Modeling Earth Systems and Environment
https://doi.org/10.1007/s40808-022-01484-6

ORIGINAL ARTICLE

A spatial copula interpolation in a random field with application in air pollution data

Debjoy Thakur[1] · **Ishapathik Das**[1] · **Shubhashree Chakravarty**[2]

Received: 20 May 2022 / Accepted: 16 July 2022

Abstract
Interpolating a skewed conditional spatial random field with missing data is cumbersome in the absence of Gaussianity assumptions. Copulas can capture different types of joint tail characteristics beyond the Gaussian paradigm. Maintaining spatial homogeneity and continuity around the observed random spatial point is also challenging. Especially when interpolating along a spatial surface, the boundary points also demand focus in forming a neighborhood. As a result, importing the concept of hierarchical clustering on the spatial random field is necessary for developing the copula model with the interface of the Expectation-Maximization algorithm and concurrently utilizing the idea of the Bayesian framework. This article introduces a spatial cluster-based C-vine copula and a modified Gaussian distance kernel to derive a novel spatial probability distribution. To make spatial copula interpolation compatible and efficient, we estimate the parameter by employing different techniques. We apply the proposed spatial interpolation approach to the air pollution of Delhi as a crucial circumstantial study to demonstrate this newly developed novel spatial estimation technique.

Keywords Von-Mises distribution · Expectation-Maximization algorithm · Hierarchical Spatial Clustering · Spatial Copula Interpolation · Bayesian Spatial Copula Interpolation

Introduction

The upward trend of Particulate Matter (*PM*) concentrations in the atmosphere and air pollution has become the greatest threat to human civilization daily. Every year, nearly 0.8 million people die due to the direct and indirect effects of air pollution, and approximately 4.6 million people endure from serious diseases such as chronic obstructive pulmonary disease (COPD), respiratory hazards, premature deaths, and so on (Auerbach and Hernandez 2012; Lim et al. 2012). It is unavoidable, that the air-pollutant concentration is estimated with greater accuracy to control air pollution. The spatial and spatio-temporal application of geostatistics is crucial during prediction.

A very interesting task in geostatistics is interpolating a target variable at a particular time stand, in an unobserved location, considering its surroundings. In this scenario, the researchers prefer to use Inverse Distance Weight (IDW), Ordinary Kriging (OK), Universal Kriging (UK), Disjunctive Kriging (DK), etc (Isaaks and Srivastava 1989; Cressie 1990). Because of significant advances in data science, many scientists prefer neural networking-based spatial and spatio-temporal interpolation techniques like Geo-Long Short Time Memory (Geo-LSTM), Random Forest Regression Kriging (RFK), and others (Ma et al. 2019; Shao et al. 2020). The previously mentioned algorithms use the variance-covariance function as a measure of dependence. The main drawback of this traditional spatial interpolation algorithm is the gaussianity assumption, is rarely met. The neural networking-based algorithms outperform, but the mathematical justification is difficult. As a result, applying this model in other cases can be challenging. These significant limitations promote the use of the copula-based spatial and spatio-temporal interpolation approach. This copula-based spatial interpolation technique is both theoretically

✉ Debjoy Thakur
debjoythakur95@gmail.com

Ishapathik Das
ishapathik@iitp.ac.in

Shubhashree Chakravarty
chasubha@gmail.com

[1] Department of Mathematics and Statistics, Indian Institute of Technology, Tirupati, India

[2] Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

Published online: 18 August 2022

and practically flexible. A spatial copula can easily capture a Geo-spatial variability. Spatial lag-based Gaussian and non-Gaussian bivariate copulas interpolate four different groundwater quality parameters in Baden-Wurttemberg (Bárdossy 2006). Following that, many advances in spatial copula are established, for example, the utilization of asymmetric copulas to measure spatial independence (Bárdossy and Pegram 2009), the use of Gaussian and non-Gaussian vine copula to derive the conditional distribution in the unobserved location (Bárdossy 2011), and employing the convex combination of archimedean copulas to kriging (Sohrabian 2021). Besides these, application of the Gaussian Copula (GC) via Bayesian framework to predict the maximum temperatures in the Extremadura region in southwestern Spain, for the period 1980-2015 (García et al. 2021), and a bivariate copula (Masseran 2021) measures the association between air pollution severity and its duration. Copula-based bias-correction method Alidoost et al. (2021) develops three multivariate copula-based quantile regression to map daily air temperature data. They (Khan et al. 2020) apply regular (C- and D-) Vine copula and the Student t-copula to explore the structure of spatial dependence of different climate variables, for instance, precipitation, and air temperature. A combination of extreme value models like the Generalized Pareto Distribution (GPD) with copulas (Masseran and Hussain 2020) illustrate a dependence between PM_{10} and a set of four major pollutant variables, namely, CO, NO_2, SO_2, and O_3. Employing extra-parameterized multivariate extreme valued copula (Carreau and Toulemonde 2020) introduces a spatial copula model. Application of D-vine copula in quantile regression (D'Urso et al. 2022) has explained the spatial and spatio-temporal behavior of COVID-19 in Italy. After capturing the seasonality and temporal dependency of the daily mean temperature a new spatial distance-based R-vine copula is introduced (Erhardt et al. 2015). Usage of spatial lag as an important dependence parameter of a vine copula (Gräler 2014; Bostan et al. 2021) introduce a spatial vine copula. With the help of the Metropolis-Hastings Algorithm (MHA), (Kazianka and Pilz 2011) they have improved spatial copula in the Bayesian framework to approximate the posterior predictive density whereas, this method is limited to the GC family. Spatial vine copula and dimension reduction transformation (Musafer and Thompson 2017) create a non-linear optimal multivariate spatial design to mitigate the prediction uncertainty of more than one variable. Introducing the copula-based semi-parametric algorithm (Quessy et al. 2015) models intrinsic stationary and isotropic spatial random fields. A pairwise composite likelihood with the help of a pair copula is defined using the generalized method of moments (Bai et al. 2014). A spatial factor copula model Krupskii et al. (2018) combines the flexibility of a copula, accountability of factor models, and the tractability of GC in higher dimensions, to fit spatial data at different temporal replicates. Extension of spatial Gaussian copula interpolation method (Gnann et al. 2018) predicts the primary variable, groundwater quality, and the categorical information of the primary variable as a secondary variable. A mixture copula explains the spatial dependency of an air temperature of a location on its Geo-spatial neighborhood (Alidoost et al. 2018). A translation process (TP) is discovered (Richardson 2021) for a non-Gaussian spatial copula interpolation process and is too effective to model in the absence of a link function. A spatio-temporal heterogeneous copula-based kriging (HSTCAK) (Wang et al. 2021) measures the space-time dependency by the copula function and mitigates the heterogeneity problem by fuzzy clustering. Crucial advancements in the spatial copula approach like tail dependency, asymmetric dependency, and extension of the linear model of coregionalization specifically model the multivariate spatial data (Krupskii and Genton 2019), and they use cross-covariance function as the measure of spatial dependence.

The research articles used copulas in the spatial interpolation very well in the literature, but there are some constraints that the previous authors have ignored. (i) To estimate parameters, they use the Maximum Likelihood estimate, which does not provide a good estimate in presence of missing data. (ii) After creating spatial copula interpolation, they fix one point and calculate the probability distribution at different lags from that point. As a result, the copula is limited within the fixed reference frame, but the reference frame is random in reality. Therefore, we consider the random points to form a cluster, prioritizing a relative distance. (iii) At the time of spatial clustering, they disregard the significance of disjoint Geo-spatial regions. Thus the intersection part is the most affected area, where the different effects of different clusters become confounded. (iv) They use conditional expectation for interpolation, but it is invalid for the extreme valued Probability Density Function (PDF).

In this study we evolve a novel spatial cluster-based copula modeling in different frameworks. We divide the entire spatial domain into k spatial clusters to get m number of spatial regions i.e. $\mathscr{L} \subseteq \mathbf{R}^{2\times 2}$ which is the class of all possible set of points in a spatial region. We create a conditional spatial random field $Y : \mathscr{L} \times \mathscr{F}_{\mu^*} \to \mathscr{M}$. Here, $\mathscr{F}_{\mu^*}$ is an induced probability space created using caratheodory's extension theorem and $\mathscr{F}_{\mu^*} = \{A \in \mathscr{L} \mid A \text{ is } \mu^* - \text{measurable } i.e.\ \mu^*(A) \leq \delta\}$ and Y is $< \mathscr{L} \times \mathscr{F}_{\mu^*}, \mathscr{M} >$ measurable random field and $\mathscr{M} \subseteq \mathbf{R}$. Our objective in this research is to predict Y at an unobserved location on $s_0 \in \mathscr{L}_i$ based upon the n distinct observed location $s_1, s_2, ..., s_n \in \mathscr{L}_i$ using spatial copula interpolation algorithm in classical and Bayesian framework.

The outline of this study is as follows: the details of the algorithm are introduced in "Method" section, the study area and the behavior of data are described in "Study area and data" section, the results and discussion regarding the case

study are summarized in "Results and discussion" section and the conclusion is made in "Conclusion" Section.

Method

Fitting marginal distribution

In this section, we illustrate how to fit an ideal univariate parametric distribution on the empirical distribution. We have divided the choice of PDF into three steps: (i) choice of a family of distributions, (ii) suitable marginal PDF of that family, and (iii) to estimate a parameter of these marginal PDF. For step (i), we use the Cullen and Frey graph of skewness-kurtosis plot. We utilize Kernel Density Estimation (KDE) for step (ii). In this step we prioritize the value of Akaike Information Criteria (AIC), and Bayesian Information Criteria (BIC), Log-Likelihood value (Log-Lik), and Kolmogrov Smirnov (KS) statistic. Although, we face a real challenge in coordination (iii) because there is missing data. Therefore, the Maximum Likelihood Estimation (MLE) of a parameter is not recommendable. For the Parametric Exponential Family distribution (PEF) we use Expectation-Maximization algorithm (EM) (McLachlan and Krishnan 2007). We use Uniformly Minimum Variance Unbiased Estimator (UMVUE) technique for the circular probability distributions. For PEF we consider the fitted distribution is Log-Normal (LN) probability distribution then, $\log W \sim \mathcal{N}(\mu, \sigma^2)$. Let, $w_i; i = 1, 2, 3, \ldots, n_1$ are the observed data points and $w_i; i = n_1 + 1, n_1 + 2, n_1 + 3, \ldots, n_2$ are the un-observed data points. The likelihood function of (μ, σ) based upon the observed data:

$$\log L_o(\mu, \sigma) = \frac{-1}{2\sigma^2} \cdot \sum_{i=1}^{n_1} (\log w_i - \mu)^2 - \sigma \cdot \sqrt{2\pi} \sum_{i=1}^{n_1} \log w_i \tag{1}$$

The complete, observed and missing data vectors are respectively, $\mathbf{x} = (w_1, w_2, w_3, \ldots, w_{n_2})^T; \mathbf{y} = (w_1, w_2, w_3, \ldots, w_{n_1})^T$ and $\mathbf{z} = (w_{n_1+1}, w_{n_1+2}, w_{n_1+3}, \ldots, w_{n_2})^T$ reveals $\mathbf{x} = \mathbf{y} \bigcup \mathbf{z}$. The complete data log-likelihood function is:

$$\log L_c(\mu, \sigma) = \frac{-1}{2\sigma^2} \cdot \sum_{i=1}^{n_2} (\log w_i - \mu)^2 - \sigma \cdot \sqrt{2\pi} \sum_{i=1}^{n_2} \log w_i \tag{2}$$

Let's consider the E-step on the $(m+1)^{\text{th}}$ iteration of the EM algorithm where $\left(\mu^{(m)}, \sigma^{(m)}\right)$ is the value after the m^{th} iteration of EM. Using the Eq. (2) we compute the conditional expectation of Log-Likelihood of the Complete data (CElikC) based upon the updated value at the m^{th} iteration, defined as $\mathcal{Q}\left((\mu, \sigma) \mid \left(\mu^{(m)}, \sigma^{(m)}\right)\right)$ in the following:

$$\begin{aligned} &\mathcal{Q}\left((\mu, \sigma) \mid \left(\mu^{(m)}, \sigma^{(m)}\right)\right) \\ &= E_{\left(\mu^{(m)}, \sigma^{(m)}\right)}\left[\log L_c(\mu, \sigma) \mid \mathbf{y}\right] \\ &= \int_{\mathbf{z}} \left(\log L_c(\mu, \sigma) \mid \vec{y}\right) \cdot f(\mathbf{z} \mid \mathbf{y}, \left(\mu^{(m)}, \sigma^{(m)}\right) \mathrm{d}\mathbf{z} \\ &= \int_{\mathbf{z}} \frac{\left(\log L_c(\mu, \sigma) \mid \mathbf{y}\right)}{f(\mathbf{y}; \left(\mu^{(m)}, \sigma^{(m)}\right)}) \cdot f(\mathbf{z}, \mathbf{y}; \left(\mu^{(m)}, \sigma^{(m)}\right) \mathrm{d}\mathbf{z} \\ &\leq E_{\left(\mu^{(m)}, \sigma^{(m)}\right)} \left[\frac{\log L_c(\mu, \sigma)}{\log L_o(\mu^{(m)}, \sigma^{(m)})}\right] \end{aligned} \tag{3}$$

