Pol. J. Environ. Stud. Vol. 34, No. 6 (2025), 6575-6588
DOI: *10.15244/pjoes/192978* ONLINE PUBLICATION DATE: 2025-01-07

Original Research

Copula-Based Spatial Model and Identification of Extremal Regions of Soil Heavy Metal Concentrations in a Mine Consolidation Area

Xiaohui Chen[1], Qiong Wang[1]*, Qinfei Yu[2], Kai Li[3]

[1]BGRIMM Technology Group, Beijing 100160, China
[2]Chinese Academy of Natural Resources Economics, Beijing 101149, China
[3]State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China

Received: 26 April 2024
Accepted: 4 September 2024

Abstract

The specification of environmental extrema is a persisting problem, especially in soil with spatial heterogeneity owing to anthropogenic activities. Using a geographic detector, a Bayesian spatial model, and a copula-based spatial model, methods of identification of extremal regions in a mine area were compared. The results are as follows. (1) All of the heavy metals in anthropogenic soil, including As, Cd, Cr, Hg, and Ni, had a weak random spatial heterogeneity, but Cd and As exhibited strong stratification spatial heterogeneity ($q = 0.21^{**}$ and 0.11^{*}, respectively). (2) The Cr, Hg, and Ni predictions are very similar for both models (the improvements in the mean absolute percentage error (MAPE) and R^2 are 5.88% at most and 3.29%, respectively). The copula-based spatial model outperformed the Gaussian spatial model in the predictions of Cd (MAPE: 12.12%; R^2: 16.67%) and As (MAPE: 4.16%; R^2: 7.89%). (3) Based on the comparison with the Gaussian spatial model using a Bayesian process, the identification of the extremal regions using the copula-based spatial model had a higher accuracy for the extreme samples. In general, the prediction obtained using the copula-based model revealed the probability of exceeding a certain threshold at a location. Moreover, it uses the copulas fitting of the samples' spatial heterogeneity obtained through maximum likelihood estimation, rather than variogram fitting, resulting in the random spatial heterogeneity summing to a nugget, which preserves more information about the samples. Thus, we conclude that the copula-based spatial model can be used to predict the heavy metal concentrations in soil with weak random spatial heterogeneity but strong stratification spatial heterogeneity.

Keywords: anthropogenic influences, spatial heterogeneity, spatial Copula, extremal region

*e-mail: wangqiong426@126.com

Introduction

Obtaining the geostatistics of soil samples can help government departments and agencies to accomplish soil surveys, but is always affected by the spatial heterogeneity caused by anthropogenic influences such as mining, soil covering, soil improvement, etc. [1–4]. Previous studies on anthropogenic soil related to mining have often focused on issues of environmental significance [5, 6]. China has recognized the importance of protecting the soil environment and has implemented the Action Plan for Prevention and Control of Soil Pollution [7]. In addition, the risk screening and intervention values for soil contamination have been confirmed through soil environmental quality risk control standards, which also guide the monitoring of land consolidation areas [8].

Both the environmental factors and physicochemical properties of soil exhibit spatial heterogeneity and autocorrelation, and a semi-variable function has always been used as the spatial model before obtaining a locally linear unbiased estimation [9, 10]. However, the modeling data, which may be either raw data or Box-Cox conversion data, need to have or approximate a Gaussian distribution. Moreover, owing to the random effects or extremal distribution, the spatial heterogeneity and unbiased estimation obtained from semi-variance do not perform well enough. For instance, researchers have proposed multifractal analysis to manage the geochemical anomalies at the boundaries between different geologic zones [11, 12]. In addition, Wang et al. developed the sandwich model for stratified mapping in scenarios in which spatial heterogeneity relies more on the spatial stratification of some influencing factors [13, 14]. For semi-variance based on stratification heterogeneity, it is difficult to fit a function due to the cross-distribution of the stratifications and the limited number of samples. Compared with the Gaussian unbiased estimation, which reduces the spatial correlation of the extremes, the copula-based spatial model may be sufficient and provide more characterization of the spatial variations in the heavy metal concentrations of the soil [15–18]. Bárdossy (2006) first proposed the use of the copula-based spatial model as a replacement for variograms and covariance functions in order to describe the spatial heterogeneity. The structure of the copula-based spatial model is composed of the dependent function and the marginal distributions; and the dependent function, i.e., the copula, is the basis, which is based on Sklar's theory [19, 20]. Some experts have used the Copula-based model to study the spatial dependence of extreme rainfall weather in the Mediterranean Sea, and there are also studies that use the spatial copula model combined with Bayesian approximation to predict extreme temperature [16, 17]. In addition, the copula has many families that can serve as alternatives for spatial model construction [21, 22], which can achieve visualization through the web app Copulatheque [23].

Soil pollution should be investigated before and after land consolidation in mining areas to identify and control the key environmental areas [24–27], especially when a certain emergency threshold may have been exceeded [28]. Although a spatial Bayesian process unifies the estimation and prediction and takes into account the uncertainties of the model parameters, which means it can acquire the uncertain information in the high-risk areas regarding heavy metals, it generally assumes that the model is a Gaussian random field [29, 30]. In addition, it is possible to use indicator kriging (IK), which does not require the data to obey the normal distribution. As a non-parametric geostatistical method, IK is robust regarding outliers [9, 31, 32]. However, one threshold means one model with one map, which is extremely unintelligent. Thus, a copula-based spatial model can provide the probability distribution function of each prediction site and can identify the extremal regions, i.e., where the heavy metal concentrations of the soil exceed particular thresholds with a 50% probability [33]. It can be concluded that the copula-based spatial model combines the optimal prediction results of the Bayesian spatial model and the non-normality assumption of IK.

In summary, we analyzed and predicted the spatial heterogeneity of the As, Cd, Cr, Hg, and Ni contents and pH value of soil in a mine consolidation area, and we identified the extremal regions of Cd pollution of the soil. First, using variogram analysis and a geographic detector, the spatial heterogeneities of these environmental factors were determined. Second, the copula-based and Gaussian spatial models were compared, and the soil environmental factors were predicted throughout the area. Finally, the extremal regions were identified using the copula-based spatial model and the Gaussian spatial model through a Bayesian process.

Materials and Methods

Study Area

The study area is located in the southern Sichuan Basin (27°41′–28°20′ N, 105°34′–106°20′ E), which has a long history of sulfur ore mining. It shares the climatic characteristics of both the Sichuan Basin and the Guizhou Plateau and is located in the subtropical humid climate zone, with mean temperature, precipitation, and relative humidity values of 16.8–18.6°C, 1000 mm, and 83%, respectively. The elevation of the study area ranges from 501 m in the southwest to 948 m in the northeast, and the main soil types are yellow soil and brunisolic soil. Owing to long-term mining and concentration, most of the study area contains harmful substances, such as sulfur waste, and the original ecological landscape has been seriously damaged. Since 2013, consolidation of abandoned industrial and mining areas has been conducted, including soil replacement in mining areas (SRM), soil replacement in concentration areas (SRC), and natural vegetation.

Sampling and Analyses

The soil environment survey for heavy metals included 175 observations throughout the entire 4 km^2 study area.

