Transactions on Environment, Energy and Earth Sciences
ISSN: 2960-1924, eISSN: 2960-2300 | Volume 5
EMCEME 2025

# Source Apportionment of Soil Heavy Metal Pollution in the Yangtze River Basin Based on the Multiscale Geographically Weighted Spatial Lag Model

Fengying Yan *

Brook Institute at Anhui University, Anhui University, Hefei, Anhui, China, 230039

* Corresponding Author Email: R22214019@stu.ahu.edu.cn

**Abstract.** As a crucial agricultural base, economic belt, and ecological barrier in China, the Yangtze River Basin is increasingly plagued by soil heavy metal pollution due to the combination of complex natural geographical conditions and intensive human activities. Based on soil heavy metal (As, Cd, Cr, Cu, Hg, Ni, Pb, Zn) content data from 906 sampling sites in the Yangtze River Basin from 2000 to 2020, this study integrated 11 explanatory variables from four categories (topography, land use, socio-economy, and meteorology) to construct three models: Geographically Weighted Regression (GWR), Multiscale Geographically Weighted Regression (MGWR), and Multiscale Geographically Weighted Spatial Lag Model (MGWR-VSLR). The spatial distribution characteristics and driving mechanisms of heavy metal pollution were systematically analyzed.Results showed MGWR - VSLR model excelled in $R^2$ and AICc, capturing heavy metal pollution's spatial features. As and Hg pollution were affected by different factors. Spatial lag term indicated cross - regional influence. Soil heavy metal pollution in Yangtze River Basin results from natural and human factors. The research aids prevention and green development.

**Keywords:** Yangtze River Basin, soil heavy metals, multiscale geographically weighted spatial lag model, spatial heterogeneity, driving mechanism.

## 1. Introduction

Soil is essential for human survival, yet economic development has led to severe soil heavy metal pollution, harming crops and human health. The Yangtze River Basin, crucial for China's development, suffers from heavy metal pollution due to industrial, agricultural, and urbanization factors. In the face of global environmental issues, exploring the multiscale geographically weighted regression model with spatial lag for predicting its soil heavy metal content is significant, aiding research and various aspects of development.

In recent years, the Geographically Weighted Regression (GWR) model has been widely used in processing heterogeneous spatial data. In the context of spatial variables, spatial correlation intuitively reflects the spatial distribution characteristics of variables, while spatial heterogeneity characterizes the instability of spatial structures. For the GWR model, not all independent variables exhibit unstable spatial structures. To address this, Cronie, O [1] et al. proposed the Multiscale Geographically Weighted Regression (MGWR) model, which considers that only some independent variables have stable spatial structures. In addition, apart from the spatial heterogeneity of independent variables, the dependent variable itself has a spatial lag effect; from the perspective of model equations, the spatial effect of independent variables affects the dependent variable. Qiao Ningning [2] further proposed the Mixed Geographically Weighted Regression with Constant Coefficient Spatial Lag (MGWR-CSLR) model. In fact, spatial heterogeneity affects both independent and dependent variables. Fosgerau, M [3] et al. and Páez [4] et al. conducted in-depth research on the regression of spatial lag terms of dependent variables based on the traditional GWR model, replacing the constant coefficient of the lag term with a variable coefficient. Brunsdon [3] et al. also pointed out that even if the differences in the coefficients of the spatial lag of the dependent variable across regional points are not significant, it cannot be assumed that a constant coefficient is applicable to all regions. Based on this, Tang Zhipeng [5] proposed the Mixed Geographically Weighted Regression with Variable Coefficient Spatial Lag (MGWR-VSLR) model.

However, there remains a research gap in the application of the MGWR-VSLR model to soil heavy metal pollution studies by incorporating a wide range of predictive variables covering natural and human factors. These predictive variables may enhance the regression performance and spatial heterogeneity of pollutants. For example, topographic factors such as slope and altitude affect the transport and spatial distribution of terrestrial pollutants; rainfall intensity may alter the runoff of pollutants into rivers [6]; socio-economic factors such as population density and gross domestic product (GDP) directly reflect the intensity of human interference on the soil environmental quality of the basin. In view of this, this study selects multiple variables such as land use type [7], socio-economy, topography, and meteorology as predictive variables, aiming to reflect the impact of these variables on the spatial distribution of related sedimentary pollutants, thereby improving the application performance of the MGWR-VSLR model.

This study aimed to fill this gap. First, it integrated 11 independent variables from four categories (topography, land use, socio - economy, meteorology) and 8 soil heavy metal dependent variables. Data from 906 sampling sites in 11 Yangtze River Basin provinces and cities from 2000 - 2020 were collected for spatial analysis. Then, data from multiple sources were gathered and tested. Based on the significant spatial agglomeration of pollution, a spatial lag model was introduced. OLS, GWR, MGWR, and MGWR - VSLR models were constructed and optimized. The MGWR - VSLR model, with spatial lag terms, inverse distance weighted spatial weight matrix, and weighted least - squares local parameter estimation, improved goodness - of - fit and accuracy. It outperformed traditional models in $R^2$ and AICc, accurately characterizing pollution spatial features. Finally, policy recommendations were made to support the "Yangtze River Protection" strategy.

## 2. Research Methods and Model Construction

### 2.1. Study Area and Data Preparation

To investigate the impact of human activities and the environment on soil heavy metal content in the Yangtze River Basin, this study collected data on soil heavy metal content from 906 sampling sites in the Yangtze River Basin from 2000 to 2020 (Figure 1). Then, a combined model of multiscale geographically weighted regression and spatial lag was used to analyze the influencing factors and spatial effects of soil heavy metal content in each region.

