Comparison of Different Spatial Regression Methods for Predicting Organic Matter Content in Soil

Xumeng Zhang
College of Resources and Environment, Shanxi Agricultural University, Taigu 030801, China
Wuping Zhang (zwping@126.com)
College of Software, Shanxi Agricultural University, Taigu 030801, China
Mingjing Huang
Dryland Agriculture Research Center, Shanxi Agricultural University, Taiyuan 030031, China;
Li Gao
College of Resources and Environment, Shanxi Agricultural University, Taigu 030801, China
Lei Qiao
College of Resources and Environment, Shanxi Agricultural University, Taigu 030801, China
Guofang Wang
College of Resources and Environment, Shanxi Agricultural University, Taigu 030801, China
Linlin Yan
College of Resources and Environment, Shanxi Agricultural University, Taigu 030801, China

Research Article

Keywords: Spatial Regression Methods, Predicting Organic Matter Content, Soil, prediction accuracy, soil organic matter, technical support

Posted Date: January 6th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1220228/v1

Comparison of Different Spatial Regression Methods for Predicting Organic Matter Content in Soil

Xumeng Zhang[1], Wuping Zhang[2, **], Mingjing Huang[3,4], Li Gao[1], Lei Qiao[1], Guofang Wang[1], Linlin Yan[1]

1. College of Resources and Environment, Shanxi Agricultural University, Taigu 030801, China.

2. College of Software, Shanxi Agricultural University, Taigu 030801, China.

3. Dryland Agriculture Research Center, Shanxi Agricultural University, Taiyuan 030031, China.

4. State Key Laboratory of Integrative Sustainable Dryland Agriculture（in preparation）, Shanxi Agricultural University, Shanxi Institute of Organic Dryland Farming，Taiyuan 030031, China

* College of Software, Shanxi Agricultural University, Taigu 030801, China. Correspondence and requests for materials should be addressed to W.P. Zhang (email:zwping@126.com)

Dynamic changes in soil organic matter content affects the sustainable supply of soil water and fertilizer and impacts the stability of soil ecological function. Understanding the spatial distribution characteristics of soil organic matter will help deepen our understanding of the differences in soil organic matter content, soil formation law; such understanding would be useful for rational land use planning. Taking terrain data, meteorological data, and remote sensing data as auxiliary variables and the ordinary Kriging (OK) method as a control, this study compares the spatial prediction accuracies and mapping effects of various models (MLR, RK, GWR, GWRK, MGWR, and MGWRK) on soil organic matter. Our results show that the spatial distribution trend of soil organic matter predicted by each model is similar, but the prediction of composite models can reflect more mapping details than that of unitary models. The OK method can provide better support for spatial prediction when the sampling points are dense; however, the local models are superior in dealing with spatial non-stationarity. Notably, the MGWR model is superior to the GWR model, but the MGWRK model is inferior to the GWRK model. As a new method, the prediction accuracy of MGWRK reached 47.72% for the OK and RK methods and 40.08% for the GWRK method. The GWRK method achieved a better prediction accuracy. The influence mechanism of soil organic matter is complex, but the MGWR model more clearly reveals the complex nonlinear relationship between

soil organic matter content and factors influencing it. This research can provide reference methods and mapping technical support to improve the spatial prediction accuracy of soil organic matter.

Soil organic matter (SOM) is a crucial component of the soil carbon pool [1-3]. It features plant growth and is heavily utilized to evaluate soil quality [4]. SOM is affected by managerial, topographic, and climatic factors, as well as others [5, 6]. Studies on the spatial distribution of SOM are helpful for understanding regional soil quality, and have significance in the improvement of spatial organic matter content and land resource management [7, 8]. The spatial distribution of SOM usually has spatial variation due to the influence of structural and random factors; that is, the existence of spatial non-stationary characteristics leads to uncertainty in the modeling process [9]. Therefore, it is difficult to accurately simulate the spatial distribution of SOM.

There are numerous methods for predicting SOM; however, the commonly used method is the ordinary Kriging (OK) method, which can simulate the distribution of soil organic matter in the entire region through limited point data. It fully considers spatial variability and theoretically interpolates errors. However, the OK method is dependent on the quality of the sampling point and is restricted without considering the influence of other factors [10]. The relationship between soil properties and environmental variables is an extremely complicated non-linear relationship [11], and the accurate

expression of this relationship vital to improving the prediction accuracy.

Therefore, auxiliary variables are introduced into the prediction model. Chen et al. [12] introduced surface albedo to participate in the spatial prediction modeling of SOM. Li et al. [13] used 11 environmental factors from climate, terrain, and vegetation combined with principal component analyses as predictors of SOM. In addition to considering quantitative auxiliary variables, Li et al. [14] integrated qualitative auxiliary variables, such as soil texture. The results proved that the introduction of auxiliary variables can significantly improve the prediction accuracy of SOM. Additionally, many studies have proved that the introduction of terrain, vegetation, and climate factors as auxiliary variables can significantly improve the prediction accuracy of soil properties [15-17].