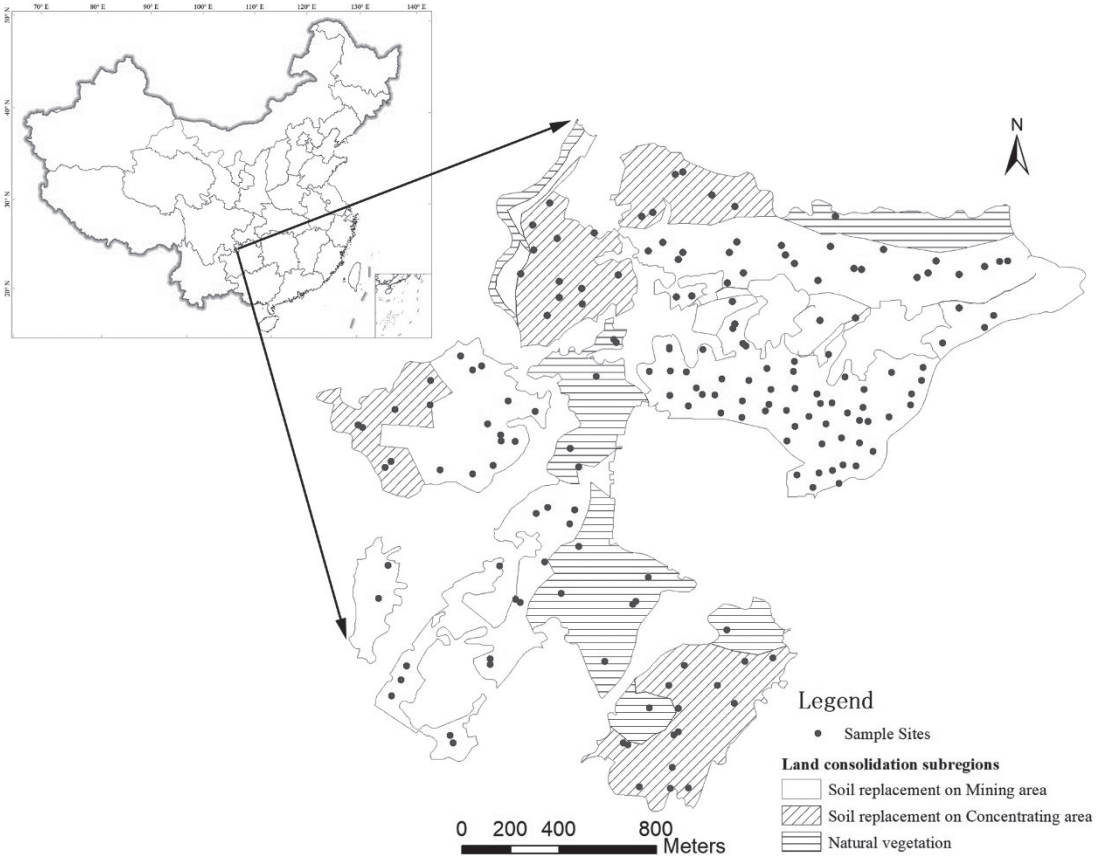

Fig. 1. Map of the study area showing the land consolidation subregions and sampling sites (n = 175).

The distribution of all of the samples is shown in Fig. 1. The soil samples were collected from the topsoil layer (0–20 cm), and they were combined into a composite sample for a 20×20 m area. The soil samples were air-dried, sieved to 2 mm, and digested using HNO_3 and H_2O_2 via Method 3050B (USEPA, 1996). The Cd, Ni, and Cr concentrations of the digestion solution were measured via inductively coupled plasma optical emission spectrometry (ICP-OES, Optima 5300DV, PerkinElmer Instrument Co., Ltd., USA), and the As and Hg concentrations were analyzed using an atomic fluorescence spectrometer (AFS-2202, Haiguang Instrument Co., Ltd., China). The pH values of the soils were determined using the potentiometric method. Standard reference soils GSS-1 and GSS-2 were obtained from the Center of National Standard Reference Material of China and were used for quality assurance and quality control.

Stratification of Spatial Heterogeneity Using the Geographic Detector

The geographic detector is a geostatistical method that can be used to explore spatial heterogeneity without the Gaussian distribution assumption. Its core idea is that if the environmental factor is dependent on one influencing factor, they may have similar spatial distributions [34, 35]. The goal of this study was to identify the environmental factors with significant spatial differences due to anthropogenic influences, which always exhibit intense randomness. In addition, the geographic detector provides the driver of environmental pollution to a certain extent. Here, we used the q value of the geographic detector:

$$q = 1 - \frac{\sum_S^L N_S \sigma_S^2}{N\sigma^2} \quad (1)$$

where s = 1, …, L is the stratification of the influencing factors; N_s and N are the number of units in each stratification and in the entire district, respectively; and σ_s^2 and σ^2 are the variances in each stratification and in the entire district, respectively.

Geostatistical Models

Gaussian Spatial Models and Bayesian Process

The Bayesian spatial model is a Gaussian spatial process, which fits the multivariate normal distribution, and the basic function, which is obtained from the semi-variogram is

$$Z(\boldsymbol{x}) = \mu(\boldsymbol{x}) + \omega(\boldsymbol{x}) + \varepsilon(\boldsymbol{x}) \quad (2)$$

where $\mu(x) = m$ is constant; the spatial effect $\omega(x) = f(\sigma^2, \varphi)$ is a function of the partial sill (σ^2) and the range (φ); and non-spatial effect $\varepsilon(x) = f(\tau^2)$ is a function of the nugget (τ^2).

The procedures of the Bayesian spatial model include establishing the model and the prior distribution, the posterior distribution based on Markov Monte Carlo (MCMC) sampling, and the spatial prediction based on the posterior distribution [36, 37].

First, based on the Bayesian hierarchical modeling approach, its data model, procedural model, and parametric model are

$$\begin{cases} Z|\theta, \omega \sim MND(\mu + \omega, \tau^2) \\ \omega|\sigma^2, \varphi \sim MND\left(0, \sigma^2 H(\varphi)\right) \\ \theta = (m, \sigma^2, \varphi, \tau^2) \end{cases} \quad (3)$$

where *MND* is the multivariate Gaussian distribution function; $H(\varphi)$ is the correlation matrix, which is represented by an exponential function, i.e., $H(\varphi) = \exp(-\varphi||x_i - x_j||)$, and $||x_i - x_j||$ is the distance between two spatial points; m, σ^2, φ and τ^2 are random variables, which are independent with each other; m is constant; and the three other variables have exponential prior distributions.

Then, the posterior distribution sampling is conducted using the Markov chain Monte Carlo (MCMC) method, with the convergence of the Geweke z value. The mathematical formula is

$$J(\theta|Z) = \frac{f(Z|\theta) \cdot J_0(\theta)}{\int f(Z|\theta) J_0(\theta) d\theta} \quad (4)$$

where $J_0(\theta)$ is the prior joint probability distribution of the four parameters m, σ^2, φ and τ^2, and $f(Z|\theta)$ is the likelihood of the maximum likelihood estimation.

Finally, the spatial prediction is realized using the following equation:

$$f(Z_i|Z) = \int f(Z_i|Z, \theta) J(\theta|Z) d\theta \quad (5)$$

In this study, the R package spBayes and the sp and geoR packages were used to obtain the Bayesian spatial prediction of the soil environmental factors [38]. The Metropolis algorithm was used in the MCMC, with 10000 iterations and Geweke z value convergence checking.

Copula-Based Geostatistical Models

The copula-based spatial model describes the spatial stochastic field based on Sklar's theory [28, 39, 40]. The basic formula is

$$\mathrm{F}(z_1, \cdots, z_n) = \mathrm{P}(Z_1 \le z_1, \cdots, Z_n \le z_n) = \mathrm{C}_{\theta,\lambda}\{F_\eta(z_1), \cdots, F_\eta(z_n)\} \quad (6)$$

where $C_{\theta,\lambda}$ is the n-dimensional copula, with a correlation structure related to θ and λ; and F_η is the marginal distribution function related to η.

The sort order of the marginal functions does not affect the results because the copula is symmetric, which means that the sort order of the observations does not affect the structure of the spatial model. Similar to the traditional method, two points far apart are known to have mutual independence, while there is a strong dependency between points close to each other. In this study, the Gaussian spatial copula with non-Gaussian margins was used for the soil environmental factors, and the maximum likelihood function estimation through $\Theta = (\theta, \lambda, \eta)$ was as follows:

$$L(\boldsymbol{\Theta}; D) = \mathrm{c}_{\boldsymbol{\theta},\boldsymbol{\lambda}}(F_{\boldsymbol{\eta}}(z_1), \cdots, F_{\boldsymbol{\eta}}(z_n)) \quad (7)$$

where $c_{\theta,\lambda}$ is the probability density function of the copula; and $D = \{z_1, \ldots, z_n\}$ are the observations.

Based on the observations and parameter estimation, the probability density distribution at location needs to be predicted as follows:

$$p(Z(x_i)|\boldsymbol{\Theta}, \mathrm{D}) = \mathrm{c}_{\boldsymbol{\theta},\boldsymbol{\lambda}}(F_{\boldsymbol{\eta}}(Z(x_i))|D) \cdot f_{\boldsymbol{\eta}}(Z(x_i)) \quad (8)$$

where $c_{\theta,\lambda}$ is the probability density function of the conditional copula. The expected value and variance are calculated respectively as follows:

$$\hat{Z}(x_i) = \int_0^1 F_{\boldsymbol{\eta}}^{-1}(\mu) \mathrm{c}_{\boldsymbol{\theta},\boldsymbol{\lambda}}(\mu|D) d\mu \quad (9)$$

$$\hat{\sigma}^2(x_i) = \int_0^1 (F_{\boldsymbol{\eta}}^{-1}(\mu) - \hat{Z}(x_i))^2 \mathrm{c}_{\boldsymbol{\theta},\boldsymbol{\lambda}}(\mu|D) d\mu \quad (10)$$

It should be noted that the direct calculation of the multivariant copula is miscellaneous and has poor accuracy, so the multivariant probability density function was decomposed into a series of copula pairs and marginal functions using the vine-copula method. In this study, the C-Vine copula model was used [41, 42]. In this study, the spcopula and VineCopula R packages were mainly used to create the copula-based spatial model [15, 40, 43].