Using the Eqs. (1), (2) we simplify the Eq. (3) and then in M-step we maximize $\mathcal{Q}\left((\mu, \sigma) \mid \left(\mu^{(m)}, \sigma^{(m)}\right)\right)$. Therefore, the updated values are $\left(\mu^{(m+1)}, \sigma^{(m+1)}\right)$ which is defined in the following:

$$\left(\mu^{(m+1)}, \sigma^{(m+1)}\right) = \text{argmax}_{(\mu,\sigma)} \mathcal{Q}\left((\mu, \sigma) \mid \left(\mu^{(m)}, \sigma^{(m)}\right)\right) \tag{4}$$

We estimate the parameter from Eq. (4). But to estimate the parameter of a circular probability distribution for example, Von-Mises (VM) distribution, we avoid the computational complexity of EM algorithm due to absence of closed form. Presence of Bessel Function ($I_n(k)$) promotes us to introduce a new theorem regarding the completeness and sufficiency of an estimator to deduce a UMVUE of the parameter of VM distribution in the following:

Theorem 1 *If $X_i \sim^{iid}$ VM then $\frac{I_0(k)\cdot\cos(x_i)}{I_1(k)}$ and $\frac{I_0(k)\cdot\sin(x_i)}{I_1(k)}$ are the UMVUE of $\cos\mu$ and $\sin\mu$ respectively and their corresponding variances are*

$$\begin{aligned} \text{var}(\cos(x_i)) &= \frac{1}{2} + \frac{I_2(k)\cdot\cos(2\mu)}{2I_0(k)} - \left(\frac{I_1(k)\cdot\cos(\mu)}{I_0(k)}\right)^2 \\ \text{var}(\sin(x_i)) &= \frac{1}{2} - \frac{I_2(k)\cdot\sin(2\mu)}{2I_0(k)} - \left(\frac{I_1(k)\cdot\sin(\mu)}{I_0(k)}\right)^2 \end{aligned} \tag{5}$$

Proof See Appendix (6) □

Using Theorem (1) and trigonometric inverse function we get the initial value and, update the parameter values of VM distribution like LN (see Appendix (A.1)).

Copula

A copula is used to model the dependence between two or more random variables, for formulating the joint multivariate distribution from the marginal Cumulative Distribution Function (CDF). Let, $X_1, X_2, \ldots, X_n$ be n Random Variables (RV) with corresponding marginal CDF s are respectively, $F_1(x_1), F_2(x_2), \ldots, F_n(x_n)$. The joint distribution function can

be defined as, $F_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$ which is the product of marginal and conditional distribution but, because of the complexity of this approach with the increasing number of random variables this approach is not applicable for the large number of random variables. Therefore, the copula function is defined to create a multivariate distribution from the n marginal distribution (Nelsen 2007; Sklar 1973) to model the dependence between the multidimensional variables in the following way: $C : [0,1]^n \rightarrow [0,1]$

$$F_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = C(F_1(X_1),F_2(X_2),\ldots,F_n(X_n)). \tag{6}$$

There are different copula families, for example, Gaussian, Archimedean, Product, etc. They behave differently in the tail part of the distribution. Compared to the other traditional multivariate, elliptical, Archimedean copulas, and Vine Copulas (VC) are more flexible to capture the inherent dependency. Under some certain regularity conditions it's possible to express the n–dimensional multivariate copula mentioned in the Eq. (6) as multiplication of pair-copulas (Aas et al. 2009) in the following iterative approach. For $n = 2$ the bi-variate probability density function (BPDF) is:

$$\begin{aligned} f(x_1,x_2) &= f_{12}(x_1 \mid x_2)\cdot f_2(x_2) = c_{12}(F_1(x_1),F_2(x_2)) \\ &\quad \cdot f_1(x_1)\cdot f_2(x_2). \end{aligned} \tag{7}$$

In the Eq. (7) $c_{12}(.,.)$ is the applicable Pair-Copula Density Function (PCDF) for $F_1(x_1)$ and $F_2(x_2)$. For $n = 3$ the Tri-Variate Probability Density Function (TPDF) is:

$$\begin{aligned} f(x_1,x_2,x_3) &= f_{123}(x_1 \mid x_2,x_3)\cdot f_{23}(x_2 \mid x_3)\cdot f_3(x_3) \\ &= c_{12}(F_1(x_1),F_2(x_2))\cdot c_{23}(F_2(x_2),F_3(x_3)) \\ &\quad \cdot c_{13|2}(F(x_1 \mid x_2),F(x_3 \mid x_2))\cdot f_1(x_1)\cdot f_2(x_2)\cdot f_3(x_3). \end{aligned} \tag{8}$$

In Eq. (8) $c_{23}(.,.), c_{13|2}(.,.)$ are the applicable PCDF and Conditional PCDF (CPCDF) respectively. Likewise, for $n = 4$ the four-variate probability density function (FPDF) is:

$$\begin{aligned} f(x_1,x_2,x_3,x_4) &= f_{4123}(x_4 \mid x_1,x_2,x_3)\cdot f_{312}(x_3 \mid x_1,x_2) \\ &\quad \cdot f_{21}(x_2 \mid x_1)\cdot f_1(x_1) = c_{12}(F_1(x_1),F_2(x_2)) \\ &\quad \cdot c_{13}(F_1(x_1),F_3(x_3))\cdot c_{14}(F_1(x_1),F_4(x_4)) \\ &\quad \cdot c_{23|1}(F(x_2 \mid x_1),F(x_3 \mid x_1))\cdot c_{24|1}(F(x_2 \mid x_1),F(x_4 \mid x_1)) \\ &\quad \cdot c_{34|2}(F(x_3 \mid x_2),F(x_4 \mid x_2))\cdot f_1(x_1)\cdot f_2(x_2)\cdot f_3(x_3)\cdot f_4(x_4) \end{aligned} \tag{9}$$

VC is a graphical approach representing an n-dimensional Multivariate PDF (MPDF) using $\binom{n}{2}$ suitable PCDF in a hierarchical manner where the dependence structures of $(n-1)$ have unconditional PCDF and that of the remaining has Conditional PCPDF (CPCDF). In this paper, we focus on C-Vine Copula (C-VC), because of its better flexibility.

A C-VC with 4-variables has 3 trees, T_j and each tree, T_j has $4-j+1$ nodes and $4-j$ edges where $j = 1,2,3$ like Fig. 1. In tree T_1 each edge between two nodes represent the PCDF. From Fig. 1 in tree T_2 the edges between each node is CPCDF where $c_{13|4}$ denotes the CPCDF of the first and third variable given the fourth variable and $c_{23|4}$ represents the CPCDF of the second and third variable given the fourth variable. In the tree T_3 each node is connected by an edge representing CPCDF (Fig. 1) of the first and second variable given third and fourth variable where $C_{12|34} = F(x_1,x_2 \mid x_3,x_4)$.

Spatial interpolation

Here, we propose two novel spatial interpolation approaches combining spatial clustering, knowledge of copula, and C-VC assuming the directional stationarity of data is defined in the following:

Spatial copula estimation

Let $\mathscr{S}$ be the spatial domain of interest for the spatial interpolation purpose. Salient features of this spatial clustering

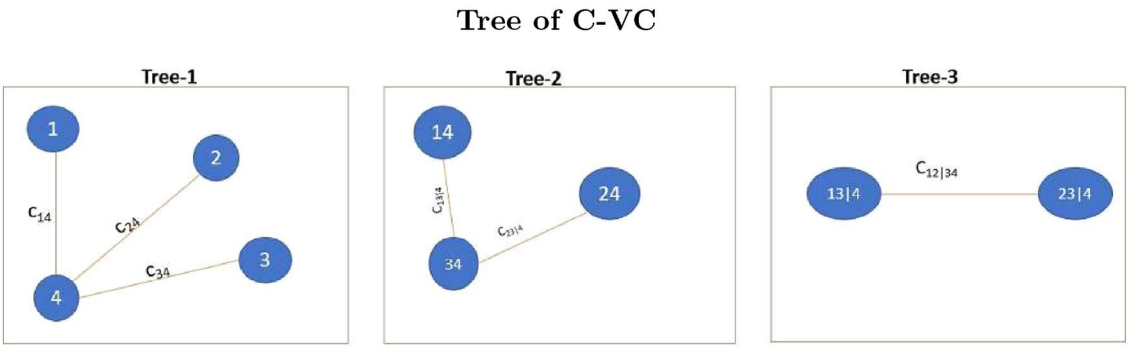

Fig. 1 Detail of the tree of the Vine Copula

technique are distance and degree of similarity between two spatial points. Hierarchical Spatial Clustering (HSC) is defined using the complete linkage method (Hubert 1974). The threshold criterion of inclusion is a cutoff value of Haversine Distance (HD) (Gade 2010) and correlogram between each pair of points. Here, the number of HSCs and HSC's radius are the two most important parameters. According to the principle of HSC, the Sum of Squares Within a Cluster (SSW) is lesser than Between the Clusters (SSB). We consider an optimal number of HSCs while SSW reaches a plateau according to the Elbow method. To determine HSC's radius, we arrange HSC's height in ascending order and consider a significant height as a radius. Therefore, in this context, we can think of the HSC as a spatial field defined in the following Eq. (10)

$$\begin{aligned} N_i = \{ & \left(\mathrm{ob}_{i_1}, \mathrm{ob}_{i_2}\right) : HD(\mathrm{ob}_{i_1}, \mathrm{ob}_{i_2}) \le HD_{\mathrm{cut}} \\ & \rho(||\mathrm{ob}_{i_1} - \mathrm{ob}_{i_2}||) \ge r_{\mathrm{cut}} \ \& \ i_1 \ne i_2 \} \\ & \bigcup \{ \left(\mathrm{ob}_{i_j}, y_{\mathrm{ij}}\right) : HD(\mathrm{ob}_{i_j}, y_{\mathrm{ij}}) \\ & \le HD_{\mathrm{cut}} \quad \rho(||\mathrm{ob}_{i_j} - y_{\mathrm{ij}}||) \ge r_{\mathrm{cut}} \} \end{aligned} \tag{10}$$

Let $N_1, N_2, \ldots, N_k$ be the k clusters, y_{ij} be the j^{th} unobserved point, and ob_{i_j} be the j^{th} observed point of the i^{th} HSC where, $j = 1, 2, \ldots, n_i$; $i = 1, 2, 3, \ldots, k$ where n_i be the number of the unobserved points in i^{th} HSC and c_i, the centre of i^{th} cluster. A threeshold HD (HD_{cut}) is chosen $\ni$ $dist(c_i, y_{\mathrm{ij}}) \le HD_{\mathrm{cut}}$ that is surrounding the centre of each HSC, a circle is constructed with radius of HD_{cut} units and r_{cut} refers the spatial auto-correlation cutoff to maintain the spatial continuity of HSC. For k clusters maximum number of distinct Spatial Regions (SR) is $2^k - 1$. Let $\vec{v}$ be the presence vector of all unobserved points $y_{\mathrm{ij}} = (\mathrm{lon}, \mathrm{lat})$, where *lon* and *lat* stand for longitude and latitude of an unobserved point respectively. Now, we create a linear map in the following:

$$f : \mathscr{R}^{2\times 1} \to \mathscr{R}^{k\times 1} \Rightarrow f(y_{ij}) = \vec{v} \tag{11}$$

In the Eq. (11) $\vec{v}$ is a binary vector of length k, where

$$\begin{aligned} & \vec{v} = [v_1, v_2, \ldots, v_k] \\ & v_l = \begin{cases} 1 & \text{if (lon,lat)} \in N_l \\ 0 & \text{otherwise} \end{cases} \\ & \quad where\ l = 1, 2, 3, \ldots, k. \end{aligned} \tag{12}$$

The maximum number of presence vectors for k clusters is $2^k - 1$. As a result, we obtain at most $2^k - 1$ distinct SR in our entire study area. Let $R_1, R_2, \ldots, R_m$ be the m SR where $m \le 2^k - 1$. Let $v_i = 0 \forall i = 1, 2, \ldots, (i-1), (i+1), \ldots, k$ and $v_i = 1$ that denotes y_{ij} is inside N_i only.

