A Spatial Copula Interpolation in a Random Field with Application in Air Pollution Data

Ishapathik Das
IIT Tirupati
Tirupati, AP, India

SGD-2023
JUNE 15, 2023

Air Pollution

- Air pollution is the biggest intimidation to public health and one of the primary reasons to cause respiratory hazards and chronic obstructive pulmonary disease (COPD) as a result risk of a patient being attacked by the COVID-19 virus is too high.

- The pollution at an alarming level attracts the researchers' recognition so much that the influence of different policies for example odd and even trials, implementation of CNG fuel are reviewed in detail.

CAUSE OF CONCERN

Air pollution exacerbate burden on workers' health

Study says over the years, air pollution in Delhi has become a significant concern

First India Bureau

New Delhi: The health effects of breathing hazardous air on a regular basis are more frequently felt by people who work outside. Confirming this, a recent study found that workplace exposure to ambient air pollution and extreme weather events exacerbate the overall burden on outdoor workers' health in the country's national capital.

Breathing difficulties or acute lung function impairment, irregular heartbeat and chest discomfort, back, shoulder, and joint pains, eye redness and irritation, and finally skin rashes, headaches, and overall weakness were assessed for three occupational groups working in New Delhi — autorickshaw drivers, sweepers, and vendors — who were exposed to poor air quality and the effects of extreme weather during peak summer and winter months.

The study, titled "Health impact assessment of Delhi's outdoor workers exposed to air pollution and extreme weather events: an integrated epidemiology approach," was carried out by doctors, scientists, and academicians from a number of reputable national institutes and used a detailed questionnaire in consultation with medical experts to understand their perceptions of the health effects of air pollution and extreme weather events on outdoor workers, as well as a pulmonary function test for reassurance.

Poor air quality, extreme weather wreaking havoc on health of Delhi's outdoor workers – https://www.nationalheraldindia.com/national/poor-air-quality-extreme-weather-wreaking-havoc-on-health-of-delhis-outdoor-workers.

Introduction

- Transport sector emission inventory for megacity Delhi has been developed for the period 2000-2005 to quantify vehicular emissions and
- Evaluate the effect of relevant policy reforms on total emissions of various air pollutants like CO2, CO, HC, NOx, $SO_{2,}$ etc. over the years to assist in future policy formulations.
- Emission factor and vehicle utilization factor based approach as recommended by IPCC [13] have been used for estimating emissions.

Introduction

- Megacity Delhi – the National Capital Region of India – is one of the most polluted cities in the world [25] having transport as major source of criteria area pollutants [12].
- In terms of emissions of various pollutants, Delhi was among the top five SO2 emitting megacities of the world in early nineties and transport sector was the prime culprit for it [11].
- Mashelkar et al. (2002) [18] state that the emission range of NOx from transport sector is 66% to 74% in Delhi.
- CPCB (Central Pollution Control Board) [4] data shows that almost 50% of the emission in Delhi is from vehicular activities, followed by domestic, industrial, and power plants.

Introduction

- According to Xie and Shah (2002) [26], diesel driven vehicles were the major source of NOx emission in Delhi, whereas least contribution was from two- and three-wheelers.
- According to ADB (Asian Development Bank) [1], diesel driven vehicles are the major contributor of PM emission among all vehicle categories in Delhi.
- As a consequence, incidence of respiratory diseases in Delhi is 12 times the national average, and 30% of Delhi's population suffers from respiratory disorders [16].
- Its poor air quality is responsible for about 18600 premature deaths per year [22].

Important methodologies in previous research work

Location	Focus area	Prime detection	Reference
Delhi	Estimation of $PM_{2.5}$	IDW and OK	Shukla et al. 2020
Guangdong province, China	Spatio-temporal estimation of $PM_{2.5}$	IDW-BLSTM performed efficiently	Ma et al. 2019
China	Spatio-temporal estimation of $PM_{2.5}$	RFSTK responded satisfactorily	Shao et al. 2020
China	Spatial and temporal pattern recognition of $PM_{2.5}$ concentration	Linear regression and grey system correlation analysis	Lu et al. 2017
Delhi	Relationship between $PM_{2.5}$ and other spatio-temporal covariates are explained	Six ML learners	Mandal et al. 2020
China	Statistical analysis of $PM_{2.5}$ concentration	Seasonal influence	San Martini et al. 2015

CO2 Emission

- The CO2 emissions increased by about 24% from 13.36 Tg in 2000 to 16.62 Tg in 2005.
- From 2000 to 2002, emissions decreased by about 11% from 13.36 Tg in 2000 to 11.92 Tg in 2002 followed by rising trend till 2005.
- Implementations of various emission norms and phasing out of old vehicles during this period might be responsible for decreased emissions during 2001 and 2002.
- Increasing vehicle population led to rising emissions between 2002 and 2005.

CO Emission

- Emissions of CO increased by 78% from 197 Gg in 2000 to 350 Gg in 2005.
- Gradual increasing trend was observed in emissions from 2000 to 2001.
- Increasing two wheeler population might be responsible for the gradual increment in CO emissions during 2000 to 2005.

HC Emission

- Hydrocarbons (HC) emission increased 1.3 times from 96 Gg in 2000 to 222 Gg in 2005.
- Highest increment in HC emission was observed from 107 Gg in 2001 to 177 Gg in 2002 (65%).
- After 2002, HC emissions showed a constant rising trend.

NOX

- NOx emission trend, in comparison to other pollutants, showed a different scenario.
- About 0.5% of decline was observed in NOx emission (105 Gg in 2000 to 104 Gg in 2005).
- During 2000 to 2002, steep decreases were observed followed by steadily rising emission trend until 2005.
- Highest increase (23%) was found during 2002 to 2003; followed by 6% annual average increase till 2005.
- Goods vehicles were the predominated source of NOx emission (51%) followed by buses (36%), cars (6%), two wheelers (6%), and autos (1%) during the study period.

TSP Emission

- Total suspended particles (TSP) emissions, increased from 8 Gg in 2000 to 10 Gg in 2005 (about 31% increment).
- Two-wheelers were responsible for high TSP emissions (40%) during 2002-2005 followed by goods vehicle (29%), buses (19%), cars (10%) and autos (2%).
- However, it is observed that TSP emissions from goods vehicles were higher before 2002.
- It shows the efficacy of CNG implementation and clean fuel related initiatives taken after 2001.

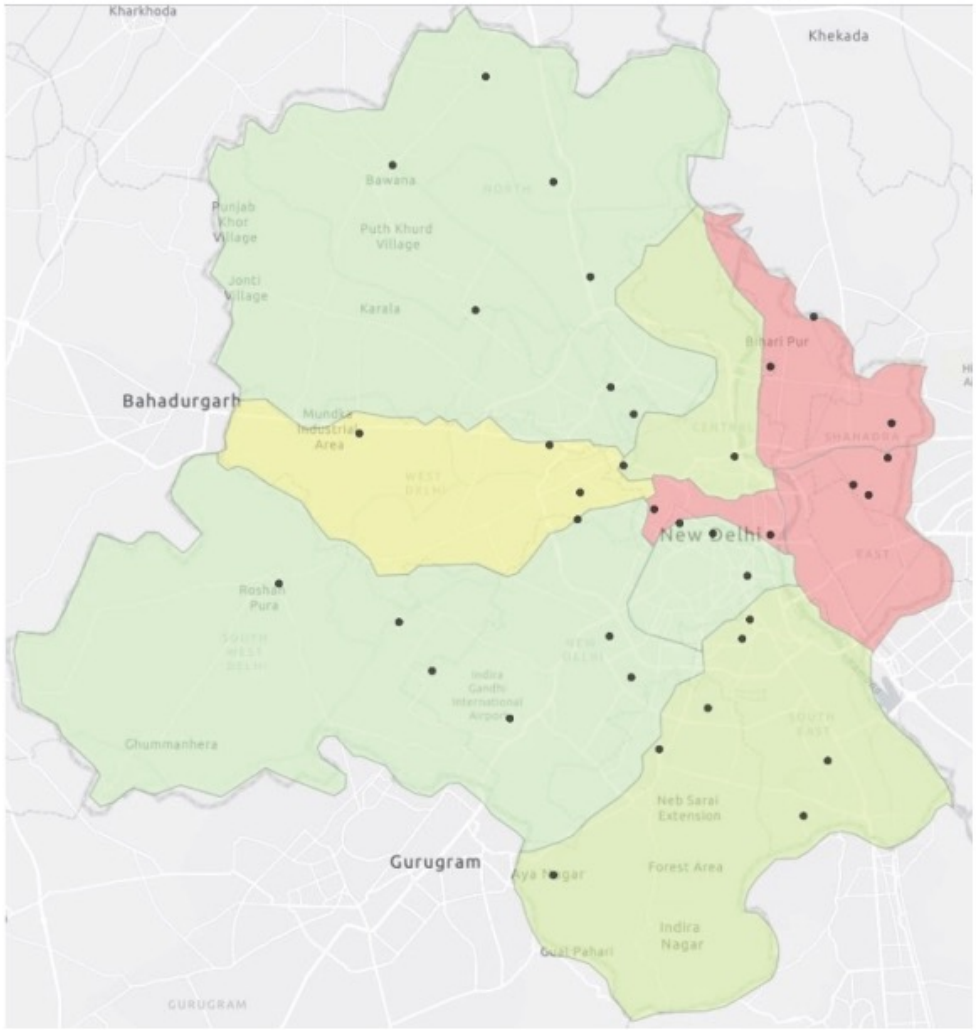
SO2 Emission

- Approximately 9% of the decrease was observed from 12 Gg in 2000 to 11 Gg in 2005, with highest decrease (32%) between 2000 and 2002.
- After 2002, emissions increased with annual average rate of 7.69% per year, but were less in comparison to those of 2000.
- Buses contributed highest amount (45%) of SO2 followed by goods vehicle (31%), cars (12%), two-wheelers (11%) and auto (1%)

Study Area

- We selected Delhi, the capital of India, to study the air pollution of BL and DL during the first wave of the Covid-19 pandemic in India.
- We considered the air pollution data collected by the monitoring stations, maintained by the Central Pollution Control Board (CPCB).
- The dataset contains 38 monitoring stations, where the data were collected over 24 hours.
- The time period was taken from 1st January 2019 to 28th February 2021.

