Spatial Mixture Copula Model For Multiple Correlated Responses With An Environmental Application

Mohomed Abraj (✉ abrajsz20@gmail.com)

Queensland University of Technology (QUT) <https://orcid.org/0000-0002-0841-385X>

M. Helen Thompson

Queensland University of Technology (QUT)

You-Gan Wang

Queensland University of Technology (QUT)

Research Article

Keywords: Mixture copula, Multivariate spatial dependence, Environmental modelling, Interpolation

Posted Date: December 29th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1169352/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Spatial mixture copula model for multiple correlated responses with an environmental application

Mohomed Abraj^{*1,2,3}, M. Helen Thompson^{1,2}, You-Gan Wang^{1,2,3}

¹School of Mathematical Sciences, Faculty of Science, Queensland University of Technology (QUT), Australia

²QUT Centre for Data Science, Queensland University of Technology, Australia

³Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)

*abrajsz20@gmail.com; mohomed.amsar@hdr.qut.edu.au

Declarations

Conflict of interest

The authors have no conflict of interest.

Availability of data and material

Research data are electronically uploaded.

Code availability

Provided on request.

Ethics approval

Not applicable.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Author contributions

Model developing and simulation: Mohomed Abraj. Data analysis and interpretation: Mohomed Abraj. Principal study supervision: M. Helen Thompson. Associate study supervision: You-Gan Wang. Manuscript drafting: Mohomed Abraj. Critical revision of the manuscript: Mohomed Abraj, M. Helen Thompson, You-Gan Wang.

Acknowledgements

The authors would like to thank the Queensland University of Technology for funding the present study. Mohomed Abraj is funded through an Australian Government Research Training Programme award and a Queensland University of Technology Higher Degree Research award.

Abstract In environmental monitoring, multiple measurements are often collected at many locations and these measurements depend on each other in complex ways, such as nonlinear dependence. In this research, a novel copula-based geostatistical modelling approach was developed to model multivariate continuous spatial random fields using mixture copulas that captures both spatial and joint dependence of multiple responses over two-dimensional locations. In a bivariate context, the mixture copulas were used to capture the joint spatial dependence of a bivariate random field and the spatial copula of the bivariate random field was constructed as the convex combination of mixture copulas. The proposed model was applied to real forest data and simulated nonlinear data. The performance of the novel method was compared with existing spatial methods, which included a univariate spatial pair-copula model, a multivariate spatial pair-copula model that utilises nonlinear principal component analysis (NLPCA), and conventional kriging. The results show that the proposed model outperforms the existing methods in the interpolation of individual responses and reproduction of their bivariate dependence.

Keywords Mixture copula. Multivariate spatial dependence. Environmental modelling. Interpolation.

1 Introduction

In environmental science, multiple attributes are often measured at a given geographical location. For instance, mining industries collect multiple measurements of minerals, such as cadmium, zinc, and copper (Goovaerts *et al.*, 1997). Forest inventories monitor and collect multiple measurements of biomass, such as bole, foliage, stump, branch, and root biomass (Finley *et al.*, 2007). These multiple measurements can depend on each other and ignoring the joint dependence of multiple variables can decrease the accuracy in the interpolation process (Musafer *et al.*, 2017). Specifically, this article focusses on the spatial modelling of multivariate continuous spatial random fields that models the joint spatial dependence of multiple spatial responses. Subsequently, the individual responses can be interpolated across locations by utilising the modelled multivariate spatial dependency.

Traditionally, kriging based geostatistical models have been widely used to capture the linear relationships of univariate continuous random fields over space, where the variogram has been used to measure the dissimilarity of the variables at different locations (Cressie and Wikle, 2011). However, kriging based methods are limited in their ability to model real-world complexity, such as nonlinear spatial dependence (Musafer and Thompson, 2016). Also, a common assumption in the variogram method is that samples are independent of each other, which means that the kriging estimation variance is independent of the sample values (Diggle *et al.*, 2003; da Rocha and Yamamoto, 2000). This assumption is over simplistic, it is more realistic that sample points do depend on each other, for example, exhibiting high correlation between points that are close to each other and with decreasing correlation the further that points are

separated. In contrast to kriging, spatial copula models can be used to describe the spatial dependence structure of variables by considering the distance dependent correlation between sample values as well as capturing complex spatial dependencies (Musafer *et al.*, 2017).

