As covariance modeling may not account for nonlinear dependence, recently, copula functions have been proposed to account for spatial dependence. For example, Bárdossy (2006) introduced copula approaches for modeling spatial dependence in groundwater quality, Gräler and Pebesma (2011) used vine copulas for more flexible spatial dependence structures, and Erhardt et al. (2015a, b) applied vine copulas to model the dependence between 73 observation stations in Germany for temperature data, where detailed statistical inference was conducted.

The main challenge for modeling the dependence among random losses at different locations is how to construct feasible and flexible dependence structures for handling the high dimensionality. If one looks at the random element at one specific location s as a univariate random variable $Y(s)$, $s \in \mathcal{D}$, then the domain of locations \mathcal{D} can be large for many real-world applications, and the question of interest is a high-dimensional and even an infinite-dimensional one (e.g., for point-referenced spatial data). Moreover, depending on the geographical features of \mathcal{D} , the dependence between $Y(s)$ could be very complex, and thus it requires that the approaches used to model such dependence structures be flexible enough to account for the complexity.

The existing methods of modeling spatial dependence belong to the following two themes. One is based on modeling covariance structures of spatial processes; for example, the notion of a variogram has been widely used as a measure of spatial dependence:

Address correspondence to Lei Hua, Division of Statistics, Northern Illinois University, DeKalb, IL 60115. E-mail: lhua@niu.edu

$2\gamma(\mathbf{h}) = \mathbb{E}[(Y(s) - Y(s + \mathbf{h}))^2]$ (see, e.g., Cressie 1993). The other theme is based on partitioning the data into several groups based on geographical locations and then constructing dependence structures among these groups. For example, Ebrahimi and Hua (2014) employed factor copulas for modeling reliability of grouped nano-components, and Erhardt et al. (2015b) applied vine copula models in accounting for a complex dependence structure.

The first theme assumes a linear dependence structure, which brings great convenience for modeling, as well as constraints on being restricted to a linear dependence structure. Methods within the second theme cannot account for spatial dependence between microscale locations, such as in the case of point-referenced data. The selection of groups is often based on criteria such as weather stations and counties that may not be natural for spatial events, and moreover it can be very cumbersome on deciding on a reasonable and complex dependence structure.

We propose to use latent factors to account for complex and high-dimensional spatial dependence. This approach does not belong to either of the above two themes. In our spatial dependence model, copula functions are used to model dependence among microscale locations. Copula models have been widely used in social, economical, and financial areas for modeling complex dependence structures. We refer to Joe (2014) for recent developments in copula theories and applications. Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ have marginal univariate cumulative distribution functions (cdf) F_1, \dots, F_n , respectively. The copula function C corresponding to the joint cdf F is defined through

$$F(y_1, \dots, y_n) = C(F_1(y_1), \dots, F_n(y_n)).$$

Sklar (1959) showed that there exists a unique copula function C when the cdf F_i are all continuous. The Copula provides flexible statistical tools for modeling dependence structures, especially the nonlinear dependence commonly seen in insurance risks.

Among several copula approaches that are suitable for high-dimensional dependence modeling, such as the vine copula (Joe 1994; Cooke 1997; Bedford and Cooke 2002; Aas et al. 2009; Kurowicka and Joe 2011) and factor copula (Krupskii and Joe 2013, 2015), we find the factor copula approach to be particularly useful and flexible for modeling spatial dependence. Here we use a one-factor copula model to briefly explain the idea, and the idea can be easily extended to multiple factors. Let V be a latent factor that is connected with each $Y(s)$, $s \in \mathcal{D}$, through a bivariate copula $C(u, v; \theta(s))$ where $\theta(s)$ denotes the dependence parameters corresponding to location s . $Y(s)$ are conditionally independent given V . Then the dependence structure among $Y(s)$ can be obtained by integrating over the support of V .

To the best of our knowledge, the proposed method is the first copula-based approach for modeling spatial dependence among point-referenced data directly. A major contribution of the present article is that, instead of modeling the dependence between many different locations directly like the way vine copulas do (Erhardt et al. 2015b), we propose to use latent auxiliary random variables to model each single location separately, so that the spatial features can be easily learned at each location. This provides ease and flexibility in modeling individual spatial locations, as opposed to modeling the connections among them. More specifically, we assume that the dependence parameters $\theta(s)$ rely on the geographical location s , so that the dependence structure between two different locations s_i and s_j can be obtained through marginalization. Thus the task of modeling the high-dimensional spatial dependence becomes an relatively easier task of estimating the function $\theta(s)$. In addition, we propose to use kernel smoothing functions, such as radial basis functions, to model $\theta(s)$. We show that the function $\theta(s) : \mathbb{R}^2 \rightarrow \text{codomain}(\theta)$ can be effectively estimated based on data via this kernel smoothing model. The proposed semiparametric approach based on factor copulas is effective in accounting for the high-dimensional spatial dependence, using which both the well-known nugget effects on microscale areas and the dependence between microscale locations can be easily estimated.

We implement the proposed method in the analysis of a thunderstorm wind loss dataset in Texas from year 2006 to 2012. The detailed modeling, implementation, and analysis as well as predictive methods for loss prediction for positive wind loss amounts in this dataset are provided in Section 3. We also discuss extensions to loss frequency and to spatio-temporal models.

The paper is organized as follows. Section 2 introduces the concept of factor copulas, and how spatial dependence can be introduced in factor copula models. The radial basis function model for the function $\theta(s)$ is also introduced here. Section 3 illustrates the application of the proposed model with a detailed analysis of the Texas thunderstorm wind damage dataset. In Section 4, extensions of the proposed model to spatio-temporal modeling and to loss frequency models are briefly discussed. Concluding remarks are provided in Section 5.

2. FACTOR COPULAS FOR SPATIAL DEPENDENCE

Copula-based models for high-dimensional spatial dependence includes the flexible vine copula model; see, for example, Erhardt et al. (2015b). In the vine copula approach, one first constructs groups within the sample by, for instance, weather stations or counties, and then models the dependence structure among different groups using vine copulas. Vine copula models thus cannot explain the dependence within the grouped data, referred to as microscale dependence here (e.g., for the so-called point-

referenced data in the spatial literature). Additionally, the choice of an ideal method to connect these weather stations or other types of clusters remains a challenge.

When the observations contain geographical coordinates, such as longitudes and latitudes in point-referenced data, we may not want to group those data based on a criterion but may choose to consider the underlying random variable at each distinct location individually instead. The random variables at these different geographical locations are assumed to be independent conditional on latent factors \mathbf{V} . Furthermore, these random variables at different locations share a similar sampling scheme, with the strength of dependence varying with the longitudes and latitudes. To this end, in order to model the complex spatial dependence structure, one only needs to conduct inference on the “similar sampling scheme.”