Validation of the Spatial Prediction Effect

In this study, first, the validation of the spatial prediction effect of the expected values for both the copula-based and Gaussian spatial models was conducted, and then, the comparison of the extremum characteristics was conducted [44, 45]. The performances of these models were compared via leave-one-out cross-validation, which was utilized to calculate the mean absolute percentage error (MAPE) and the goodness fit (R^2). For each observation site x_i, the distribution of Z_i was conditional upon all of the observed data, except when $z(x_i)$ was calculated

and the predicted probability that $z < z(x_i)$ was extracted. The cross-validation MAPE and R^2 were computed as follows:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{Z}_{-i}(x_i)-z(x_i)}{z(x_i)}\right| \quad (11)$$

$$R^2 = 1 - \frac{\sum(\hat{Z}_{-i}(x_i)-z(x_i))^2}{\sum(\bar{z}(x_i)-z(x_i))^2} \quad (12)$$

where $\hat{Z}_{-i}(x_i)$ is the prediction at the location x_i calculated using all of the observed values, except $z(x_i)$; $\bar{z}(x_i)$ is the mean of $z(x_i)$. The closer MAPE is to zero, the better it is; and, an R^2 value closer to 1 is better.

The extremum analysis of the copula-based and Bayesian spatial models was based on the distribution of the probability of the predicted locations in order to obtain the extremal regions [18]. First, the observations exceeding the specific threshold were identified. Then, the predicted regions exceeding the threshold with a 50% or 75% probability were identified, and the coverage of the observations was checked.

Results and Discussion

Descriptive Statistics and Variogram Analysis

The descriptive statistics of the heavy metal concentrations and pH values of the topsoil in the mining area after consolidation are presented in Table 1. The As, Cd, Cr, Hg, and Ni concentrations and pH values of the soil throughout the study area were 4.19–32.92 mg·kg^{-1}, 0.09–6.29mg·kg^{-1},95.26–556.53mg·kg^{-1},0.03–0.47mg·kg^{-1}, 21.19–166.89 mg·kg^{-1}, and 2.78–8.52 mg·kg^{-1}, respectively. Based on the soil screening values (SVs), the Cd concentrations of the soil seriously exceeded the regulation standard, and the mean concentrations of the other heavy metals were all below the SVs. According to the coefficients of variation (CVs) for the heavy metals, there was considerable variation within the range of 31.17% to 83.72%, and the CVs of the pH had the lowest value. After standardization using the SVs, Fig. 2 shows whether the concentrations of the heavy metals in the three consolidation types exceeded the regulatory limit. It shows that the standardized Cd value is much greater for the natural vegetation than for the SRM and SRC, while the values for the other heavy metals are similar for the different consolidation types.

The variogram analysis of the heavy metal concentrations and pH values of the soil based on their spatial distributions is presented in Fig. 3. Fig. 3 shows that the variograms for As, Cd, and Hg hold for a large distance (at least 600 m) based on the spatial autocorrelation. Cr has a range of 385.72 m, while those of Ni and pH are 60.26 m and 76.41 m, respectively. Several studies have suggested that the nugget/sill ratio n/(n+p) represents the ratio of the spatial heterogeneity, which obtains the spatial heterogeneity from the randomness of the variable, while the 1−n/(n+p) value reflects the structural spatial heterogeneity [31, 46, 47]. Thus, this range can be considered to be the spatial heterogeneity scale, and the larger the nugget/sill ratio is, the stronger the random spatial heterogeneity is. A ratio of greater than 75% indicates strong random spatial heterogeneity; values of 25–75% indicate moderate random spatial heterogeneity; and values of < 25% indicate weak random spatial heterogeneity. Overall, in the study area, the nugget/sill ratios of As, Cd, Cr, and Hg were 16.00%, 22.00%, 0%, and 20.69%, respectively, and in the different types of consolidation land, the ratios also indicate weak random spatial heterogeneity. Moreover, Ni and pH also exhibited weak random spatial heterogeneity.

Based on the first law of geography, Cr, Ni, and pH exhibited very weak random spatial heterogeneity, which was calculated from the variograms. Weak random spatial heterogeneity is equivalent to strong autocorrelation. The emergence of this situation can be attributed to the soil replacement in most regions of the study area. However, the spatial heterogeneity of As, Cd, and Hg are somewhat different, including a larger range and stronger random spatial heterogeneity. Thus, we conclude that several influencing factors continue to affect the spatial distributions of the As, Cd, and Hg concentrations.

Table 1. Descriptive statistics of heavy metal concentrations and pH values of the soil samples.

	Minimum (mg·kg^{-1})	Maximum (mg·kg^{-1})	Mean (mg·kg^{-1})	SD	CV (%)	SV (mg·kg^{-1})
As	4.19	32.92	15.89	6.53	41.10	30
Cd	0.09	6.29	1.00	0.84	83.72	0.30
Cr	95.26	556.53	184.91	79.45	42.97	200
Hg	0.03	0.47	0.22	0.10	43.91	2.40
Ni	21.19	166.89	72.13	22.48	31.17	100
pH	2.78	8.52	6.56	1.20	18.33	7

Note: SD denotes standard deviation; CV denotes the coefficient of variation; SV denotes screening value for heavy metal, and the pH's standardized values are calculated based on 7.

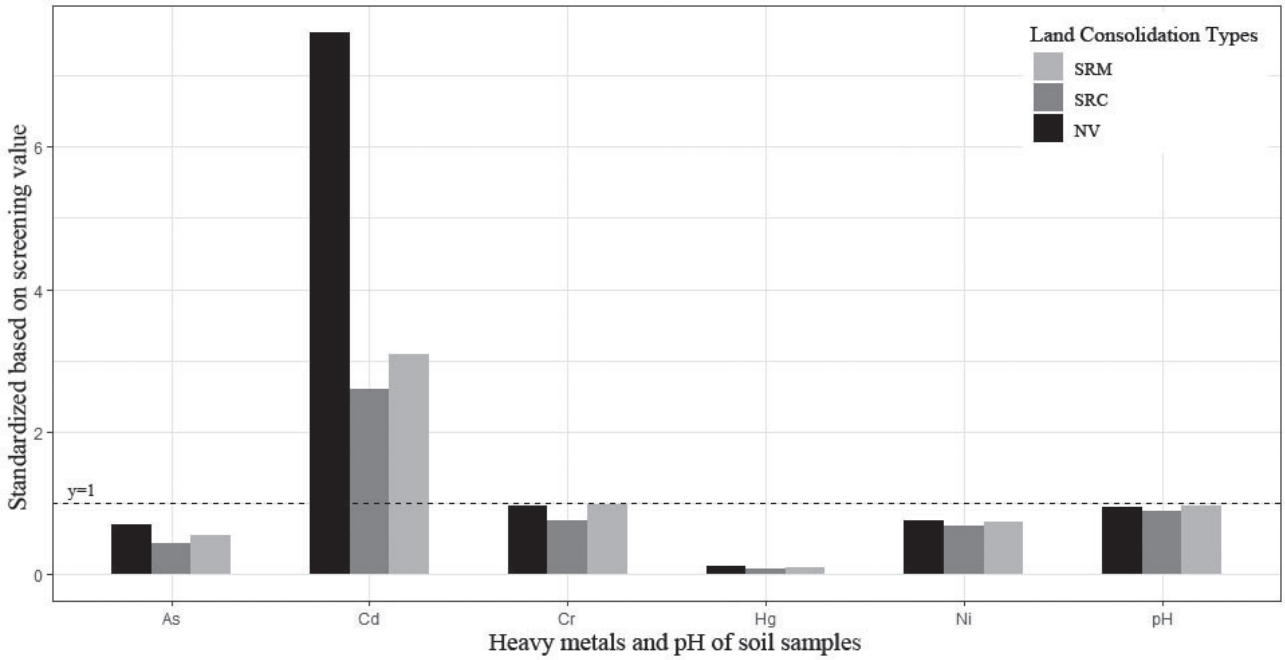

Fig. 2. Standardized values for heavy metals and pH based on screening values for different land consolidation types.