Copulas provide a distribution free approach to model multiple random variables by capturing their nonlinear and multivariate dependence (Nelsen, 2007). The copula method has been widely applied in many fields, such as finance, economics, and civil engineering (D’Amico and Petroni, 2018; Patton, 2012; Li *et al.*, 2018), to describe the dependence structure of two or more random variables, and subsequently used in the prediction process. In geostatistics, Bárdossy (2006) introduced an univariate spatial copula framework to predict groundwater quality parameters, such as pH, sulphate, nitrate, and chloride, at unobserved locations. Bárdossy’s (2006) spatial copula method divides the distance over which spatial dependence exists into equally spaced intervals, also referred to as distance classes or spatial bins, and requires the use of the same family of copulas to be fitted for each spatial bin. Bárdossy and Li (2008) subsequently developed a method for interpolation from Bárdossy’s (2006) spatial copula model. Gräler (2014) proposed a more flexible copula based spatial model that permits copulas from different families to be fitted across the distance classes. They also developed a method for spatial interpolation using the pair-copula construction of Aas *et al.* (2009), to predict measurements at unsampled locations. The added flexibility of Gräler’s (2014) model, over Bárdossy’s (2006) model, permits increased accuracy in modelling and prediction. More recently, Sohrabian (2021) modified Gräler’s (2014) spatial model with mixture copulas to jointly model upper and lower tail dependence within each spatial bin. However, these spatial copula methods model only univariate continuous spatial random fields and are not capable of jointly modelling multiple dependent continuous spatial random variables.

Existing spatial models for multivariate random fields include coregionalisation kriging models (Cressie and Wikle, 2011), which ignore nonlinear spatial dependence and fail to reproduce the individual responses across locations (Leuangthong and Deutsch, 2003). Also, Gnann *et al.* (2018) proposed a method to model multivariate spatial dependence using Gaussian copula, which was not capable of modelling non-Gaussian spatial random fields. A solution to model non-Gaussian multivariate spatial random fields, Musafer *et al.* (2017) proposed a multivariate pair-copula spatial model, whereby the spatial variables are transformed into uncorrelated factors using NLPCA prior to fitting univariate spatial copula models to the independent factors. Subsequent back transformation is required to transform interpolated values to the scale of the original variables and to re-inject correlation. Musafer *et al.*’s (2017) method indirectly models the joint dependence between spatial variables through a black-box transformation. We propose a novel method to model multivariate spatial random fields that jointly models the spatial variables via a white-box mixture copula. In the application and simulation study presented, this multivariate spatial mixture copula model more accurately reproduced individual responses across locations and successfully

reproduced the joint dependence between variables.

2 Methodology

The proposed approach is presented in a bivariate context. However, the method can be extended to more than two responses with a trivial generalisation of the bivariate setting. The methodology consists of two essential components: modelling each spatial variable separately using Gräler's (2014) univariate spatial copula; then joining the univariate spatial copulas using Sohrabian's (2021) mixture copula. We also use the proposed model for interpolation : the joint dependence between variables is interpolated; then Nelsen's (2007) inverse conditional approach is used to obtain the conditionally interpolated values of an individual variable. A summary of each component of the methodology, and a description of the modelling and interpolation method are provided in the following.

2.1 Copulas

Copulas are joint cumulative distribution functions (CDFs) of two or more random variables with uniformly distributed margins on the interval $[0, 1]$. Let Y_1 and Y_2 be continuous random variables with marginal CDFs F_1 and F_2 , then there exists a unique copula $C : [0, 1]^2 \rightarrow [0, 1]$ such that,

$$\begin{aligned} C(u, v) &= P[F_1(Y_1) \leq u, F_2(Y_2) \leq v], \\ &= C(F_1(y_1), F_2(y_2)), \end{aligned} \tag{1}$$

where u and v are any selected quantiles on the interval $[0, 1]$.

There are various types of copula families that describe a variety of dependence structures. Commonly used bivariate copulas include Gaussian, Student- t , Clayton, Frank, Gumbel, and Joe copulas (Nelsen, 2007).