The factor copula fits into the above idea very well, and it overcomes the need to model complex connections directly. The factor copula approach allows us to model each specific location separately, and the interdependence among these locations can be automatically taken care of by the latent factors.

2.1. The Model

We introduce the factor copula model by using three random variables, and the idea easily extends to high-dimensional cases. Figure 1 demonstrates how a trivariate one-factor copula model can be constructed based on three bivariate copulas, for which one needs to estimate only the $\theta(s_i)$ functions that are location-wise properties.

Let $Y(s_1)$, $Y(s_2)$, and $Y(s_3)$ be random variables representing the losses at three different locations s_1 , s_2 , and s_3 , respectively, with a joint cumulative distribution function (cdf) $F(\cdot, \cdot, \cdot)$, joint density function $f(\cdot, \cdot, \cdot)$, marginal cdf $F_i(\cdot)$, and marginal density function $f_i(\cdot)$, $i = 1, 2, 3$. Let V be a latent factor with cdf $F_V(\cdot)$ and density $f_V(\cdot)$. Without loss of generality, V is assumed to follow a Uniform(0, 1) distribution in what follows. The dependence between $Y(s_i)$ and V is then modeled by a copula $C_i(\cdot, \cdot)$; that is, the joint cdf of $(Y(s_i), V)$ is $F_{iV}(y_i, v) = C_i(F_i(y_i), v)$. Conditional on the latent factor V , the $Y(s_i)$ are assumed to be independent. The joint density function of $Y(s_1)$, $Y(s_2)$, and $Y(s_3)$ can then be written as

$$f(y_1, y_2, y_3) = \int_0^1 \prod_{i=1}^3 c_i(F_i(y_i), v) f_i(y_i) dv, \quad (1)$$

where we have used the fact that $f_V(v) = 1$ for $0 \leq v \leq 1$, and $c_i(\cdot, \cdot)$ is the density function for copula $C_i(\cdot, \cdot)$.

The copulas C_i 's can belong to different families, but for the convenience of statistical inference, in what follows, we assume that C_i 's are within the same parametric family, indexed by the dependence parameters θ . This assumption does not bring any obvious constraint as we allow the dependence parameters to change among locations, and thus the overall dependence of $Y(s_i)$ can still be very flexible. Write $C_i(\cdot, \cdot) \equiv C(\cdot; \theta(s_i))$, where we assume the dependence parameters θ vary smoothly in s_i , as it is reasonable to assume that random variables close to each other should interact with the other locations and contribute to the overall dependence in a similar way. If there are k dependence parameters such that $\theta(s) = (\theta_1(s), \theta_2(s), \dots, \theta_k(s))$, $s \in \mathcal{D}$, we can estimate them separately. Therefore, it suffices to estimate the smooth function $\theta_j(s)$ for each $j = 1, 2, \dots, k$.

The main advantages of the proposed approach are the following.

1. It provides flexibility in modeling dependence structures and marginals as opposed to the commonly assumed linear dependence structure.

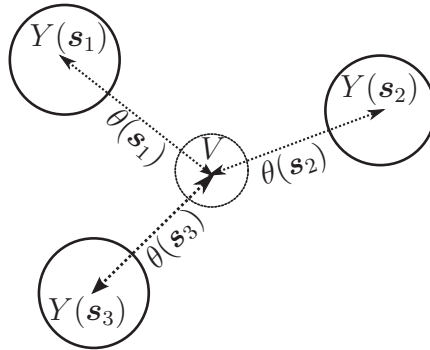


FIGURE 1. One-Factor Copula Model for Spatial Dependence.

2. The task of modeling complex high-dimensional spatial dependence becomes a relatively easier task of estimating the function $\theta(s)$.
3. The complex spatial dependence can be explicitly modeled by the latent factors, and the model can be calibrated based on observed data.
4. Unlike vine copula approaches, it can capture the dependence between any microscale geographical locations.
5. The model can be easily extended to a spatio-temporal model by letting s include both the coordinates of the geographical location and the time t .

If a one-factor model is not sufficient for capturing the complex dependence structure, then two-factor or more general p -factor models can be applied. The following is an expression for spatial dependence based on a two-factor copula model (see Krupskii and Joe 2013):

$$\begin{aligned}
 C(u_1, \dots, u_n; \theta_1, \theta_2) &= \int_0^1 \int_0^1 \prod_{i=1}^n \mathbb{P}[U_i \leq u_i | V_1 = v_1, V_2 = v_2; \theta_1(s_i), \theta_2(s_i)] dv_1 dv_2 \\
 &= \int_0^1 \int_0^1 \prod_{i=1}^n \frac{\partial \mathbb{P}[U_i \leq u_i, V_2 \leq v_2 | V_1 = v_1; \theta_1(s_i), \theta_2(s_i)]}{\partial v_2} dv_1 dv_2. \\
 &= \int_0^1 \int_0^1 \prod_{i=1}^n C_{i|V_2; V_1}(C_{i|V_1}(u_i | v_1; \theta_1(s_i)) | v_2; \theta_2(s_i)) dv_1 dv_2.
 \end{aligned}$$

For each geographical location s_i , there are two bivariate copulas involved: $C_{i|V_1}(\cdot, \cdot; \theta_1(s_i))$ and $C_{i|V_2}(\cdot, \cdot; \theta_2(s_i))$. Therefore, one needs to estimate only the two smooth functions $\theta_1(s)$ and $\theta_2(s)$.

The integration for the factor copula can be performed by numerical methods. For instance, we can use Gaussian quadrature to approximate the integration. When there are not many factors involved, the Gaussian quadrature method can lead to accurate and fast approximations. 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 Joe (2014) for R implementations of factor copulas.

2.2. Gaussian Copulas as Link Copulas

The bivariate link copulas should belong to the same parametric family to make a smooth function $\theta(s)$ reasonable in accounting for spatial dependence. For instance, the bivariate copula family can be one of the bivariate Gaussian, Frank's, and Student's t copulas.