In Fig. 2 we divide the entire spatial domain into some HSC grounded on the HD of monitoring stations and the

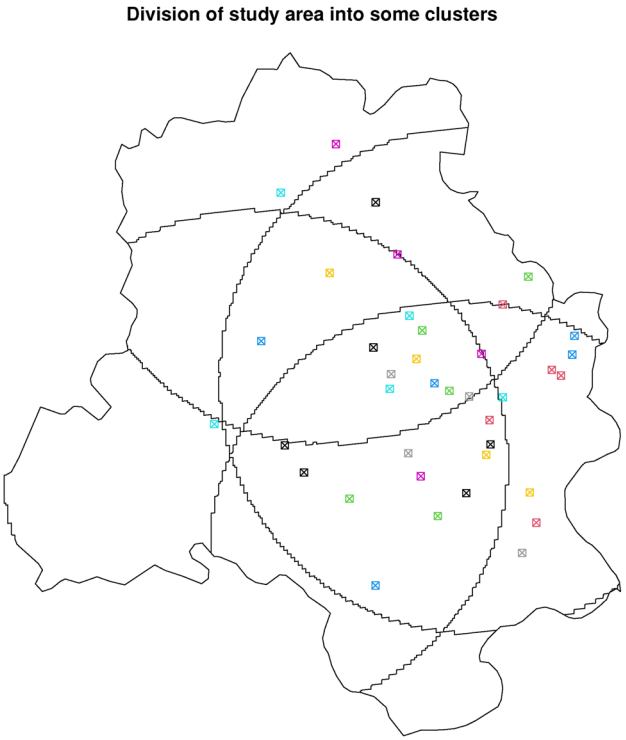

Fig. 2 The entire spatial domain, Delhi is divided into four clusters and the corresponding monitoring stations

degree of homogeneity where each circle denotes an HSC, and the dotted points represent the observed monitoring stations contained in that HSC. As a result, the whole area is split into some disjoint SR and the dotted points represent the monitoring stations, included in that SR (Fig. 3) are salient points to interpolate along the surface of each SR.

Utilizing the concept of the copula, we transform an HSC into a Spatial Random Field (SRF) to predict the values on the unobserved location. Therefore, we concentrate on the inclusion probability of the latitude (ω_i) and longitude (β_i) of an observed location in R_i using univariate Marginal PDF (MDF). The corresponding MDFs are respectively $P(\{\omega : \omega \in R_i\})$ and $P(\{\beta : \beta \in R_i\})$. Then, we evaluate bivariate PDF (BDF) of inclusion of latitude and longitude in the i^{th} SR, and Kendall's τ as the measure of association between two RVs. So, the joint BDF is as follows:

$$H(x_1, x_2) = C(F_\omega(X_{1\omega}), F_\beta(X_{2\beta})) \text{ where } C : [0,1]^2 \to [0,1].$$

To evaluate the CDF of that spatial random process (SRP), a composite function of two RVs i.e.,

$Y(.,.) \equiv \{Y(\omega, \beta) : \omega \in R_i, \beta \in R_i\}$ we implement the KDE to get the MDF of Y i.e., $F(y)$ and making use of copula we deduce the joint Tri-variate probability distribution (TDF) as follows:

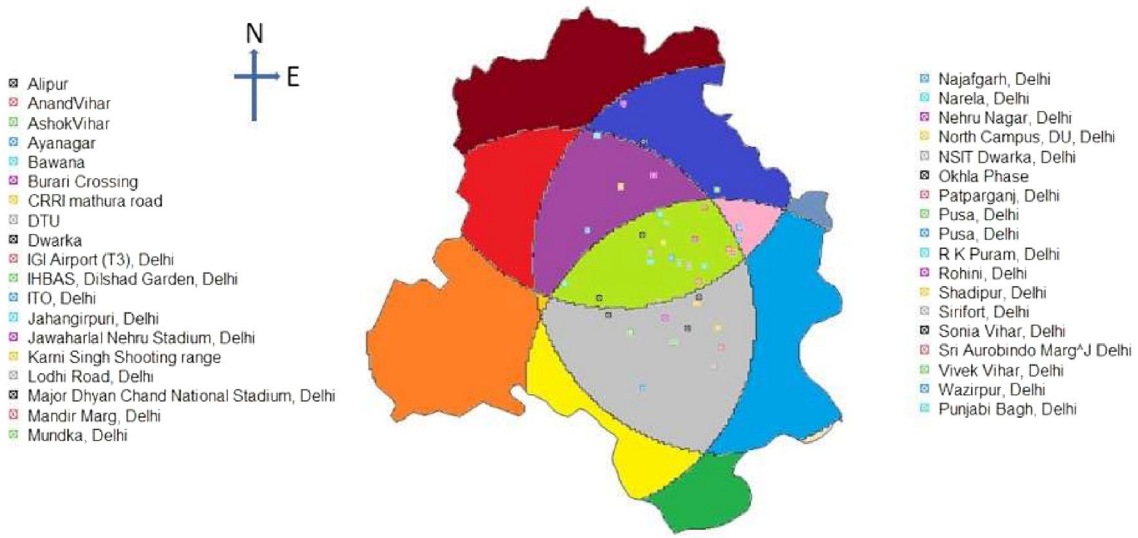

Fig. 3 The entire geographical area is split up into some disjoint regions which are shaded by different colour and the corresponding observed points belonging into each region

$H(x_1, x_2, y) = C(F_\omega(X_{1\omega}), F_\beta(X_{2\beta}), F(Y))$ where $C : [0,1]^3 \to [0,1]$.

After getting, the BDF ($H(x_1, x_2)$) and TDF ($H(x_1, x_2, y)$) using copula we find out the conditional PDF (CPDF) defined in the following two equations:

$$f(y_1 \mid x_1, x_2) = \frac{\frac{\partial^3 H(x_1, x_2, y)}{\partial x_1 \partial x_2, \partial y}}{\frac{\partial^2 H(x_1, x_2)}{\partial x_1 \partial x_2}} \tag{13}$$

Here, we assume two RVs, latitude (X_1) and longitude (X_2) follow Uniform $[a_1, b_1]$ and Uniform $[a_2, b_2]$ respectively. Now we generate random points (x_1, x_2) along, $\mathscr{S}$ to measure the CDF of SRP i.e., $Y(x_1, x_2)$ having different CDF for each geographical position. We employ EM algorithm to estimate the parameter at the time of fitting MDF. Applying copula we get the joint TDF of $Y(x_1, x_2), X_1, X_2$ is defined as $F(y, x_1, x_2)$. Next, we will introduce the HSC-based SI namely, Spatial Copula interpolation (SC). We split $\mathscr{S}$ in m number of regions making use of the HSC algorithm, discussed earlier. Let's consider, R_1 has three observed locations ob_1, ob_2, ob_3. In that region we can generate a number of gridded points not necessarily of uniform size, out of those unobserved points we consider one unobserved point, defined as $un^j_{R_1}$ which is the j^{th} point in R_1. Applying the conditional copula (from Eq. (13)) we establish the Conditional Copula-based Probability Distribution Function (CCDF) for each un-observed point in a SR i.e. $F_{\text{un}^j_{R_1}}(y)$.

$$\begin{aligned} F_{\text{un}^j_{R_1}}(y) &= P[Y(x_{1,\text{un}^j_{R_1}}, x_{2,\text{un}^j_{R_1}}) \le y \mid X_1 \\ &\quad = x_{1,\text{un}^j_{R_1}}, X_2 = x_{2,\text{un}^j_{R_1}}] \end{aligned} \tag{14}$$

From the Eq. (14) we get the CCDF of j^{th} un-observed point included in the first SR. That lets us calculate the CCDF of SRP, Y at the unobserved centroid of a cluster, and making the use of CCDF we can calculate the conditional copula-based probability density function (CCPDF). The mathematical formulation is described in the following:

$$\begin{aligned} F_{\text{un}^j_{R_1}}(y) &= \sum_{i \in R_1} \alpha_{ij} \cdot P[Y(x_{1,\text{ob}_i}, x_{2,\text{ob}_i}) \\ &\le y \mid X_1 = x_{1,\text{ob}_i}, X_2 = x_{2,\text{ob}_i}] \\ &\Rightarrow f_{\text{un}^j_{R_1}}(y) = \sum_{i \in R_1} \alpha_{\text{ij}} \cdot f_{\text{ob}_i}(y \mid x_1, x_2) \\ &\Rightarrow \text{argmax}_y f_{\text{un}^j_{R_1}}(y) = \sum_{i \in R_1} \alpha_{ij} \cdot \text{argmax}_y f_{\text{ob}_i}(y \mid x_1, x_2) \end{aligned} \tag{15}$$

In the Eq. (15) the weights are defined as α_{ij}. These weights are proportional to the spatial auto correlation function (ACF) but inversely proportional to the degree of separation. So the required α_{ij} is defined in the following assuming the fact that, $\text{ob}_i, \text{un}_j \in R_1$

$$\alpha_{\text{ij}} = \frac{\text{d}_{\text{ij}} \cdot \rho(|| \text{ ob}_i - \text{un}^j ||)}{\sum_{i \in R_1} \text{d}_{\text{ij}} \cdot \rho(|| \text{ ob}_i - \text{un}^j ||)} \tag{16}$$

In the Eq. (16), d_{ij} is the degree of seperation established upon the HD and the degree of departure of two PDFs defined on the SRF in the following Eq. (17)

$$
\begin{aligned}
d_{ij} &= \epsilon + e^{-\left(\left[\int_0^1 |f_{ob_i}(y|x_1,x_2) - f_{un^j}(y|x_1,x_2)|^p\right]^{1/p} dy\right)} \cdot \\
&\quad e^{-\left(\sin^{-1}\left[\sqrt{A + \cos(x_{1,ob_i}) \cdot \cos(x_{2,un^j}) \cdot B}\right]\right)} \\
&\quad where,\ A = \sin^2\left(\frac{x_{1,ob_i} - x_{1,un^j}}{2}\right) \\
&\quad B = \sin^2\left(\frac{x_{2,ob_i} - x_{2,un^j}}{2}\right)
\end{aligned}
\tag{17}
$$

In the Eq. (17) ϵ is included for the computational adjustment (Machuca-Mory and Deutsch 2013) along the boundary points of each SR. $e^{-\left(\left[\int_0^1 |f_{ob_i}(y|x_1,x_2) - f_{un^j}(y|x_1,x_2)|^p\right]^{1/p} dy\right)}$ specifying modified gaussian distance kernel and here, as a distance we apply the degree of separation between two Conditional Copula-based Spatial PDF (CCSPDF) to capture the probabilistic spatial dissimilarity, and the last part is HD. For $\rho(||\ ob_i - un^j\ ||)$ in the eq. (16) choice of suitable covariance function is necessary. Therefore, we choose the suitable covariance function among well-defined variogram clouds, for example, Exponential, Gaussian and Spherical, Cressie (1990) etc. Applying this concept we adopt the Algorithm (1) to interpolate over the entire spatial surface:

Algorithm 1 Algorithm of SC Interpolation

Require: $0 < m \leq 2^k - 1;\ R_i \cap R_j = \phi$

$\mathscr{S} \leftarrow \bigcup_{i=1}^{k} N_i = \bigcup_{j=1}^{m} R_j$

$Gen_j = \{(lon_j, lat_j)\}$ ▷ Set of randomly generated points

for each $j \in Gen_j$ **do**

  $\vec{v} \leftarrow Presence(Gen_j)$ ▷ $Presence(.)$ is a binary vector like $f(.)$ in Equation(11)

  if $freq(\vec{v}) = 1$ **then** ▷ $freq(.)$ returns the sum of $\vec{v}$

    $index \leftarrow Index(\vec{v})$ ▷ $Index(.)$ returns position of 1 in $\vec{v}$

    $\{ob_1, ob_2, ..., ob_r\} \leftarrow$ observed location in $index^{th}$ HSC

    $SR(j) \leftarrow$ Choose p^{th} closest SR from $\{ob_1, ob_2, ..., ob_r\}$ close to (lon_j, lat_j)

    $SC_j \leftarrow \sum_{i \in SR(j)} \alpha_{ij} argmax_y f_{ob_j}(y \mid lon, lat)$

  else

    $\vec{Ind} \leftarrow Index(\vec{v})$

    for each $q \in \vec{Ind}$ **do**

      $\{ob_1, ob_2, ..., ob_r\} \leftarrow$ observed location in q^{th} HSC

      $S \leftarrow S.append(\{ob_1, ob_2, ..., ob_r\})$ ▷ $X.append(.)$ adds the argument to existing X values

    end for

    $SR(j) \leftarrow Unique(S)$ ▷ $Unique(.)$ removes the duplicate elements from its argument