Validation of the Performances of the Spatial Prediction Models

Similar to the above variogram analysis, for the copula-based spatial model, a correlation function related to the spatial distance also needs to be established. In this study, the rank correlation for each bin was calculated, which was obtained from the point pairs in each step, to generate a function through polynomial regression (Fig. 4). Similarly, Ni and pH also had shorter ranges regarding their spatial heterogeneity, and the strengths of the correlations for As, Cd, Cr, and Hg gradually decreased as the distance increased, with good fitting conditions. However, their fitting functions could not be used to build an indicator similar to the nugget/sill ratio, so we explored the stratification of the spatial heterogeneity based on the different land consolidation types using a geographic detector. The geographic detector analysis generated q values (Table 2), and the larger the q value is, the stronger the stratified spatial heterogeneity is. It was found that the q value of Cd was the largest and it was extremely significant, which means the spatial distribution of Cd was highly dependent on the consolidation type. In addition, the q value of As was significant, while those of the others were not significant. Thus, the spatial distribution of Cd had a strong stratified spatial heterogeneity, and As had a moderate stratified spatial heterogeneity. However, based on the variogram analysis, the distributions of all of the heavy metals exhibited weak random spatial heterogeneity. Based on previous studies [34, 35], we conclude that the stratified spatial heterogeneity is not affected by the random spatial heterogeneity. Moreover, the former relies on spatial stratification, while the latter relies on samples, which means that they are calculated using different geostatistical scales but are complementary in terms of exploring spatial heterogeneity.

The copula-based spatial model constructs the different copulas based on the marginal distribution of the observations in each step. Cd is discussed as an example (Fig. 5). Because the joint probability distribution is only related to the distance between the spatial points, all of the copulas are symmetric. There are four types of copulas, including the frankCopula, tCopula, normalCopula, and claytonCopula. The selection of these copulas and their parameters is determined via maximum likelihood estimation. The two-dimensional frankCopula and claytonCopula are both Archimedean copulas with a single parameter [48–50]. The frankCopula is characterized by the symmetry of the upper and lower tails, while the claytonCopula is characterized by a light upper tail and heavy lower tail. The frankCopulas of Cd are located at mean distances of 40.86 m, 101.67 m, 227.27 m, and 422.84 m. In addition, as the parameter decreases from 5.43 to 2.38, the tails of the frankCopulas become lighter, which indicates stronger randomness throughout the probability distribution. The claytonCopulas of Cd are located at mean distances of 489.41 and 552.12 m, and as the parameter decreases from 0.148 to 0.058, the lower tails become lighter. The upper and lower tails of the normalCopula are also symmetrically distributed. Compared with the other symmetrical copulas, the tCopula always has heavy upper and lower tails. When the tails of the copulas with mean distances of greater than 227.27 m become light, the tCopula with a mean distance of 164.41 m has the heaviest tail, which means that the extremal regions appear within a radius of nearly 200 m around the extreme samples with a high probability.

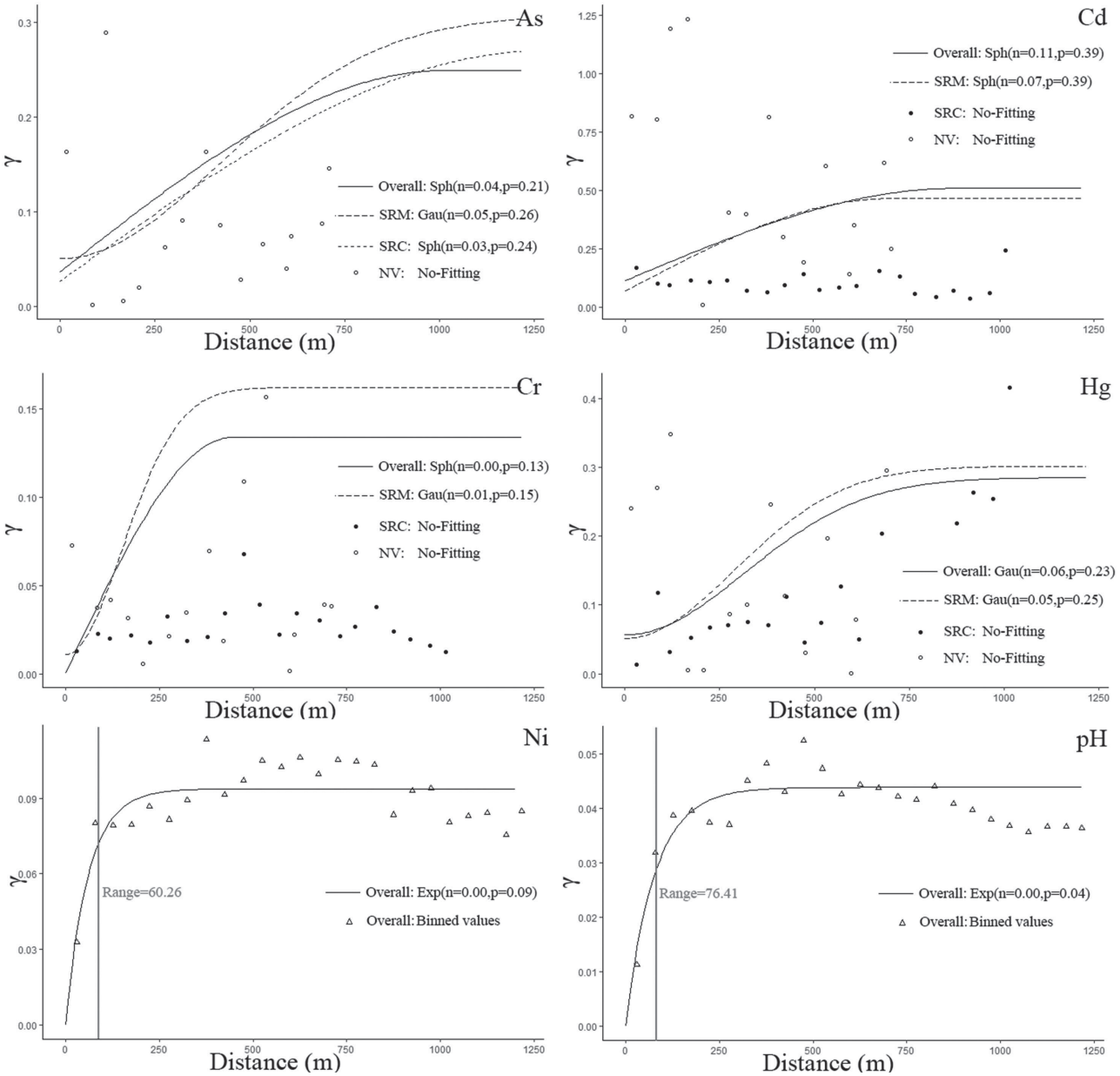

Fig. 3. Variogram analysis throughout the study area and for the different land consolidation types (Note: n stands for nugget and p stands for partial sill, which were calculated from the variogram model, including spherical (Sph), exponential (Exp), and Gaussian (Gau). No-fitting indicates that no variogram function can be fitted to the specific data).

As was previously discussed, the copula-based spatial model does not have an indicator similar to the ratio of the heterogeneity from the variogram. The random error of the variogram fitting is finally summed up as the nugget, but the copula-based spatial model distributes the randomness to every copula in the form of probability through maximum likelihood estimation. Thus, the predicted results exhibit the characteristic of raw data rather than a Gaussian distribution.

After establishing the variogram, the Gaussian spatial model was used to conduct a locally optimal unbiased estimation for the predicted locations. If the raw data did not have a Gaussian distribution, a logarithmic or Box-Cox transformation was used. However, the copula-based spatial model does not require the data to obey a Gaussian distribution. The cross-validation of the spatial predictions obtained using both the Gaussian and copula-based spatial models is presented in Table 2. It reveals that the MAPE values of Cr, Ni, and pH, for both the Gaussian and copula-based spatial models, are much smaller than those of the other factors. This could be due to the weak random spatial heterogeneity of these three factors. In addition, the R^2 value of Cr is the largest, which may be because it has the largest range among these three factors, which means more useful data for the local estimation of the locations. In general, the prediction effects of both models are very similar for Cr, Ni, and pH. In addition, the prediction of Cd based on the Gaussian spatial model has the largest MAPE