2.2 Spatial copulas for univariate spatial random fields

Whilst a non-spatial copula models the dependence between two or more non-spatial continuous random variables, a spatial copula describes the joint spatial dependence of an univariate spatial variable at any two spatial locations.

Let $Z(x)$ denote an univariate second-order stationary spatial random field Z that is observed at two-dimensional location $x \in \mathcal{X}$, and let $X = (x_1, x_2, \dots, x_n)$ be the set of existing observed locations in

the given spatial domain $\mathcal{X}$. The spatial copula that describes the dependence of a variable at any two locations separated by distance h is,

$$\begin{aligned} C_h(u, v) &= P[F(Z(x)) \leq u, F(Z(x+h)) \leq v], \\ &= C_h(F(Z(x)), F(Z(x+h))), \end{aligned}$$

here F is the CDF of Z and is assumed to be same at each location x .

In calculating the empirical bivariate copula distribution, the distances between every pair $x_i - x_j = h$, $i \neq j$, $\forall i, j = 1, 2, \dots, n$ are calculated, then each pair $\{F(Z(x_i)), F(Z(x_j))\}$ is placed into the relevant distance class from the following: $[0, h_1), [h_1, h_2), \dots, [h_{K-1}, h_K)$, where h_K is the maximum distance at which significant dependence is observed. Given the pairs of points for each distance class, the bivariate copulas for the each distance class are calculated, and the distance dependent spatial copula is formed as a convex combination of copulas of each distance class as follows,

$$C_h(u, v) = \begin{cases} C_{1,h}(u, v), & 0 \leq h < h_1, \\ (1 - \lambda_2)C_{1,h}(u, v) + \lambda_2 C_{2,h}(u, v), & h_1 \leq h < h_2, \\ \vdots & \vdots \\ (1 - \lambda_K)C_{K-1,h}(u, v) + \lambda_K C_{K,h}(u, v), & h_{K-1} \leq h < h_K, \\ 1, & h_K \leq h, \end{cases} \quad (2)$$

where $\lambda_k = \frac{\bar{h}_k - h_{k-1}}{\bar{h}_k - h_{k-1}}$ for $k = 2, 3, \dots, K$, $\bar{h}_k$ is the mean separation distance corresponding to each spatial bin, and $h_1, h_2, \dots, h_K$ denote the upper limits for different spatial bins (Gräler, 2014).

2.3 Mixture copula

The mixture copula is a weighted linear combination of two or more copulas on the interval $[0, 1]$. Let C_1 and C_2 be the bivariate copulas, then there exists a bivariate mixture copula $C^m : [0, 1]^2 \rightarrow [0, 1]$ such that,

$$C^m(u, v) = (1 - w)C_1 + wC_2, \quad (3)$$

where $0 \leq w \leq 1$.

2.4 Spatial mixture copula construction for multivariate spatial random fields

Let $\mathbf{Z}(x) = [Z_1(x), Z_2(x)]$ be a bivariate second-order stationary spatial random field that is observed at two-dimensional location x . Fig. 1 shows the basic steps for modelling and interpolation in a bivariate context. However, the method can be extended to spatial random field with more than two responses.

Fig. 1 Spatial mixture copula construction of multiple responses.

Modelling

Step 1: For each $Z_l, l = 1, 2$, fit the univariate marginal probability distributions of Z_l that gives the maximum log-likelihood among competing distributions, e.g., Gamma, Weibull, Normal, Log-normal, and let F_l denote the best CDF of Z_l .

Step 2: The maximum distance h_K is referred as cut-off distance, which can be determined by visual inspection of the correlogram that is a plot of Kendall's tau (τ) against the mean distance ($\bar{h}_k$) of each bin. The $\bar{h}_k$ for bin k is calculated as the average of the distances between the pairs belonging to that bin, and each pair $F_l(Z_l(x_i)), F_l(Z_l(x_j))$ is placed into K equally spaced bins $[0, h_1), [h_1, h_2), \dots, [h_{K-1}, h_K)$, then τ is calculated for each bin. Fig. 2 depicts an example correlogram

Fig. 2 An example correlogram. The blue dashed line indicates the cut-off distance at which pairs of points are no longer considered to be spatially dependent.

In Fig. 2, nearby point pairs are highly correlated, the correlation decreases as the distance increases between point pairs, and far points tend to independence at the cut-off distance of 800 km.