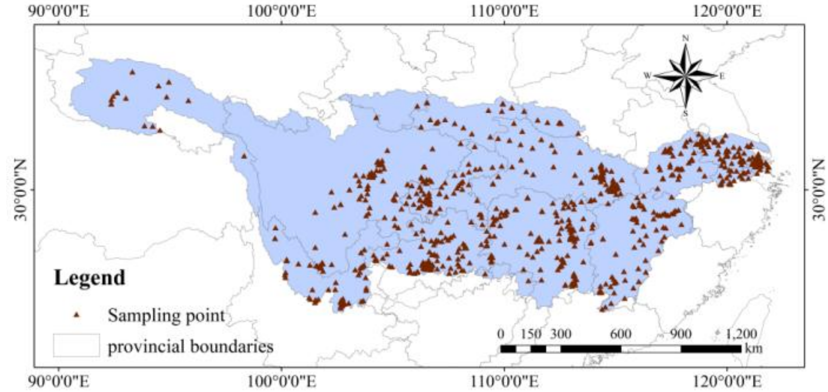



**Figure 1.** Research Sampling Sites.

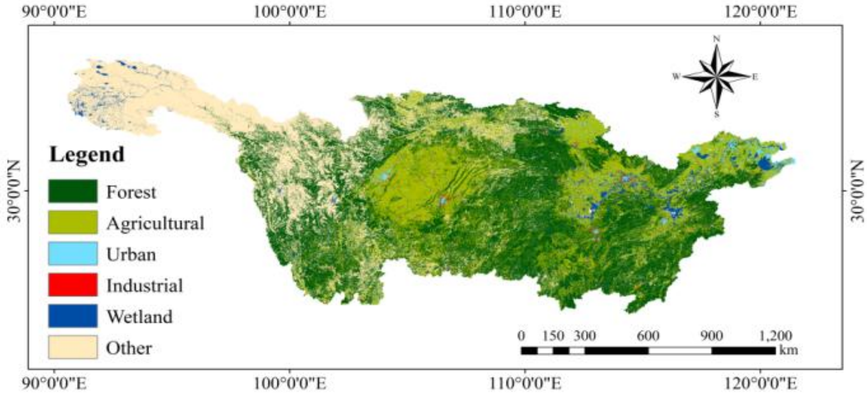



**Figure 2.** Land Use Map of the Yangtze River Basin.

In the regression analysis of soil pollutants, this study comprehensively included four categories of potential explanatory variables (Table 1), including land use (Figure 2), topography, socio-economy, and meteorological indicators. For topographic predictors, altitude and slope were selected. Based on a 30-meter resolution Digital Elevation Model (DEM), the altitude (EL) of each sampling site was extracted, and the average slope (SL) of each county was calculated using the zonal statistics tool in ArcGIS 10.8 [8]. For land use predictors, 2020 Landsat8 remote sensing image data were used, and the proportions of five main land use types (forest land (For), agricultural land (Agric), urban land (Urban), industrial and mining land (Ind), and wetland (Wetl)) in each county were calculated using the analysis tools in ArcGIS 10.8 [9]. For socio-economic variables, 2020 county-level average population and average gross product (GPP) data [10] were collected to estimate the population density and gross product corresponding to each sampling site. Meteorological data, including annual average temperature (TMP) and annual rainfall (RA) in counties of the Yangtze River Basin from 2020 to 2022, were collected from the Resource and Environmental Science Data Center of the Chinese Academy of Sciences (https://www.resdc.cn/). By integrating multiple types of variables, this study can analyze the distribution characteristics and evolution laws of pollutants from a more comprehensive spatial perspective, significantly enhancing the explanatory power and predictive performance of the regression model, and laying a reliable foundation for subsequent research.

**Table 1.** Explanatory Variables for Geospatial Regression of Soil Heavy Metal Content in the Yangtze River Basin.

| Data Type | Independent Variable | Unit | Spatial Resolution | Temporal Resolution | Source |
|---|---|---|---|---|---|
| Topography | Altitude | m | 30m | - | http://www.gscloud.cn |
| | Slope | m | 30m | - | http://www.gscloud.cn |
| Land Use Type | Urban, Industrial, Forest, Wetland, Agricultural | - | 30m | 2020 | https://www.resdc.cn/ |
| Socio-Economy | Population Density | person/km² | County-level | 2020 | 2020 County Statistical Yearbooks |
| | Gross Product (GPP) | 10,000 yuan | Provincial level | 2020 | 2020 County Statistical Yearbooks |
| Meteorology | Temperature | ℃ | Station-level | 2020-2022 annual data | https://www.resdc.cn/ |
| | Precipitation | mm | Station-level | 2020-2022 annual data | https://www.resdc.cn/ |

Spatial autocorrelation tests were also conducted. Spatial autocorrelation refers to the correlation of the same variable at different spatial locations, which is a measure of the aggregation degree of attribute values in spatial units. In the study of soil pollutants, the concentration of soil pollutants in one region may be related to that in surrounding regions. Therefore, it is necessary to conduct spatial autocorrelation tests to determine whether to introduce the multiscale geographically weighted spatial lag model. First, the global Moran's I test was performed on the collected data using the global Moran's I formula:

$$I = \frac{n}{S_0} \cdot \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{1}$$

Where: n is the number of spatial units; $w_{ij}$ is the element of the spatial weight matrix; $x_i$ and $x_j$ are the observed values at locations i and j; $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the mean of all observed values; $S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$ is the sum of the spatial weight matrix.

In screening regression variables, Python software was used to optimize Ordinary Least Squares (OLS) regression using the stepwise elimination method [11]. The specific process was in each round of regression analysis, based on the results of Analysis of Variance (ANOVA), the variable with the largest probability p-value corresponding to the F-value was removed until the p-values of all retained variables were less than the significance level $\alpha = 0.05$.

## 2.2. Classical Geographically Weighted Regression Model

The Geographically Weighted Regression (GWR) model, a regional regression model proposed by Cronie, O [1] et al. in 1996, is widely used in exploring the dynamic relationships between dependent and explanatory variables. The following details GWR with relevant formulas.

The GWR model can be expressed as:

$$y_j = \beta_0(u_j, v_j) + \sum_{i=1}^{p} \beta_i(u_j, v_j)x_{ij} + \varepsilon_j \tag{2}$$

Where: $(u_j, v_j)$ represents the geographical coordinates of the $j$-th point; $\beta_0(u_j, v_j)$ is the intercept, i.e., the constant term of the model; $\beta_i(u_j, v_j)$ is the slope, i.e., the regression coefficient; $\varepsilon_j$ is the regression residual, which is the difference between the observed and predicted values of the dependent variable; $p$ is the number of explanatory variables; $y_j$ is the observed value at the $j$-th point, i.e., the dependent variable; $x_{ij}$ is the $i$-th independent variable at the $j$-th point, which can be used to estimate environmental variables such as precipitation.