If the bivariate link copulas are not Gaussian, then an integration has to be involved in the representation of factor copulas, which can be handled numerically for a small number of latent factors. However, if all bivariate link copulas are Gaussian, then based on Krupskii and Joe (2013), factor copulas have Gaussian mixture stochastic representations. For example, with the one-factor model,

$$Y(s_i) = \rho_{i1}W + \sqrt{1 - \rho_{i1}^2}\epsilon_i, \quad i = 1, \dots, n,$$

where $W, \epsilon_1, \dots, \epsilon_n$ are i.i.d. $N(0, 1)$ random variables. It can be easily verified that the correlation coefficient between $Y(s_i)$ and $Y(s_k)$ is $\rho_{i1}\rho_{k1}$. Since we specify the correlation coefficient between $Y(s_i)$ and V as a function of the location s_i , ρ_{i1}^2 can then be interpreted as the nugget effect; see Journel and Huijbregts (1978) for the concept of the nugget effect. For the p -factor Gaussian factor copula,

$$Y(s_i) = \sum_{j=1}^p \rho_{ij}W_j + \sqrt{1 - \sum_{j=1}^p \rho_{ij}^2} \cdot \epsilon_i, \quad i = 1, \dots, n$$

where $W_1, \dots, W_p, \epsilon_1, \dots, \epsilon_n$ are i.i.d. $N(0, 1)$ random variables, and the marginal correlation between $Y(s_i)$ and $Y(s_k)$ is $\sum_{j=1}^p \rho_{ij}\rho_{kj}$; the nugget effect is $\sum_{j=1}^p \rho_{ij}^2$ for location i . To facilitate the numerical optimization, one often use Fisher's z-transformation of ρ as a link function, that is, $\nu := (1/2)\log[(1 + \rho)/(1 - \rho)]$, where $-1 < \rho < 1$. So $\rho = (\exp(2\nu) - 1)/(\exp(2\nu) + 1)$.

In Section 3.4, estimates based on a real dataset will be presented using bivariate Gaussian link copulas, where both positive and negative spatial dependence structures are captured by the proposed model.

2.3. Radial Basis Functions for Strength of Dependence

In this section, we discuss estimation of $\theta(s)$ based on observations $y(s_i)$, $i = 1, \dots, n$, where the arguments in s are spatial coordinates such as longitude and latitude. One can use kernel smoothing functions to approximate the vector of functions $\theta(s)$. In this article, we will focus on radial basis functions, and this family consists of many useful kernel smoothing functions, such as Gaussian kernels and inverse multiquadratic kernels.

Denote m as the dimension of the spatial index s . For a real-valued continuous function $\theta : \mathbb{R}^m \rightarrow \mathbb{R}$, we can approximate it by the form

$$\theta(s; \mathbf{w}, K, \gamma) \approx \sum_{k=1}^K w_k \phi(\|s - \mathbf{e}_k\|), \quad (2)$$

where $\mathbf{w} = (w_1, w_2, \dots, w_K)$, with $w_k \in \mathbb{R}$ being the weights that can be estimated by solving linear models using methods such as the maximum likelihood approach, $\phi : \mathbb{R} \rightarrow \mathbb{R}$, is the basis function, γ is a shape parameter for the basis function ϕ , and \mathbf{e}_k are the K prespecified points where the function θ is evaluated. The norm $\|\cdot\|$ is usually the Euclidean norm, but other types are also allowed. The choice of radial basis function depends on the problem of interest. Among many others, some commonly used basis functions (see, e.g., Mongillo 2011) are the following:

$$\begin{aligned} \text{Gaussian: } \phi(x) &= \exp(-\gamma x^2); \\ \text{Inverse multi-quadratic: } \phi(x) &= \sqrt{x^2 + \gamma^2}; \\ \text{Thin plate spline: } \phi(x) &= x^2 \ln(x). \end{aligned}$$

Note that, with a sufficiently large K , one can approximate $\theta(s; \mathbf{w}, K, \gamma)$ with arbitrary precision. From the viewpoint for statistical inference, however, large K values may lead to overfitting and poor prediction.

There are several ways of estimating the weights $\mathbf{w} = (w_1, w_2, \dots, w_K)$. For example, one can construct a linear system as

$$\sum_{k=1}^K w_k \phi(\|s_i - \mathbf{e}_k\|) = y_i, \quad i = 1, \dots, n, \quad (3)$$

for each distinct observation (s_i, y_i) . When $K \leq n$, the weights \mathbf{w} can be easily solved as the least square estimates of the regression coefficients from linear regression.

The number K cannot be very large in order to avoid a possible overfit, such as an extremal case of the saturated model when $K = n$. One can choose some representative centers instead, and a possible option is to select the K centers and group the data based on certain clustering algorithms, such as the K-means clustering. 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. The function $\theta(s; \mathbf{w}, K, \gamma)$ can then be approximated at those K centers of the clusters. The K-means method assigns those K centers so that the within-cluster sum of squares is minimized. More specifically, let S_K be the partition, then

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

where $\|\cdot\|$ is the Euclidean norm, and \sqcup is the union operation of mutually exclusive and exhaustive subsets of S . For our later applications, s correspond to the longitudes and latitudes associated with the samples.

We remark that the K-means method is just one way to partition the data. What we need is a partition over which to approximate the function $\theta(s; \mathbf{w}, K, \gamma)$, so we hope that data can be clustered roughly evenly based on geographical information only. To this end, the K-means approach is a reasonable data preprocessing approach, and it works well for our real application.

In what follows, we will use the radial basis function method twice: one for normalizing the data for accounting for geographical effects in order to model spatial heterogeneity, and the other for spatial dependence, that is, for approximating the dependence parameter $\theta(s; \mathbf{w}, K, \gamma)$ as a function of the geographical location s . We refer to Mongillo (2011) for more information on how to choose the type of the basis function, the number of clusters K , and the shape parameter γ in the basis function.

TABLE 1
Summary of Loss Amounts (Thousands)

	Sample size	Min.	1st quantile	Median	Mean	3rd quantile	Max.
Training	6765	0.01	4.34	11.39	89.26	32.88	54100
Test	433	0.20	4.06	7.10	42.97	20.29	5073

3. A CASE STUDY WITH THUNDERSTORM LOSS DATA

The thunderstorm wind loss dataset contains property damage losses due to thunderstorm winds in Texas, United States, from 1996 to 2012, obtained from the National Climatic Data Center (NCDC) of the National Oceanic and Atmospheric Administration (NOAA). For our analysis, we used loss data from 1996 to 2011 as training data for fitting the model, and the 2012 data for evaluating out-of-sample forecasting. We adjusted the loss amounts by the consumer price index (CPI) to the 2013 level to minimize the inflation effect. The data contain many records that have zero loss amounts. In this section, we describe loss severity modeling of nonzero losses. Loss frequency modeling including zeros can also be conducted with our proposed approach, which will be briefly discussed in the following section.