Step 3: For each bin k , calculate the empirical bivariate copula $C_{l,k}$ of Z_l ,

$$\begin{aligned} C_{l,k}(u, v) &= P[F_l(Z_l(x_i)) \leq u, F_l(Z_l(x_j)) \leq v], \\ &= C_k(F_l(Z_l(x_i)), F_l(Z_l(x_j))). \end{aligned}$$

Step 4: For each bin k , construct the bivariate mixture copula C_k^m ,

$$C_k^m(u, v) = (1 - w)C_{1,k} + wC_{2,k},$$

where $0 \leq w \leq 1$.

Step 5: The spatial mixture copula for the multivariate spatial random field $\mathbf{Z}$ is constructed as the convex combination of mixture copulas as follows

$$C_h^m(u, v) = \begin{cases} C_{1,h}^m(u, v), & 0 \leq h < h_1, \\ (1 - \lambda_2)C_{1,h}^m(u, v) + \lambda_2 C_{2,h}^m(u, v), & h_1 \leq h < h_2, \\ \vdots & \vdots \\ (1 - \lambda_K)C_{K-1,h}^m(u, v) + \lambda_K C_{K,h}^m(u, v), & h_{K-1} \leq h < h_K, \\ 1, & h_K \leq h, \end{cases} \quad (4)$$

where $\lambda_k = \frac{\bar{h}_k - h_{k-1}}{\bar{h}_k - \bar{h}_{k-1}}$ for $k = 2, 3, \dots, K$, $\bar{h}_k$ is the mean separation distance, and $h_1, h_2, \dots, h_K$ denote upper limits of the chosen distances for the spatial bins.

Interpolation

Interpolation of Z_2 at location x conditional on the known given value of Z_1 can be generated at the same location x , using the copula CDF of C_h^m . The interpolation of Z_1 , given Z_2 , is described by simply switching the subscripts 1 and 2 in the interpolation steps.

Step 6: Obtain the joint CDF values of $\mathbf{Z}$ using C_h^m , and let T be the vector with joint CDF values, where $T = (t_1, t_2, \dots, t_n)$.

Step 7: Obtain the marginal CDF values of Z_1 using F_1 , and let S be the vector with marginal CDF values, where $S = (s_1, s_2, \dots, s_n)$.

Step 8: Derive the conditional distribution of T , given $S = s$, using the partial derivative of C_h^m as follows,

$$\begin{aligned} C_{h,s}^m(T|S=s) &= P[T \leq t|S=s], \\ &= \frac{\partial}{\partial s} C_h^m(t, s), \end{aligned}$$

let $r = (C_{h,s}^m)^{-1}(T|S=s)$ be the interpolated value of Z_2 at location x .

Step 9: Take, $Z_2 = F_2^{-1}(r)$.

3 Application

The proposed method was applied to model real forest data that was taken from georeferenced forest inventory plots in the US Department of Agriculture Forest Service Bartlett Experimental Forest (BEF) in Bartlett, New Hampshire (Finley *et al.*, 2007). The attributes of interest were forest-wide biomass estimations within the area of 1,053 hectares. In this study, only foliage biomass (Z_1) and bole biomass (Z_2) were used that observed at 335 two-dimensional locations. The interpolation of bole biomass can be used for carbon accounting purposes, and the interpolation of foliage biomass can be used to identify regions with high values of foliage biomass (Musafer *et al.*, 2017).

Table 1 gives the summary statistics of the data. Fig. 3a and 3b show the spatial distributions of Z_1 and Z_2 at observed locations. The marginal distributions are illustrated in Fig. 3c and 3d respectively. Fig. 3e shows a strong bivariate nonlinear relationship between Z_1 and Z_2 .

Table 1 Summary statistics of Z_1 and Z_2 .