The regression coefficients are calculated as:

$$\hat{\beta}(u_i, v_i) = (X^T W_i X)^{-1} X^T W_i Y \tag{3}$$

Where: $\hat{\beta}(u_i, v_i)$ is the local regression coefficient at the estimation location; $X$ and $Y$ are the vectors of independent and dependent variables, respectively; $W_i$ is the weight matrix. Weights are calculated using the following formula:

$$W_{ij} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{h}\right)^2\right]^2, & d_{ij} \le h \\ 0, & d_{ij} > h \end{cases} \tag{4}$$

Where: $W_{ij}$ is the weight of the observed value at the $j$-th point when estimating the coefficient at location $i$; $d_{ij}$ is the Euclidean distance between the $j$-th point and the adjacent observed value $i$; $h$ is the adaptive threshold of the kernel function bandwidth, defined as the $k$-nearest neighbor distance.

First, the kernel function is used to estimate the weight matrix (i.e., W_i in the above formula). In the GWR model, four kernel functions are available: fixed Gaussian, adaptive Gaussian, fixed bi-square, and adaptive bi-square.

Second, the selection criterion is used to determine the value of the adaptive bandwidth (i.e., h in the above formula). Four common selection criteria are available: AIC (Akaike Information Criterion), AICc (small sample bias corrected AIC), BIC (Bayesian Information Criterion), and CV (Cross Validation). In this study, the adaptive bi-square function was selected as the kernel function, and AICc was used as the selection criterion.

### 2.3. Multiscale Geographically Weighted Regression Model

The Multiscale Geographically Weighted Regression model is an important improvement over the classical GWR model. The GWR model uses a unified bandwidth for all variables when processing spatial data, making it difficult to accurately characterize the spatial characteristics of different variables. In contrast, the MGWR model overcomes the limitation of bandwidth selection by allowing different variables to select different bandwidth values according to their characteristics, thereby more effectively presenting the spatial heterogeneity of variables and significantly improving the accuracy of regression analysis.

The expression of the MGWR model is as follows:

$$Y_i = \beta_{0,i} + \sum_{k=1}^{p} \beta_{k,i} X_{ki} + \varepsilon_i, i = 1,2,\cdots,n \tag{5}$$

Where: $\beta_{k,i}$ represents the regression coefficient of the $k$-th variable under a specific bandwidth. Compared with the GWR model, the parameters of the MGWR model are usually solved using the back-fitting algorithm [11][12]. The specific steps are as follows:

First, parameter initialization is performed using the GWR model to obtain initial regression coefficients. Second, let $f_k$ represent the regression value of the $k$-th variable ($f_k = \beta_{k,i} X_{ki}$), and calculate the initial residual $\varepsilon = y - \sum_{k=1}^{p} f_k$. Then, starting from the first independent variable, add this residual to the regression value of the first variable, i.e., $\varepsilon + f_1$, use it as the response variable, and perform geographically weighted regression with the independent variable $X_1$ to obtain the optimal bandwidth and new estimated parameters for the first independent variable. Subsequently, update the corresponding estimated parameters in the MGWR model and calculate new residuals. Next, add the new residual to the regression value of the second independent variable $f_2$, and repeat the above steps until the optimal bandwidths and regression coefficient estimates for all independent variables are calculated. Finally, after each cycle, check whether the convergence condition is met: if met, exit the cycle and terminate the iteration; otherwise, start from the first independent variable and continue iterating to improve the regression coefficient estimates.

A commonly used convergence criterion is to calculate the rate of change (score of change, SOC) of the residual sum of squares (RSS):

$$SOC_{RSS} = \frac{|RSS_{new} - RSS_{old}|}{RSS_{new}} \tag{6}$$

Where $RSS_{old}$ is the residual sum of squares from the previous cycle, and $RSS_{new}$ is the residual sum of squares from the current cycle. When $SOC_{RSS}$ is less than a given control parameter, the algorithm determines that the iteration meets the convergence condition; otherwise, it is deemed not to meet the condition.

### 2.4. Multiscale Geographically Weighted Regression Model+ Spatial Lag Model

The study of soil pollution's spatial distribution and factors is key in environmental - geography intersection. Traditional spatial econometric models have limitations. This study introduces the multiscale geographically weighted spatial lag model. Based on MGWR, it combines spatial lag terms.

But the traditional constant - coefficient model has flaws, so the study uses MGWR - VSLR for more accurate spatial distribution modeling.

(1) Model Formula

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ik} + \rho(u_i, v_i) \cdot SpatialLag_i + \epsilon_i \tag{7}$$

where: $y_i$ is the value of the dependent variable at the $i$-th sample point; $\beta_0(u_i, v_i)$ is the intercept related to the spatial location $(u_i, v_i)$; $\beta_k(u_i, v_i)x_{ik}$ represents the $k$-th independent variable $x_{ik}$ and its coefficient $\beta_k(u_i, v_i)$ that varies with spatial location, reflecting spatial non-stationarity; $\rho(u_i, v_i) \cdot SpatialLag_i$ is the spatial lag term, reflecting spatial dependence, where $\rho(u_i, v_i)$ is the coefficient and $SpatialLag_i$ is the neighborhood influence value; $\epsilon_i$ is the model error, including unconsidered factors and random fluctuations.

（2）Parameter Estimation Process

The spatial weight matrix is a core element of spatial econometric models, used to characterize the intensity of spatial association between geographical units. The specific steps are as follows:

Distance calculation: Calculate the Euclidean distance $d_{ij}$ between sample points $i$ and $j$ to construct a distance matrix.

Threshold setting: Determine the maximum distance $d_{max}$ of $k$-nearest neighbors using an adaptive method as the threshold for spatial connection (if $d_{ij} \le d_{max}$, the two points are considered to have spatial association).