The steps for our modeling and analysis of these data are as follows: (1) The training and test datasets are first prepared. This step includes adjusting the payment amounts to reflect inflation, identifying outliers that might significantly affect the analysis, etc. (2) Model selection is performed for the univariate marginals, after accounting for potential effects from spatial heterogeneity, seasonality, and other covariates. At this step, we consider only the marginal models; the spatial dependence modeling is considered at the next step. (3) The proposed factor copulas model is then used to account for the spatial dependence. The key for this step is to estimate the bivariate function $\theta(s)$ that encodes the spatial dependence structures, where we use radial basis functions for approximating $\theta(s)$. (4) This step focuses on maximum likelihood estimation (MLE) of the parameters of the model. The likelihood function of the model includes both contributions from the marginal models for spatial heterogeneity and the factor copula model for spatial dependence. We first obtain estimates of the parameters in the univariate marginal models which are then used as initial values in the numerical optimization process for obtaining the MLEs of the overall model. (5) At this final step, the performance of the proposed model is evaluated and compared with other existing approaches.

3.1. The Data

There are 254 counties in Texas and county-level population density is used as a covariate in the model we fit. Population density data are obtained from U.S. Census Bureau (<http://www.census.gov/>) and are calculated as the population in the corresponding year divided by the area of the county.

Table 1 is a summary of the monthly inflation adjusted loss amounts for the training set and for the test set, respectively. Figure 2 presents inflation adjusted loss amounts by month. From Figure 2, we observe that there are clear seasonality patterns in the loss amounts, and we may use trigonometric functions such as $\sin(\cdot)$ and $\cos(\cdot)$ to account for the pattern.

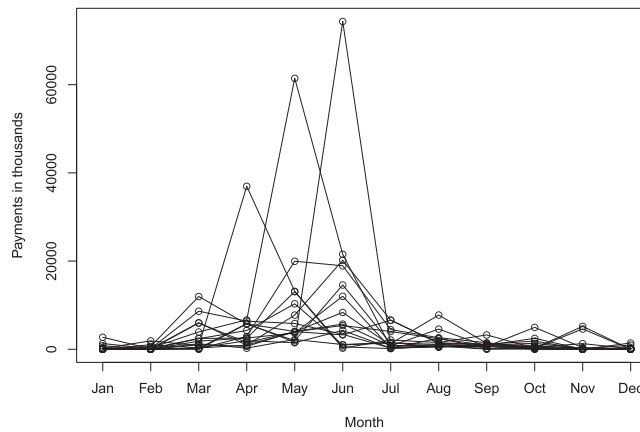


FIGURE 2. Seasonality of Loss Amounts (Thousands) from Year 1996 to 2011.

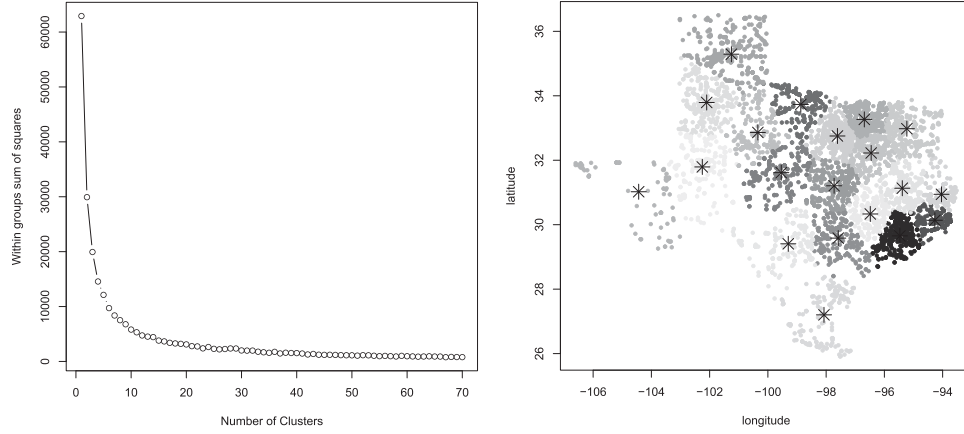


FIGURE 3. Within-Group Sum of Squares vs. Number of Clusters (Left), and 20 Chosen Clusters (Right).

On February 10, 1998, parts of Texas experienced a series of rare winter storms. As Figure 2 illustrates, there were almost no other February losses in the observed data when we exclude this rarely extreme datum. In preliminary analyses, we have explored fitting our proposed and comparator models both including and excluding this extreme event. Including this datum affects the fit of the models in the training data, but we did not observe significant effects on predictive performance in test data. Whether including the extreme event or not, the proposed model outperforms its counterparts for forecasting. In the analysis that follows, we consider loss data from 1996 to 2011 in the training data excluding this most extreme event.

3.2. The Spatial Heterogeneity Model

We propose to model both the spatial heterogeneity as well as the spatial dependence of the thunderstorm loss data. The spatial heterogeneity is modeled by a linear regression model. The model diagnostics support the normal distribution assumption for the natural logarithm of the losses. Therefore, we let Y_i be the logarithm of the loss amounts adjusted by inflation indexes, where $i = 1, \dots, n$ with the sample size $n = 6765$. We specify the model as

$$\mu = \mathbb{E}[Y|s, t, x] = \eta(s) + \beta_0 + \beta_1 t + \beta_2 \sin(wt) + \beta_3 \cos(wt) + \zeta x, \quad (4)$$

where s contains the longitude and latitude of the location; $\eta(s)$ is a function used to explain the geographical effects at s , and it can be estimated directly by least square estimates as discussed, and therefore, one can look at $Y(s) - \hat{\eta}(s)$ as normalized response variables for further analysis; $t = 1, \dots, 12$ is the month indicator; $w = 2\pi/12$; x is the population density of the county where s is located.

Considering that there are 254 counties in Texas, instead of using a fixed effect for each county that requires too many degrees of freedom, we use a smooth function $\eta(s)$ to explain the effect of a specific location s . In order to estimate $\eta(s)$, we can use kernel-smoothing functions. The Gaussian kernels worked well for estimating $\eta(s)$. We choose the number of clusters with the K-means algorithm based on how much variability is explained by those clusters. The left panel of Figure 3 shows the relationship between the within-group sum of squares and the number of clusters. We observe that when there are about 10 or more clusters, the total within-group variability is dramatically decreased. In other words, those 10 or more clusters are representative.

We chose 20 clusters as illustrated by the right panel of Figure 3, and the Gaussian basis function was chosen with the shape parameter γ determined based on root mean square errors (RMSE), that is, $\arg \min_{\gamma} \sqrt{\sum_{i=1}^n (\hat{y}_i(\gamma) - y_i)^2 / n}$.