Statistics	Z_1	Z_2
n	335	335
Mean	0.334	0.119
Standard deviation	0.219	0.115
Minimum	0.200	0.010
First quartile Q_1	0.140	0.030
Median	0.300	0.090
Third quartile Q_3	0.510	0.160
Maximum	0.820	0.560

The best marginal distributions were selected based on the maximum log-likelihood (ML) values. The Weibull distribution was achieved as the best distribution for Z_1 based on the ML values, 65.03, **69.43**, 33.26, 44.03, and the Gamma distribution was achieved as the best distribution for Z_2 based on the ML values, **378.86**, 378.28, 250.20, 377.51, of Gamma, Weibull, Normal, Log-normal distributions respectively. Then, the CDF values of Z_1 and Z_2 were calculated using the corresponding CDFs. The following steps for the modelling is only incorporated the CDF values of Z_1 and Z_2 .

Fig. 3 BEF data. Spatial distributions of (a) Z_1 and (b) Z_2 , histogram and marginal distributions of (c) Z_1 and (d) Z_2 , and (e) scatter plot between Z_1 and Z_2 .

The distance between each pair of sample locations was calculated and the cut-off distance was selected as 800 km using the correlograms of Z_1 and Z_2 . Ten equally spaced spatial bins were created and the pairs of points were placed into relevant spatial bins based on the distance calculated. The empirical bivariate copulas for each bin of Z_1 and Z_2 were fitted and the ML values were compared among competing bivariate copulas, Gaussian, Student's t, Clayton, Frank, Gumbel, and Joe. Table 2 shows the best fit copulas and the estimated copula parameters θ , where $C_{1,k}$ and $C_{2,k}$ are the fitted bivariate copulas of Z_1 and Z_2 respectively.

Table 2 The best bivariate copulas for each bin.

Bins (km)	$C_{1,k} = \text{Copula } (\theta)$	$C_{2,k} = \text{Copula } (\theta)$
0-80	$C_{1,1} = \text{Joe (1.71)}$	$C_{2,1} = \text{Joe (1.48)}$
80-160	$C_{1,2} = \text{Gumbel (1.31)}$	$C_{2,2} = \text{Gaussian (0.29)}$
160-240	$C_{1,3} = \text{Gumbel (1.16)}$	$C_{2,3} = \text{Frank (1.12)}$
240-320	$C_{1,4} = \text{Gumbel (1.10)}$	$C_{2,4} = \text{Clayton (0.19)}$
320-400	$C_{1,5} = \text{Gumbel (1.06)}$	$C_{2,5} = \text{Clayton (0.13)}$
400-480	$C_{1,6} = \text{Joe (1.09)}$	$C_{2,6} = \text{Joe (1.08)}$
480-560	$C_{1,7} = \text{Joe (1.07)}$	$C_{2,7} = \text{Gumbel (1.03)}$
560-640	$C_{1,8} = \text{Clayton (0.09)}$	$C_{2,8} = \text{Clayton (0.06)}$
640-720	$C_{1,9} = \text{Clayton (0.08)}$	$C_{2,9} = \text{Clayton (0.05)}$
720-800	$C_{1,10} = \text{Joe (1.05)}$	$C_{2,10} = \text{Gumbel (1.03)}$

The joint dependence of Z_1 and Z_2 was modelled using the mixture copulas with the equal weight ($w = 0.5$). The mixture copulas were constructed as the weighted sum of the bivariate copulas in Table 2, and Table 3 shows the mixture copulas of each bin.

Table 3 The mixture copulas of each bin.

Bins	Mean distance ($\bar{h}_k$)	Mixture copula (C_k^m)
0-80	68	$C_1^m = 0.5C_{1,1} + 0.5C_{2,1}$
80-160	105	$C_2^m = 0.5C_{1,2} + 0.5C_{2,2}$
160-240	209	$C_3^m = 0.5C_{1,3} + 0.5C_{2,3}$
240-320	284	$C_4^m = 0.5C_{1,4} + 0.5C_{2,4}$
320-400	368	$C_5^m = 0.5C_{1,5} + 0.5C_{2,5}$
400-480	436	$C_6^m = 0.5C_{1,6} + 0.5C_{2,6}$
480-560	518	$C_7^m = 0.5C_{1,7} + 0.5C_{2,7}$
560-640	606	$C_8^m = 0.5C_{1,8} + 0.5C_{2,8}$
640-720	678	$C_9^m = 0.5C_{1,9} + 0.5C_{2,9}$
720-800	758	$C_{10}^m = 0.5C_{1,10} + 0.5C_{2,10}$