Weight transformation: Convert the original distance into a weight matrix using the inverse distance weighting method:

$$w_{ij}^* = \begin{cases} \frac{1}{d_{ij}^2}, & d_{ij} > 0 \\ 0, & d_{ij} = 0 \end{cases} \tag{8}$$

When $i = j$ (self-distance), the weight is forced to 0.

Normalization: Perform row normalization on the neighbor weights of each sample point to ensure $\sum_{j=1}^{n} w_{ij}^* = 1$, eliminating the impact of differences in the number of neighbors on the model and adapting to the calculation of spatial lag terms.

Based on the normalized weight matrix, calculate the spatial lag term of the dependent variable:

$$Wy_i = \sum_{j=1}^{n} w_{ij}^* y_j \tag{9}$$

Merge the original independent variable matrix $X$ with the spatial lag term $Wy$ to construct an extended dataset matrix $Z = [X, Wy]$ as the model input.

The bandwidth parameter controls the spatial influence range of local regression; a larger bandwidth includes more adjacent points in the fitting. For the MGWR model, each explanatory variable corresponds to an independent optimal bandwidth $h_k$, determined by minimizing the corrected AIC criterion (AICc):

$$AICc = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n \cdot \frac{n + tr(S)}{n - 2 - tr(S)} \tag{10}$$

Where $S$ is the smoothing matrix, and $tr(S)$ represents the number of effective parameters. Optimizing this criterion balances local fitting accuracy and model complexity.

Use Weighted Least Squares (WLS) to estimate parameters for the local neighborhood of each sample point:

$$\hat{\beta}(x_i, y_i) = (Z_i^T W_i Z_i)^{-1} Z_i^T W_i y_i \tag{11}$$

Where $W_i$ is a diagonal matrix of kernel weights based on bandwidth $h_k$, generated by Gaussian or bi-square kernel functions, assigning higher fitting weights to adjacent points.

## 3. Comparison and Analysis of Model Results

### 3.1. Data Processing Results

According to Sun, Y.'s [13] [14] study on soil heavy metal content at 906 sampling sites in the Yangtze River Basin, the concentrations of heavy metals in the soil vary significantly [15]. The concentration ranges of lead (Pb), mercury (Hg), arsenic (As), cadmium (Cd), chromium (Cr), nickel (Ni), copper (Cu), and zinc (Zn) are 1-7168, 0.01-205.2, 0.02-4495.865, 0.01078-135.5, 0.2874-3715, 0.285-704.2, 0.064-2878, and 0.38-19737 mg/kg, respectively (Table 2). Before using soil heavy metal concentrations as dependent variables in model analysis, normality and spatial autocorrelation tests were conducted on the concentrations [16]. Non-normal variables were transformed to normality using box-cox transformation before model application. Taking As and Cr as examples, SPSS software was used to draw normal Q-Q plots, confirming that the transformed data conform to a normal distribution.

**Table 2.** Descriptive Statistics of Heavy Metal Concentrations in Soil.

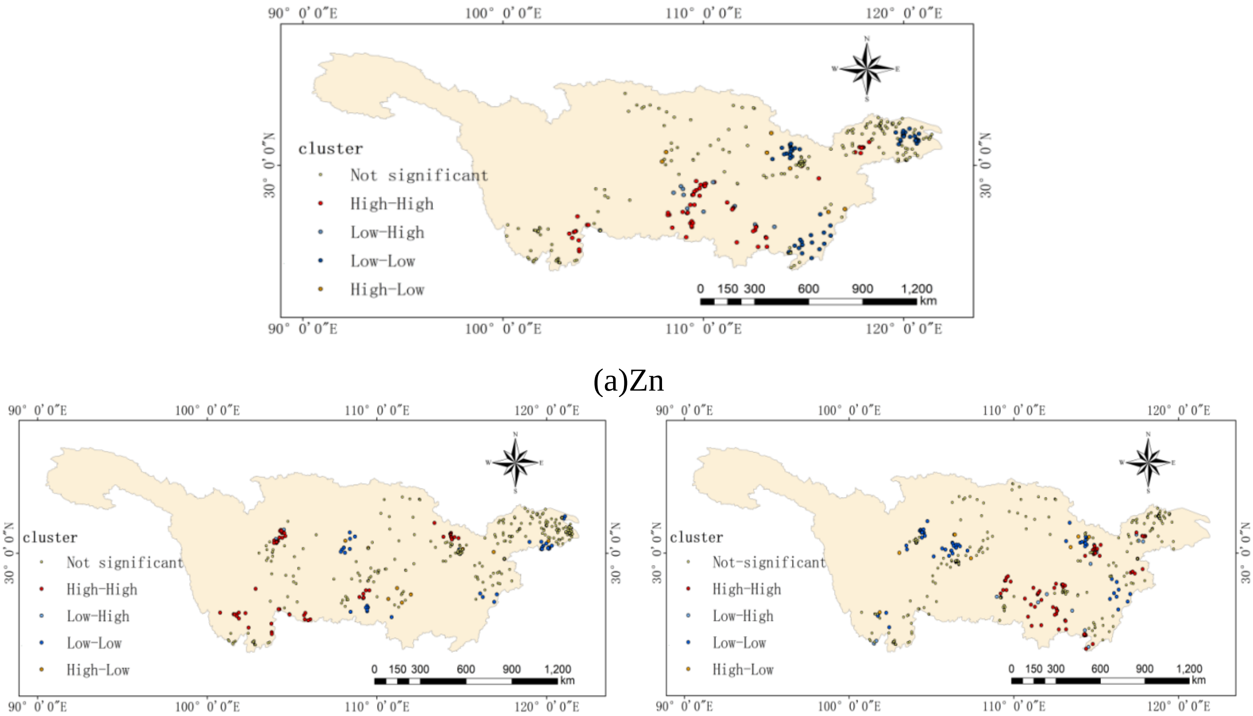
| Heavy Metal | Minimum | Maximum | Median | Mean | Standard Deviation |
|---|---|---|---|---|---|
| As | 0.02 | 4495.865 | 13 | 35.10731 | 200.3255 |
| Cd | 0.01078 | 135.5 | 0.3445 | 3.704906 | 12.73963 |
| Cr | 0.2874 | 3715 | 71.9 | 89.9425 | 164.3416 |
| Cu | 0.064 | 2878 | 34.2 | 78.59577 | 186.5087 |
| Hg | 0.01 | 205.2 | 0.15 | 1.365766 | 12.66671 |
| Ni | 0.285 | 704.2 | 31.48 | 55.11148 | 79.7457 |
| Pb | 1 | 7168 | 37.023 | 104.3403 | 372.6935 |
| Zn | 0.38 | 19737 | 101 | 248.485 | 976.627 |