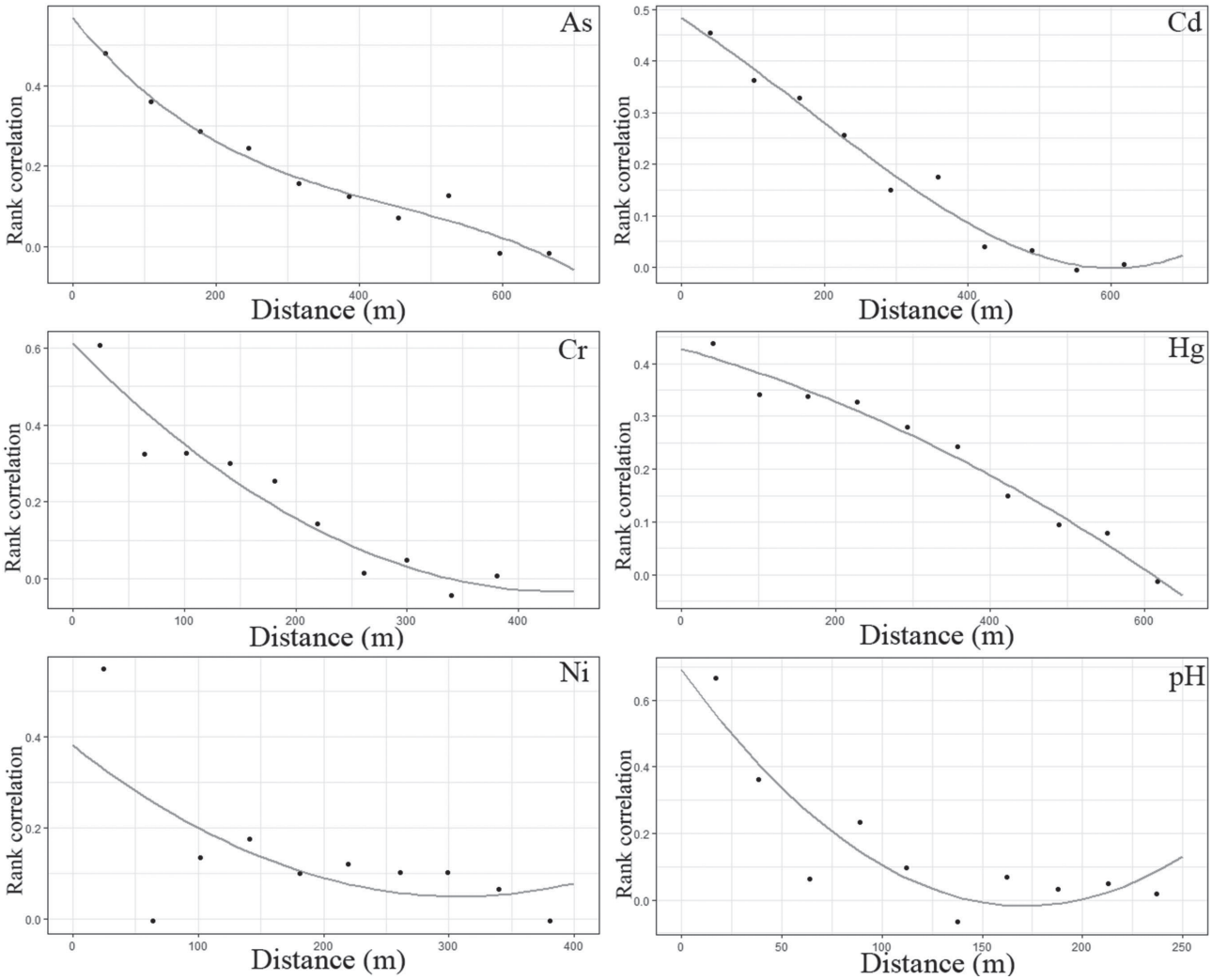

Fig. 4. Strengths of the correlation functions calculated from the optimally estimated copulas per lag.

Table 2. Performances of the different spatial models based on different q values from the geo-detector.

Items	Q	Gaussian spatial model		Copula-based spatial model	
		MAPE	R^2	MAPE	R^2
As	0.11*	0.24	0.38	0.23	0.41
Cd	0.21**	0.33	0.36	0.29	0.42
Cr	0.06	0.17	0.56	0.16	0.53
Hg	0.05	0.27	0.51	0.28	0.53
Ni	0.01	0.20	0.28	0.20	0.29
pH	0.03	0.17	0.24	0.17	0.32

Note: * denotes a significant difference between the different land consolidation types and ** denotes an extremely significant difference.

value, which is largely because it has the greatest random spatial heterogeneity. However, among these heavy metals, Cd's prediction effect was the most improved through the use of the copula-based spatial model, that is, its MAPE decreased from 0.33 to 0.29 and its R^2 increased from 0.36 to 0.42. Moreover, the prediction of As was also improved through the use of the copula-based spatial model. Relative to the MAPE and R^2 of the Gaussian spatial model, the improvements for Cr, Hg, and Ni were 5.88% (MAPE of Cr decreased from 0.17 to 0.16) and 3.92% (R^2

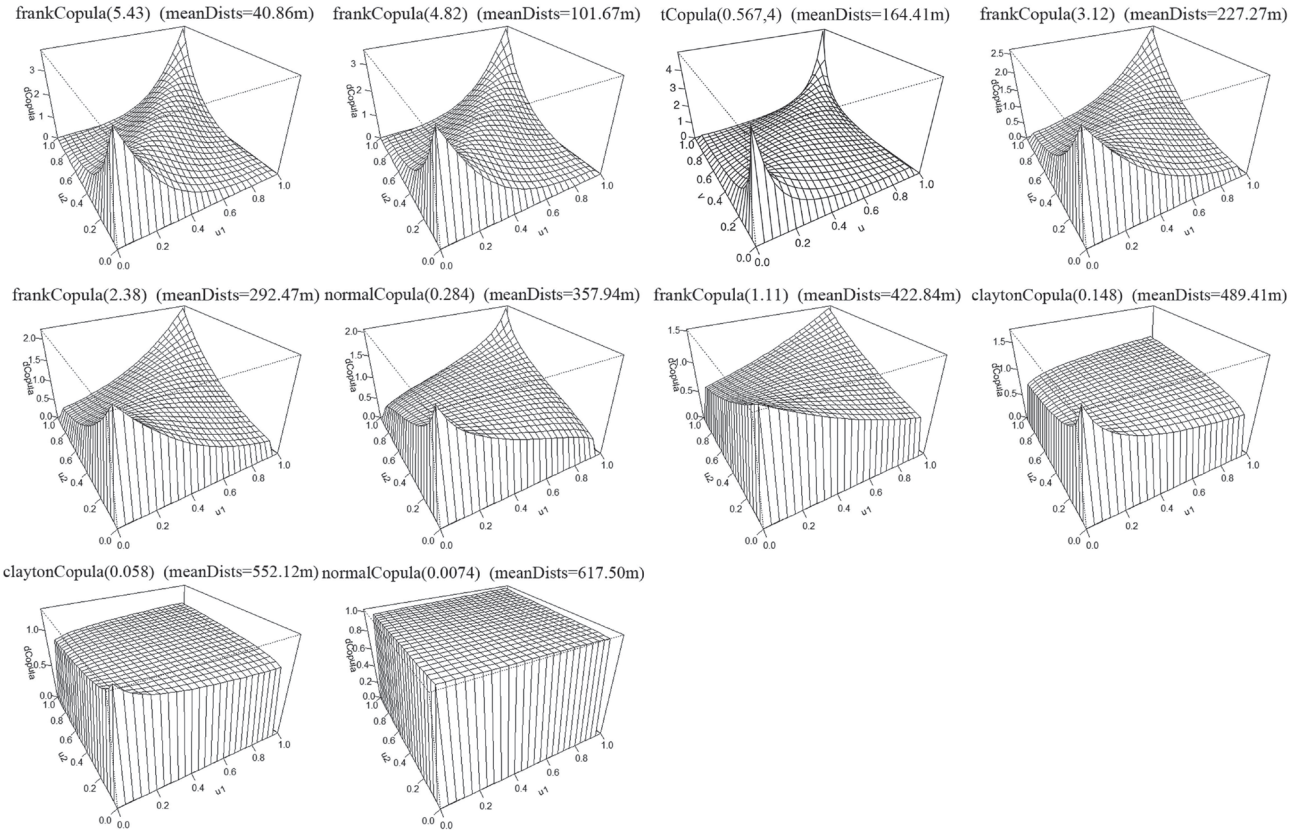

Fig. 5. 3D scatter plots of the optimally estimated copulas per lag for Cd.

of Hg increased from 0.51 to 0.53) at most, respectively. However, the improvements of Cd's MAPE and R^2 values were 12.12% and 16.67%, respectively, which are far better than those of the other heavy metals. In conclusion, the soil environmental factors of the consolidation land always exhibited weak random spatial heterogeneity because of the soil replacement in most of the study area, and the spatial model based on the Gaussian assumption satisfied the prediction requirements. However, some factors, such as Cd and As, in the study area may have strong stratified spatial heterogeneity, and their prediction effects were improved using the copula-based spatial model.

Identification of Extremal Regions for Heavy Metal Concentrations of the Soil

The prediction results obtained using the copula-based spatial model are the probability distribution for each location, so the Gaussian spatial model using the Bayesian process was used for comparison. In addition, the phenomenon of the Cd concentration of the soil in the study area exceeding the standard overwhelming occurred, so Cd was taken as the object of the extremal region identification study. Based on both observation statistics and regulation, two thresholds were set, including the 95% quantile of the samples (2.47 mg/kg) and the risk control value based on the regulation (3.00 mg/kg). These values are abbreviated as q.95 and ConV, respectively.