List of monitoring stations in Delhi



- CPCB monitoring stations

Major activity of monitoring stations

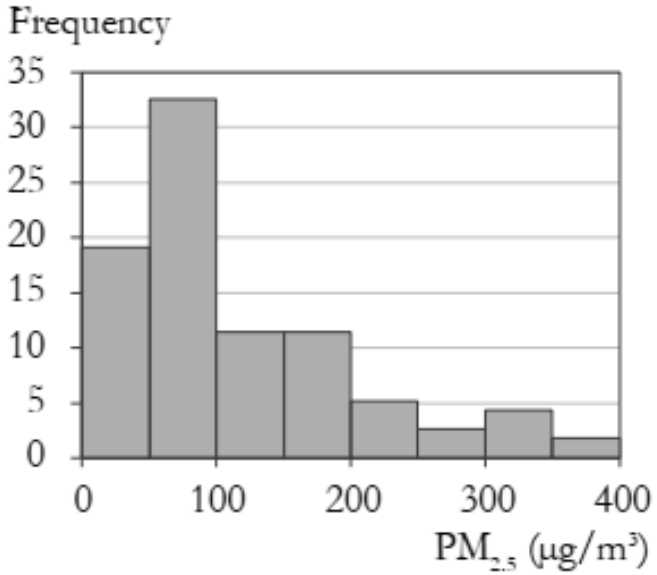
S.No.	Monitoring station	Major activity
1	AnandVihar	Transport
2	IGI Airport, T3	Transport
3	ITO, Delhi	Transport
4	R K Puram	Residential
5	Punjabi Bagh	Residential
6	Ayanagr	Residential
7	Burari Crossing	Residential
8	Sirifort	Residential
9	Mandirmarg	Commercial
10	Lodhi Road	Commercial
11	Shadipur	Commercial
12	CRRI, Mathura Road	Institutional
13	DTU	Institutional
14	NSIT, Dwaraka	Institutional
15	North Campus DU	Institutional
16	Pusa	Institutional

Descriptive statistical measurements of 24 hour $PM_{2.5}$ emissions during the entire time period

Summary	Value, ($\mu g/m^3$)
Minimum	10.22
First quartile (Q_1)	41.88
Median (Q_2)	67.77
Mean	94.12
Third quartile (Q_3)	122.14
Maximum	715.04
Mode	45.044
Standard deviation	75.18

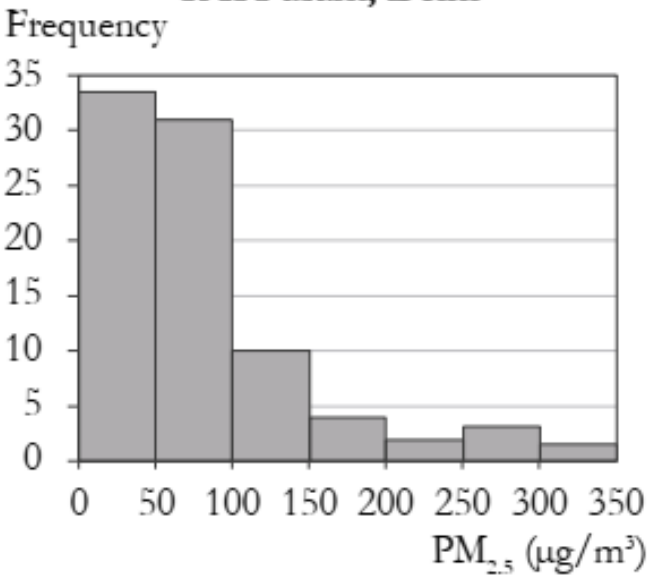
Histogram of $PM_{2.5}$ of monitoring stations

AnandVihar



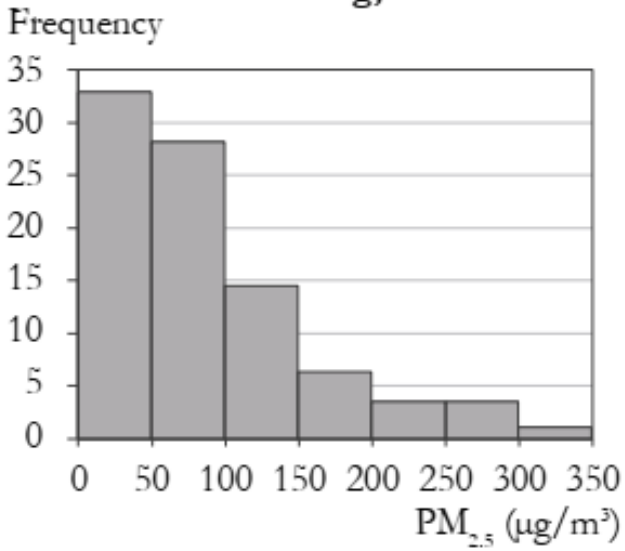
Minimum: 20.78, First quartile: 53.57,
Median: 86.25, Mean: 112.83,
Third quartile: 152.83, Maximum: 373.46,
Standard deviation: 85.23, Skewness: 0.94
Kurtosis: 4.28

R K Puram, Delhi



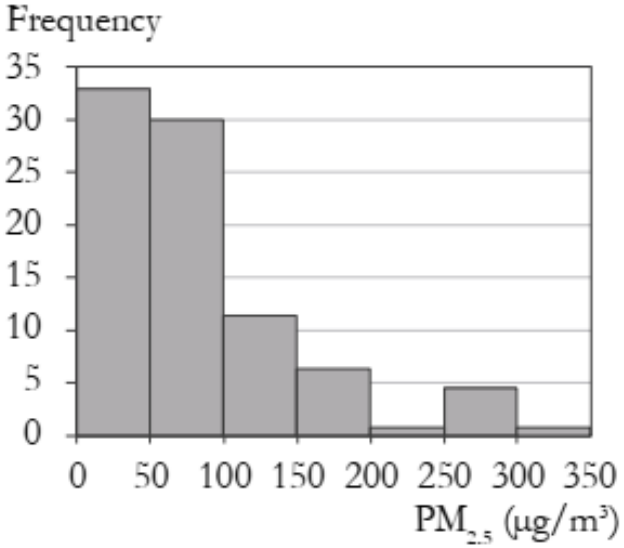
Minimum: 11.38, First quartile: 35.37,
Median: 57.75, Mean: 78.33,
Third quartile: 97.52, Maximum: 318.06,
Standard deviation: 64.79, Skewness: 0.95
Kurtosis: 6.04

Mandir Marg, Delhi



Minimum: 15.65, First quartile: 39.37,
Median: 67.97, Mean: 87.02,
Third quartile: 110.31, Maximum: 311.65,
Standard deviation: 67.55, Skewness: 0.85
Kurtosis: 4.94

CRRI mathura road



Minimum: 15.31, First quartile: 43.68,
Median: 69.08, Mean: 91.29,
Third quartile: 122.29, Maximum: 307.3,
Standard deviation: 70.53, Skewness: 0.94
Kurtosis: 4.91

Inverse Distance Weight (IDW)

- $$Z(x_0) = \sum_{j=1}^{n} \lambda_j Z(x_j) / \sum_{j=1}^{n} \lambda_j$$

  Where, $\lambda_j = 1/d(x_0, x_j)^p$; j $= 1, ..., n$.

- Find out p on the basis of minimum RMSE.

RMSE and R^2 changing with the variation of p in IDW

p	RMSE	R^2	p	RMSE	R^2
0.1	21.713862	0.915102	1.1	21.41463	0.917632
0.2	21.5611777	0.916301	1.2	21.585	0.9163575
0.3	21.4079816	0.9175161	1.3	21.76469	0.9149997
0.4	21.2656589	0.9186298	1.4	21.947072	0.9136144
0.5	21.14805	0.9589295	1.5	22.12609	0.9122447
0.6	21.0694	0.959249	1.6	22.29762	0.9109228
0.7	21.0408	0.9593693	1.7	22.4589	0.9096708
0.8	21.06682	0.9202081	1.8	22.608	0.9085022
0.9	21.144	0.9196403	1.9	22.74565	0.9074232
1.0	21.264	0.9187535	2.0	22.8706	0.9064344

Choice of p

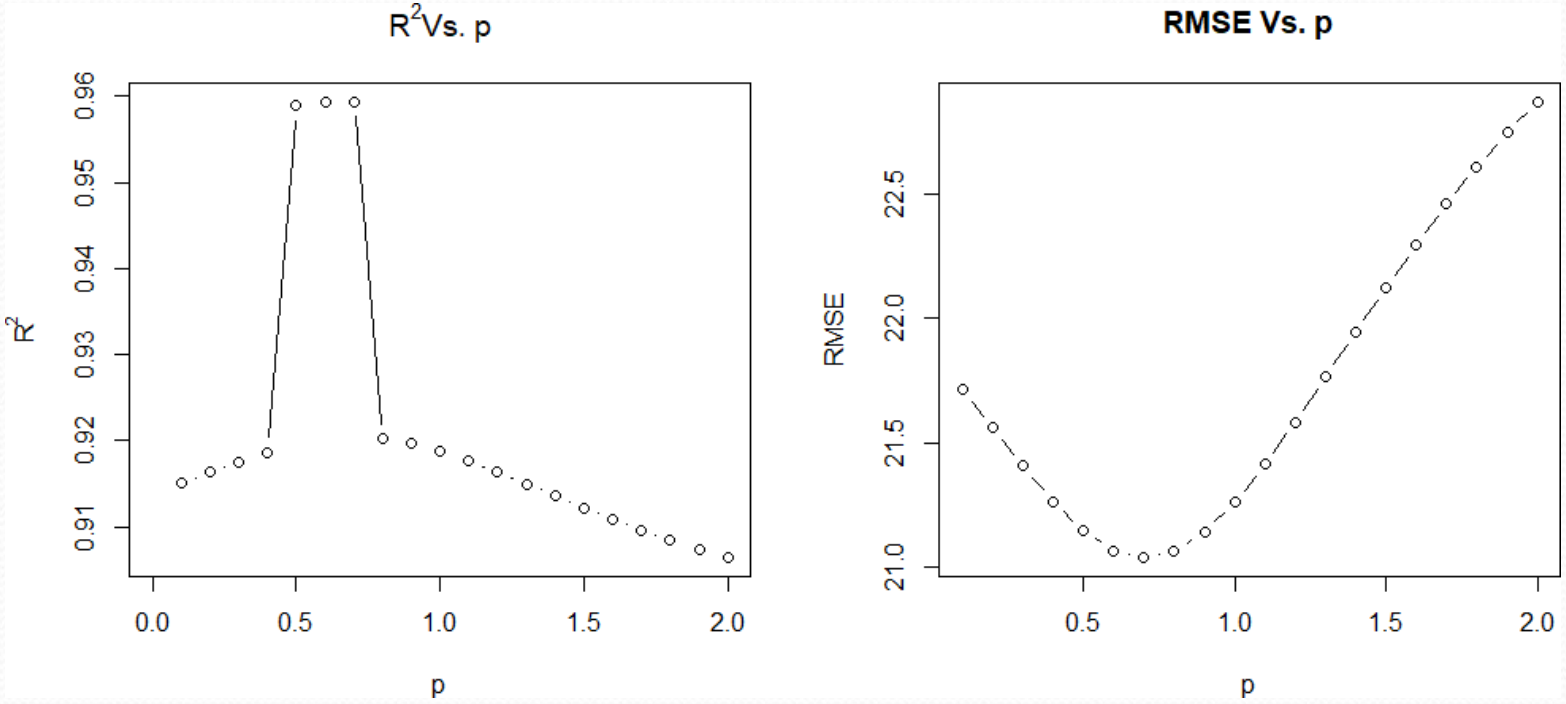


Figure: R^2 and RMSE vs. p in IDW

IDW Contour Plot-I

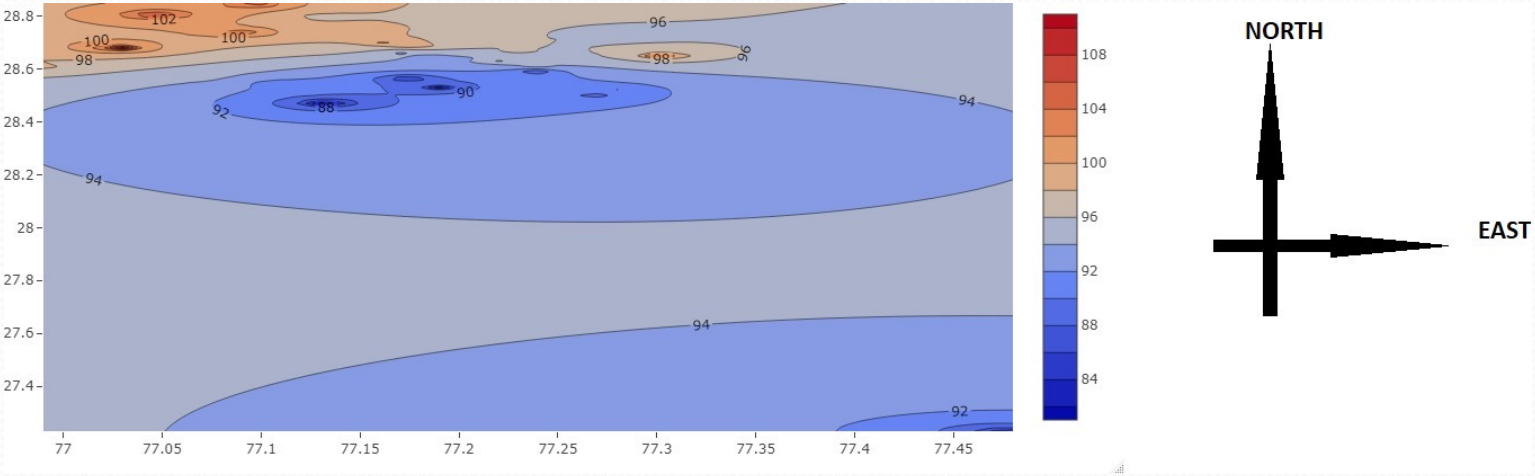


Figure: IDW for BL.