    $SR(j) \leftarrow$ Choose p^{th} closest SR from $\{ob_1, ob_2, ..., ob_r\}$ close to (lon_j, lat_j)

    $SC_j \leftarrow \sum_{i \in SR(j)} \alpha_{ij} argmax_y f_{ob_j}(y \mid lon, lat)$

  end if

end for

Spatial Bayesian Vine-Copula estimation

We introduce spatial vine copula estimation based upon the Bayesian statistical approach (SBVC). Under the square error loss function employing MHA we do the posterior estimate of the parameter in the following way:

$$\pi(\vec{\theta}) = \frac{f_{X_1,X_2|X_3,X_4}(x_1,x_2,x_3,x_4 \mid \vec{\theta}) \cdot p(\vec{\theta})}{\int f_{X_1,X_2|X_3,X_4}(x_1,x_2,x_3,x_4 \mid \vec{\theta}) \cdot p(\vec{\theta}) d\vec{\theta}} \tag{18}$$

In Eq. (18) $f(. \mid .)$ defines the Conditional Copula-based PDF (CCPDF) applying the inherited concept of Fig. 1. $p(\vec{\theta})$ denotes the prior PDF of $\vec{\theta}$, and $\pi(\vec{\theta})$ defines the Posterior PDF (PPDF) of $\vec{\theta}$. Using the PPDF we'll calculate the posterior estimation of $\vec{\theta}$ under the absolute error loss function. After getting the most updated values $\vec{\theta}$ using MHA we find out the conditional bayesian prediction of two variables in the following:

$$\begin{aligned} &E(\hat{\vec{\theta}} \mid X_1,X_2,X_3,X_4) = \int \vec{\theta} \cdot \pi(\vec{\theta}) d\vec{\theta} \\ &E(X_1,X_2 \mid X_3,X_4) \\ &\quad = \int \int x_1 \cdot x_2 \cdot f_{X_1,X_2|X_3,X_4}(x_1,x_2,x_3,x_4 \mid \hat{\vec{\theta}}) \mathrm{d}x_1 \mathrm{d}x_2 \end{aligned} \tag{19}$$

Using Eq. (19) we interpolate the target variables on target locations. Here, we assume two SRPs, $Y_1(x_1,x_2)$ and $Y_2(x_1,x_2)$ and apply the concept of tail-dependency of a bivariate copula to measure their hidden reliance. Utilizing VC we find the CCPDF i.e., $F(y_1,y_2 \mid x_1,x_2;\vec{\theta}) = P[Y_1(x_1,x_2) \leq y_1, Y_2(x_1,x_2)] \leq y_2 \mid X_1 = x_1, X_2 = x_2;\vec{\theta}]$. Regarding parameter estimation, during fitting MDF, we use UMVUE, EM etc, but to estimate the parameter of the copula family we only consider the posterior estimate. Then using the conditional expectation technique we estimate the Y_1 & Y_2 values all the randomly generated gridded points and interpolate the values.

Model validation

The accuracy of the models are validated by the following three methods where, $Y(\vec{s_i})$ is the observed data points and, $\hat{Y}(\vec{s_i})$ is the predicted value:

1. Mean of Absolute Error (MAE)

$$MAE = \frac{\sum_{i=1}^{n} \mid Y(\vec{s_i}) - \hat{Y}(\vec{s_i}) \mid}{n} \tag{20}$$

2. Root of Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(Y(\vec{s_i}) - \hat{Y}(\vec{s_i})\right)^2}{n}} \tag{21}$$

3. K-fold CV

Then, using Eqs. (21) and (20) we measure the accuracy of the proposed model. K-fold CV assumes K as 10. The 10-folded CV indicates that the data set is divided into the 10 random sub-sets, and among these data sets, 9 sub-sets as training data set, and the rest 1 is a test data set, termed as one-leave-one out CV (OLOCV). It helps compare the MAE of proposed and old models

Study area and data

To demonstrate the SC, SBVC, and to compare with OK we take Delhi-air pollution as a circumstance study. Delhi, the capital of India is the most polluted due to, rapid urbanization, boosting amounts of traffic, increasing population, and energy consumption at an alarming level. Sometimes, the level of $PM_{2.5}$ concentration in the air has reached up to 999 $\mu g/m^3$ (Mukherjee et al. 2018) and, among all other air pollutants, it affects public health (Zheng et al. 2015) badly. Boosting levels of automobiles, cars, etc cause higher pollutant concentrations in the air (Samal et al. 2013). We look at the air pollution data collected by the monitoring stations, maintained by the Central Pollution Control Board (CPCB), Delhi Pollution Control Committee (DPCC), and the Indian Institute of Tropical Meteorology (IITM). To get the research goal, we collected data on several air pollutants, such as $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x and wind direction (WD), from the CPCB websites. To map the Spatio-temporal distribution of air quality and deduce the effect of WD on the air pollution in Delhi, these data play an important role. The data were collected over 24 hours, and the period was taken from $_{1st}$ February 2017 to $_{31st}$ December 2021. The Fig. 5 depicts the temporal variability of daily $PM_{2.5}$ emission which is cyclic after a fixed time stand. There is always a higher concentration witnessed from almost the end of November to the end of December (Fig. 5) around 400 $\mu g/m^3$ and sometimes it grows up to 800 $\mu g/m^3$ which is very much alarming for the human life, primarily during winter due to burning of firecrackers, agricultural crop burning, etc.

There are 38 monitoring stations in this data set, as shown in Fig. 3. According to Fig. 4, we detect that the Northern part of Delhi is very much sensitive to pollution whereas the Central, East, and West parts of Delhi are less sensitive regarding the pollutant concentrations in the air. According to Fig. 4, the $PM_{2.5}$ concentration in the Northern part of Delhi can reach up to 220 $\mu g/m^3$ whereas in the Central,

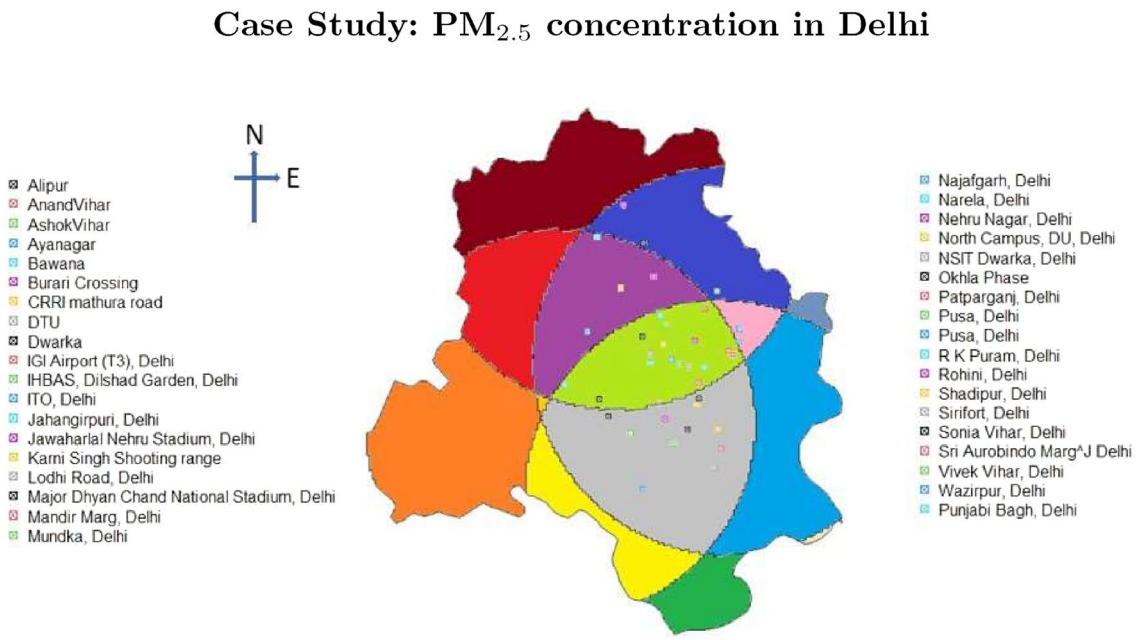

Fig. 4 The interpolated $PM_{2.5}$ values of November in the year, 2019

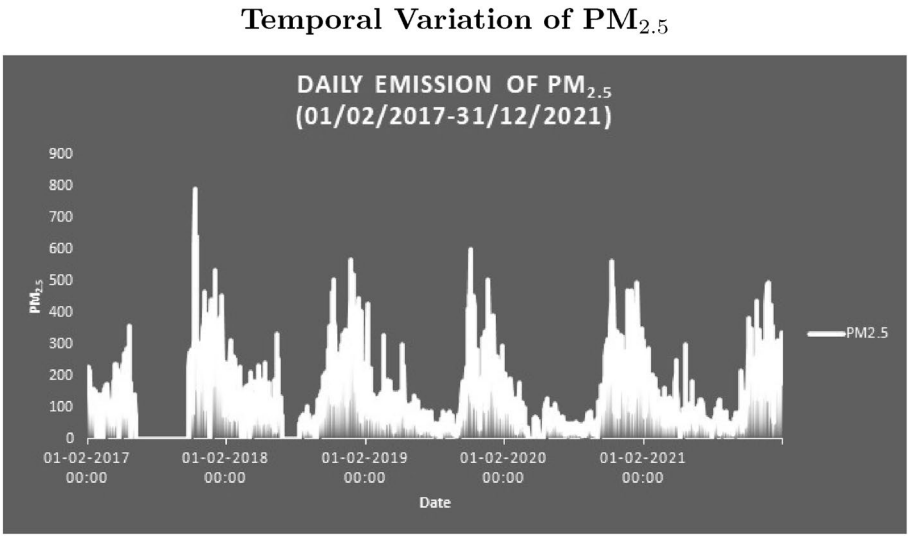

Fig. 5 The time series plot of daily $PM_{2.5}$ emission from 1st Feburary, 2017 to 31st December, 2021 in the study area, Delhi

East, and West part of Delhi that is limited into 190 to 200 $\mu g/m^3$. As a result, the Spatio-temporal variability in air pollutant concentrations is visible. However, there are two shortcomings to applying spatial interpolation techniques to interpolate (i) the Delhi NCR region is far away from other monitoring stations in Delhi, (ii) The missing Data.

We can easily conclude from the Fig. 6 that the observed frequency distribution of daily $PM_{2.5}$ emission during this period is positively skewed which gives an idea of how to fit the positively skewed distribution such as log-Normal, Gamma, Exponential, Weibull, etc depending upon the tail distribution.