During the calculation of the Gaussian spatial model using a Bayesian process, the posterior distributions of the parameters (m, τ^2, σ^2 and φ) and were obtained via the MCMC method. The mean of m is 0.01 (confidence interval of 90% is [-0.07, 0.05]), the mean of τ^2 is 0.12 (confidence interval of 90% is [0.09, 0.14]), the mean of σ^2 is 0.39 (confidence interval of 90% is [0.38, 0.40]), and the mean of φ is 615.47 (confidence interval of 90% is [601.12, 645.11]).

The results of both the copula-based spatial model and the Gaussian spatial model using a Bayesian process have four types of extremal regions, including the Cd concentration of the soil exceeding q.95=2.47 mg/kg with a 50% probability, the Cd concentration of the soil exceeding q.95=2.47 mg/kg with a 75% probability, the Cd concentration of the soil exceeding ConV=3.00 mg/kg with a 50% probability, and the Cd concentration of the soil exceeding ConV=3.00 mg/kg with a 75% probability (Figs. 6 and 7). The red dots in the figures are based on whether the samples exceeded the thresholds. The recognition accuracy was obtained based on whether the red dots were located within the regions above (Table 3). It was found that the recognition accuracies of the copula-based spatial model for regions with Cd concentrations exceeding q.95=2.47 mg/kg with a 50% probability and Cd concentrations exceeding ConV=3.00 mg/kg with a 50% probability were significantly higher than those of the Gaussian spatial model using a Bayesian process. In addition, Fig. 8 shows that the probability density functions

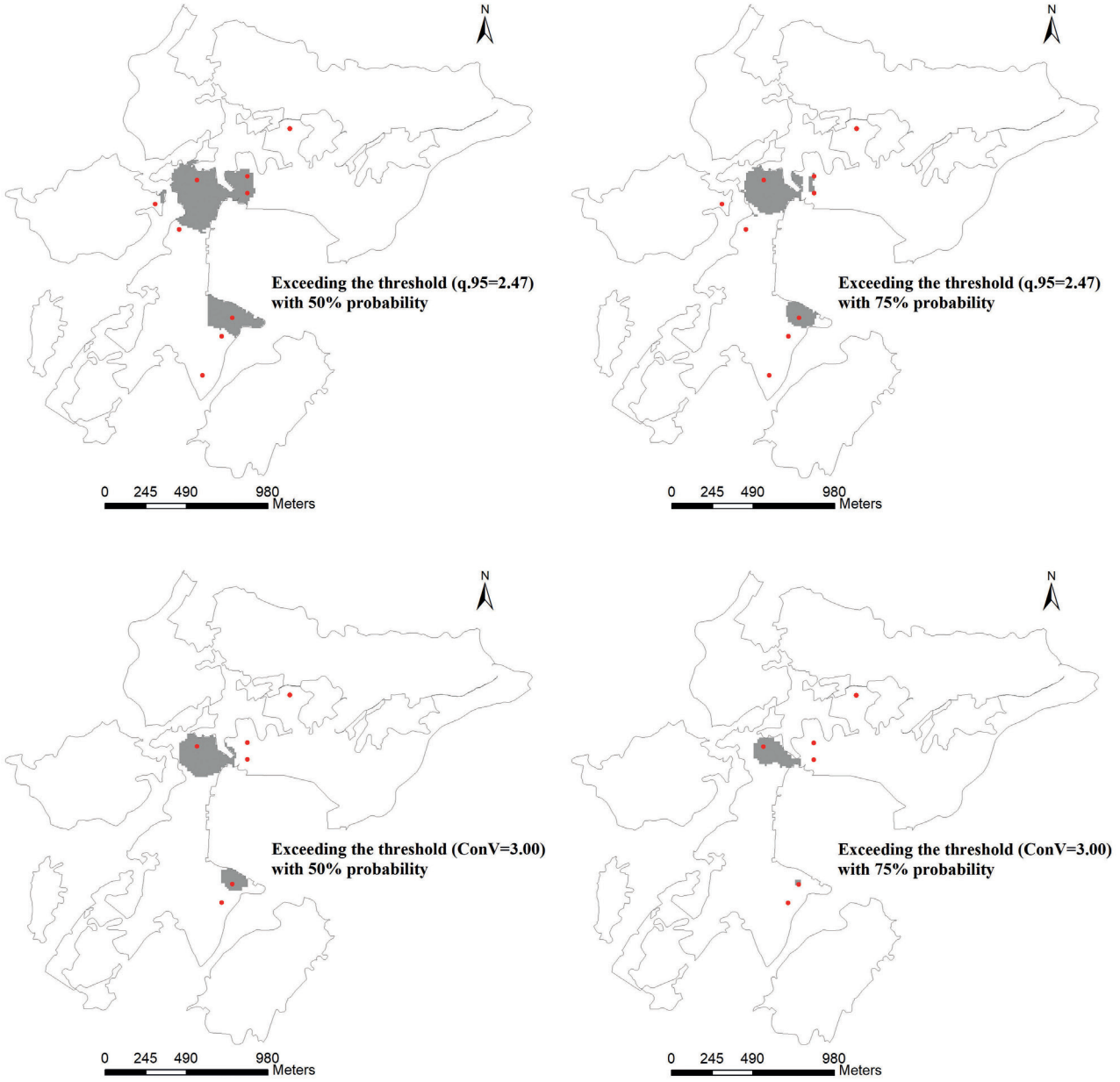

Fig. 6. The areas exceeding the specific thresholds (q.95 and ConV) with a 50% or 75% probability identified by the Gaussian spatial model using a Bayesian process.

(PDFs) of the predicted locations obtained using the copula-based model are closer to the characteristics of the samples, whereas the results of the Bayesian spatial model may produce negative values.

Conclusions

In summary, in this study, the advantages of the copula-based spatial model were demonstrated, and the spatial heterogeneities of the heavy metal concentrations of anthropogenic soil related to mining were determined.

(1) After land consolidation, the heavy metal concentrations of the soil environment in the mining area always exhibit weak random spatial heterogeneity because of soil replacement. Among these heavy metals, several have strong stratified spatial heterogeneity, including Cd and As, in the study area. For this situation, e.g., for Cd and As, the copula-based spatial model may perform better than the spatial model based on the Gaussian assumption.

(2) The spatial heterogeneities of the heavy metal concentrations of anthropologic soil should be explored on two scales, including samples and stratified statistics. The combination of random and stratified spatial