IDW Contour Plot-II

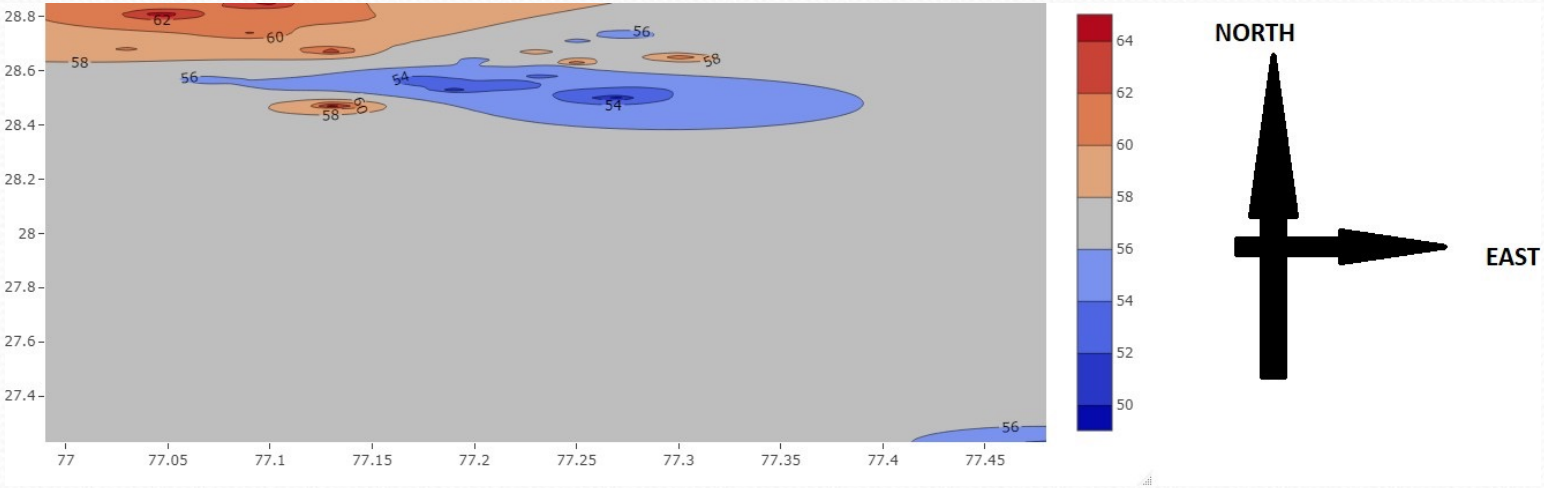


Figure: IDW for DL

Ordinary Kriging (OK)

- Suppose the process value in the location x_0 is defined as $Z^*(x_0)$ where

$$Z^*(x_0) = \sum_{i=1}^{n} w_i Z(x_i)$$

and w_i s are the weights.

- Here the constraint is $\sum_{i=1}^{n} w_i = 1$.

Minimizing Error Variance

$$\underset{}{\mathbf{C}} \cdot \mathbf{w} = \mathbf{D}$$

$$\underbrace{\begin{bmatrix} \tilde{C}_{11} & \cdots & \tilde{C}_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \tilde{C}_{n1} & \cdots & \tilde{C}_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}}_{(n+1)\times(n+1)} \cdot \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_n \\ \mu \end{bmatrix}}_{(n+1)\times 1} = \underbrace{\begin{bmatrix} \tilde{C}_{10} \\ \vdots \\ \tilde{C}_{n0} \\ 1 \end{bmatrix}}_{(n+1)\times 1}$$

$$\mathbf{w} = \mathbf{C}^{-1} \cdot \mathbf{D}$$

OK Contour Plot-I

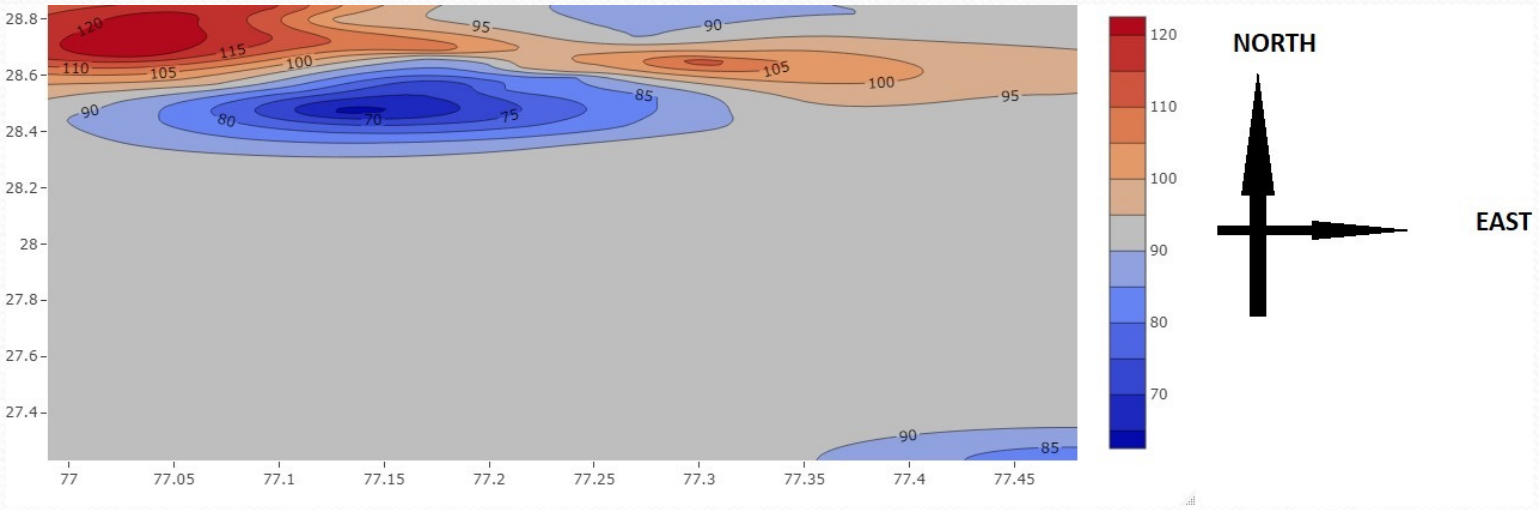


Figure: OK for BL.

OK Contour Plot-II

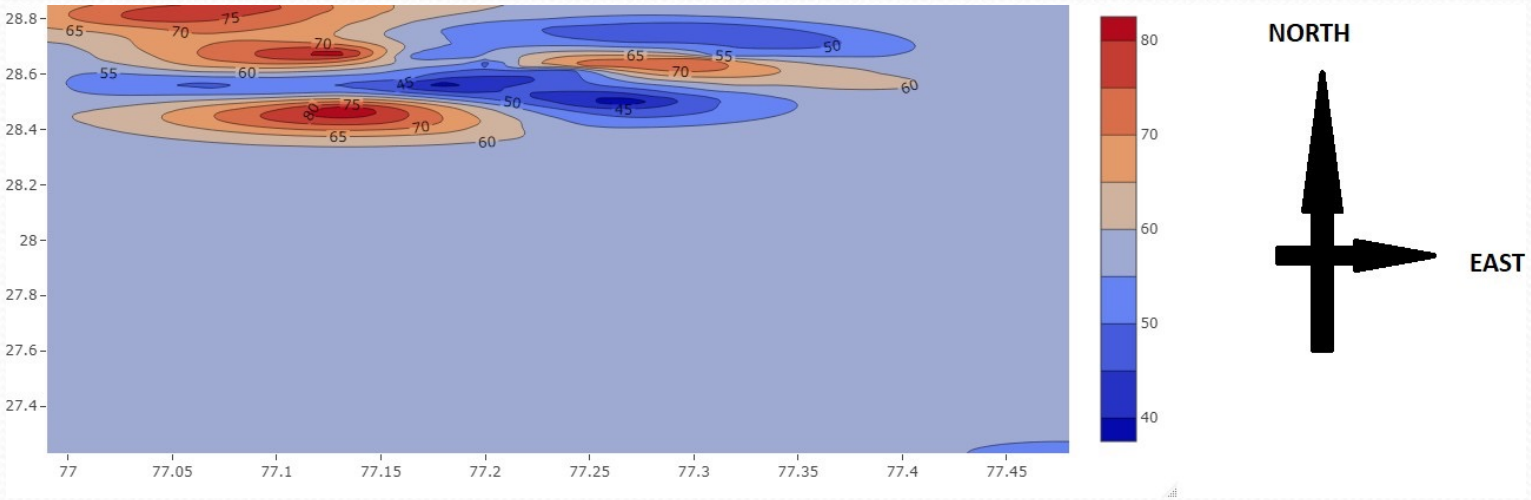


Figure: OK for DL.

Random Forest Regression Kriging (RFK)

$$\hat{Z}(\vec{x_0}) = \sum_{i=1}^{m} \hat{b}_i \cdot a_i(\vec{x_0}) + \sum_{i=1}^{n} w_i \cdot e(\vec{x_i})$$

(*i*) RF to fit the explanatory variables.

(*ii*) OK to fit the OOB errors with expectation 0.