**Table 3.** Global Moran's I Values for Each Heavy Metal.

| Heavy Metal | Moran's I Value |
|---|---|
| As | 0.4283 |
| Cr | 0.2791 |
| Cd | 0.3786 |
| Cu | 0.2239 |
| Hg | 0.4378 |
| Ni | 0.1972 |
| Pb | 0.3169 |
| Zn | 0.487 |

The global Moran's I values for each heavy metal were calculated (Table 3), with all P-values = 0.0000, passing the significance test. Thus, it can be concluded that the pollutant distribution in this region has spatial correlation, and the spatial lag term was introduced to establish the multiscale geographically weighted spatial lag model.

However, it remains unclear which areas have spatial agglomeration. Therefore, local Moran's I was introduced to draw LISA cluster maps, with Zn, Cr, and Cd shown as examples (Figure 3).




(a)Zn

(b) Cr (c) Cd

**Figure 3.** LISA Cluster Maps of Zn, Cr, and Cd.

For example, the LISA cluster map of Zn (Figure 3a) shows that red "High-High" clusters are concentrated in the western and central regions, indicating spatial agglomeration of high Zn content in these areas; blue "Low-Low" clusters are mainly located in the eastern region, forming contiguous cold spots of low Zn content.

Based on the stepwise elimination method, the final selected variables are as follows (Table 4).

**Table 4.** Finally Selected Significant Variables.

| Dependent Variable | Significant Variables ($p < 0.05$) |
|---|---|
| As | Altitude, agricultural land, urban land, wetland, precipitation |
| Cd | Population density, altitude, average slope, urban land, temperature, precipitation |
| Cr | Altitude, average slope, wetland |
| Cu | Altitude, average slope, forest land, urban land, precipitation |
| Hg | Gross regional products, urban land, precipitation |
| Ni | Altitude, forest land, temperature |
| Pb | Altitude, average slope, urban land, precipitation |
| Zn | Population density, altitude, temperature, precipitation |

### 3.2. Analysis of Model Regression Results

**Table 5.** Comparison of OLS, GWR, MGWR, and MGWR-VSLR Models.

| | R2 | | | | AICc | | | |
|---|---|---|---|---|---|---|---|---|
| Heavy Metal | OLS | GWR | MGWR | MGWR-VSLR | OLS | GWR | MGWR | MGWR-VSLR |
| As | 0.188 | 0.542 | 0.569 | 0.597 | 1389.678 | 1263.101 | 1213.815 | 1193.545 |
| Cd | 0.085 | 0.496 | 0.517 | 0.528 | 1289.371 | 1187.045 | 1143.068 | 1103.878 |
| Cr | 0.097 | 0.176 | 0.36 | 0.371 | 1216.878 | 1206.759 | 1145.672 | 1021.57 |
| Cu | 0.046 | 0.451 | 0.461 | 0.493 | 1328.512 | 1256.107 | 1211.87 | 1207.511 |
| Hg | 0.096 | 0.57 | 0.576 | 0.585 | 874.997 | 750.147 | 735.135 | 720.872 |
| Ni | 0.058 | 0.374 | 0.411 | 0.457 | 1065.887 | 999.788 | 988.693 | 976.217 |
| Pb | 0.077 | 0.417 | 0.432 | 0.468 | 1016.110 | 951.042 | 930.364 | 921.214 |
| Zn | 0.187 | 0.453 | 0.554 | 0.571 | 1335.844 | 1319.968 | 1252.866 | 1184.945 |

The significant variables for each heavy metal were used to fit models, and the results are as follows (Table 5). In regression analysis, a higher goodness - of - fit ($R^2$) and lower AICc and residual sum of squares signify better model fitting and accuracy. As shown in Table 5, for different heavy metals, the MGWR - VSLR model generally has higher $R^2$ values than GWR and MGWR models, and significantly higher than the OLS model. For instance, for As, the MGWR - VSLR model has an $R^2$ of 0.597, higher than GWR's 0.542 and MGWR's 0.569. Hg shows a similar trend. Moreover, the AICc values of the MGWR - VSLR model for each heavy metal are usually smaller. For example, for As, the MGWR - VSLR model has an AICc of 1193.545, less than GWR's 1263.101 and MGWR's 1213.815; for Cd, it's 1103.878, smaller than OLS's 1289.371, GWR's 1187.045, and MGWR's 1143.068. Based on $R^2$ and AICc values, the MGWR - VSLR model has higher accuracy and a better fitting effect than GWR and MGWR models in analyzing soil heavy metal content in the Yangtze River Basin. It is more suitable for studying the relationship between soil heavy metal content and related factors, providing a more reliable basis for exploring heavy metal sources, migration, and transformation.

### 3.3. Analysis of Influencing Factors of Soil Heavy Metal Content

This study used the multiscale geographically weighted spatial lag model to systematically reveal spatial differentiation characteristics and multiscale influencing factors of heavy metal pollution such as As, Cr, and Cd. Compared with traditional Geographically Weighted Regression (GWR), MGWR-VSLR not only quantifies the spatial dependence (spatial lag effect) of pollutants but also identifies differences in the scale of action of each driving factor by introducing spatial lag terms and variable-specific bandwidth parameters.