Precisely there is a higher concentration in the interval from $30-40\,\mu g/m^3$. Figure 7 provides a brief overview of the variability and a five-point summary of the pollutants and WD which establishes the fact that PM_{10} and $PM_{2.5}$ have higher variability compared to other pollutants and the variance of WD also sensitive. We can easily conclude from the Fig. 6 that the observed frequency distribution of daily $PM_{2.5}$ emission during this period is positively skewed which gives an idea of how to fit the positively skewed distribution such as log-Normal, Gamma, Exponential, Weibull, etc depending upon the tail distribution. Precisely there is a higher

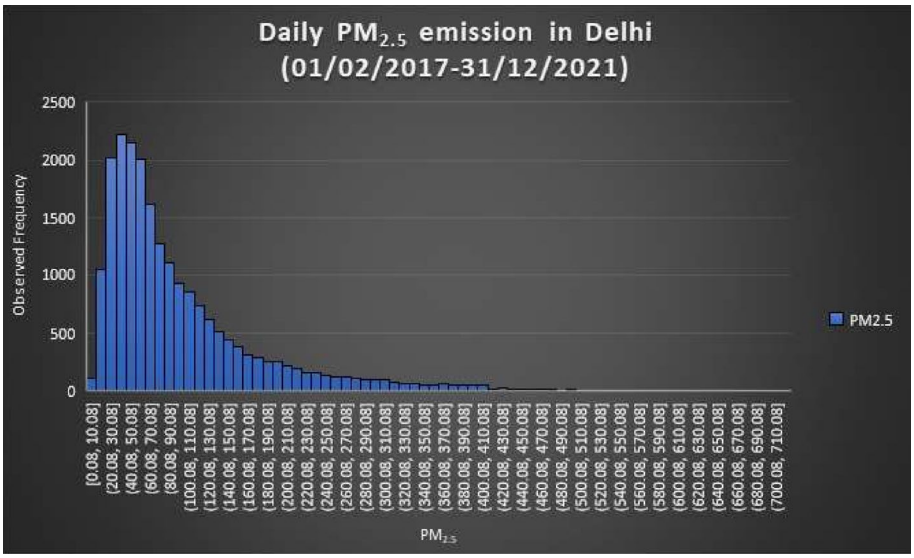

Fig. 6 The Box-Plot of daily $PM_{2.5}$ emission from 1^{st} Feburary, 2017 to 31^{st} December, 2021 in the study area, Delhi

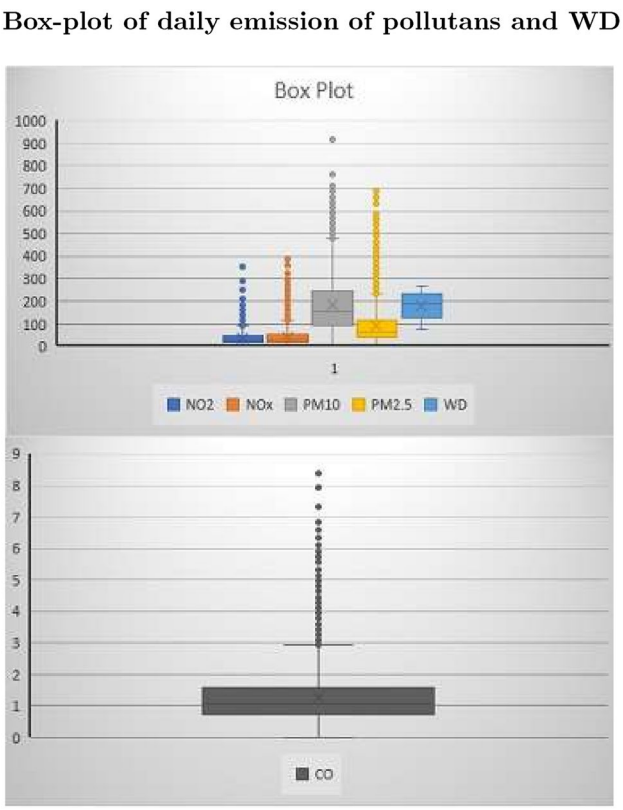

Fig. 7 The Box-Plot of daily $PM_{2.5}$, PM_{10}, NO_2, NO_x emission and WD from 1^{st} Feburary, 2017 to 31^{st} December, 2021 in the study area, Delhi

concentration in the interval from $30 - 40\,\mu g/m^3$. Figure 7 provides a brief overview of the variability and a five-point summary of the pollutants and WD which establishes the fact that PM_{10} and $PM_{2.5}$ have higher variability compared to other pollutants and the variance of WD also sensitive.

Results and discussion

This section goes over how to compare two new models, SC (Algorithm (1)) and SBVC ("Spatial Bayesian Vine-Copula estimation" Section) to other well-known spatial models step by step. Following that, we will attempt to provide a brief overview of pollutant concentrations in the future, as well as discuss how an important meteorological parameter can affect pollution concentrations mathematically.

We fit the parametric marginal CDF and PDF on the empirical CDF and PDF of an RV based on the AIC, BIC value, and KS test statistic value in Fig. 8. Because the empirical PDF is positively skewed, we only consider well-known positively skewed distributions such as Weibull, Log-normal (LN), Gamma, and Exponential distributions, and Table 1 shows that the LN distribution is suitable to fit based on the lowest AIC, BIC, and KS test statistic value. Similarly, we fit the circular family distributions on WD and discover that the VM distribution is the best PDF to fit.

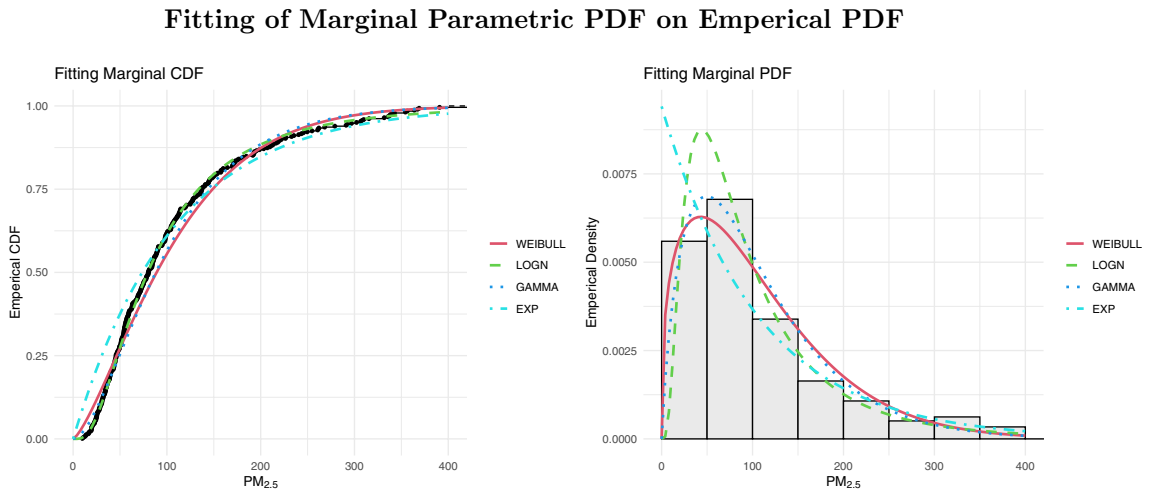

Fig. 8 The emperical marginal CDF is fitted by the marginal positively skewed parametric CDF and the fitting of marginal PDF

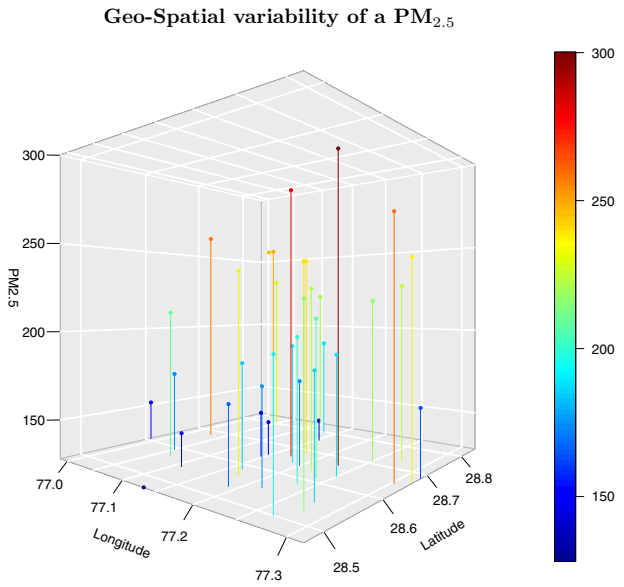

Fig. 9 The nature of spatial variation of a RV

Table 1 The value of KS statistic, AIC and BIC to determine the feasible marginal parametric PDF

Test Criteria	Weibull	Log-normal	Gamma	Exponential
KS	0.06884229	0.02849193	0.06276518	0.1478233
AIC	365.360	322.296	346.187	413.547
BIC	373.098	330.035	353.926	417.416

The next step is to estimate the parameter of the MDF. We already discussed the disadvantages of using MLE to estimate the parameter in "Fitting marginal distribution" section. As a result, we can use the EM algorithm to obtain the updated shape and scale parameters of the LN distribution. We discuss how the LogLik value converges to a fixed value after a certain number of iterations in Fig. 14. The required number of iterations for the EM algorithm in this case study is 223, after which the difference between the two LogLik values is negligible. In Fig. 9 depicts how the $PM_{2.5}$ value varies with respect to latitude and longitude. While the Longitude (Lon) ranges from 77.0 to 77.1 and the Latitude (Lat) varies, from 28.5 to 28.6, the $PM_{2.5}$ concentration is generally within $100 - 150\,\mu g/m^3$ but if Lon varies from 77.15 to 77.3, the $PM_{2.5}$ concentration becomes high and it ranges from $150 - 200\,\mu g/m^3$. Similarly, while Lat is varying from $28.6 - 28.7$ then the most spatial variability of $PM_{2.5}$ is detected in every interval of Lon and sometimes reaches up to $300\,\mu g/m^3$ while the Lat varies from $28.7 - 28.8$ the spatial variation is identified and the variation of $PM_{2.5}$ is almost lying between $200 - 250\,\mu g/m^3$. However, during fitting VM distribution we use the concept of UMVUE which is mentioned in the Theorem (1) in the "Fitting marginal distribution" section to get the shape and scale parameter of the VM distribution with better accuracy. In the following Table 2 we discuss the shape and scale parameters of LN and VM PDF and corresponding the last updated LogLik values.

Table 2 Details and updated values of shape and scale parameter and the corresponding Log-likelihood values

PDF	Shape	Scale	LogLik
LN	4.3764856	0.7701984	-1959.331124
VM	3.583	1.908	-32.41559

Our goal now is to run the two novel spatial interpolation algorithms mentioned in "Spatial copula estimation"

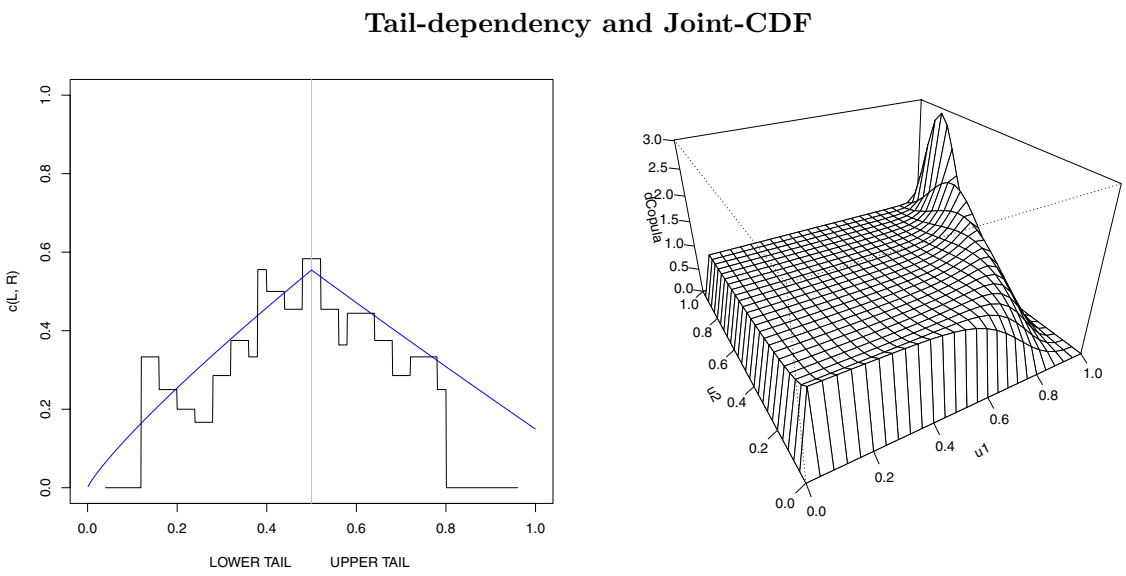

Fig. 10 In the *left* part the discussion regarding the lower tail and upper tail dependency of two RV and in the *right* part the joint CDF of two RV

and "Spatial Bayesian Vine-Copula estimation" sections and compare them to other spatial interpolation approaches. Using the threshold criteria mentioned in "Spatial copula estimation" section we divide the entire spatial domain into 4 HSC (Fig. 18) and consider the cutoff radius is 18026m as shown in Fig. 17. In the cluster dendrogram, the height represents the HD, and in the optimal number of clusters section, we plot SSW along the Y-axis and the optimal number of HSC along the X-axis. The following section focuses on the tail dependence of two RVs, as shown in Fig. 10. These two RVs in this case study are $PM_{2.5}$ and WD.

$$\begin{aligned} d_u &= \lim_{u\to 1} P\left[Y \geq F_y^{-1}(u) \mid X \geq F_x^{-1}(u)\right] \\ &= \lim_{u\to 1} \frac{C(1-u, 1-u)}{(1-u)} \\ d_l &= \lim_{u\to 0} P\left[Y \leq F_y^{-1}(u) \mid X \leq F_x^{-1}(u)\right] \\ &= \lim_{u\to 0} \frac{C(u,u)}{u} \end{aligned} \tag{22}$$