RFK Contour Plot-I

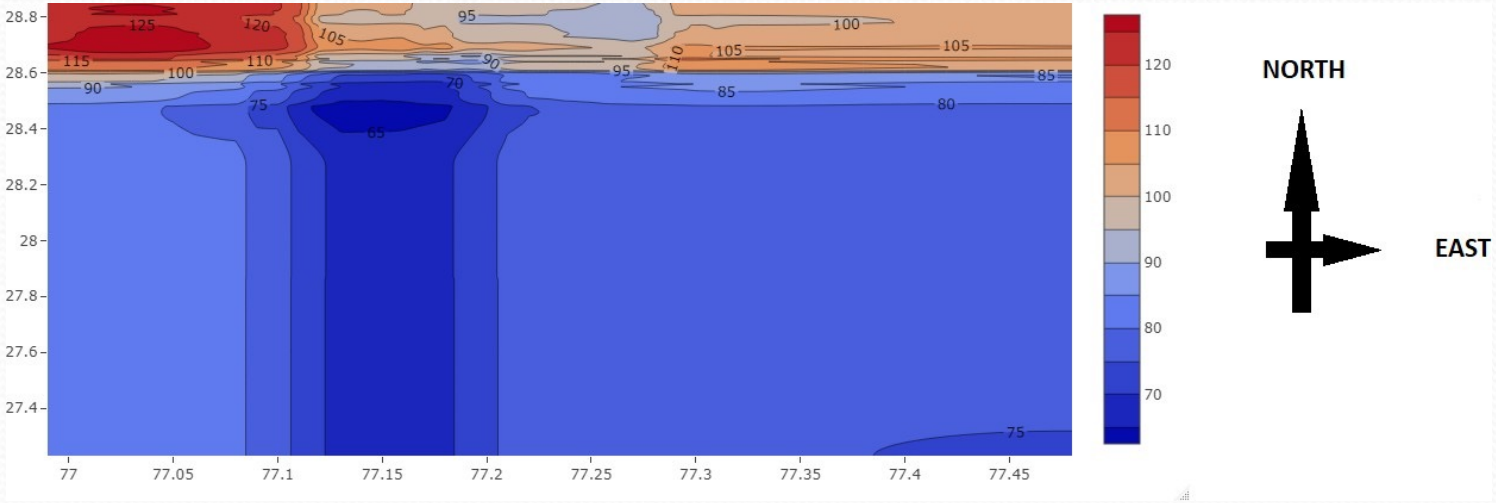


Figure: RFK for BL.

RFK Contour Plot-II

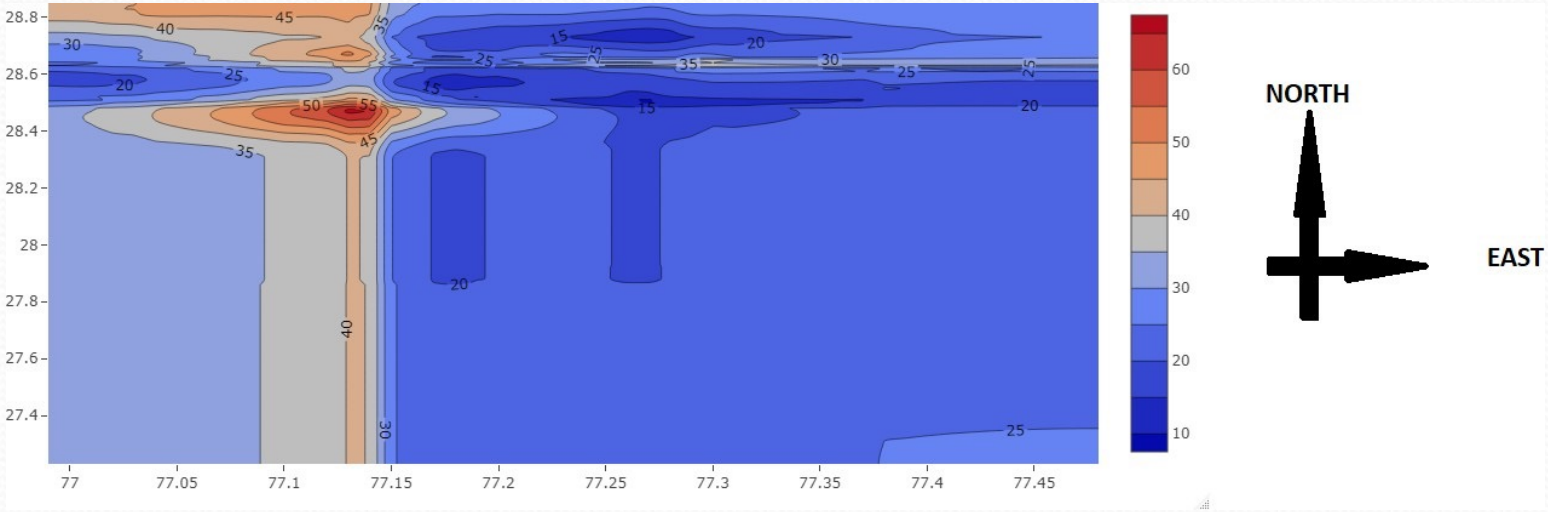


Figure: RFK for DL.

AQI Contour Plot-I

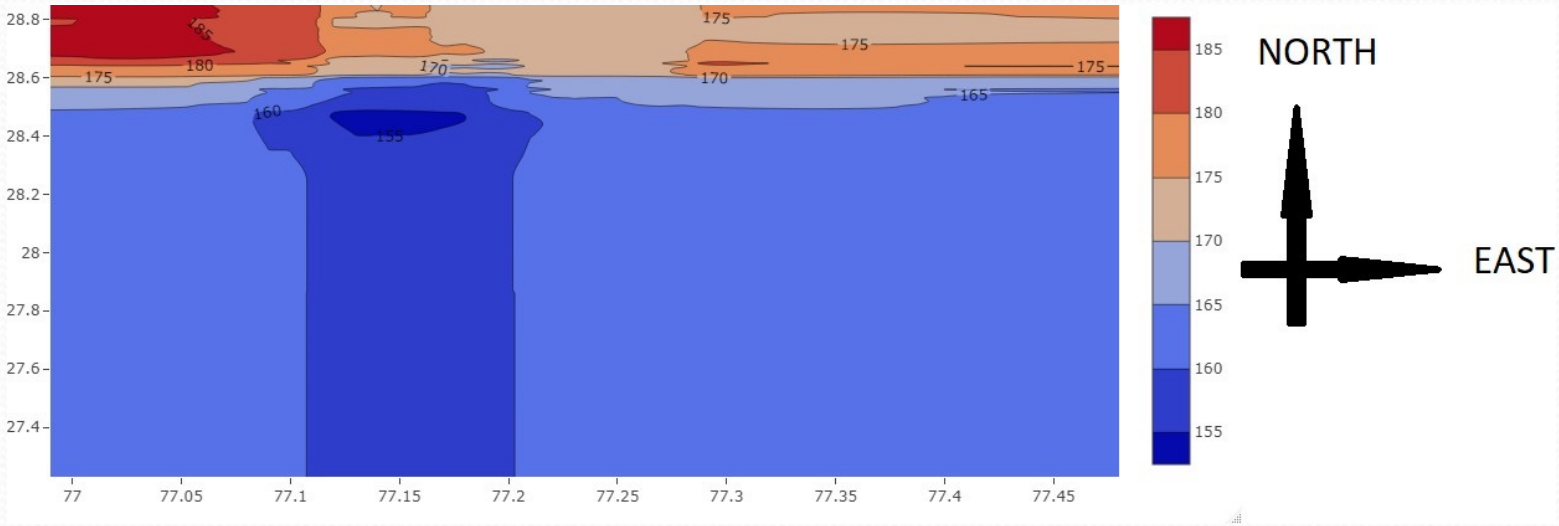


Figure: Contour of AQI for BL

AQI Contour Plot-II

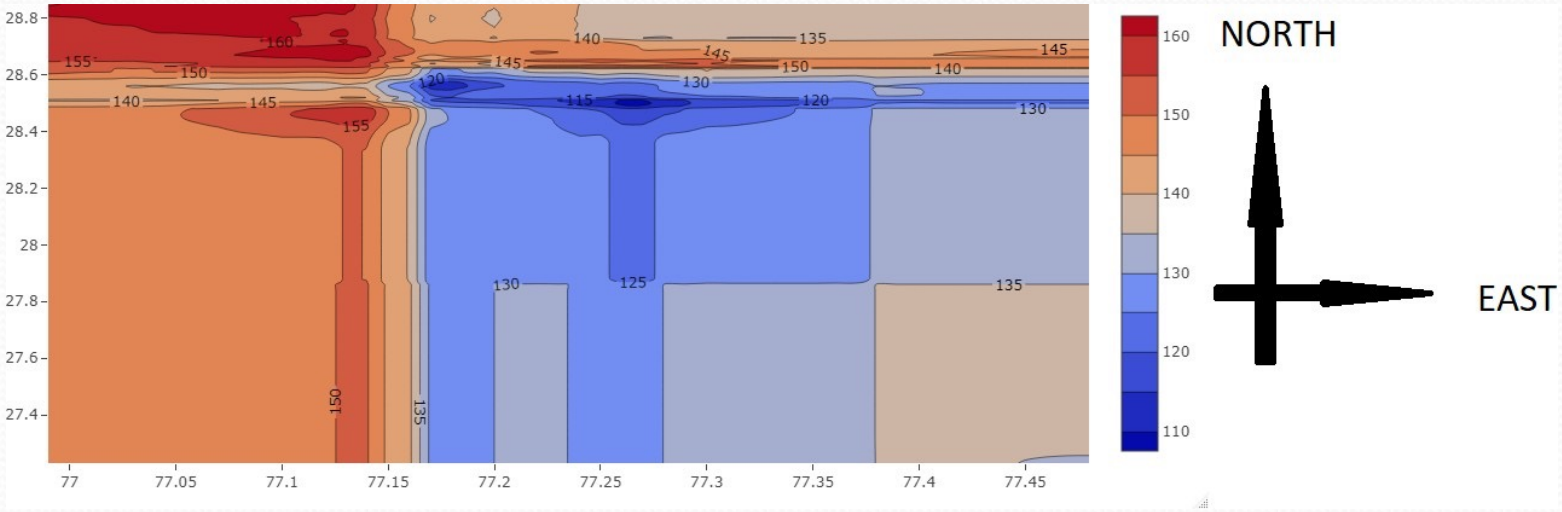


Figure: Contour of AQI for DL

Comparison between IDW, OK, and RFK corresponding to the spatially autocorrelated weeks in 2019 and 2020 on the basis of RMSE

Sl.No.	Week	Year	IDW	OK	RFK
1	12	2020	15.46	12.53	0.030
2	13	2019	16.73	15.93	0.071
3	13	2020	9.90	9.64	0.011
4	14	2019	18.14	16.32	0.135
5	14	2020	11.91	11.45	0.043
6	15	2019	16.05	14.52	0.069
7	16	2019	9.90	9.53	0.024
8	17	2019	15.87	14.78	0.009
9	18	2019	13.24	14.10	0.069
10	19	2019	28.13	25.38	0.409
11	20	2019	16.40	13.15	0.138
12	20	2020	12.72	11.78	0.442

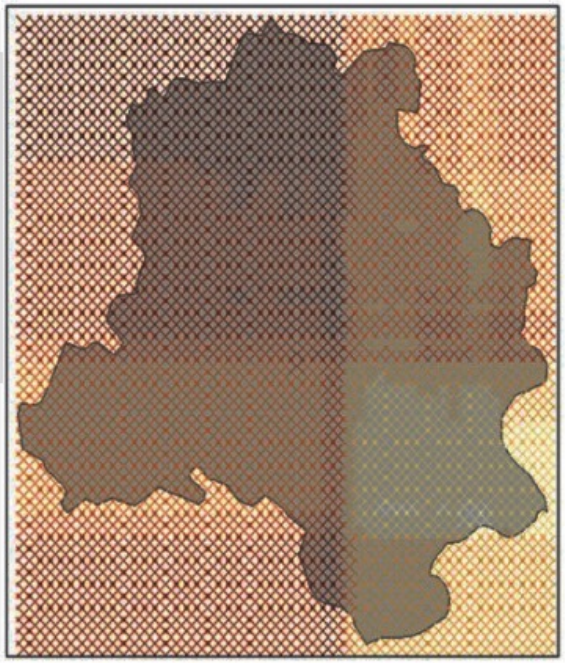
Spatial trend for weekly $PM_{2.5}$ emission of 20th week

A) Before lockdown (i.e. the year 2019) B) During lockdown (i.e. the year 2020)

$PM_{2.5}$

- x 70 to 80
- x 81 to 90
- x 91 to 100
- x 101 to 110
- x 111 to 120

Higher $PM_{2.5}$ emission



$PM_{2.5}$

- x 40 to 45
- x 46 to 50
- x 51 to 55
- x 56 to 60
- x 61 to 65
- x 66 to 70
- x 71 to 75

Higher $PM_{2.5}$ emission

Note: The green line indicates the weekly higher $PM_{2.5}$ emissions from weekly lower $PM_{2.5}$ emissions.