Figure 4 illustrates how the value of γ affects the RMSE. Note that, when there are 20 clusters, the ideal γ is about 0.03, and we used these two specific values in estimating the function $\eta(s)$.

3.3. The Spatial Dependence Model

For spatial dependence, we apply the proposed factor copula approach. We consider only geographical locations as the arguments of $\theta(s)$ to demonstrate the method. If one believes that there are some other covariates that may affect the dependence structure, then those covariates can be included as additional arguments of the function $\theta(\cdot)$. However, for the particular data, we assume that the spatial locations are the only factors that may influence the spatial dependence. Therefore, the radial basis function $\theta(s; w, K, \gamma)$ is a singular term that accounts for how each location contributes to the overall dependence structure through the latent factor.

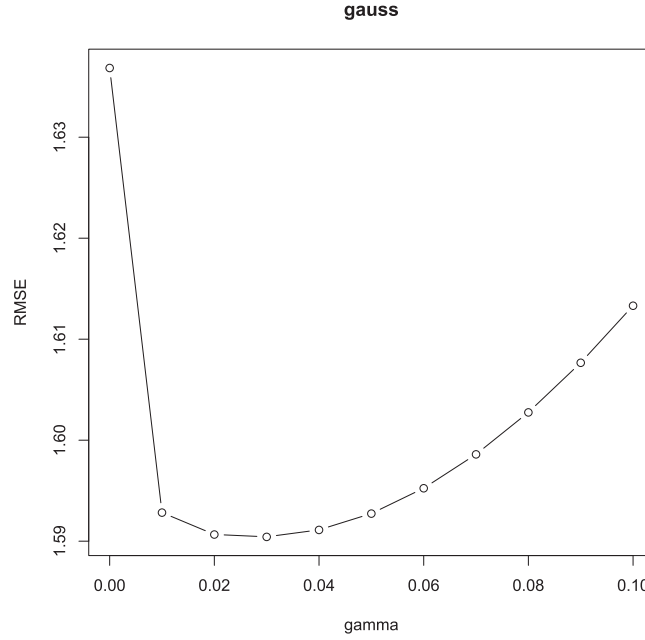


FIGURE 4. RMSE vs. Shape Parameter of Basis Function.

The choice of radial basis functions determines the speed of decay of either the spatial heterogeneity in the previous subsection, or spatial dependence for the current subsection that evolves along geographical locations. The type of radial basis functions can be chosen based on overall model-fitting performance, such as AIC. We used both Gaussian basis and inverse multiquadratic basis functions, where the former has an exponential decay and the latter has a power decay. The Gaussian radial basis function outperforms according to AIC, and therefore we choose to use it for the current analysis.

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, with the complex dependence structures explained by the function $\theta(s; \mathbf{w}, K, \gamma)$, which can be very flexible and can be estimated based on data.

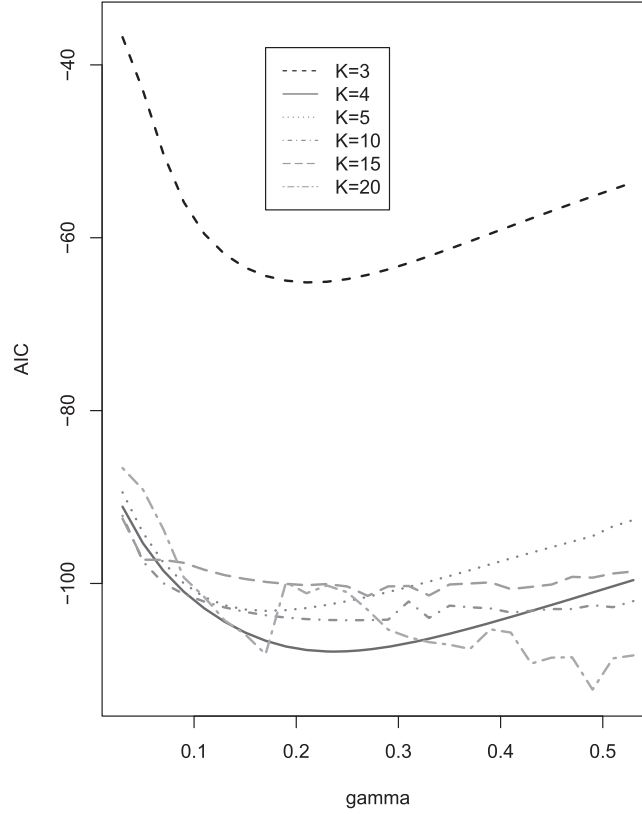
One can first choose different copula families to check which family leads to a better fit based on certain criteria such as AIC. For the current data, we conducted model fitting by using bivariate link copulas such as Gaussian copula, Frank's copula, Student's t copula, and Gumbel's copula, respectively. We found that, first of all, the link copula has to be able to account for both positive and negative dependence. Due to the lack of this feature, Gumbel's copula always leads to the worst performance. Gaussian, Frank's, and Student's t copulas all have both positive and negative dependence, and we found that Gaussian and Student's t copulas are generally better than Frank. Moreover, for Student's t copulas, the estimated degree of freedom was very large, and the corresponding AIC was comparable to that from Gaussian copula models. Therefore, we chose the bivariate Gaussian copula as the link copula, and the results reported in what follows are based on this copula.

3.4. Parameter Calibration

The model for the storm data contains two parts: the proposed factor copula approach for high-dimensional spatial dependence, and a regression model for spatial heterogeneity in the marginals. For parameter calibration, we use MLE based on the joint likelihood function containing both the marginal spatial heterogeneity models and the spatial dependence model.

Note that, as we use the logarithm of the loss amount as the response variable, the normal assumption works well for the current data. Hence, for approximating $\eta(s)$ in Equation (4), the 20 weights for the radial basis function are obtained as the estimated regression coefficients from linear regression. After approximating the $\eta(s)$ in (4), we fixed their weights and used the estimated $\hat{\eta}(s)$ to normalize the response variable. Then the rest of the parameters were estimated altogether by the maximum likelihood approach. The overall likelihood function is

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

FIGURE 5. Tuning Parameters of K and γ .

where $c(\cdot)$ is the n -dimensional copula density,

$$f_i(y_i; \boldsymbol{\beta}, \zeta, \sigma) = \phi(y_i), \quad \phi : \text{normal density with} \\ \text{mean } \beta_0 + \beta_1 t_i + \beta_2 \sin(\omega t_i) + \beta_3 \cos(\omega t_i) + \zeta x_i; \\ \text{variance } \sigma^2,$$

and $\theta(s; \mathbf{w}, K, \gamma)$ is the dependence parameter of the copula and the value θ is determined by the geographical location s through the radial basis function given in Equation (2).