The mixture copulas in Table 3 were used to develop the distance dependent convex combination of mixture copulas as given in the Equation 4, which was the proposed spatial mixture copula of the

multivariate spatial random field $\mathbf{Z}$, is given by

$$C_h^m(u, v) = \begin{cases} C_1^m, & \text{if } 0 \leq h < 80, \\ 0.69C_1^m + 0.31C_2^m, & 80 \leq h < 160, \\ 0.39C_2^m + 0.61C_3^m, & 160 \leq h < 240, \\ 0.45C_3^m + 0.55C_4^m, & 240 \leq h < 320, \\ 0.40C_4^m + 0.60C_5^m, & 320 \leq h < 400, \\ 0.55C_5^m + 0.45C_6^m, & 400 \leq h < 480, \\ 0.51C_6^m + 0.49C_7^m, & 480 \leq h < 560, \\ 0.42C_7^m + 0.58C_8^m, & 560 \leq h < 640, \\ 0.52C_8^m + 0.48C_9^m, & 640 \leq h < 720, \\ 0.52C_9^m + 0.48C_{10}^m, & 720 \leq h < 800, \\ 1, & \text{if } 800 \leq h, \end{cases}$$

where $\lambda_2 = \frac{105-80}{160-80} = 0.31$, $\lambda_3 = 0.61, \dots, \lambda_{10} = 0.48$.

The proposed spatial mixture copula method was used to interpolate Z_1 and Z_2 using the inverse conditional approach as described in the steps 6–9. The interpolated values were compared with the interpolation based on the existing pair-copula, co-kriging, and NLPCA spatial methods. Subsequently, the actual values and interpolated values at sampled locations were used to calculate the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Also, the accuracy in the reproduction of the bivariate relationship of Z_1 and Z_2 was evaluated based on the mean square error from the kernel density estimation (KDE MSE). The KDE MSE is calculated by taking the mean of the squared differences between the bivariate KDEs of the actual data and interpolated data. Table 4 shows the model validation results.

Table 4 Model validation in interpolation of Z_1 and Z_2 .

Method	Z_1			Z_2			KDE MSE
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	
Pair-copula	0.20	0.17	1.12	0.11	0.08	1.98	3.61
Co-kriging	0.19	0.16	0.54	0.11	0.08	0.75	3.71
NLPCA	0.29	0.24	1.64	0.14	0.10	2.21	12.40
Mixture copula	0.14	0.13	0.56	0.06	0.05	0.63	1.34

According to Table 4 almost all the RMSE, MAE, and MAPE values are the lowest for the Z_1 and Z_2 interpolations based on the spatial mixture copula. The MAPE value of co-kriging method is the smallest for the Z_1 interpolation that is very close for the spatial mixture copula. Thus, it is clear that the proposed method outperformed in the interpolation of Z_1 and Z_2 across the observed locations. Also, the proposed method accurately reproduces the bivariate relationship in terms of the minimum value of KDE MSE. Fig. 4 shows the bivariate relationship of Z_1 and Z_2 between actual and interpolated values.

Fig. 4 Reproduction of bivariate relationship using various methods. Actual (red), interpolated (black), Z_1 given Z_2 (green), and Z_2 given Z_1 (blue).

The pair-copula and co-kriging methods are not accurately reproduced cross-validated interpolated values in the lower half of the distribution. The NLPCA method does not interpolate values in the upper end of the distribution. However, the novel mixture copula method more accurately produces cross-validated interpolated values as the actual values of the distribution.

4 Simulation study

A simulation study was proposed to assess the performance of the novel spatial mixture copula method. In this simulation study, an existing ordinary kriging simulation method was modified to simulate non-Gaussian and non-linear bivariate random fields. The summary of the simulation study is given in steps a–d.

- a) Create a 100 by 100 two-dimensional grid of data locations, denote coordinates belong to domain Ω
- b) Create an exponential variogram model with artificial parameters as given in Fig. 5,

Fig. 5 Exponential variogram with nugget (E_0), partial sill (E), and range (a).

The semi-variance $\gamma(h)$ with separation distance h is calculated as

$$\gamma(h) = E_0 + E \left[1 - \exp\left(-\frac{h}{a}\right) \right],$$

and ordinary kriging can be used to predict an unobserved measurement at an unsampled location x_0 as follows

$$\hat{Z}(x_0) = \sum_{i=1}^n \beta_i Z(x_i),$$

where $Z(x)$ is a univariate spatial random field and β_i are weights that determined from the exponential variogram.