(1) Altitude and Slope

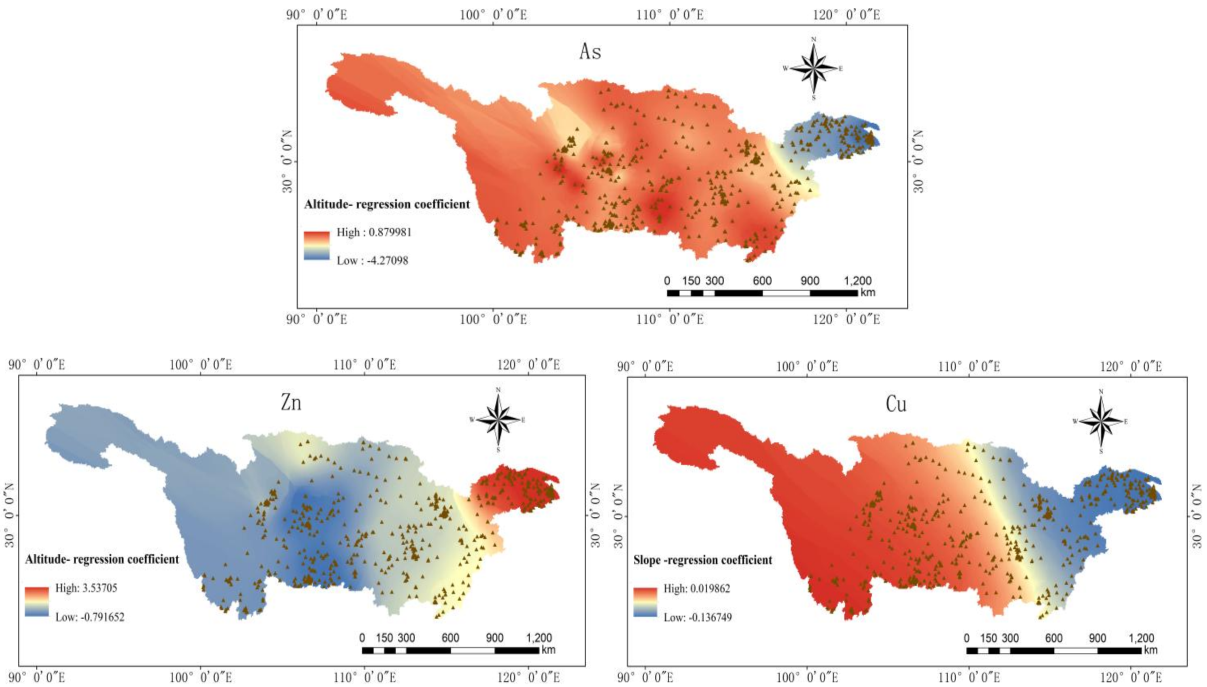



**Figure 4.** Regression Coefficients of As, Zn-Altitude Variables and Cu-Average Slope Variable.

The OLS model (see Figure 4) shows a single negative correlation for altitude - As, indicating its limitation. The MGWR - VSLR reveals altitude's spatial non - stationarity on As. As content is "east - high west - low", opposite to Zn. Their distribution trends with altitude are reverse. For Cu, slope's impact is bidirectional, with different effects in west and east [17].

(2) Precipitation

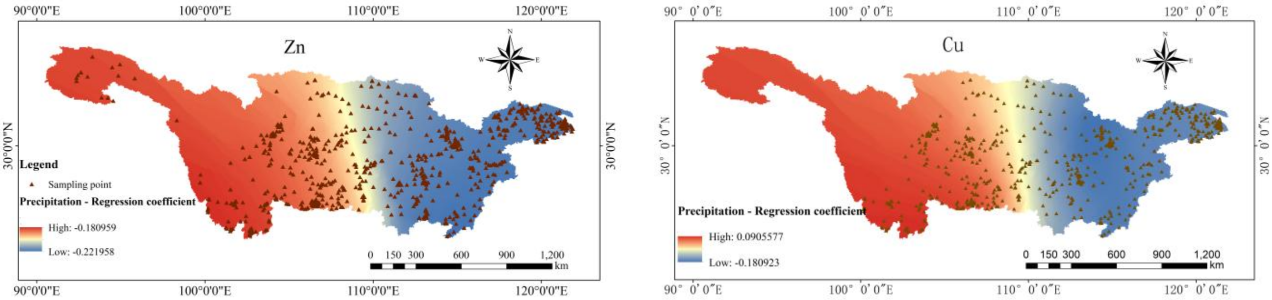



**Figure 5.** Regression Coefficients of Zn, Cu-Precipitation Variables.

OLS regression's precipitation coefficients can't explain spatial heterogeneity. For Cu and Zn, OLS shows altitude impact coefficients of β = 0.0039 and β = -0.0041(see Figure 5). MGWR - VSLR shows Cu's precipitation regression is positive in the west (max 0.0906) and negative in the east (min -0.1809). Precipitation has an overall negative correlation with Zn, with significant east - west differences (-0.222 to -0.181), and its leaching effect on $Zn^{2+}$ [18] weakens from west to east.