Unlike the case of $\eta(s)$ where we approximated the function through a separate procedure, the dependence parameter $\theta(s; \mathbf{w}, K, \gamma)$ is embedded in the likelihood and cannot be solved directly through the relatively easier linear regression approach. Instead, we can estimate the parameters in Equation (2) by the maximum likelihood method. The parameters K and γ for the radial basis functions can be viewed as tuning parameters. Although γ can also be treated as a generic parameter to be estimated, the direct involvement of an unknown γ in the likelihood function, based on our experience, would lead to a much slower and less unstable process in getting the MLEs. Hence, it is suggested that the number of clusters K and the shape parameter γ for the radial basis function be treated as tuning parameters.

We compared the Gaussian kernel and the inverse multiquadratic kernel, and the former did a better job based on AIC. In what follows, we use the Gaussian kernel to demonstrate the data analysis with the proposed method. We tuned the values of K and γ by comparing the AICs based on the training set under various combinations of K and γ . The residuals from the marginal regression models were used as the pseudo data for tuning K and γ , and for choosing the dependence structure of the copula C . As mentioned above, the function $\theta(s; \mathbf{w}, K, \gamma)$ to be estimated is embedded in the nonlinear likelihood function. Hence, the speed of computation should also be taken into consideration when choosing a candidate combination of K and γ . We refer to Mongillo (2011) for more discussions about tuning the parameters K and γ .

Figure 5 displays the choices of the tuning parameters K and γ based on AIC. Note that here the listed AIC values are based only on the residuals used to tune the parameters. These are different than those to be reported in Table 2, where the AIC values are

TABLE 2
MLEs of Models, $AIC = -2 \times \text{Log Likelihood} + 2 \times \text{Number of Parameters}$

		Independence	S.E.	Proposed model	S.E.
Loss amounts	Intercept	9.55e+00	4.16e-02	9.46e+00	3.98e-02
	t	-1.11e-03	3.50e-04	-1.34e-03	3.43e-04
	$\sin(t)$	6.75e-02	3.02e-02	1.04e-01	2.90e-02
	$\cos(t)$	1.14e-01	3.59e-02	7.68e-02	3.34e-02
	Pop. density	8.85e-05	3.40e-05	1.30e-04	2.89e-05
	σ	1.63e+00	1.41e-02	6.29e-01	1.52e-02
Dependence	w_0			-2.64e-01	8.96e-02
	w_1			-2.90e-01	9.82e-02
	w_2			-9.07e-01	8.68e-02
	w_3			1.12e+00	1.02e-01
	w_4			-3.50e-01	1.29e-01
AIC		25846.22		25536.42	

based on the overall model including the likelihood due to the marginals and dependence when applicable. The computation was conducted in software R, and the initial values for optimizing the objective functions were chosen randomly with some reasonable constraints to ensure that the iterations converge. We note that when K is relatively larger, say, $K > 10$, the AIC values become unstable, with the speed of calculation greatly decreased for a larger K . Based on Figure 5, we finally chose $K = 4$ and $\gamma = 0.23$ for the candidate model for spatial dependence. The MLEs are reported in Table 2, including those for the marginals without considering the effect of the spatial dependence.

Table 2 indicates that, the proposed spatial dependence model leads to a smaller AIC and thus can improve the model fitting. Both models suggest that, over the years, the average property damage loss amount due to thunderstorm wind in Texas has been slightly decreasing, after adjusting for inflation. The population density certainly affects the loss amount in such a way that a higher population density leads to a higher average loss amount. In Section 3.5, we will show that the introduction of spatial dependence leads to better out-of-sample performance as well.

For a given set of values of covariates, such as longitudes, latitudes, population density, and months, risk assessment can be directly conducted based on the model in Equation (4) and the estimates of parameters from Table 2. The uncertainty in the risk assessment can be quantified based on the covariance matrix obtained for the MLEs. That is, assume that the set of parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \xi)$ in Equation (4) have an approximate multivariate Normal distribution with the mean being the MLEs and the variance-covariance matrix being approximated by the inverse of the estimated Hessian matrix and denoted as $\hat{\Sigma}$. Then the standard error of the predicted loss $\mu = \mathbb{E}[Y|s, t, x]$ can be easily calculated. That is, $\hat{\sigma}_\mu^2 := \widehat{\text{var}}(\mathbb{E}[Y|s, t, x]) = \mathbf{x}' \hat{\Sigma} \mathbf{x}$, where the column vector $\mathbf{x} = (1, t, \sin(wt), \cos(wt), \text{pop.den})^\top$ contains a given set of covariates. On the original scale, the risks $\exp\{Y\}$ are highly skewed to the right, and thus the median will be a more robust measure of risk severity at the particular location. Hence, the risk severity can be assessed as $\exp\{\hat{\mu}\}$, with μ defined in (4), and a $(1 - \alpha)$ confidence interval (C.I.) of the risk $\exp\{\mu\}$ can be calculated by inversion as $(\exp\{\hat{\mu} + \hat{\sigma}_\mu \Phi^{-1}(\alpha/2)\}, \exp\{\hat{\mu} - \hat{\sigma}_\mu \Phi^{-1}(\alpha/2)\})$, where $\Phi^{-1}(\cdot)$ is the inverse of the cdf of the standard Normal distribution. Note that we treat the $\eta(s)$ in (4) as fixed effects for spatial heterogeneity, so it will not contribute to the uncertainty in estimating $\mathbb{E}[Y|s, t, x]$. On the other hand, taking $\eta(s)$ as fixed effects will not affect but greatly simplify the spatial dependence modeling. In Table 3, we give some examples of risk assessments (on the original scale) for four major cities in Texas for July 2012 based on the estimated model from the training set, with 95% C.I.s of the median risks reported for reference.

The microscale correlation between any two geographical locations s_i and s_j can be estimated by $\hat{\rho}(s_i)\hat{\rho}(s_j)$, while the nugget effect based on spatial correlations at s_i can be estimated as $\hat{\rho}(s_i)^2$. Based on Figure 6, one can observe negative dependence between two major cities in Texas: one is Houston and the other is Dallas. This interesting pattern was automatically identified by the radial basis function. The pattern is consistent with the phenomena that for the months when there were more losses in Houston there were usually fewer losses in Dallas for the same period, and vice versa. In the surrounding areas of Houston, the spatial dependence is significantly positive among different locations, and the same pattern appears in the surrounding areas of Dallas. The spatial dependence among other areas is very mild, as is suggested by the contour plot.