Here, we used the ordinary kriging method to simulate 10000 univariate Gaussian random fields over the 100 by 100 girded locations, which was clearly explained in Bárdossy (1997).

- c) Next, randomly select 100 samples from each Gaussian random field. Algorithm 1 shows the basic steps of simulating non-Gaussian bivariate spatial responses using the simulated univariate Gaussian random fields from the step b.

Algorithm 1 Algorithm for simulating non-Gaussian bivariate spatial responses.

Definition and Notation:

Let $\mathbf{S}$ be the matrix with 10000 simulated Gaussian random fields with sample size 100

i is the index of first response

j is the index of second response

$S_{1,i} = NULL$ # vector to store the first non-Gaussian responses

$S_{2,j} = NULL$ # vector to store the second non-Gaussian responses

$S_{1,i}(\omega)$ is the i^{th} simulated response at location $\omega \in \Omega$

$S_{2,j}(\omega)$ is the j^{th} simulated response at location $\omega \in \Omega$

$S_{1,i}(\omega)$ and $S_{2,j}(\omega)$ are vectors with sample size 100, i.e., $\omega = (\omega_1, \omega_2, \dots, \omega_{100})$

Calculation:

while (100 non-Gaussian bivariate spatial responses are simulated){

for i in 1 to 10000

while ($S_{1,i}$ with desired spatial correlation structure is obtained - see step 2 in the method)

$S_{1,i} = \mathbf{S}[i] + \text{noise}$ # e.g., normal distribution with mean = 2 and standard deviation = 0.01

Calculate the best marginal distribution of $S_{1,i}$ (a non-Gaussian marginal distribution)

Create the spatial bins of $S_{1,i}$ (see step 2 and here only seven bins were considered)

end while

end for

for j in 1 to 10000

while ($S_{2,j}$ with desired spatial correlation structure is obtained)

$S_{2,j} = \frac{(S_{1,i})^7}{10^5} + \text{noise}$ # to obtain a strong non-linear bivariate relationship

Calculate the best marginal distribution of $S_{2,j}$ (a non-Gaussian marginal distribution)

Create the spatial bins for $S_{2,j}$ (seven bins)

end while

end for

} end while

- d) Fit the proposed spatial mixture copula C_h^m for each simulated bivariate non-Gaussian random field, and simultaneously predict each response at location ω using the inverse conditional method explained in the steps 6–9.

Table 5 gives the cross-validation results of the simulation study, which compares the mean RMSE, KDE MSE, and standard error (SE) in the interpolation based on 100 simulated samples with the existing methods. Fig. 6 shows a graphical summary of the bivariate relationship of the simulation study.

Table 5 Model validation of simulation.

Method	RMSE		KDE MSE			
	Mean (S_1)	SE	Mean (S_2)	SE	Mean	SE
Pair-copula	0.809	0.010	78.370	2.020	0.000	0.000
Co-kriging	0.780	0.010	78.360	1.960	0.000	0.000
NLPCA	1.130	0.020	104.905	3.220	0.003	0.008
Mixture copula	0.630	0.010	47.980	1.220	0.000	0.000

Fig. 6 Relationship between first simulated response (S_1) and second simulated response (S_2), 100 actual simulated values (red) overlay with 100 interpolated simulated values (black).

The mixture copula method more accurately interpolates each response across locations that achieves the minimum mean RMSE. Also, all these methods are pretty good at reproducing the bivariate relationship in terms of KDE MSE. However, the pair-copula and the co-kriging methods fail to predict extremes of the actual data and predict a small number of data outside of the trend the actual simulated data.

5 Conclusions

This article proposed a new geostatistical method for modelling spatially correlated multiple responses using mixture copulas. The novel method models multiple spatial responses without any normalisation of the original variables. The method was applied to model bivariate nonlinear spatial responses Z_1

(foliage biomass) and Z_2 (bole biomass). Also, 100 nonlinear bivariate random fields were simulated and the method was applied to each simulation. The model performance was assessed in the cross-validation of actual versus interpolated values at sampled locations. The results showed that the proposed spatial mixture copula model outperformed the existing pair-copula, co-kriging and NLPCA methods in terms of the minimum values of RMSE, MAE, MAPE, and KDE MSE.