(3) Temperature

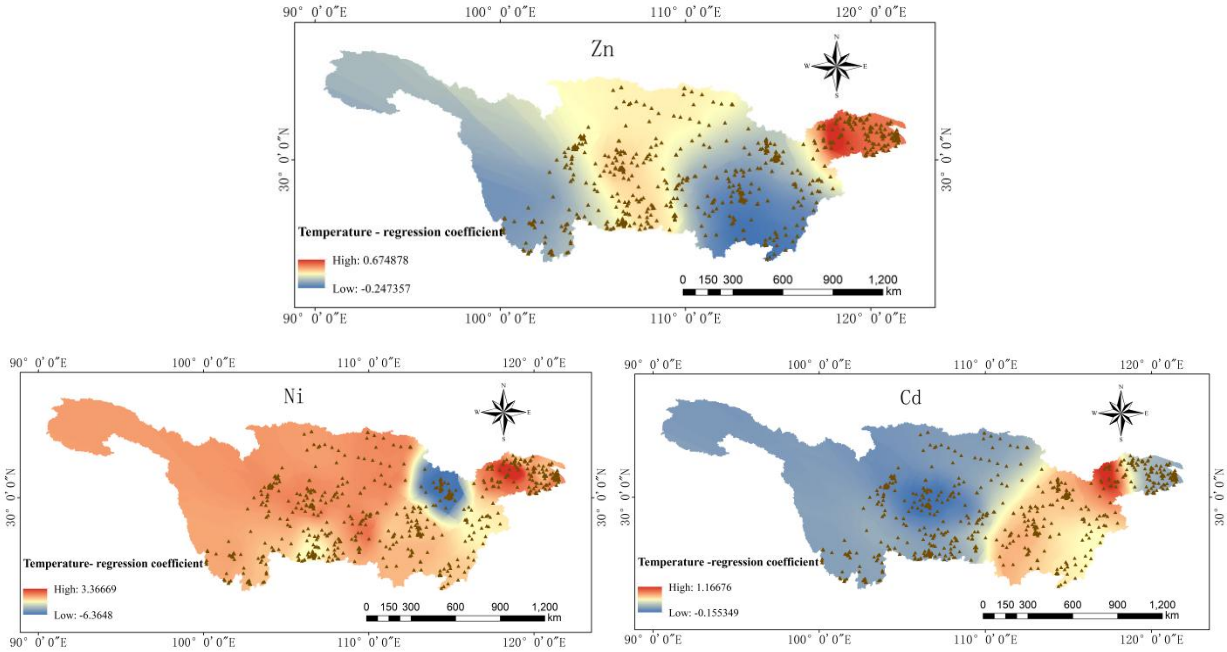



**Figure 6.** Regression Coefficients of Zn, Ni, Cd-Temperature Variables.

For Zn, Ni, and Cd soil heavy metal contents, temperature regression coefficients are highest in the Jianghuai region (see Figure 6). Its location in the climate transition zone with large temperature fluctuations may enhance the impact on heavy metal content, as temperature affects microbial activity and heavy metal behavior.

(4) Proportions of Forest Land and Urban Land

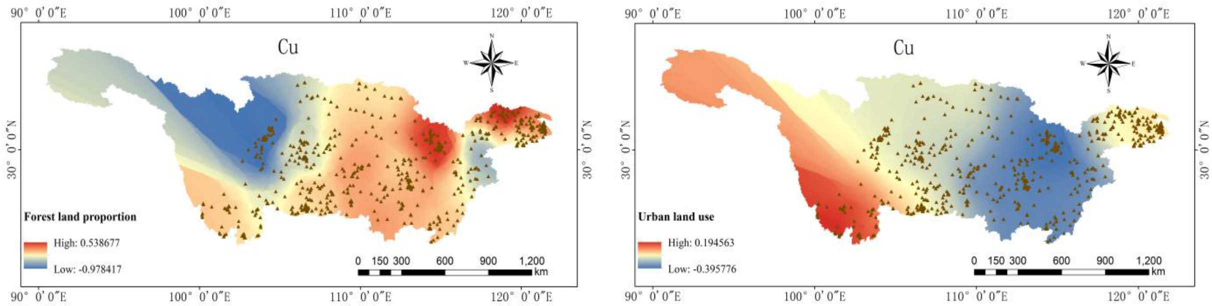



**Figure 7.** Regression Coefficients of Cu-Forest Land and Urban Land Variables.

Forest and urban land show a complementary relationship (see Figure 7). Urban land proportion (-0.360 to 0.195) has bidirectional effects: positive in traditional industrial areas and negative in emerging urban areas. Forest land proportion (0.539 to 0.878) is positively correlated with Cu content, likely due to overlap with mining areas.

(5) Farmland and Wetland

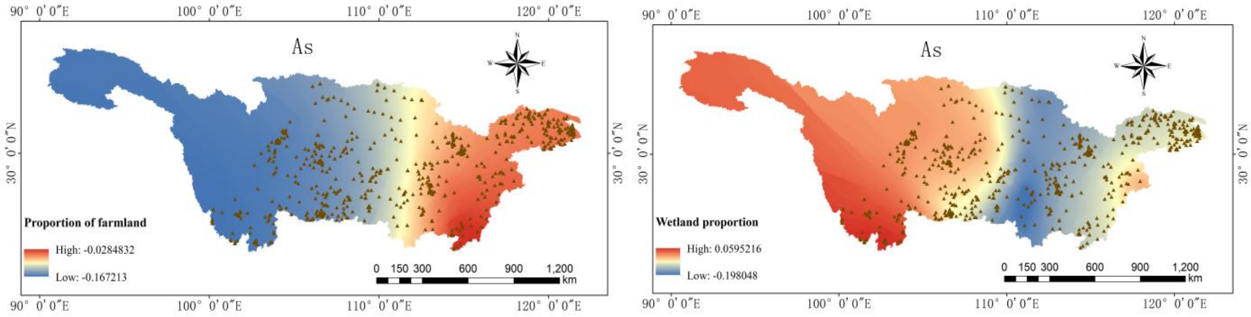



**Figure 8.** Regression Coefficients of As-Farmland and Wetland Proportion Variables.

The As - farmland proportion regression coefficient range is -0.167213 to -0.0284832(see Figure 8). Higher farmland proportion may reduce heavy metal pollution as modern farming uses less heavy - metal - laden organic fertilizers and some crops absorb heavy metals [19]. The coefficient increases eastward. Wetlands(-0.199 to 0.060) have dual effects, with eastern wetlands reducing pollution.

(6) Population Density and Gross Regional Product.