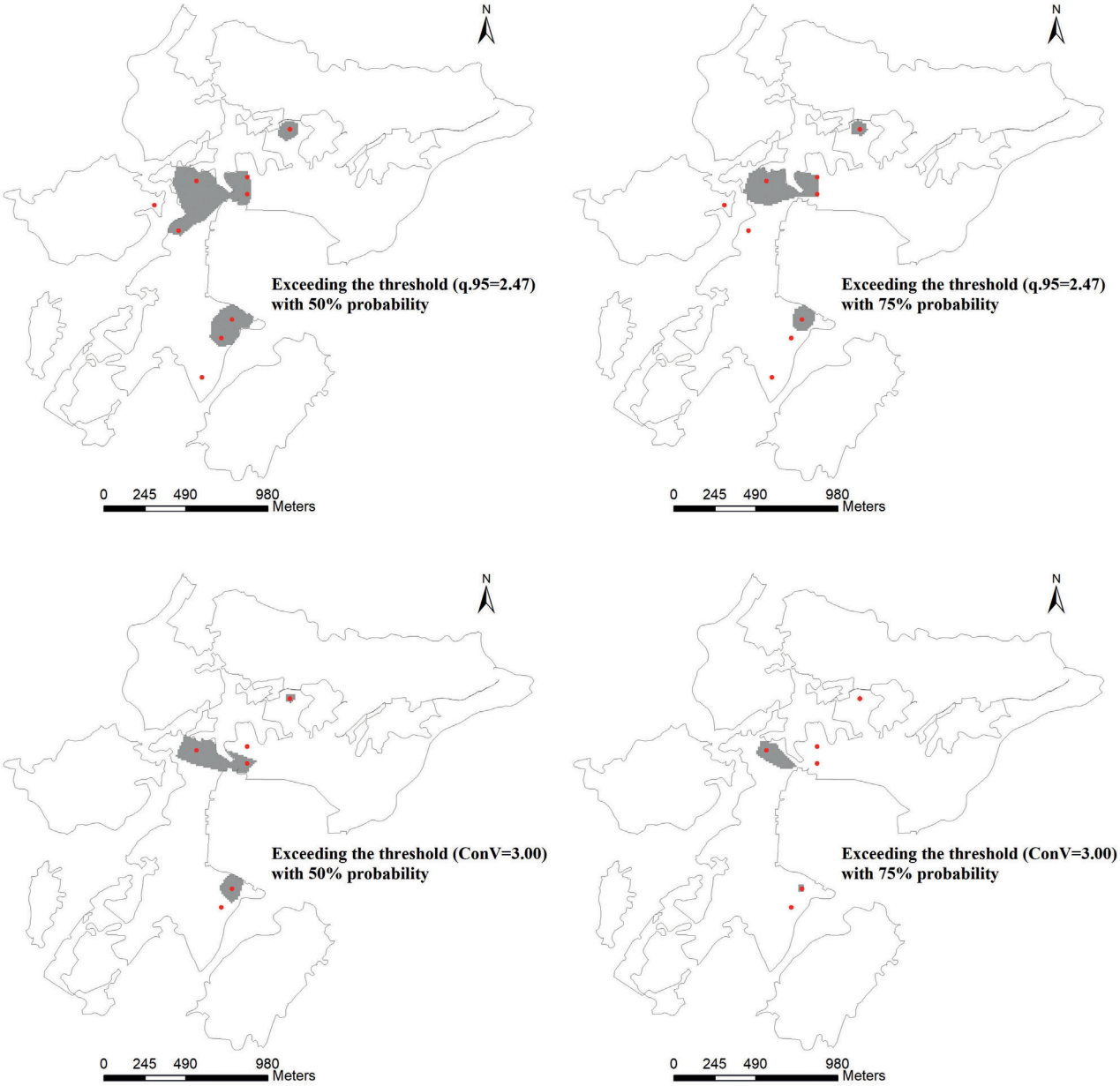

Fig. 7. The areas exceeding the specific thresholds (q.95 and ConV) with a 50% or 75% probability identified by the copula-based spatial model.

Table 3. Recognition accuracies of the extremal regions for the different models.

	Gaussian spatial model using a Bayesian process		Copula-based spatial model	
Threshold value	50%	75%	50%	75%
q.95=2.47	0.44	0.44	0.78	0.44
Control value=3.00	0.33	0.33	0.67	0.33

heterogeneity can depict the spatial characteristics of anthropological soil well. If some of the factors have strong stratified heterogeneity, the copula-based spatial model may be the better model to use.

(3) There are many copulas that can be used to fit the spatial rank correlation of the point pairs. The frankCopula, normalCopula, and tCopula have two heavy tails, while the claytonCopula and gumbelCopula

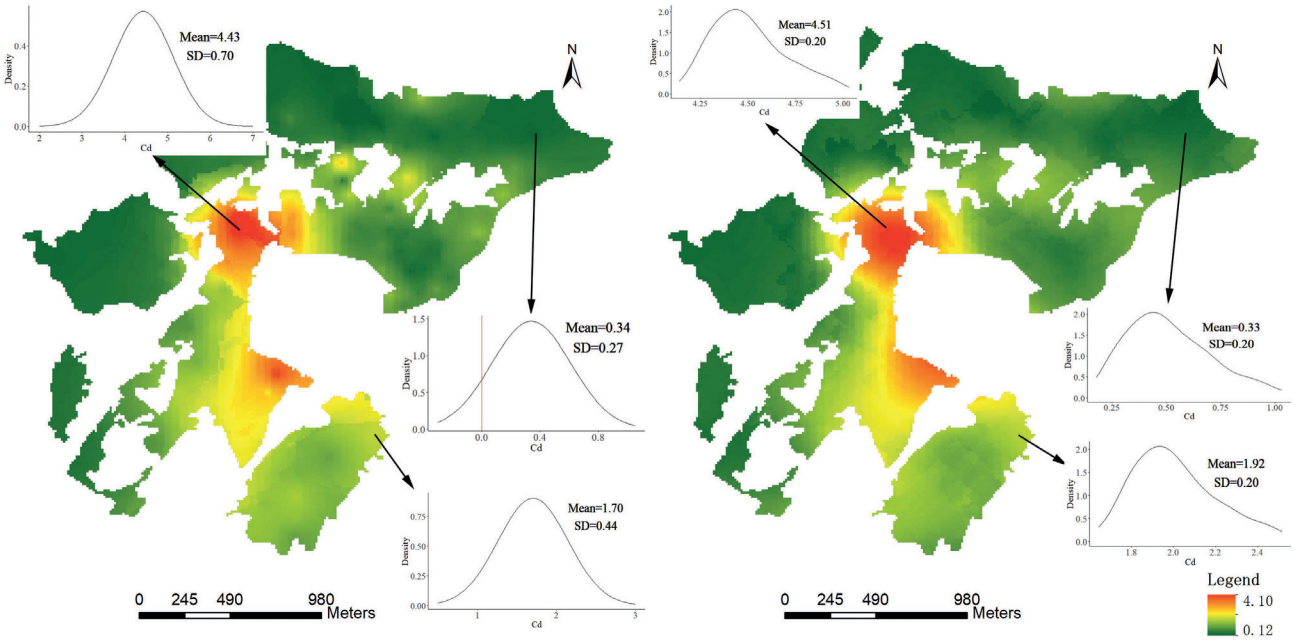

Fig. 8. Gaussian spatial distribution of Cd based on a Bayesian process (left) and the copula-based spatial distribution of Cd (right). The predicted PDFs of the three particular points are also shown.

have a single heavy tail. These types of copulas can adequately describe the correlation between the marginal distributions because the correlation between the spatial locations depends only on the distance. In addition, the copula-based model always produces better visual results. In the study area, by assessing the copulas for Cd, we determined that the extremal regions appear within a radius of nearly 200 m around the extreme samples with a high probability. In addition, the random error of the variogram fitting was finally summed up as a nugget, but the copula-based spatial model distributes the randomness to every copula in the form of a probability through maximum likelihood estimation. Thus, the predicted results exhibit the characteristics of the raw data rather than a Gaussian distribution.

(4) Through identification of the extremal regions of the Cd concentration in the study area, it was determined that the recognition accuracy of the copula-based spatial model was higher than that of the Gaussian spatial model using a Bayesian process. Moreover, the results of the copula-based spatial model are more reasonable (non-negative).

Acknowledgments

This research was supported by the National Key Research and Development Program of China (Grant No. 2019YFC1805000). Gratitude is extended to all the participants and anonymous reviewers, who supported our study during the field survey, discussion, and paper writing.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. BORUVKA L., KOZAK J. Geostatistical investigation of a reclaimed dumpsite soil with emphasis on aluminum. Soil & Tillage Research. **59** (3–4), 115, **2001**.
2. ARKOC O., UCAR S., OZCAN C. Assessment of impact of coal mining on ground and surface waters in Tozaklı coal field, Kırklareli, northeast of Thrace, Turkey. Environmental Earth Sciences. **75** (6), **2016**.
3. ROSEMARY F., VITHARANA U.W.A., INDRARATNE S.P., WEERASOORIYA R., MISHRA U. Exploring the spatial variability of soil properties in an Alfisol soil catena. Catena. **150**, 53, **2017**.
4. WANG Y.Z., DUAN X.J., WANG L. Spatial distribution and source analysis of heavy metals in soils influenced by industrial enterprise distribution: Case study in Jiangsu Province. Science of the Total Environment. **710**, 134953, **2020**.
5. HOWARD J. Classification of Anthropogenic Soils. In Anthropogenic Soils. Progress in Soil Science. **6**, 95. **2017**.
6. NAETH M.A., LESKIW L.A., BRIERLEY J.A., WARREN C.J., KEYS K., DLUSSKIY K., WU R.G., SPIERS G.A., LASKOSKY J., KRZIC M., PATTERSON G., BEDARD-HAUGHN A. Revised proposed classification for human modified soils in Canada: Anthroposolic order. Canadian Journal of Soil Science. **103** (1), 81, **2023**.
7. ZHANG F., LI G. China released the Action Plan on Prevention and Control of Soil Pollution. Frontiers of Environmental Science & Engineering. **10** (4), 19, **2016**.