TABLE 3
Risk Severity Assessment for Four Major Cities in Texas for July 2012

	Austin	Dallas	Houston	San Antonio
Median risks	10056	12558	12085	10535
95% C.I. (upper)	11098	14763	14017	11734
95% C.I. (lower)	9111	10682	10419	9458

3.5. Model Comparison

The performance of the proposed model can be compared with that of existing spatial models for point-referenced data. For this particular dataset, the Gaussian copulas are reasonable link copulas for the proposed factor copulas approach. Hence, a reasonable alternative model will be Gaussian Markov Random Fields (GMRFs) for the spatial dependence between residuals, with the same regression model for the mean process $\mu(s)$ for the spatial heterogeneity.

The models were fitted using data from year 2006 to 2011, with the model performance compared using the forecasts for year 2012 versus the actual values. We use root mean square errors to compare the accuracy of forecasting. Table 4 lists the RMSE when assuming no spatial dependence, GMRF, and the proposed factor copula structure with different combinations of tuning parameters. The corresponding γ values were selected for each of the K values, based on AIC values plotted in Figure 5. These combinations of the tuning parameters are included for comparing the corresponding models in terms of out-of-sample performance.

From Table 4, we observe that, for the current test set, the out-of-sample performance of the proposed model is better, regardless of the value of K when $K \leq 10$. When $K > 10$, the models might overfit the training set, leading to unstable predictions on the test data. We should remark that the computational cost is dramatically lower and the results are much more stable for smaller values of K . Hence, we decided to choose $K = 4$ and $\gamma = 0.23$ for the candidate model, based on in-sample performance and out-of-sample performance, as well as computational speed and stability.

For point-referenced data, the GMRF approach may fail to work (due to the computational difficulty of inverting a large-scale covariance matrix) when there are a large number of locations. This happened for the current study where the number of locations is 6765 for the training set. Hence, we employed the typical approach of aggregating the data into a smaller number of locations and modeling the dependence among the aggregated locations with GMRF. We used the K-means method to divide the data into 1000 clusters, with the mean of the longitudes and latitudes within each group used as the geographical coordinates for each cluster, and the mean of the residuals within each group as the residual. We then modeled the spatial dependence between these 1000 clusters with GMRF. The joint likelihood contributed from both the GMRF and univariate marginals was used to calibrate the parameters for the overall model. For the GMRF, we use the R package `geoR` to obtain the likelihood, with exponential functions assumed to

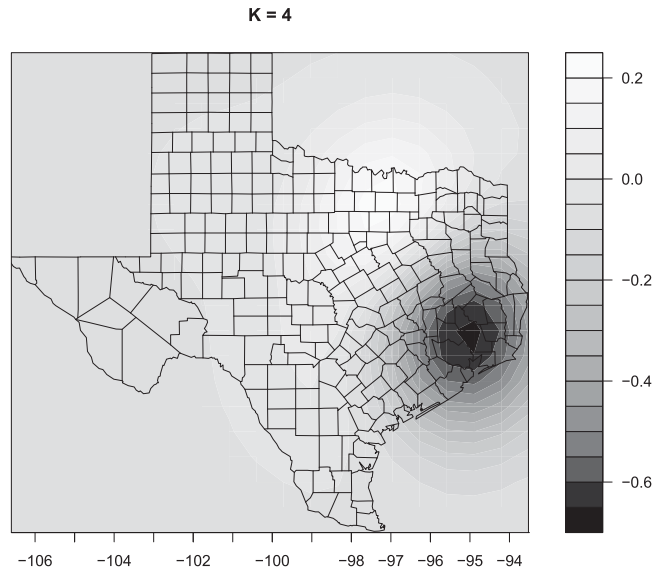


FIGURE 6. Estimated Spatial Dependence Parameters $\hat{\rho}(s)$.

TABLE 4
Root Mean Square Errors (RMSEs) for Forecasting

ind.	GMRF	$(K, \gamma) : (3, 0.21)$	$(4, 0.23)$	$(5, 0.17)$	$(10, 0.25)$	$(15, 0.27)$	$(20, 0.43)$
1.472	1.474	1.455	1.457	1.458	1.456	1.467	1.466

Note: The first column is for the model without considering spatial dependence, the second column is the case when spatial dependence is modeled by GMRF, and the other columns are for the proposed model with different tuning parameters.

represent how covariance decays with distance. We used the same training and test sets as those used by the other methods reported in Table 4.

For GMRF, one can also choose other spatial covariance functions and/or a different number of locations for the aggregated areal data. Nevertheless, the proposed model is still superior than the GMRF approach in the sense that it can easily handle high dimensional spatial dependence without aggregating the point-referenced spatial data, while accounting for much more flexible dependence structures with Gaussian and various non-Gaussian link copulas.

4. EXTENSIONS

In Section 3, a factor copula approach was applied for modeling spatial dependence in the continuous monthly loss amounts, with the spatial dependence assumed to be homogeneous over time. The proposed method can be easily extended to spatio-temporal models. Discrete data such as loss frequencies can also be easily modeled by the proposed approach. In what follows, we will briefly discuss the relevant extensions, and the efforts needed for such extensions are actually very mild.

4.1. Spatio-temporal Models

Constructing a spatio-temporal model is very straightforward with the proposed approach. The sampling scheme $\theta(s)$ can be used for both spatial and temporal dependence, by letting s be trivariate including the longitude, latitude, and time t . To partition the domain of $\theta(s)$, a partition for the time period $[0, t_{\max}]$ is also needed, where t_{\max} is the maximum time t considered. Therefore, the K for the radial basis functions becomes $K = K_l \times K_t$, where K_l is the number of partitions for the geographical area, and K_t is the number of partitions along the time.

We will not discuss details about the spatio-temporal model based on factor copulas; that is out of the scope of this article. We believe that the proposed model will be very promising in handling both spatial and temporal dependence, while the difficulties will be mainly the computational efficiency and speed for statistical inference.

4.2. Extension to Loss Frequencies

In insurance loss frequency modeling, the data usually contain many zero losses. One can use zero-inflated models, such as hurdle models, to account for the excess zeros in addition to those assumed by usual count processes. The proposed model can be easily extended to modeling spatial dependence for loss frequencies. We refer to Nikoloulopoulos and Joe (2013) for factor copulas for discrete data. In what follows, we give an example with hurdle models.