The upper tail dependence and lower tail dependence are defined in the Eq. (22). It describes the relationship between two RVs when one goes to extreme values and what the behavior of the other one is (Czado and Nagler 2022). We can conclude from this Fig. 10 that after 0.8 the upper tail of their distributions is independent and lesser than 0.1, the lower tail of their distributions is independent. As a result, we can say that higher values of $PM_{2.5}$ concentration are unaffected by WD because there is a very low concentration at that point but where the marginal PDF of $PM_{2.5}$ is moderate, there is a significant tail dependence on WD. The joint CDF of $PM_{2.5}$ and WD are plotted in the Fig. 10 on

Table 3 Two-way ANOVA to explain the dependence of $PM_{2.5}$ emission on WD and SC

Treatment	Df	SS	MS	F Ratio	*P*-Value
WD	21	20599.2864	980.9184	3.696	0.02404_*
Cluster	3	3623	1207.8	4.550	0.0334^*
Cluster˙ WD	3	1283	427.7	1.611	0.2543
Residuals	9	2389	265.4		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

right applying *BiCopSelect*() function in R, where the fitted copula is Rotated Twan Type-2 Copula with estimated Kendall's $\tau = 0.1341$ and the LogLik value is -1.204 which is the highest of any copula family, including GC, t-Copula, Frank, Clayton, Joe, and so on. We apply our novel copula-based spatial interpolation algorithm (SC) in a Bayesian framework after fitting the copula using *CDVineCondFit*() function in R. The posterior distribution and posterior estimate of the parameter are critical in this context. MHA is used in this context to obtain the posterior estimate of the parameters. According to this Fig. 15 we use the concept of Bayesian Inference to give the posterior estimate, assuming that the parameter prior distributions are uniform and truncated normal distributions. Following that, we use MHA to obtain the posterior estimate under the MSE loss function, which is 0.04898261 and -13.61893, respectively. The rate of convergence of two parameters is plotted in the Fig. 15 depicting that the rate of convergence of parameter 1 is faster than that of Parameter 2. Now we will look at how WD and

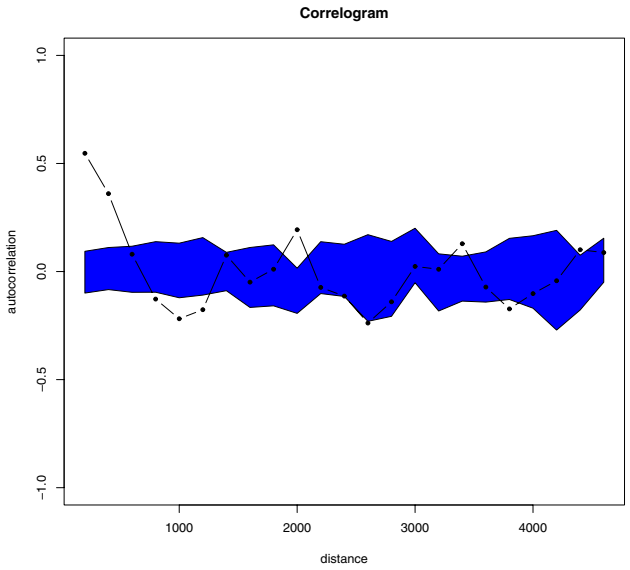

Fig. 11 Spatial ACF corresponding to every spatial lag, we plot the lag distance along the X-axis and the ACF along the Y-axis

spatial clustering affect the variance of $PM_{2.5}$ in Table 3 and the Fig. 16.

In the Two-Way Analysis Of the Variance model (Two way ANOVA), we consider $PM_{2.5}$ as a dependent variable and WD and clusters as independent variables. In the columns of Table 3, we represent Treatments, Degrees of freedom (Df), Sum of square (SS), Mean Square (MS), F-Ratio, and *P*-value, and along the rows, we represent WD, Cluster, their interaction effect, and residuals. We can see from Table 3 that there is a significant impact of WD and clustering on $PM_{2.5}$ emission at the 0.05 level of significance. However, the interaction effect of WD and Clusters has no significant impact on $PM_{2.5}$ emission. To aid comprehension, we present a graphical representation of these ANOVA tables in Fig. 16 in "Appendix C two-way ANOVA" section where WD is represented along the X-axis, $PM_{2.5}$ is represented along the Y-axis, and each spatial cluster is used as a panel. In the SC interpolation method, we investigate another factor, spatial ACF, which is employed as an important weight to counteract spatial variability across all lags.

As a result, in the Fig. 11, we depict the variation of ACF concerning the spatial lag. In this Fig. 11, we notice that the value of ACF is comparatively higher for nearby stations than for stations far away. We use the blue shaded region to give a brief idea of the interval of variation of ACF values. In this case study the fitted variogram model is Matern variogram model with nugget: 0; sill: 617; range: 0.02 and kappa: 0.09. Utilizing this value and the other distance weights in the Eq. (15) we calculate CCPDF in every unobserved location. We assume ϵ is 0.4224 and $p = 2$ in the Eq. (15).

The entire framework is now ready to execute the new spatial copula interpolation (SC) described in "Spatial copula estimation" section and Bayesian Spatial-Vine Copula (SBVC) described in "Spatial Bayesian Vine-Copula

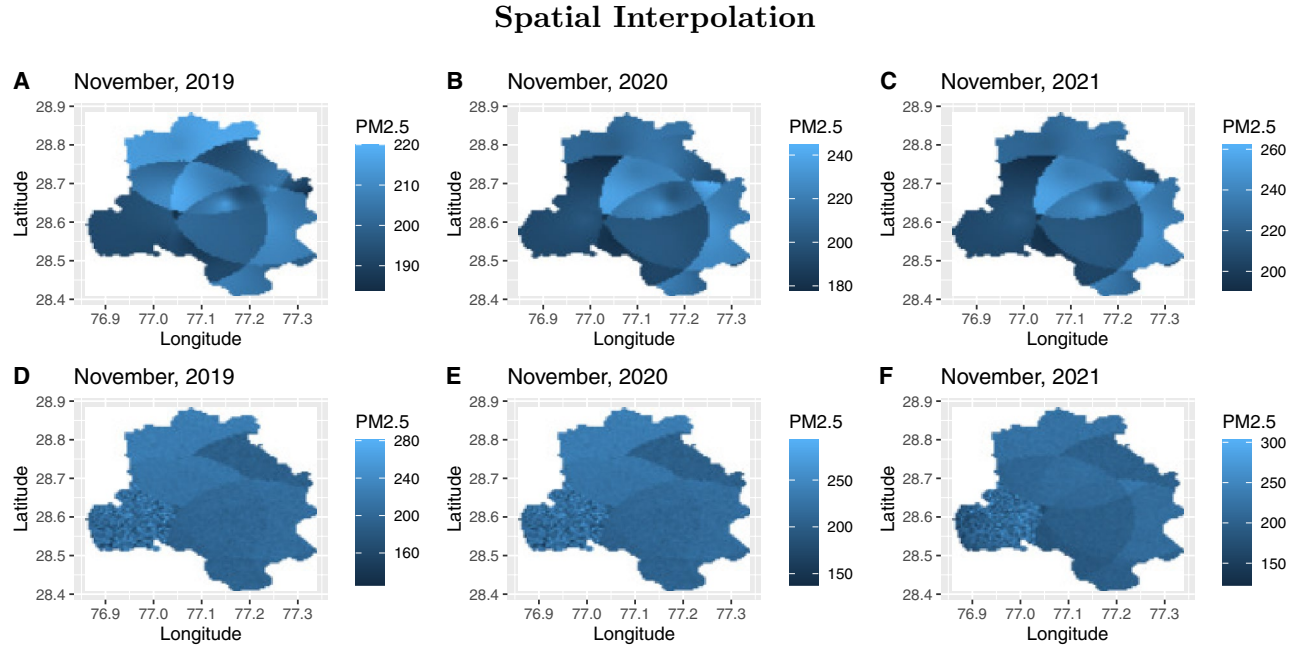

Fig. 12 Spatial Interpolation of $PM_{2.5}$ during the month of November, in 2019, 2020 and 2021. In *Top* the spatial interpolation technique SC is used and in *below* the SBVC algorithm is used as a spatial interpolation algorithm. Along X-axis we plot Longitude, along Y-axis we plot latitude and along the whole surface we plot the $PM_{2.5}$

estimation" section. As a result, we create an SRF within each HSC and focus on the spatial region between them. For SC, we assume that Lat and Lon have a bivariate uniform distribution, $PM_{2.5}$ has an LN distribution, and the suitable copula is Clayton Copula among other copula families like Gaussian, t-copula, archimedean-copulas, based on AIC, BIC, and LogLik values, to find their joint CDF using *mvdc*() function in R, with a parameter of 0.01697 and a dimension of 3. Then, using the Eq. (23), we obtain the required CCDF.

$$F(y_1 \mid x_1, x_2) = \frac{k_1 \cdot (x_1 x_2)^{-k_2} \cdot \int_{-\infty}^{y_1} \left(1 + \theta \cdot F_{Y_1}(t)\right)^{-k_3} \cdot (F_{Y_1}(t))^{k_3 - k_2} \cdot f_{Y_1}(t) dt}{c(x_1, x_2)} \quad (23)$$

Using the Eqs. (15) and (16), and (17) we get the CCPDF of each unobserved location. Then using the Algorithm (1) we get the interpolated values.

In the Fig. 12 we plot the monthly $PM_{2.5}$ emission during November for the three years, 2019, 2020, and 2021. According to this Fig. 12 we detect that using SC in November, 2019, the $PM_{2.5}$ emission varies from $180 - 220\,\mu g/m^3$. Using SBVC it ranges from $120 - 280\,\mu g/m^3$ (Fig. 12). A similar pattern is carrying on in 2020 and 2021 as well. We detect from Fig. 12 that the variation of SBVC is greater than that of SC. The northern and southeast part of Delhi is highly sensitive. In the western part of Delhi, the SBVC is ineffective to interpolate. As a result the $PM_{2.5}$ emission is random (Fig. 12). We illustrate the relationship between the observed and predicted values of three methods: SC, SBVC, and Ok in Fig. 13. We follow that there is a strong relationship between the observed and predicted values in SC, followed by SBVC, and lastly OK. Thus we conclude, that the power of explainable variation in SC is greater than SBVC and better than OK. MAE, RMSE of SC is lesser than SBVC, and lastly OK in Fig. 19.

Although the SC method outperforms the other two, there are some areas where improvements are possible, such as: (i) We assume the rate of inclusion of geo-spatial points in a cluster is constant, during clustering but this can vary in practice. (ii) We ignore the effect of extreme values during interpolation. (iii) Degree of departure of characteristics between observed and unobserved points sometimes contradicts the concept of spatial continuity but we ignore that. (iv) We do not pay enough attention to its temporal stationarity. SBVC accepts the same drawbacks with the incapability of exploration of spatial trends.

Conclusion

The proposed models' SC and SBVC are extensions of the previous spatial copula-based models that majorly addressed issues such as bin selection, usage of MLE to estimate the parameter in missing data sets, and so on. When compared to other geostatistical models, the proposed SC and SBVC are very effective and provide nearly accurate results (from Fig. 19). The SC model produces better results for spatially skewed spatial random fields and provides a mathematical argument for selecting essential covariates. This study provides an idea of alternative distance weights and distance functions that are very effective in capturing spatial

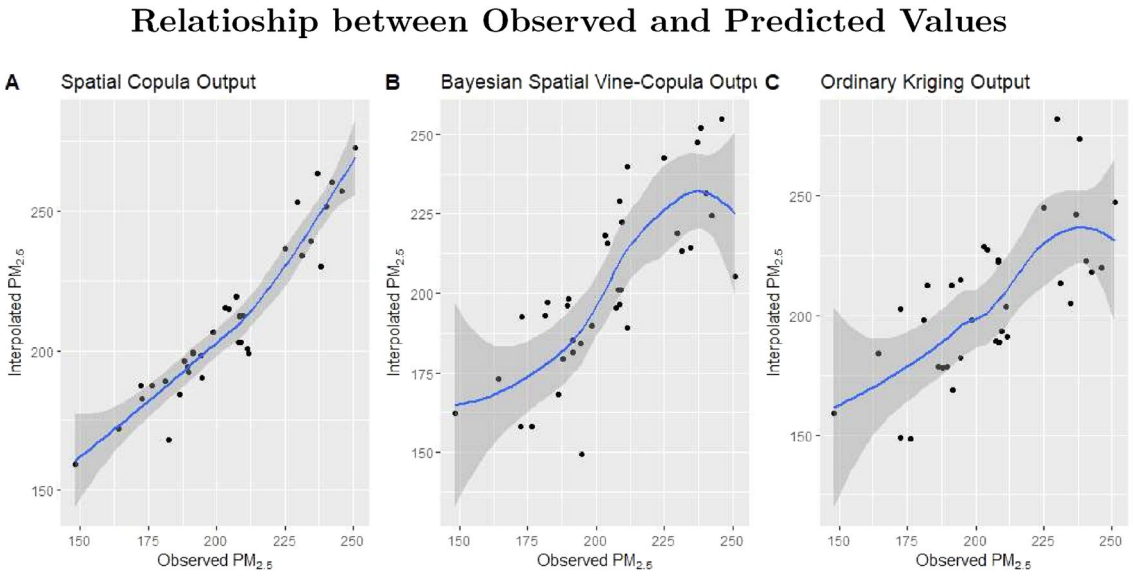

Fig. 13 Relationship between the observed and predicted values of three methods: SC, SBVC, and OK

variation. A temporal extension of this algorithm is possible, which motivates further research. This model is explained in this study using a real-world data set of PM concentrations in the air. Still, this algorithm can be used in other scenarios such as mining, temperature modeling, meteorological modeling, and so on. This algorithm may be more advantageous than other spatial estimation models because it makes no assumptions about Gaussian distribution, intrinsic stationarity, dynamic behavior, or skewed data sets.