Percentage of average decrease of $PM_{2.5}$, PM_{10}, and AQI in diferent major-activity zones because of lockdown

(%)

Components	Transport	Residential	Institutional	Commercial
$PM_{2.5}$	–63.79	–61.31	–61.38	–62.13
PM_{10}	–62.19	–54.84	–49.83	–45.73
AQI	–11.11	–29.05	–22.71	–26.76

Spatial Copula

- The copula ($C_s(.,.)$) is first employed to explain the spatial dependency by Bárdossy (2006) in 2006, to measure the groundwater quality assuming that the spatial random function $Y(s)$ having marginal distribution invariant with spatial translation and the random field is isotropic.

$$C_s(h, q_1, q_2) = \mathbb{P}\Big[Y(s) \leq q_1; Y(s+h) \leq q_2\Big] = C(F(Y(s)), F(Y(s+h))) \quad (4.1)$$

- In the previous article, to solve the limitation of interpolation and neighborhood creation in 2008 Bárdossy and Li (2008) first introduced an interpolation algorithm with the help of a spatial copula and focus on the neighborhood of a spatial point.

Convex Combination of Copulas

- Gräler (2014) defined the generalized form of the spatial copula family instead of concentrating on the particular copula family like other researchers. Therefore, they consider a convex combination of bi-variate copulas.

$$C_h(u,v) = \begin{cases} C_h^1(u,v); & 0 \le h < h_1^* \\ (1-\lambda_2)\cdot C_h^1(u,v) + \lambda_2 \cdot C_h^1(u,v); & l_1 \le h < l_2 \\ . & . \\ . & . \\ . & . \\ (1-\lambda_k)\cdot C_h^{k-1}(u,v) + \lambda_k \cdot C_h^k(u,v); & l_{k-1} \le h < l_k \\ 1 & l_k \le h \end{cases} \tag{4.2}$$

In the Equation (4.2) the weights in convex combination is $\lambda_j = \frac{h-l_{j-1}}{l_j - l_{j-1}}$ where h denoting the spatial lag.

Spatial Copula

- In 2020 Sohrabian (2021) proposed a geo-statistical estimation through a convex combination of the archimedean copula.

$$C^*\left(\frac{p}{n}, \frac{q}{n}\right) = \frac{cardinality\{(y_{1i}, y_{2i}) | y_{1i} < y_{1(p)}; y_{2i} < y_{2(q)}\}}{n} \quad (4.6)$$

In the Equation (4.6) the p^{th} and q^{th} order statistics are respectively, $y_{1(p)}$ and $y_{2(q)}$.

$$C^{Archemedian}_{convex}(u, v) = p \cdot C^{Archemedian}(u, v) + (1 - p) \cdot C^{Archemedian}(u, v) \quad (4.7)$$

In this Equation (4.7) $p = p(|s_i - s_0|) = p(h_i) \in [0, 1]$ and $(u, v) \in [0, 1]^2$.

Spatial Copula

- In the year 2017 Alidoost, Stein, and Su (2018) employed this interpolation algorithm to explain an important real-life event. They have used this interpolation algorithm to predict air temperature at Qazvin Plain, Iran.
- They estimated suitable copula in each spatial bin based on the lowest AIC values, and those selected copulas are Student-t-Clayton and Student-t-Gumbel and created spatial neighborhood by C-Vine copula.
- Explaining spatial variability by gaussian variogram, they estimate kriging and Co-Kriging parameters and compared the results with spatial copula interpolation.

Spatial Copula

- In 2019 Alidoost et al. (2021) obtained bias-corrected air temperature values at the unobserved location along the whole spatial surface via copula-based quantile mapping.
- Their copula-based quantile mapping of spatial dependency mainly illustrates three types of dependencies, such as the dependency between air temperature and the corresponding elevation at a particular location, between air temperature in a specific area and its nearest neighborhood, and the combination of the first two.
- With the help of rank order transformation, they deduced the empirical probability to exhibit copula, and employing the inverse CDF transformation technique, they calculated the predicted bias-corrected values at the unobserved spatial location.

Spatial Copula

- In 2019 Brunner, Furrer, and Favre (2019) added one more aspect to define spatial dependence via Fisher copula.
- To validate the spatial dependency underlying this flood event, they measure Kendall's τ for each pair of stations. To choose the suitable copula family, they cross-checked five different copula families: Gaussian, Student-t, Gumbel, R-Vine, and Fisher copula.
- Fisher copula is advantageous because of modeling multiple gauge stations, non-vanishing upper tail dependence, and radial asymmetricity.
- They have fitted Gaussian, Student-t, and Gumbel copula with the help of maximum pseudo-likelihood estimation, R-vine copula by the automated model selection, and Fisher copula with two-step pseudo maximum likelihood estimation.

Limitations

- To estimate parameters, they use the Maximum Likelihood estimate, which does not provide a reasonable estimate in the presence of missing data.
- 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.
- They use conditional expectation for interpolation, but it is invalid for the extreme valued Probability Density Function (PDF).

Proposed Advancements

- We emphasize more on local stationarity of a spatial random field (SRF) more than the global stationarity of that SRF. Therefore we would like to apply a hierarchical spatial clustering algorithm.
- We choose the neighborhood points by C-Vine copula and derive the conditional probability distribution of the unobserved points with the help of a modified gaussian distance kernel.
- We estimate the parameter of marginal distribution using the EM algorithm and that of copula by the MCMC scheme.
- We derive the tail dependency of the covariates with the help of a copula, and to maximum pollutant concentration in a month, we employ blended extreme valued probability distribution.
- Later, we extend this spatial copula in different temporal replicates to the Spatio-Temporal copula interpolation with the help of the Spatio-Temporal influence ratio.

Spatial Random Field (SRF)

- The spatial domain is divided into k spatial clusters to get m number of spatial regions i.e. $\mathscr{L}$, the class of all possible sets of points in a spatial region.
- A conditional spatial random field is defined as:

$$Y : \mathscr{L} \rightarrow \mathscr{M} \subset \mathbf{R}$$

- This article introduces a spatial cluster-based C-vine copula and a modified Gaussian distance kernel to derive a novel spatial probability distribution.
- The proposed spatial interpolation approach is validated by considering Delhi air pollution data.

Estimating Marginal PDFs and Parameters

- Cullen-Frey graph of skewness and kurtosis determines the suitable family of PDF and Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), Log-likelihood value (LogLik), and Kolmogorov-Smirnov (KS) values determine Log-Normal (LN) for $PM_{2.5}$, and Von-Mises (VM) for wind direction (WD).
- The algorithms used to estimate the parameter for LN, VM, and Copula are respectively, EM, Uniformly Minimum Variance Unbiased Estimator (UMVUE), and MCMC.

UMVUE of VM Distributions

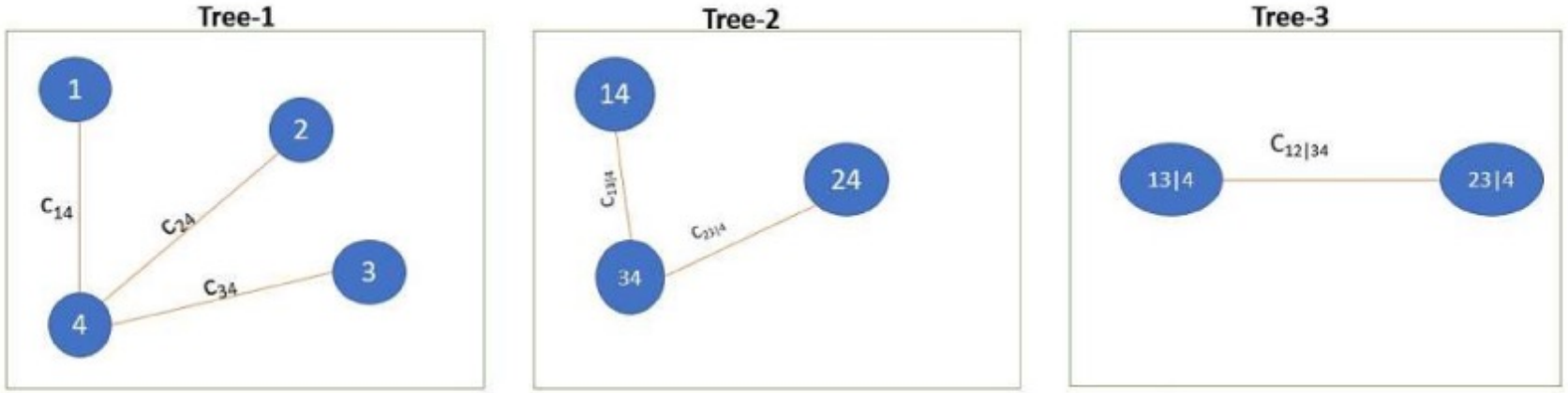
Theorem

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*

$$var(cos(x_i)) = \frac{1}{2} + \frac{I_2(k)\cdot cos(2\mu)}{2I_0(k)} - \left(\frac{I_1(k)\cdot cos(\mu)}{I_0(k)}\right)^2$$

$$var(sin(x_i)) = \frac{1}{2} - \frac{I_2(k)\cdot sin(2\mu)}{2I_0(k)} - \left(\frac{I_1(k)\cdot sin(\mu)}{I_0(k)}\right)^2$$