Let $M(s)$ be the number of losses at location s with $M(s) \in \{0, 1, 2, \dots\}$, and write $M(s) = I(s)J(s)$ with $I(s)$ being a Bernoulli random variable with the probability of success p that is independent of $J(s)$. A hurdle model can be represented as

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

We can assume that $J(s)$ follows a positive discrete distribution. For instance, assume that $J(s)$ follows a shifted Poisson distribution

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

Then the spatial dependence model based on one-factor copulas can be applied to model the dependence among $J(s)$. Therefore, the overall likelihood is

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

where F_j is the cdf of $J(s)$, p is the parameter for the Bernoulli random variable $I(s)$, α are the regression coefficients for $\lambda(s)$, and θ are the dependence parameters that can be written as a function of s ; that is, $\theta(s)$. Hence, the proposed approach for estimating $\theta(s)$ can then be used.

5. CONCLUDING REMARKS

Spatial dependence modeling has been an important topic in geostatistics or spatial statistics, where linear dependence models have been dominating the relevant literature. For actuarial applications, especially for various loss modeling, the dependence patterns observed are largely nonlinear, and thus the copula model is a natural candidate for modeling high-dimensional spatial dependence. For instance, the vine copula approach provides a successful tool for modeling potentially flexible and complex spatial dependence. The vine copula approach encodes the location information by choosing the nodes to be connected, and thus a great effort shall be put on constructing the vine itself.

To the best of our knowledge, our proposed approach is the first copula-based method for modeling spatial dependence among point-referenced data directly. It transforms the challenging task of constructing complex dependence relationships to inference of each single location. Thus a complex spatial dependence structure can be naturally represented by a smooth function whose arguments are the geographical coordinates. Therefore, the main task becomes estimating the smooth function, which is relatively easier than constructing a complicated network for spatial dependence. Hence, the microscale dependence structure in point-referenced data can be easily accounted for, which seems to be impossible with other copula-based methods such as vine copulas.

The main limitations of the proposed model are the following: The function $\theta(s)$ is embedded in the bivariate link copulas, and it does not seem easy to estimate the function directly from the data. Therefore, one way we adopted is to maximize the overall likelihood while approximating the function $\theta(s)$ with smoothing functions for which the associated parameters can be estimated. The MLE method works well for relatively smaller areas, such as the whole state of Texas. However, to model a large area, the number of clusters required for approximating $\theta(s)$ can be very large, which brings computational difficulties for obtaining the MLEs. Under this situation, new computational approaches need to be considered for conducting statistical inference with the proposed model. Feasible choices include composite likelihood methods and applying sequential estimations to choose better initial values for obtaining MLEs. We should also notice that, similar to the vine copula approach, the proposed model does not guarantee that a longer distance leads to weaker dependence, which in our opinion is not necessarily a must-have property of spatial dependence models.

Although the data we used to demonstrate the methodology suggest the Gaussian copula as a better candidate bivariate link copula, we believe that our model will be much more flexible and useful compared to existing methods, especially for modeling nonlinear spatial dependence, which can happen between extremal events or between events that have nonexchangeable dependence structures. This will be a very interesting topic for further investigations, and a comparison to the existing methods for nonlinear spatial dependence might be even more illuminating for understanding how useful the proposed model is.

ACKNOWLEDGMENTS

The research conducted is supported by the Casualty Actuarial Society and the Society of Actuaries through an Individual Grants Competition. The authors are thankful to the societies and to the reviewers of the grant for their precious time contributed. We also thank two anonymous reviewers and the editor for their constructive suggestions and comments.

ORCID

Lei Hua  <http://orcid.org/0000-0002-8825-0180>

REFERENCES

- Aas, K., C. Czado, A. Frigessi, and H. Bakken. 2009. Pair-Copula Constructions of Multiple Dependence. *Insurance: Mathematics and Economics* 44(2): 182–198.
- Bárdossy, A. 2006. Copula-Based Geostatistical Models for Groundwater Quality Parameters. *Water Resources Research* 42(11): 1–12. W11416.
- Bedford, T., and R. M. Cooke. 2002. Vines—A New Graphical Model for Dependent Random Variables. *Annals of Statistics* 30(4): 1031–1068.
- Cooke, R. M. 1997. Markov and Entropy Properties of Tree-and Vine-Dependent Variables. In *Proceedings of the ASA Section of Bayesian Statistical Science*, Vol. 27. American Statistical Association.
- Cressie, N. 1993. *Statistics for Spatial Data*. New York: John Wiley & Sons.
- Cressie, N., and C. K. Wikle. 2011. *Statistics for Spatio-temporal Data*. New York: John Wiley & Sons.
- Ebrahimi, N., and L. Hua. 2014. Assessing the Reliability of a Nanocomponent by Using Copulas. *IIE Transactions* 46(11): 1196–1208.
- Erhardt, T. M., C. Czado, and U. Schepsmeier. 2015a. R-Vine Models for Spatial Time Series with an Application to Daily Mean Temperature. *Biometrics* 71(2): 323–332.
- Erhardt, T. M., C. Czado, and U. Schepsmeier. 2015b. Spatial Composite Likelihood Inference Using Local c-Vines. *Journal of Multivariate Analysis* 138: 74–88.

- Gräler, B., and E. Pebesma. 2011. The Pair-Copula Construction for Spatial Data: A New Approach to Model Spatial Dependency. *Procedia Environmental Sciences* 7: 206–211.
- Joe, H. 1994. Multivariate Extreme-Value Distributions with Applications to Environmental Data. *Canadian Journal of Statistics* 22(1): 47–64.
- Joe, H. 2014. *Dependence Modeling with Copulas*. New York: Chapman & Hall.
- Journel, A. G., and C. J. Huijbregts. 1978. *Mining Geostatistics*. San Diego: Academic press.
- Krupskii, P., and H. Joe. 2013. Factor Copula Models for Multivariate Data. *Journal of Multivariate Analysis* 120: 85–101.
- Krupskii, P., and H. Joe. 2015. Structured Factor Copula Models: Theory, Inference and Computation. *Journal of Multivariate Analysis* 138: 53–73.
- Kurowicka, D., and H. Joe. 2011. *Dependence Modeling: Vine Copula Handbook*. Singapore: World Scientific.
- Mongillo, M. 2011. Choosing Basis Functions and Shape Parameters for Radial Basis Function Methods. *SIAM Undergraduate Research Online* 4: 190–209.
- Nikoloulopoulos, A. K., and H. Joe. 2013. Factor Copula Models for Item Response Data. *Psychometrika* 80(1): 126–150.
- Sklar, A. 1959. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris* 8: 229–231.

Discussions on this article can be submitted until October 1, 2017. The authors reserve the right to reply to any discussion. Please see the Instructions for Authors found online at <http://www.tandfonline.com/uaaj> for submission instructions.