Appendix A Proof of Theorem (1)

Proof If $X \sim VM(k,\mu)$ then we know the corresponding charachteristic function of X is $\phi_n(x) = E\left[e^{inx}\right]$

$$\begin{aligned} E\left[e^{\mathrm{inx}}\right] &= \int_0^{2\pi} \frac{e^{\mathrm{inx}} \cdot e^{k\cdot\cos(x-\mu)}}{2\pi I_0(k)}\mathrm{d}x \\ &= \frac{I_{|n|}(k)\cdot e^{in\mu}}{I_0(k)} \end{aligned} \tag{A1}$$

In the Eq. (A1) the term $I_n(k) = \frac{\int_0^{\pi} e^{k\cdot\cos(\mathrm{x})\cos(\mathrm{nx})}\mathrm{dx}}{\pi}$. In Eq. (A1) putting $n = 1$ we get,

$$E(e^{\mathrm{ix}}) = \frac{I_1(k)\cdot e^{i\mu}}{I_0(k)} \tag{A2}$$

and putting $n = -1$ we get,

$$E(e^{-\mathrm{ix}}) = \frac{I_1(k)\cdot e^{-i\mu}}{I_0(k)} \tag{A3}$$

Adding and subtracting Eqs. (A2) and (A3) we get

$$\begin{aligned} E\left(\frac{I_0(k)\cdot\cos(\mathrm{x})}{I_1(k)}\right) &= \cos\mu \\ E\left(\frac{I_0(k)\cdot\sin(\mathrm{x})}{I_1(k)}\right) &= \sin\mu \end{aligned} \tag{A4}$$

Therefore, from the Eq. (A4) the statistic $T_1(x) = \frac{I_0(k)\cdot\cos(\mathrm{x})}{I_1(k)}$ and $T_2(x) = \frac{I_0(k)\cdot\sin(\mathrm{x})}{I_1(k)}$ are the unbiased estimators of $\cos\mu$ and $\sin\mu$ respectively.

Here, the PDF of X is denoted as $f(x)$ and the parameter space is defined as Φ and support is defined as $\mathcal{X}$.

$$\begin{aligned} f(x) &= \frac{e^{k\cdot\cos(\mathrm{X}-\mu)}}{2\pi\cdot I_0(k)} \\ &= \exp\left[k\cdot\cos(\mathrm{x}-\mu) - \log(2\pi I_0(k))\right] \\ &= \exp[k\cdot\cos(\mathrm{x})\cos(\mu) + k\cdot\sin(\mathrm{x})\sin(\mu) \\ &\quad - \log(2\pi) - \log(I_0(k))\big] \end{aligned} \tag{A5}$$

From the Eq. (A5) we write the likelihood function as product of two terms moreover this VM distribution satisfying the following properties:

1. $\mathcal{X}$ is [0,2π] therefore, it is independent upon the parameter.
2. $\Phi = \{(\mu,k) : \mu \in \mathscr{R}; k > 0\}$ which indicating it is an open interval.
3. Here $\{1, \cos(\mathrm{x}), \sin(\mathrm{x})\}$ and $\{1, cos(\mu), sin(\mu)\}$ are Linearly Independent (LIN).

Therefore, we tell that the PDF is belonging Two-PEF. Therefore, cos(x) and sin(x) are complete and sufficient statistic of cos(μ) and sin(μ). Using Lehmann-Scheffe Theorem they are the UMVUE. Moreover, using Eq. (A1) replacing $n = 2$ we get

$$\begin{aligned} E[\cos 2\mathrm{x}] &= \frac{I_2(k)}{I_0(k)} \\ \Rightarrow \mathrm{E}(\cos^2(\mathrm{x})) &= \frac{1}{2} + \frac{I_2(k)\cdot\cos(2\mu)}{2I_0(k)} \\ \Rightarrow \mathrm{E}(\sin^2(\mathrm{x})) &= \frac{1}{2} - \frac{I_2(k)\cdot\sin(2\mu)}{2I_0(k)} \end{aligned} \tag{A6}$$

Using Eq. (A6) we get

$$\begin{aligned} \mathrm{var}(\cos(\mathrm{x})) &= \frac{1}{2} + \frac{I_2(k)\cdot\cos(2\mu)}{2I_0(k)} - \left(\frac{I_1(k)\cdot\cos(\mu)}{I_0(k)}\right)^2 \\ \mathrm{var}(\sin(\mathrm{x})) &= \frac{1}{2} - \frac{I_2(k)\cdot\sin(2\mu)}{2I_0(k)} - \left(\frac{I_1(k)\cdot\sin(\mu)}{I_0(k)}\right)^2 \end{aligned} \tag{A7}$$

□

A.1 EM algorithm estimation of parameters of VM distribution

Likewise LN distribution $\mathcal{Q}\left((\mu,k) \mid \left(\mu^{(m)}, k^{(m)}\right)\right)$ is the updated CElikC at $_{m^{\mathrm{th}}}$ iteration is:

$$\begin{aligned} &\mathcal{Q}\left((\mu,k) \mid \left(\mu^{(m)}, k^{(m)}\right)\right) \\ &= \mathrm{E}^{c}_{\left(\mu^{(m)},k^{(m)}\right)}\left[\frac{\sum_{i=1}^{n_2} k\cos(w_i-\mu) - n_2\cdot\log(2\pi I_0(k))}{\sum_{i=1}^{n_1} k\cos(w_i-\mu) - n_1\cdot\log(2\pi I_0(k))}\right] \end{aligned} \tag{A8}$$

Using the Eq. (A8) we complete M-step and find the most updated values.

Appendix B Convergence rate of MHA and EM

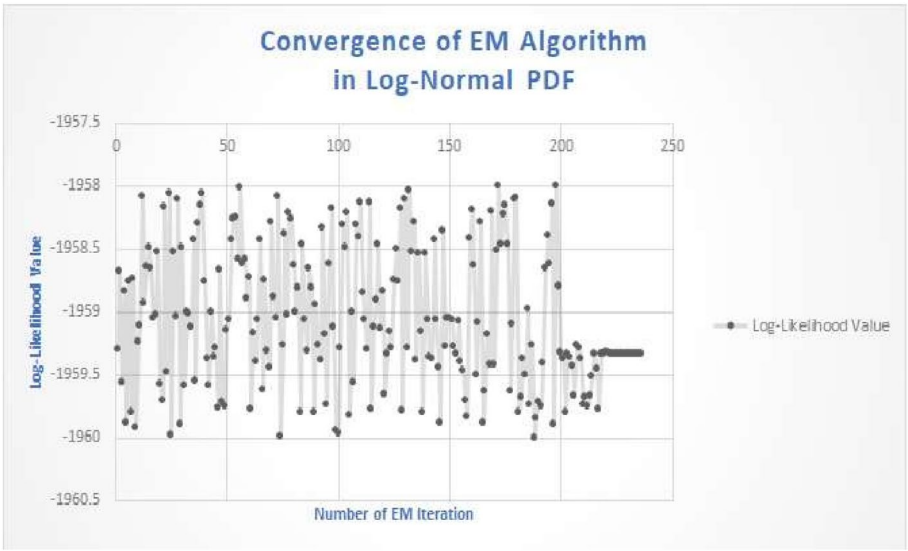

Fig. 14 The convergence of the Log-Likelihood value after updating the value of the parameters in each iteration using EM algorithm

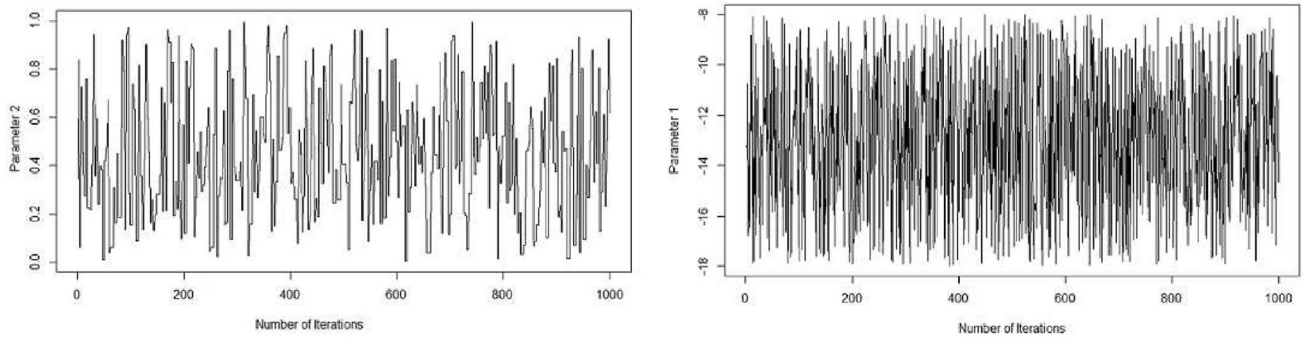

Fig. 15 The rate of convergence of the two parameters of Rotated Tawn Type-2 copula family. In (*left*) the first parameter of the copula family and in (*right*) the second parameter of the copula family are estimated

Appendix C two-way ANOVA

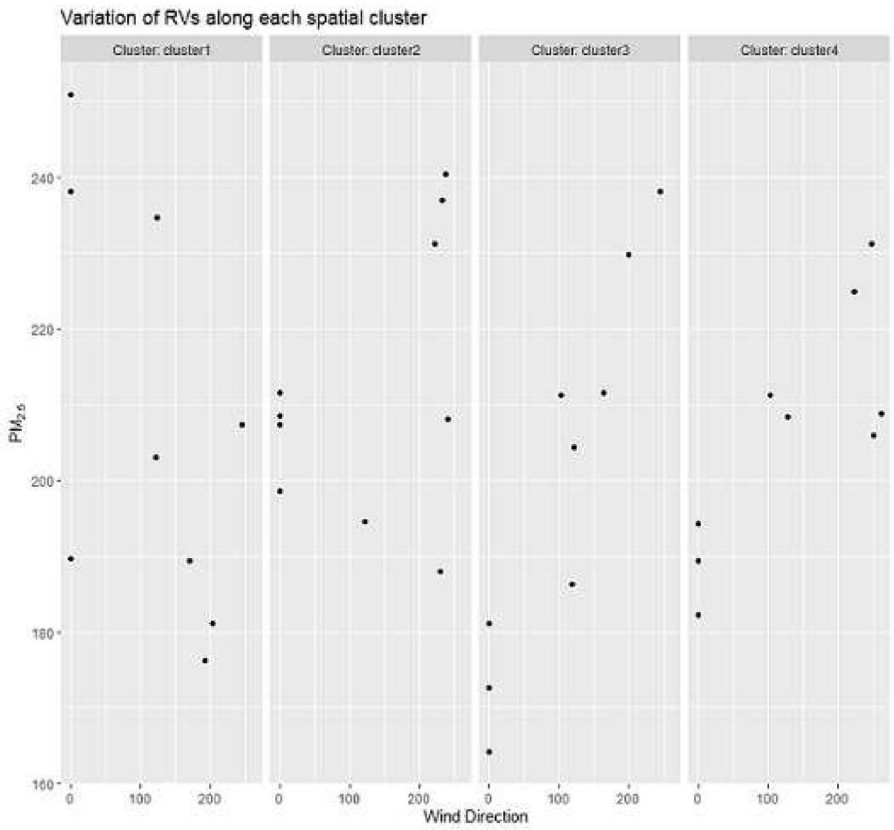

Fig. 16 How WD and Spatial cluster make an impact on $PM_{2.5}$ in this case study