Further improvement to the spatial mixture model is the optimal weights selection of each response in the mixture copula modelling. For instance, one spatial response may have a strong spatial dependence across locations than the other responses, and the optimal weights selection may increase the prediction accuracy of each response across locations. Moreover, the novel method can extend to model multiple spatial responses with time, where the temporal dependence can incorporate in the spatial mixture copula modelling with the multivariate spatial random fields.

References

- Aas K, Czado C, Frigessi A, Bakken H (2009). “Pair-copula constructions of multiple dependence.” *Insurance: Mathematics and Economics*, **44**(2), 182–198. URL <https://doi.org/10.1016/j.insmatheco.2007.02.001>.
- Bárdossy A (1997). “Introduction to geostatistics.” *Institute of Hydraulic Engineering, University of Stuttgart*.
- Bárdossy A (2006). “Copula-based geostatistical models for groundwater quality parameters.” *Water Resources Research*, **42**(11), 1–12. URL <https://doi.org/10.1029/2005wr004754>.
- Bárdossy A, Li J (2008). “Geostatistical interpolation using copulas.” *Water Resources Research*, **44**(7). 10.1029/2007wr006115.
- Cressie N, Wikle CK (2011). *Statistics for Spatio-Temporal Data*. Wiley.
- da Rocha MM, Yamamoto JK (2000). *Natural Resources Research*, **9**(3), 223–235. 10.1023/a:1010195701968.
- D’Amico G, Petroni F (2018). “Copula based multivariate semi-Markov models with applications in high-frequency finance.” *European Journal of Operational Research*, **267**(2), 765–777. 10.1016/j.ejor.2017.12.016.
- Diggle PJ, Ribeiro PJ, Christensen OF (2003). “An introduction to model-based geostatistics.” In “Spatial Statistics and Computational Methods,” pp. 43–86. Springer.

- Finley AO, Banerjee S, Carlin BP (2007). “spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models.” *Journal of Statistical Software*, **19**(4), 1. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074178/>.
- Gnann SJ, Allmendinger MC, Haslauer CP, Bárdossy A (2018). “Improving copula-based spatial interpolation with secondary data.” *Spatial Statistics*, **28**, 105–127. 10.1016/j.spasta.2018.07.001.
- Goovaerts P, *et al.* (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press.
- Gräler B (2014). “Modelling skewed spatial random fields through the spatial vine copula.” *Spatial Statistics*, **10**, 87–102. URL <https://doi.org/10.1016/j.spasta.2014.01.001>.
- Leuangthong O, Deutsch CV (2003). “Stepwise conditional transformation for simulation of multiple variables.” *Mathematical Geology*, **35**(2), 155–173. URL <https://link.springer.com/article/10.1023/A:1023235505120>.
- Li F, Zhou J, Liu C (2018). “Statistical modelling of extreme storms using copulas: A comparison study.” *Coastal Engineering*, **142**, 52–61. 10.1016/j.coastaleng.2018.09.007.
- Musafer GN, Thompson MH (2016). “Non-linear optimal multivariate spatial design using spatial vine copulas.” *Stochastic Environmental Research and Risk Assessment*, **31**(2), 551–570. 10.1007/s00477-016-1307-6. URL <https://doi.org/10.1007/s00477-016-1307-6>.
- Musafer GN, Thompson MH, Wolff RC, Kozan E (2017). “Nonlinear Multivariate Spatial Modeling Using NLPCA and Pair-Copulas.” *Geographical Analysis*, **49**(4), 409–432. 10.1111/gean.12126.
- Nelsen RB (2007). *An introduction to copulas*. Springer Science & Business Media. URL <https://doi.org/10.1007/978-1-4757-3076-0>.
- Patton AJ (2012). “A review of copula models for economic time series.” *Journal of Multivariate Analysis*, **110**, 4–18. 10.1016/j.jmva.2012.02.021.
- Sohrabian B (2021). “Geostatistical prediction through convex combination of Archimedean copulas.” *Spatial Statistics*, **41**, 100488. URL <https://doi.org/10.1016/j.spasta.2020.100488>.