C-Vine Copula



Spatial Copula Interpolation

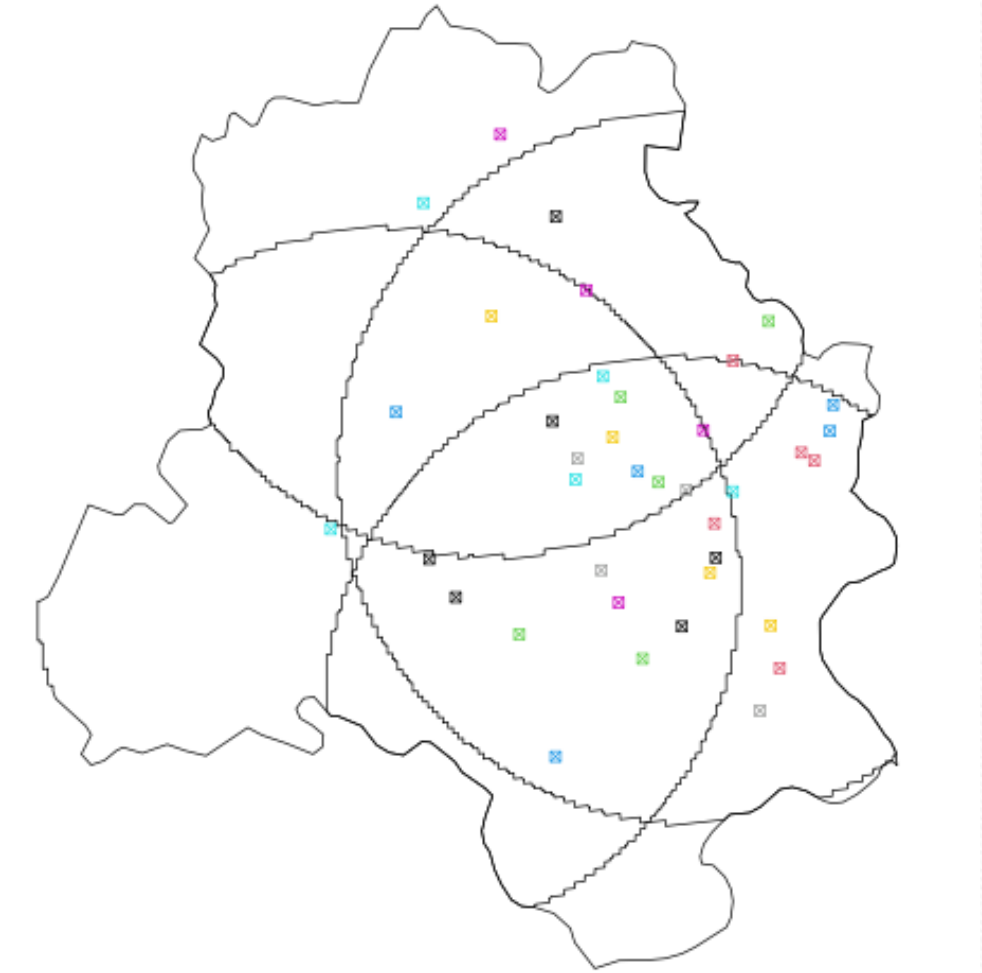
- **Step-1:** Employ the HSC algorithm along the whole spatial surface.
- **Step-2:** Devide disjoint spatial regions (SR).
- **Step-3:** Creation of Spatial Random Field (SRF) within each HSC.
- **Step-4:** Conditional copula-based probability distribution function (CCDF) and Conditional copula-based probability density function (CCPDF) for each unobserved point in an SR.
- **Step-5:** Spatial Interpolator derived from CCDF

Hierarchical Spatial Clustering (HSC)

$$
\begin{aligned}
N_i &= \{(ob_{i_1}, ob_{i_2}) : \\
HD(ob_{i_1}, ob_{i_2}) &\leq HD_{cut} \;\&\; \rho(\| ob_{i_1} - ob_{i_2} \|) \geq r_{cut} \;\&\; i_1 \neq i_2\} \\
&\bigcup \{\left(ob_{i_j}, y_{ij}\right) : \\
HD(ob_{i_j}, y_{ij}) &\leq HD_{cut} \;\&\; \rho(\| ob_{i_j} - y_{ij} \|) \geq r_{cut}\}
\end{aligned}
$$

- Here, $N_1, N_2, ..., N_k$ 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, $j = 1, 2, ..., n_i$; $i = 1, 2, 3, ..., k$. n_i be the number of the unobserved points in i^{th} HSC and c_i, the centre of i^{th} cluster.

HSC of Study Area



Spatial Region (SR)

- Let **v** be the presence vector of all unobserved points $y_{ij} = (lon, lat)$, where *lon* and *lat* stand for the longitude and latitude of an unobserved point respectively.
- $$f : \mathscr{L} \rightarrow \{0,1\}^{k \times 1} \Rightarrow f(y_{ij}) = \mathbf{v} \qquad (4.8)$$

  In the Equation (4.8) **v** is a binary vector of length k, where

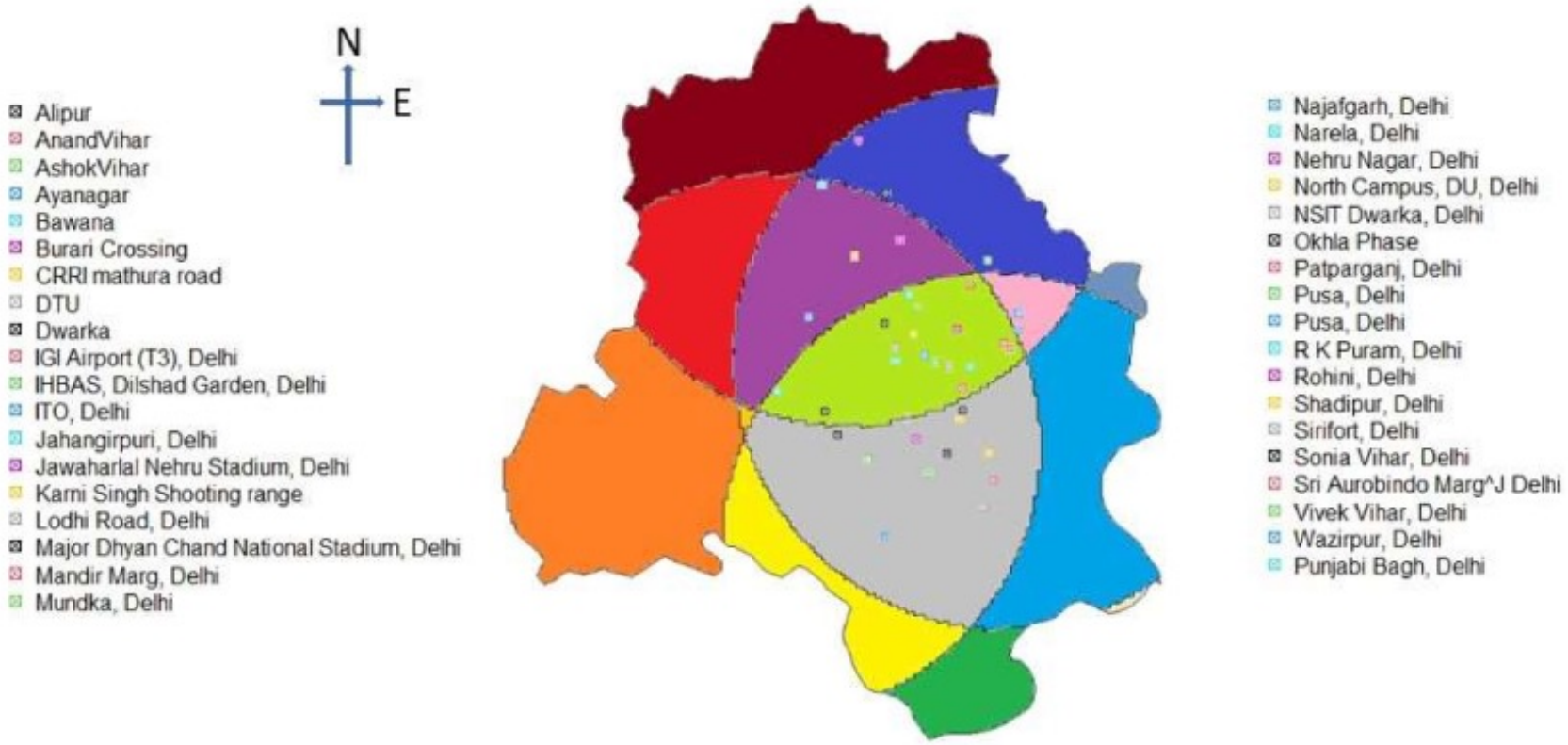
$$\mathbf{v} = [v_1, v_2, ..., v_k]$$

$$v_l = \begin{cases} 1 & \text{if } (lon, lat) \in N_l \\ 0 & \text{otherwise} \end{cases} \qquad (4.9)$$

$$where\ l = 1, 2, 3, ..., k.$$

- Let $R_1, R_2, ..., R_m$ be the m distinct SR where $m \leq 2^k - 1$.

SR of Study Area



Conditional Spatial Copula

- The inclusion probability of the latitude (ω_i) and longitude (β_i) of an observed location in R_i is $H(x_1, x_2) = C(F_\omega(X_{1\omega}), F_\beta(X_{2\beta}))$. These two RVs, latitude (X_1) and longitude (X_2) follow Uniform$[a_1, b_1]$ and Uniform$[a_2, b_2]$ respectively.
- The CDF of that spatial random process (SRP), a composite function of two RVs i.e., $Y(.,.) \equiv \{Y(\omega, \beta) : \omega \in R_i, \beta \in R_i\}$ is $H(x_1, x_2, y) = C(F_\omega(X_{1\omega}), F_\beta(X_{2\beta}), F(Y))$.
- The conditional PDF (CPDF) measures the CDF of SRP, defined in the following two equations:

$$f(y_1 \mid x_1, x_2) = \frac{\frac{\partial^3 H(x_1, x_2, y)}{\partial x_1 \partial x_2, \partial y}}{\frac{\partial^2 H(x_1, x_2)}{\partial x_1 \partial x_2}} \tag{4.10}$$

CCDF of Unobserved Random Points

- Applying the conditional copula (from Equation (4.10)) we establish the Conditional Copula-based Probability Distribution Function (CCDF) for j^{th} un-observed point in the first SR is $F_{un^j_{R_1}}(y)$.

$$F_{un^j_{R_1}}(y) = \sum_{i \in R_1} \alpha_{ij} \cdot P[Y(x_{1,ob_i}, x_{2,ob_i}) \leq y | X_1 = x_{1,ob_i}, X_2 = x_{2,ob_i}] \tag{4.11}$$

- $$\alpha_{ij} = \frac{d_{ij} \cdot \rho(||ob_i - un^j||)}{\sum_{i \in R_1} d_{ij} \cdot \rho(||ob_i - un^j||)}$$

$$d_{ij} = \epsilon + \exp - \left(\left[\int_0^1 |f_{ob_i}(y|x_1, x_2) - f_{un^j}(y|x_1, x_2)|^p \right]^{1/p} dy \right) \cdot$$

$$\exp - \left(\sin^{-1} \left[\sqrt{A + \cos(x_{1,ob_i}) \cdot \cos(x_{2,un^j}) \cdot B} \right] \right)$$

Spatial Interpolation

- The mathematical formulation of spatial interpolator derived from the Equation (4.11) is described in the following:

$$f_{un^j_{R_1}}(y) = \sum_{i \in R_1} \alpha_{ij} \cdot f_{ob_i}(y \mid x_1, x_2)$$

$$\Rightarrow argmax_y f_{un^j_{R_1}}(y) = \sum_{i \in R_1} \alpha_{ij} \cdot argmax_y f_{ob_i}(y \mid x_1, x_2) \quad (4.12)$$

Algorithm

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 (SBVC)