Appendix D Optimal number of Spatial HSC and its size

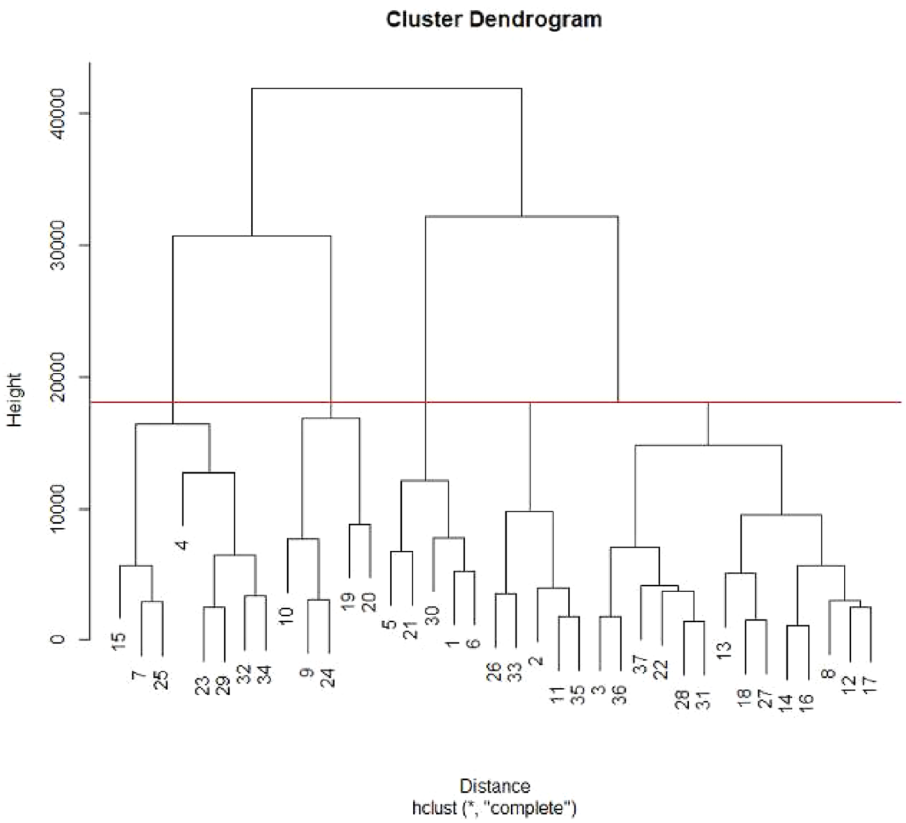

Fig. 17 Optimal HSC size and its cutoff HD

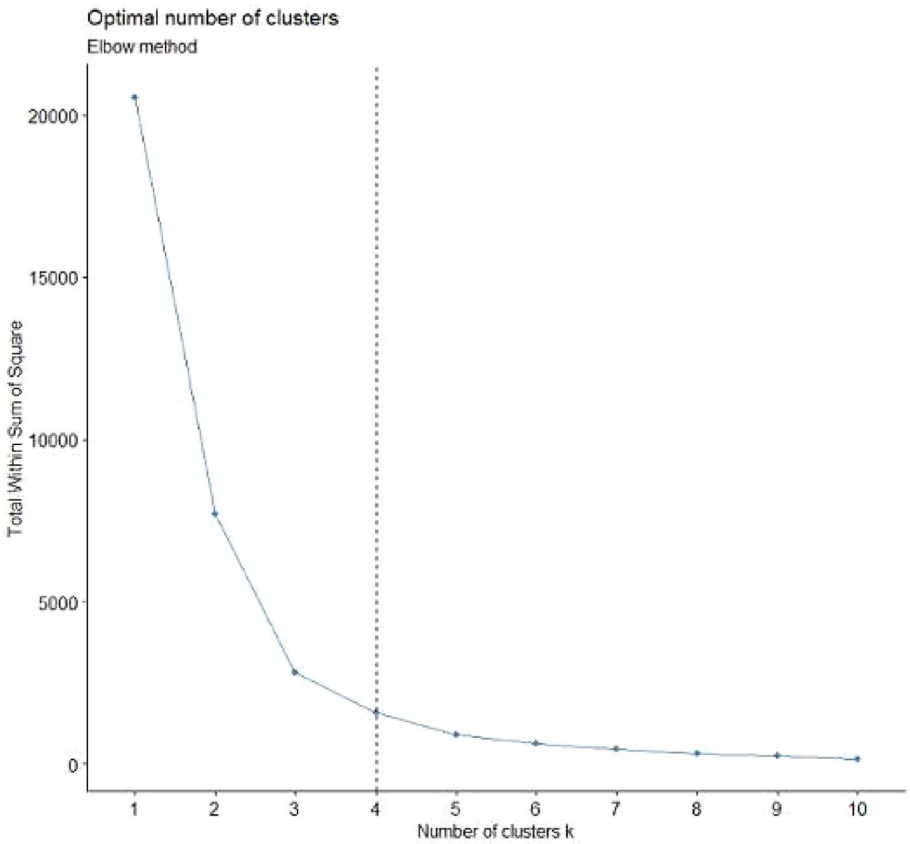

Fig. 18 Optimal number of HSC using Elbow Method

Appendix E RMSE and MAE of SC, SBVC, and OK

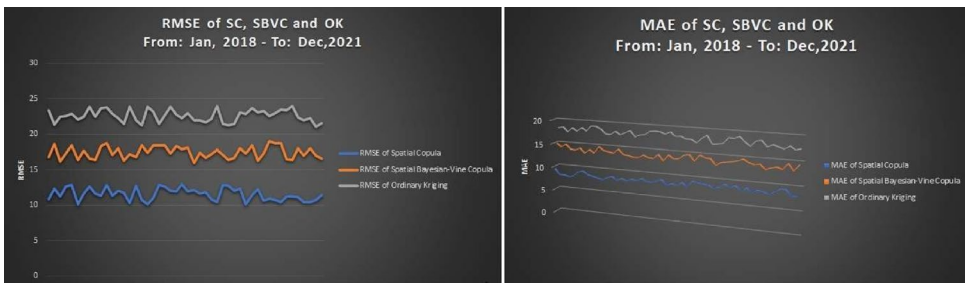

Fig. 19 Comparison of the performance of three methods: SC, SBVC, and OK. Along Y-axis we plot the RMSE and MAE of the four years from January, 2018 to December, 2021

Acknowledgements We wish to thank Editor-in-chief, honorable reviewers for giving helpful suggestions and the entire team of this journal for enough support.

Author Contributions The three authors have contributed to this paper equally, and directly. DT develops the model, writes the entire manuscript, and reviews the important literature. ID guides and motivates during the entire journey and gives important suggestions during writing. SC collects the data, outlines the algorithm, and executes the model on the case study.

Availability of data and materials The Data is available upon valid request to the corresponding author.

Declarations

Conflict of interests Authors state no conflict of interest.

References

Aas K, Czado C, Frigessi A et al (2009) Pair-copula constructions of multiple dependence. Insur Math Econ 44(2):182–198

Alidoost F, Stein A, Su Z (2018) Copula-based interpolation methods for air temperature data using collocated covariates. Spat Stat 28:128–140

Alidoost F, Stein A, Su Z et al (2021) Multivariate copula quantile mapping for bias correction of reanalysis air temperature data. Journal of spatial science 66(2):299–315
Auerbach A, Hernandez ML (2012) The effect of environmental oxidative stress on airway inflammation. Curr Opin Allergy Clin Immunol 12(2):133
Bai Y, Kang J, Song PXK (2014) Efficient pairwise composite likelihood estimation for spatial-clustered data. Biometrics 70(3):661–670
Bárdossy A (2006) Copula-based geostatistical models for groundwater quality parameters. Water Resour Res 42(11)
Bárdossy A (2011) Interpolation of groundwater quality parameters with some values below the detection limit. Hydrol Earth Syst Sci 15(9):2763–2775
Bárdossy A, Pegram G (2009) Copula based multisite model for daily precipitation simulation. Hydrol Earth Syst Sci 13(12):2299–2314
Bostan P, Stein A, Alidoost F et al (2021) Minimum temperature mapping with spatial copula interpolation. Spat Stat 42(100):464
Carreau J, Toulemonde G (2020) Extra-parametrized extreme value copula: Extension to a spatial framework. Spat Stat 40(100):410
Cressie N (1990) The origins of kriging. Math Geol 22(3):239–252
Czado C, Nagler T (2022) Vine copula based modeling. Annu Rev Stat Appl 9(1):453–477
D'Urso P, De Giovanni L, Vitale V (2022) A d-vine copula-based quantile regression model with spatial dependence for covid-19 infection rate in italy. Spatial statistics p 100586
Erhardt TM, Czado C, Schepsmeier U (2015) R-vine models for spatial time series with an application to daily mean temperature. Biometrics 71(2):323–332
Gade K (2010) A non-singular horizontal position representation. J Navig 63(3):395–417
García JA, Pizarro MM, Acero FJ et al (2021) A bayesian hierarchical spatial copula model: An application to extreme temperatures in extremadura (spain). Atmosphere 12(7):897
Gnann SJ, Allmendinger MC, Haslauer CP et al (2018) Improving copula-based spatial interpolation with secondary data. Spat Stat 28:105–127
Gräler B (2014) Modelling skewed spatial random fields through the spatial vine copula. Spat Stat 10:87–102
Hubert L (1974) Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. J Amer Statist Assoc 69(347):698–704
Isaaks EH, Srivastava MR (1989) Applied geostatistics. 551.72 ISA
Kazianka H, Pilz J (2011) Bayesian spatial modeling and interpolation using copulas. Comput Geosci 37(3):310–319
Khan F, Spöck G, Pilz J (2020) A novel approach for modelling pattern and spatial dependence structures between climate variables by combining mixture models with copula models. Int J Climatol 40(2):1049–1066
Krupskii P, Genton MG (2019) A copula model for non-gaussian multivariate spatial data. J Multivar Anal 169:264–277
Krupskii P, Huser R, Genton MG (2018) Factor copula models for replicated spatial data. J Amer Statist Assoc 113(521):467–479
Lim SS, Vos T, Flaxman AD et al (2012) A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. Lancet 380(9859):2224–2260
Ma J, Ding Y, Cheng JC et al (2019) A temporal-spatial interpolation and extrapolation method based on geographic long short-term memory neural network for pm2. 5. J Clean Prod 237:117,729
Machuca-Mory DF, Deutsch CV (2013) Non-stationary geostatistical modeling based on distance weighted statistics and distributions. Math Geosci 45(1):31–48
Masseran N (2021) Modeling the characteristics of unhealthy air pollution events: A copula approach. Int J Environ Res Public Health 18(16):8751
Masseran N, Hussain SI (2020) Copula modelling on the dynamic dependence structure of multiple air pollutant variables. Mathematics 8(11):1910
McLachlan GJ, Krishnan T (2007) The EM algorithm and extensions, vol 382. John Wiley & Sons
Mukherjee T, Asutosh A, Pandey SK et al (2018) Increasing potential for air pollution over megacity new delhi: A study based on 2016 diwali episode. Aerosol Air Qual Res 18(9):2510–2518
Musafer GN, Thompson MH (2017) Non-linear optimal multivariate spatial design using spatial vine copulas. Stoch Environ Res Risk Assess 31(2):551–570
Nelsen RB (2007) An introduction to copulas. Springer Science & Business Media
Quessy JF, Rivest LP, Toupin MH (2015) Semi-parametric pairwise inference methods in spatial models based on copulas. Spat Stat 14:472–490
Richardson R (2021) Spatial generalized linear models with non-gaussian translation processes. J Agric Biol Environ Stat pp 1–18
Samal CG, Gupta D, Pathania R et al (2013) Air pollution in micro-environments: A case study of india habitat centre enclosed vehicular parking, new delhi. Indoor Built Environ 22(4):710–718
Shao Y, Ma Z, Wang J, et al. (2020) Estimating daily ground-level pm2. 5 in china with random-forest-based spatiotemporal kriging. Sci Total Environ 740:139,761
Sklar A (1973) Random variables, joint distribution functions, and copulas. Kybernetika 9(6):449–460
Sohrabian B (2021) Geostatistical prediction through convex combination of archimedean copulas. Spat Stat 41(100):488
Wang J, Wang Z, Deng M et al (2021) Heterogeneous spatiotemporal copula-based kriging for air pollution prediction. Trans GIS 25(6):3210–3232
Zheng S, Pozzer A, Cao C et al. (2015) Long-term (2001–2012) concentrations of fine particulate matter (pm 2.5) and the impact on human health in beijing. china. Atmos Chem Phys Discuss 15(10):5715–5725