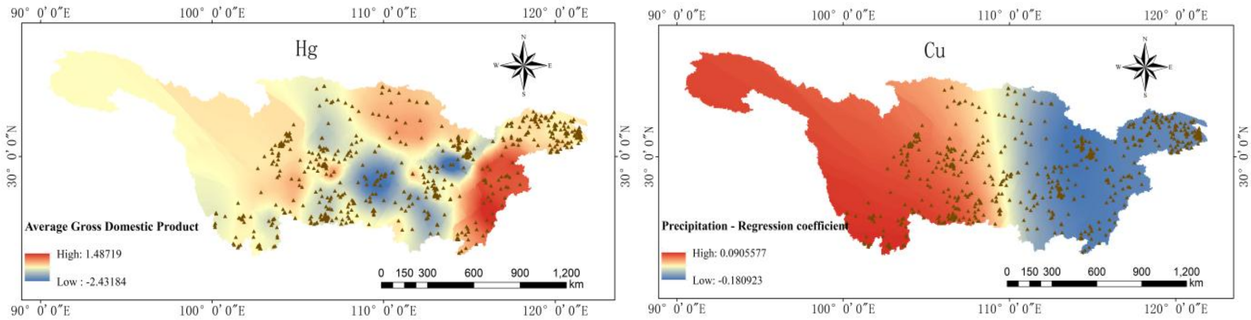



**Figure 9.** Regression Coefficients of Hg-GDP and Cd-Population Density Variables.

The GDP regression coefficient for Hg pollution (-2.4318 to 1.4872) shows an “east - south high and northwest low” trend, with higher Hg in the southeastern developed regions (see Figure 9). The population density coefficient for Cd pollution (-0.2346 to 0.00584) indicates lower Cd in some high - population - density areas, possibly due to stricter environmental governance.

## 4. Conclusions

Based on the MGWR - VSLR model, this study revealed the spatial differentiation and driving mechanisms of soil heavy metal pollution in the Yangtze River Basin. The model, with spatial lag terms and variable - specific bandwidths, improved goodness - of - fit and prediction accuracy. Results have theoretical and practical significance. They enable accurate pollution risk prediction, guide optimal resource allocation, and offer scientific reference for territorial planning and industrial layout, promoting high - quality development and ecological civilization construction.

## References


[1] Cronie, O., van Lieshout, M. N. M. Non - parametric adaptive bandwidth selection for kernel estimators of spatial intensity functions[J]. Advances in Data Analysis and Classification, 2023, (7): 1 - 21.

[2] Liu, L. L., Li, F. Y., Cheng, J. H., et al. Distribution characteristics of lead - zinc deposits in China[C]//Chinese Geophysical Society. Proceedings of the First National Conference on Mineral Exploration. School of Geography,

Nanjing Normal University; Key Laboratory of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University; Jiangsu Collaborative Innovation Center for Geographical Information Resource Development and Application, 2021: 345 - 351.

[3] Fosgerau, M., Horváth, P., Raveaux, P. Modelling Nonstationary Spatial Lag Models with Hidden Markov Random Fields[J]. Computational Statistics & Data Analysis, 2021, 159: 107196.

[4] Páez, A., Uchida, T., Miyamoto, K. A general framework for estimation and inference of geographically weighted regression models: 2. Spatial association and model specification tests[J]. Environment and Planning A: Economy and Space, 2002, 34(5): 883 - 904.

[5] Tang, Z. P., Wu, Y., Xiong, S. F., et al. A mixed geographically weighted regression model with variable coefficient spatial lag[J]. Journal of University of Chinese Academy of Sciences (Chinese and English), 2024, 41(03): 345 - 356.

[6] Li, H., Fu, P., Yang, Y., et al. Exploring spatial distributions of increments in soil heavy metals and their relationships with environmental factors using GWR[J]. Stoch Environ Res Risk Assess, 2021, 35: 2173 - 2186.

[7] Sun, Y., Guo, Q., Liu, J., et al. Scale Effects on Spatially Varying Relationships Between Urban Landscape Patterns and Water Quality[J]. Environmental Management, 2014, 54: 272 - 287.

[8] Wang, Y., Su, B. D., Jiang, T., et al. Changes in nitrogen and phosphorus loads in Fuhe River Basin under climate and socio - economic changes[J]. Yangtze River, 2024, 55(02): 65 - 76.

[9] Brunsdon, C., Fotheringham, A. S., Charlton, M. Some notes on parametric significance tests for geographically weighted regression[J]. Journal of Regional Science, 1999, 39(3): 497 - 524.

[10] Yang, X. Y., Zheng, F. W., Lei, X., et al. Distribution characteristics and sources of heavy metals in sediments of Fuhe River Basin in the middle and lower reaches of the Yangtze River[J]. Journal of Nanchang Institute of Technology, 2023, 42(06): 29 - 37.

[11] Huang, Y. H., Yang, X. M., Yue, J. T. Analysis of urban road cycling flow based on multiscale geographically weighted regression model[J]. Journal of Tsinghua University (Natural Science Edition), 2022, 62(07): 1132 - 1141.

[12] Zhai, H. Y. Study on spatiotemporal changes of vegetation coverage and driving factors in Fuhe River Basin[D]. East China University of Technology, 2020.

[13] Mei, C. L., Cheng, J. Y., Xu, Q. X. Estimation of mixed geographically weighted regression model based on coefficient average and identification of coefficient types[J]. Chinese Journal of Engineering Mathematics, 2025, 42(02): 199 - 219.

[14] Sun, Y., Liu, D., Xie, L., Gao, Z., Zhang, Q., Wang, L., et al. A georeferenced dataset of heavy metals occurrence in the soils of the Yangtze River Basin, China[J]. Geoscience Data Journal, 2025, 12: e280.

[15] Gao, Y., Wang, G. L., Jin, Z. H., et al. Spatial distribution, risk and influencing factors of heavy metals in water - sediments in the lower reaches of Qianhe River[J]. Environmental Science, 2021, 42(11): 5333 - 5345.

[16] Bhandari, R., et al. Sedimentary contamination study in the Great Chao Phraya River watershed, Thailand, using geographically weighted regression[J]. International Journal of Digital Earth, 2024, 17(1).

[17] Li, R., Lu, J., Xu, W. Y., et al. Overview of research on absorption and enrichment mechanisms of heavy metals by plant roots[J]. Modern Agricultural Research, 2025, 31(03): 1 - 7.

[18] Ren, J. Study on the source of acid wastewater from alum mines and the migration law of heavy metals in soil[D]. Xi'an University of Technology, 2021.

[19] Hu, R. N. Study on pollution levels and adsorption - desorption behaviors of heavy metals and microplastics in sediments of typical shallow lakes[D]. Hebei University, 2023.