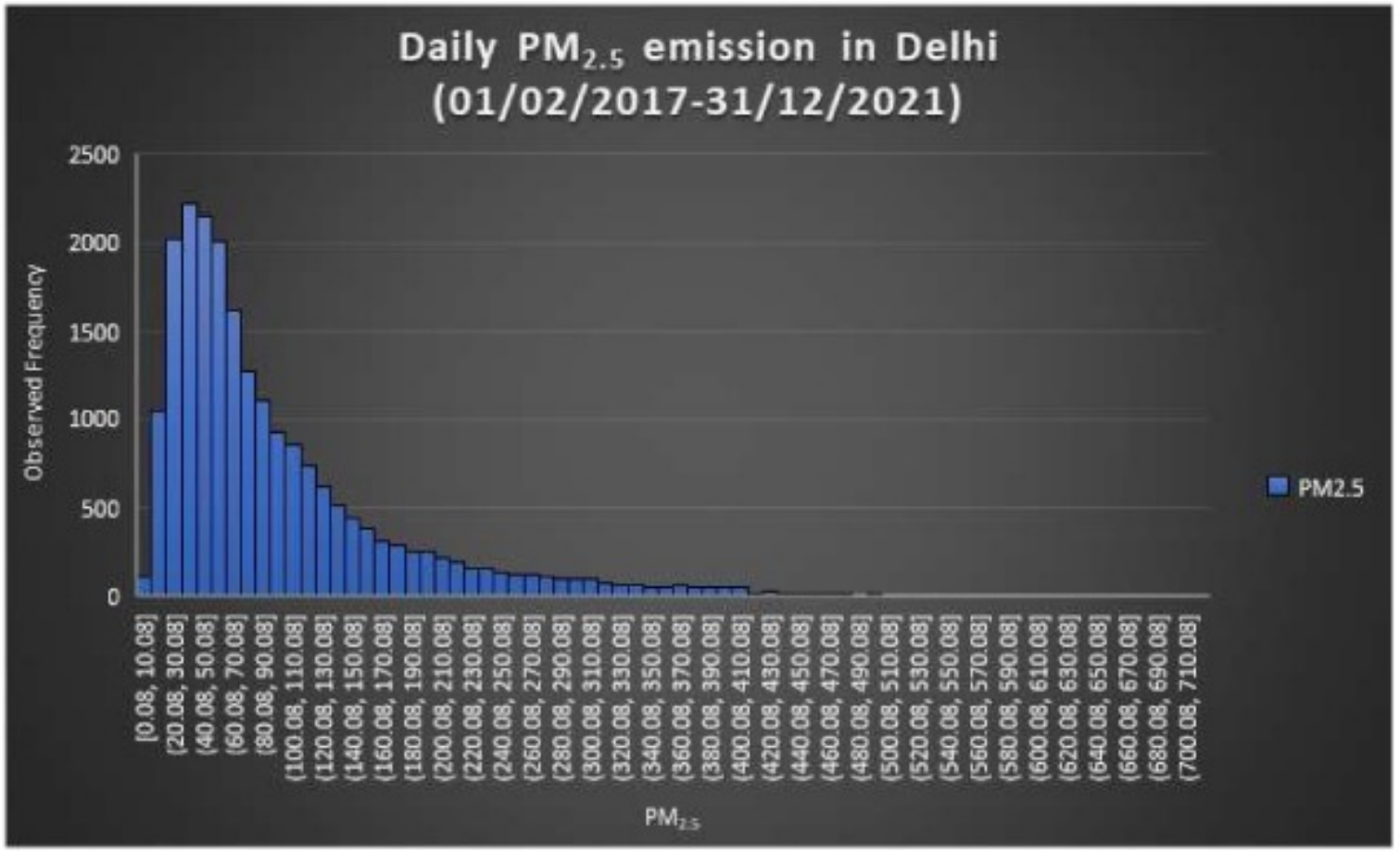
$$\pi(\theta) = \frac{f_{X_1,X_2|X_3,X_4}(x_1, x_2, x_3, x_4|\theta) \cdot p(\theta)}{\int f_{X_1,X_2|X_3,X_4}(x_1, x_2, x_3, x_4|\theta) \cdot p(\theta) d\theta}$$

$$E(\hat{\theta}|X_1, X_2, X_3, X_4) = \int \theta \cdot \pi(\theta) d\theta \qquad (4.13)$$

$$E(X_1, X_2|X_3, X_4) = \int \int x_1 \cdot x_2 \cdot f_{X_1,X_2|X_3,X_4}(x_1, x_2, x_3, x_4|\hat{\theta}) dx_1 dx_2 \qquad (4.14)$$

$$F(y_1, y_2|x_1, x_2; \theta) = P[Y_1(x_1, x_2) \leq y_1, Y_2(x_1, x_2) \leq y_2|X_1 = x_1, X_2 = x_2 \qquad (4.15)$$

Histogram


Daily PM2.5 emission in Delhi
(01/02/2017-31/12/2021)
Observed Frequency
2500
2000
1500
1000
500
0
[0.08, 10.08]
(20.08, 30.08]
(40.08, 50.08]
(60.08, 70.08]
(80.08, 90.08]
(100.08, 110.08]
(120.08, 130.08]
(140.08, 150.08]
(160.08, 170.08]
(180.08, 190.08]
(200.08, 210.08]
(220.08, 230.08]
(240.08, 250.08]
(260.08, 270.08]
(280.08, 290.08]
(300.08, 310.08]
(320.08, 330.08]
(340.08, 350.08]
(360.08, 370.08]
(380.08, 390.08]
(400.08, 410.08]
(420.08, 430.08]
(440.08, 450.08]
(460.08, 470.08]
(480.08, 490.08]
(500.08, 510.08]
(520.08, 530.08]
(540.08, 550.08]
(560.08, 570.08]
(580.08, 590.08]
(600.08, 610.08]
(620.08, 630.08]
(640.08, 650.08]
(660.08, 670.08]
(680.08, 690.08]
(700.08, 710.08]
PM2.5
PM2.5
PM2.5

Fitting of Marginal Parametric PDF on Empirical PDF

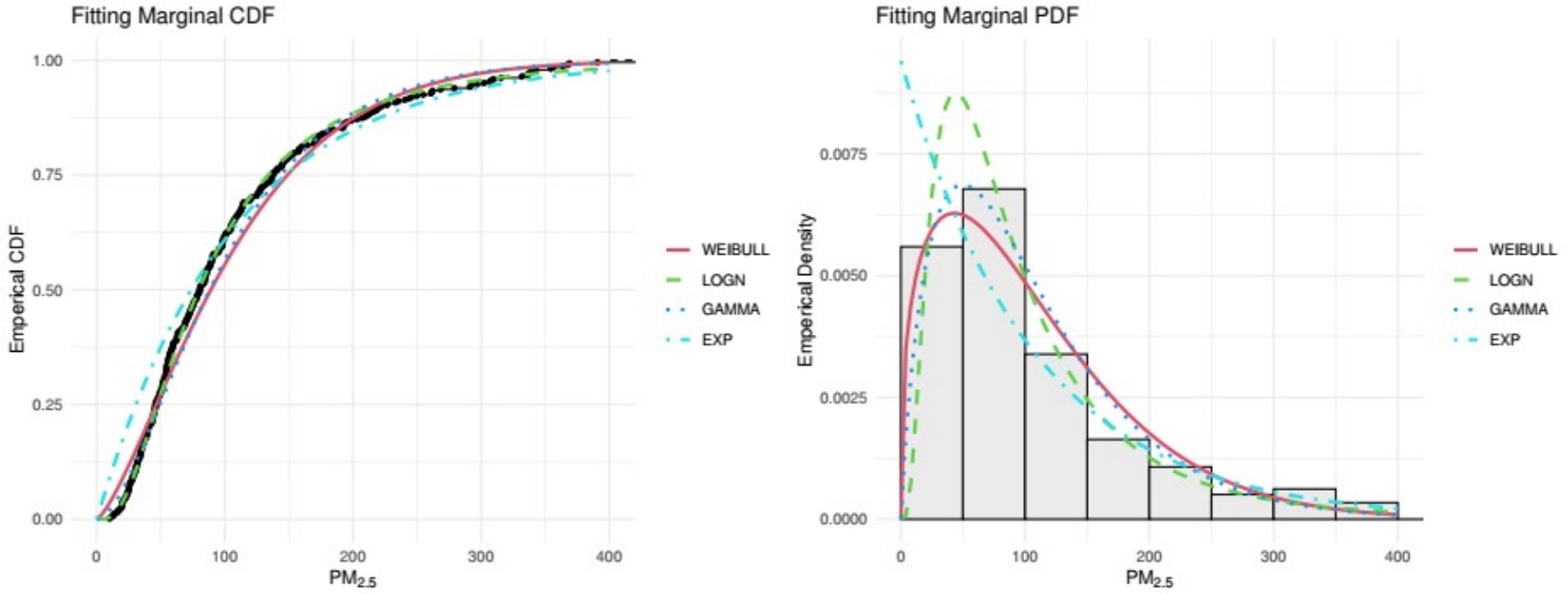
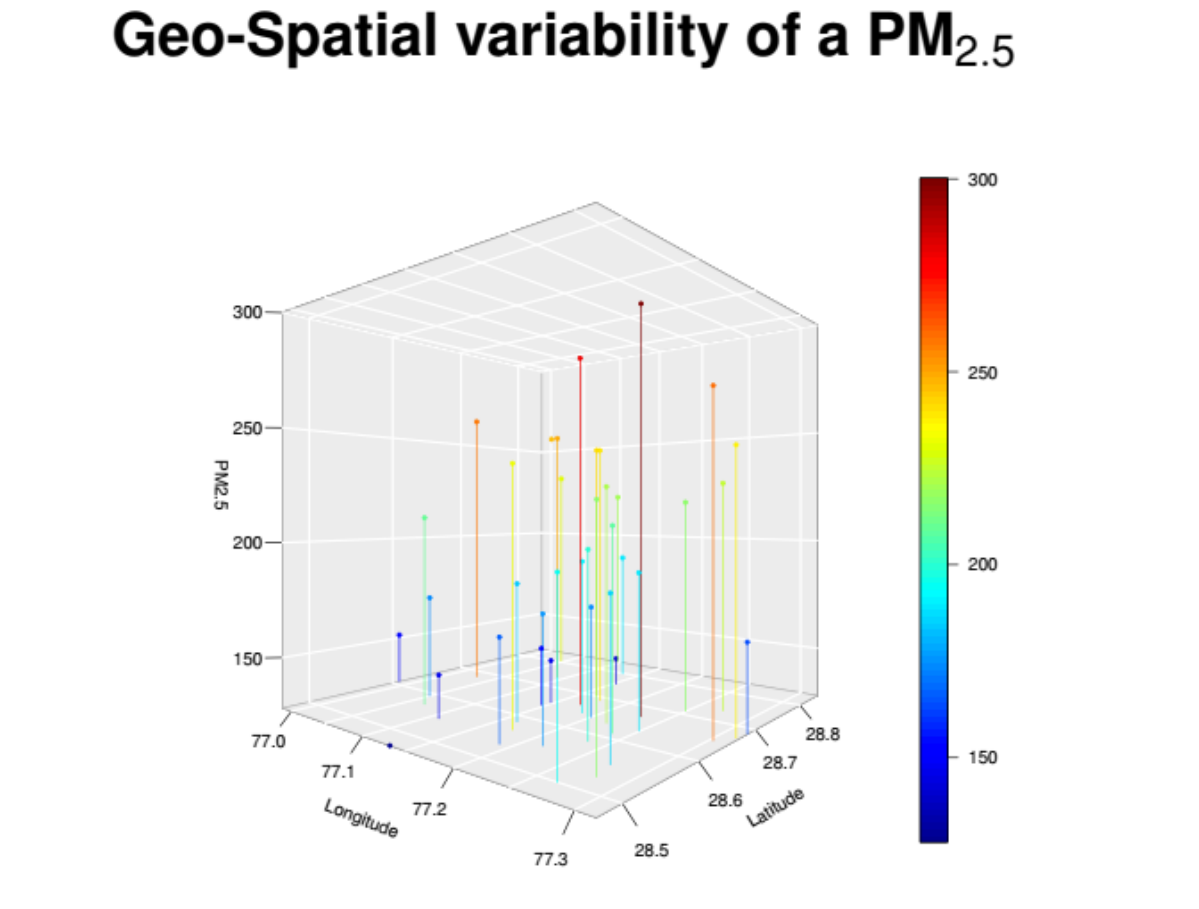


Figure: The empirical marginal CDF is fitted by the marginal positively skewed parametric CDF and the fitting of marginal PDF.