8. Ministry of Ecology and Environment of China. Soil environmental quality, Risk control standard for soil contamination of agricultural land (Standard Specification). Available onlines: http://www.mee.gov.cn/ ywgz/fgbz/bz/bzwb/trhj/201807/t20180703_446029.shtml (accessed on May 1, 2024). [In Chinese]
9. FABIJAŃCZYK P., ZAWADZKI J., MAGIERA T. Magnetometric assessment of soil contamination in problematic area using empirical Bayesian and indicator kriging: A case study in Upper Silesia, Poland. Geoderma. **308**, 69, **2017**.
10. WANG J.F., HAINING R., ZHANG T.L., XU C.D., HU M.G., YIN Q., LI L.F., ZHOU C.H., LI G.Q., CHEN H.Y. Statistical Modeling of Spatially Stratified Heterogeneous Data. Annals of the American Association of Geographers. **114** (3), **2024**.
11. LI C., LIU B.L., GUO K., LI B.B., KONG Y.H. Regional Geochemical Anomaly Identification Based on Multiple-Point Geostatistical Simulation and Local Singularity Analysis-A Case Study in Mila Mountain Region, Southern Tibet. Minerals. **11** (10), 1037, **2021**.
12. RIBEIRO B.O.L., BARBUENA D., DE MELO G.H.C. Geochemical multifractal modeling of soil and stream sediment data applied to gold prospectivity mapping of the Pitangui Greenstone Belt, northwest of Brazil. Geochemistry. **83** (2), **2023**.
13. WANG J.-F., HAINING R., LIU T.-J., LI L.-F., JIANG C.-S. Sandwich Estimation for Multi-Unit Reporting on a Stratified Heterogeneous Surface. Environment and Planning A: Economy and Space. **45** (10), 2515, **2013**.
14. LIU T., WANG J., XU C., MA J., ZHANG H., XU C. Sandwich mapping of rodent density in Jilin Province, China. Journal of Geographical Sciences. **28** (4), 445, **2018**.
15. GRÄLER B. Modelling skewed spatial random fields through the spatial vine copula. Spatial Statistics. **10**, 87, **2014**.
16. CARREAU J., TOULEMONDE G. Extra-parametrized extreme value copula : Extension to a spatial framework. Spatial Statistics. **40**, 100410, **2020**.
17. GARCíA J.A., PIZARRO M.M., ACERO F.J., PARRA M.I.A Bayesian Hierarchical Spatial Copula Model: An Application to Extreme Temperatures in Extremadura (Spain). Atmosphere. **12** (7), 897, **2021**.
18. PALACIOS-RODRIGUEZ F., DI BERNARDINO E., MAILHOT M. Smooth copula-based generalized extreme value model and spatial interpolation for extreme rainfall in Central Eastern Canada. Environmetrics. **34** (3), **2023**.
19. SKLAR A. Fonctions de Repartition a n Dimensions et Leurs Marges. Publications de l'Institut de statistique de l'Université de Paris. **8**, 229, **1959**.
20. GRALER B., PEBESMA E. The pair-copula construction for spatial data: a new approach to model spatial dependency, 1st International Conference on Spatial Statistics – Mapping Global Change; Enschede, NETHERLANDS, **2011**.
21. LEE W., KIM M., AHN J.Y. On structural properties of an asymmetric copula family and its statistical implication. Fuzzy Sets and Systems. **393**, 126, **2020**.
22. SRISOPA S., LUAMKA P., RATTANAWAN S., SOMTRAKOON K., BUSABABODHIN P. Analyzing Spatial Dependence of Rice Production in Northeast Thailand for Sustainable Agriculture: An Optimal Copula Function Approach. Sustainability. **15** (20), **2023**.
23. GRÄLER B. Copulatheque: a small shiny app that illustrates a couple of copula families implemented in the copula (Program). https://copulatheque.shinyapps.io/copulas/ (accessed on May 1, **2024**).
24. FENG Y., WANG J.M., BAI Z.K., READING L. Effects of surface coal mining and land reclamation on soil properties: A review. Earth-Science Reviews. **191**, 12, **2019**.
25. LING Q., DONG F., YANG G., HAN Y., NIE X., ZHANG W., ZONG M. Spatial distribution and environmental risk assessment of heavy metals identified in soil of a decommissioned uranium mining area. Human and Ecological Risk Assessment: An International Journal. **26** (5), 1149, **2020**.
26. XU H., CROOT P., ZHANG C. Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis. Environment International. **151**, 106456, **2021**.
27. SHENG W.K., HOU Q.Y., YANG Z.F., YU T. Spatial Distribution, Migration, and Ecological Risk of Cd in Sediments and Soils Surrounding Sulfide Mines-A Case Study of the Dabaoshan Mine of Guangdong, China. Water. **15** (12), **2023**.
28. SOHRABIAN B. Geostatistical prediction through convex combination of Archimedean copulas. Spatial Statistics. **41**, **2021**.
29. PILZ J., KAZIANKA H., SPÖCK G. Some advances in Bayesian spatial prediction and sampling design. Spatial Statistics. **1**, 65, **2012**.
30. YARALI E., RIVAZ F., KHALEDI M.J. A Bayesian nonparametric spatial model with covariate-dependent joint weights. Spatial Statistics. **51**, **2022**.
31. LI F., YANG Y.J., SHANG Z.K., LI S. Y., OUYANG H.B. Kriging-assisted indicator-based evolutionary algorithm for expensive multi-objective optimization. Applied Soft Computing. **147** (2), 110736, **2023**.
32. JANG C.S. Probabilistic assessment of spatiotemporal fine particulate matter concentrations in Taiwan using multivariate indicator kriging. Stochastic Environmental Research and Risk Assessment. **38** (2), **2024**.
33. MARCHANT B.P., SABY N.P.A., JOLIVET C.C., ARROUAYS D., LARK R.M. Spatial prediction of soil properties with copulas. Geoderma. **162** (3), 327, **2011**.
34. WANG J., XU C. Geodetector: Principle and prospective. Acta Geographica Sinica. **72**, 116, **2017**.
35. FEI X., LOU Z., XIAO R., REN Z., LV X. Contamination assessment and source apportionment of heavy metals in agricultural soil through the synthesis of PMF and GeogDetector models. Science of The Total Environment. **747** (3), 141293, **2020**.
36. LU P., LERMUSIAUX P.F.J. Bayesian learning of stochastic dynamical models. Physica D: Nonlinear Phenomena. **427**, 123003, **2021**.
37. DING X., ZHANG H., ZHANG W., XUAN Y. Non-uniform state-based Markov chain model to improve the accuracy of transient contaminant transport prediction. Building and Environment. **245**, 110977, **2023**.
38. FINLEY A.O., BANERJEE S. Bayesian spatially varying coefficient models in the spBayes R package. Environmental Modelling & Software. **125** (9341), 104608, **2020**.
39. DAVISON A.C., PADOAN S.A., RIBATET M. Statistical Modeling of Spatial Extremes. Statistical Science. **27** (2), 161, **2012**.
40. THOMAS N., ULF S., JAKOB S., BRECHMANN E.C., BENEDIKT G., ERHARDT T., ALMEIDA C., MIN A., CZADO C., HOFMANN M., KILLICHES M., JOE H., VATTER T. VineCopula: Statistical Inference of Vine Copulas (R package version 2.4.3) (Program). https://

CRAN.R-project.org/package=VineCopula (accessed on May 1, **2024**).
41. EL ADLOUNI S. Quantile regression C-vine copula model for spatial extremes. Natural Hazards. **94** (1), 299, **2018**.
42. SAHIN Ö., CZADO C. Vine copula mixture models and clustering for non-Gaussian data. Econometrics and Statistics. **21**, 2452, **2021**.
43. HOFERT M., KOJADINOVIC I., MAECHLER M., YAN J., NEŠLEHOVÁ J.G., MORGER R. Copula: Multivariate Dependence with Copulas (R package version 1.0–1) (Program). https://CRAN.R-project.org/package=copula (accessed on May 1, **2024**).
44. ZHANG S., LIU H., LUO M., ZHOU X., LEI M., HUANG Y., ZHOU Y., GE C. Digital mapping and spatial characteristics analyses of heavy metal content in reclaimed soil of industrial and mining abandoned land. Scientific reports. **8** (1), 17150, **2018**.
45. LI C., LIU B., GUO K., LI B., KONG Y. Regional Geochemical Anomaly Identification Based on Multiple-Point Geostatistical Simulation and Local Singularity Analysis — A Case Study in Mila Mountain Region, Southern Tibet. Minerals. **11** (10), **2021**.
46. HASSAN M.M., ATKINS P.J. Application of geostatistics with Indicator Kriging for analyzing spatial variability of groundwater arsenic concentrations in Southwest Bangladesh. Journal of Environmental Science and Health, Part A. **46** (11), 1185, **2011**.
47. LIU C., LI W., WANG W., ZHOU H., LIANG T., HOU F., XU J., XUE P. Quantitative spatial analysis of vegetation dynamics and potential driving factors in a typical alpine region on the northeastern Tibetan Plateau using the Google Earth Engine. Catena. **206**, **2021**.
48. GENEST C., NEŠLEHOVÁ J., QUESSY J.-F. Tests of symmetry for bivariate copulas. Annals of the Institute of Statistical Mathematics. **64** (4), 811, **2012**.
49. XIE J., LIN F., YANG J. On a generalization of Archimedean copula family. Statistics & Probability Letters. **125**, 121, **2017**.
50. ARNOLD S., MOLCHANOV I., ZIEGEL J.F. Bivariate distributions with ordered marginals. Journal of Multivariate Analysis. **177**, **2020**.