Spatial Variation



Choice of PDFs

Table: 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

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

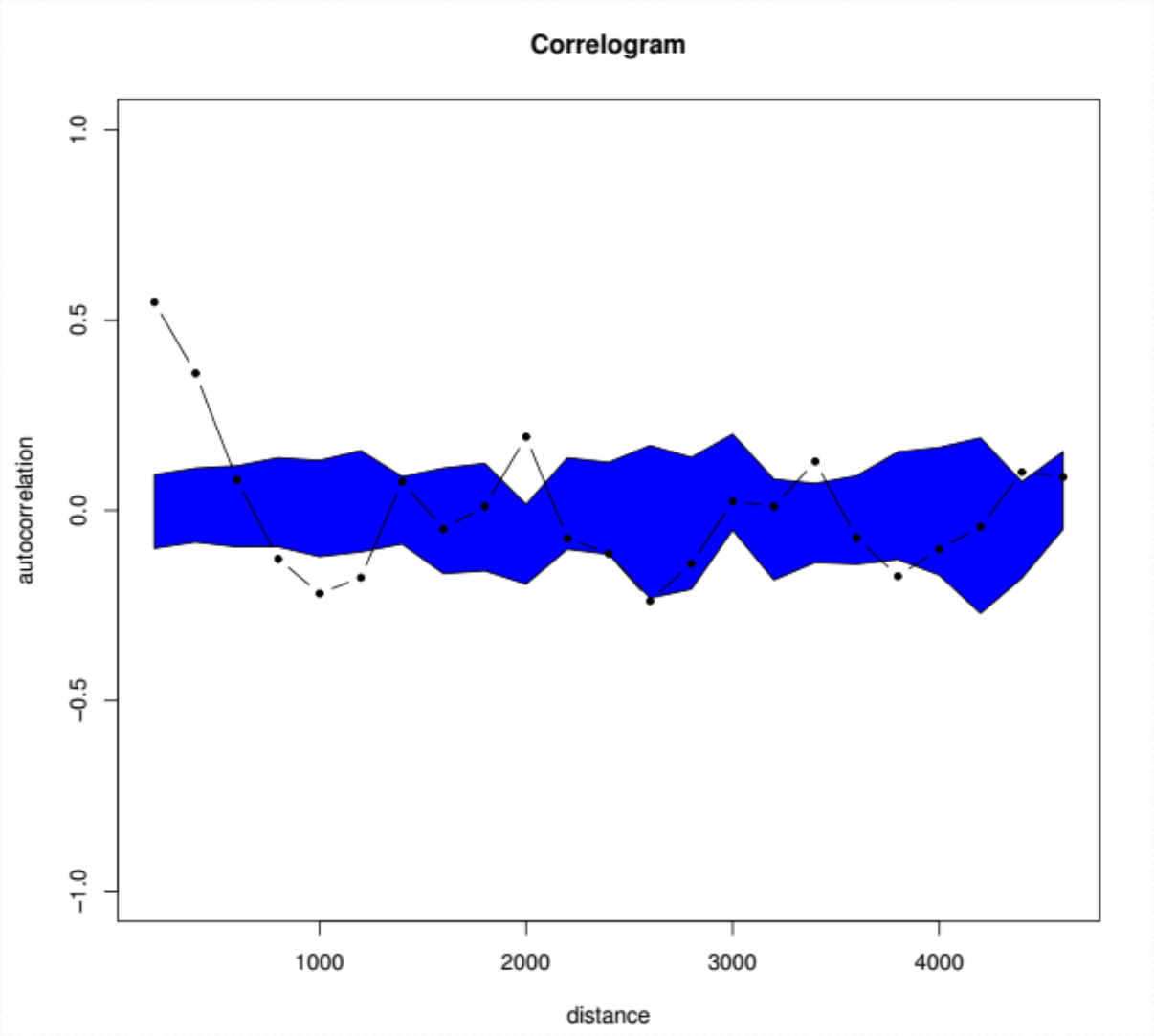
PDF	Shape	Scale	LogLik
LN	4.3764856	0.7701984	-1959.331124
VM	3.583	1.908	-32.41559

Two-way ANOVA

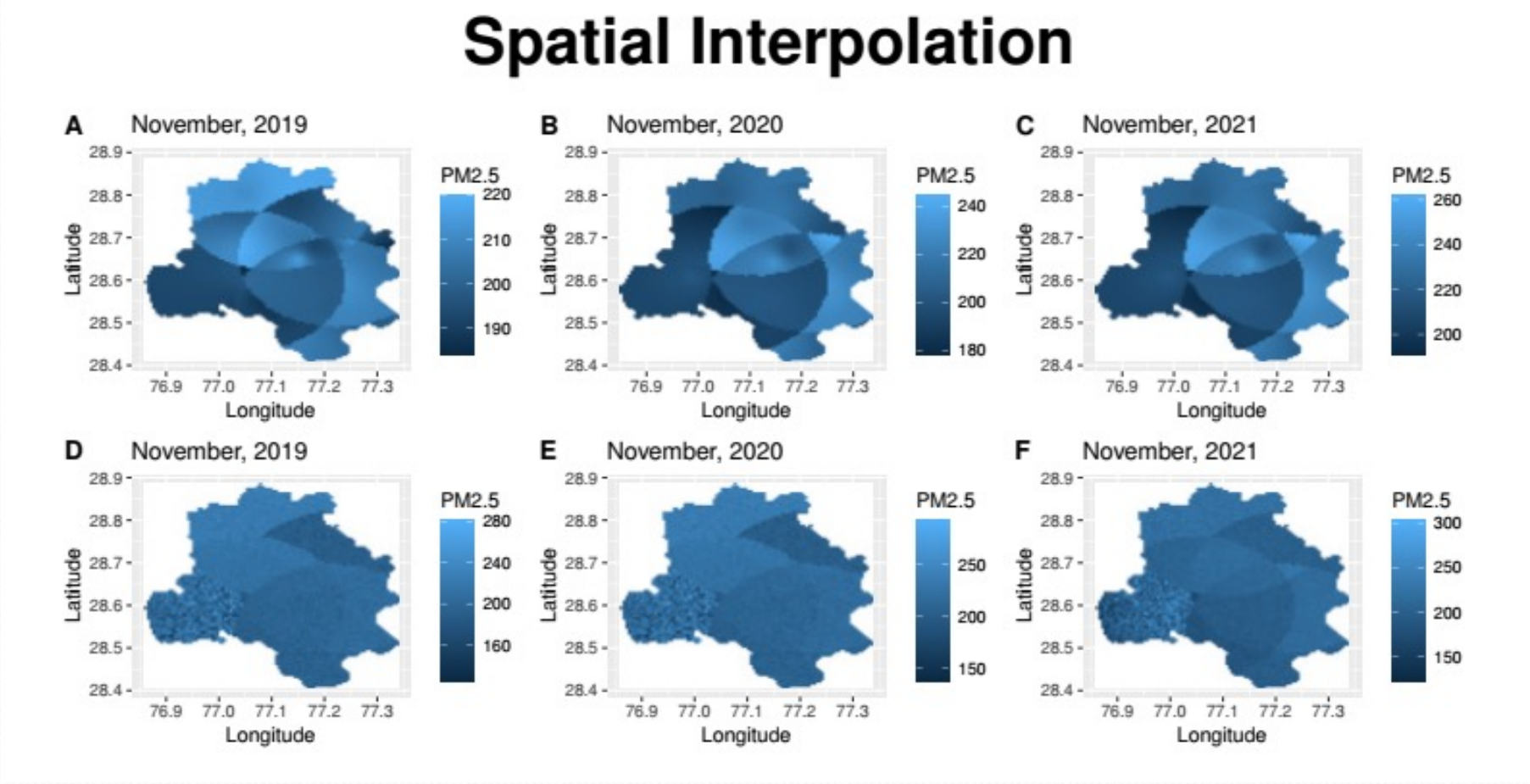
Table: Two-way ANOVA to explain the dependence of $PM_{2.5}$ emission on WD and SC

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		

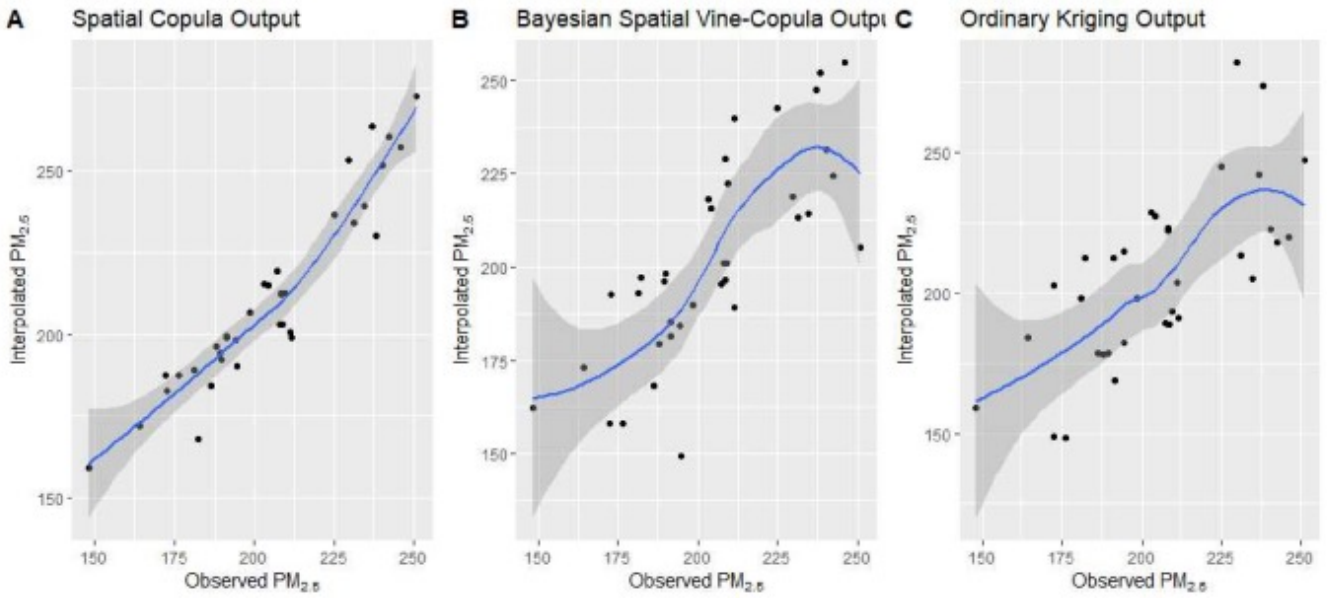
Spatial Autocorrelation



Spatial Interpolation using SC and SBVC



Relationship between Observed and Predicted Values



Conclusions

- 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.

- Compare to other geostatistical models, the proposed SC and SBVC are very effective and provide nearly accurate results.

- The SC model produces better results for spatially skewed spatial random fields and provides a mathematical argument for selecting essential covariates.

- 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.

References

- Bárdossy A (2006) Copula-based geostatistical models for groundwater quality parameters. Water Resour Res 42(11).
- Gräler B (2014) Modelling skewed spatial random fields through the spatial vine copula. Spat Stat 10:87–102.
- Sohrabian B (2021) Geostatistical prediction through convex combination of archimedean copulas. Spat Stat 41(100):488.
- 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.

Thank You!!!

Questions ??