Among various prediction methods with auxiliary factors, Lopez [18] used the multiple linear regression (MLR) method to combine the blue bands of remote sensing images and soil properties to predict the spatial distribution of SOM, pH, and K. Duan et al. used the CoKriging method combined with auxiliary variables to predict SOM [19]. The results showed that the correlation coefficient (r) increased from 0.1 to 0.70 and 0.69. Hengl et al. showed that regression Kriging (RK) combined with the digital elevation model(DEM) was more accurate than ordinary Kriging (OK) in predicting SOM and tillage thickness [20]. However, the spatial non-stationary characteristics of soil attributes limit the accuracy of global models such as the ordinary least

squares (OLS) model. The geographically weighted regression (GWR) model, a local model widely used in the field of economic geography, has been introduced into the study of soil attribute heterogeneity, and has gradually become a vital tool for such research.

Niu et al. proved that the introduction of a GWR model greatly improved the interpretation of driving factors compared with the traditional OLS model [21]. In a study of prediction methods for farmland SOM in the Weibei dryland area, Yu et al. indicated that the correlation coefficient R^2 increased to 0.902 using the GWR model [22]. Ding et al. used the GWR model to explore the spatial distribution of soil organic carbon (SOC) and its influencing factors in the Yihe River basin [23]; these results showed that the GWR model simulated the spatial distribution of SOC and analyzed the different effects of different factors on different terrain units.

Geographically weighted regression kriging (GWRK) is a combination of the GWR and OK models, and is an improvement of the RK model. GWRK optimizes the residual error of the GWR model and reflects local details masked by global models such as RK. Yang et al. [24] studied the spatial distribution of SOM based on the GWRK model. Compared to the RK model, the GWRK model considered the location of the sample points, refined the mapping effect, and improved the overall accuracy. Liu et al. [25] compared several different spatial geographic technologies to predict SOC reserves and concluded that the GWRK model significantly reduces the error of the model

and is an effective method for mapping SOC storage in a local range. Ye conducted research that indicated that GWRK can be regarded as the best method for predicting SOC when the data samples are sufficient [26].

The multi-scale geographic weighted regression model (MGWR) was proposed by STWART et al. in 2017 [27]. Each variable in the model has a multi-scale optimal bandwidth. MGWRK is analogous to the GWRK model, which not only considers the multiple scales of the MGWR model, but also optimizes the regression residuals of the model. The model is worth exploring in terms of predicting soil properties and mapping.

In previous studies, it was necessary for scholars to explore the application of different models in soil spatial attributes, which provides mentalities for the development of spatial prediction and mapping technology, but it is still confined to the selection of environmental factors, the construction of models, and the spatial scale of different variables [28-30]. The complexity of the external environment determines that the applicability of each model is different. Under different environments, the prediction performance conclusions of different SOM models are not universal. For the spatial prediction of research objects, we should explore it and select the most appropriate method. In this study, ten factors in three categories—terrain data, meteorological data, and remote sensing data—were selected as environmental covariates for modeling, and different spatial prediction methods (OK, MLR, RK, GWR, GWRK, MGWR, and MGWRK) were

compared to provide technical support and a theoretical basis to accurately predict the spatial distribution of SOM and reasonable improvement of soil organic matter content in cultivated land.

Materials and Methods

Study area. Lingshi County is in the central part of Shanxi Province, the southernmost tip of Jinzhong City (110° 20′ to 112° E, 36° 40′ to 37° N). It is part of the Loess Plateau, with undulating mountains and ravines. The county covers an area of 1206 km^2. The region consists of mountains on the East and West and the Fenhe Valley in the middle, which naturally forms a three-level ladder: the middle mountain area, hilly area, and Pingchuan Valley, of which the hilly area accounts for the largest proportion of the total area (57.6 %). The study area has a warm temperate continental climate, which is windy in the spring, hot and rainy in the summer, cool and refreshing in the autumn, and cold and a little snowy in the winter. The average annual temperature is 10.8 °C, the average annual precipitation is 473.5 mL, and the frost-free period is 156-day-long. The entire territory has 246 km^2 of cultivated land and is abundant in biological and mineral resources.

Soil Data Collection and Auxiliary Variables Sources. A total of 2245 sample points were obtained before crops were planted in the autumn of 2009. According to the “Rules for soil quality survey and assessment” (NY/t1634-2008), the “s” method was used to arrange points in each field. The

sampling depth was 0–20 cm, and the sampling points are shown in Figure 1. SOM was determined using the external heating potassium dichromate method [31]. The digital elevation model (DEM) was obtained from the Geospatial Data Cloud website (https://www.gscloud.cn) and each terrain factor was extracted from the DEM. Annual mean precipitation (PRE) and annual mean temperature (TEM) were extracted from meteorological datasets, and the annual net primary productivity (NPP) and normalized differential vegetation index (NDVI) were extracted from remote sensing datasets, both of which were obtained from the Resource and Environment Science and Data Center (https://www.resdc.cn). The environmental covariates are presented in Table 1.

Methods for estimating SOM. *Ordinary Kriging (OK)*. OK is based on the spatial autocorrelation and second-order stationary hypothesis [32], using the semivariance function for calculation and analysis according to the principle of minimum error to realize the optimal unbiased linear estimation of the unknown region by using the data of known sampling points in a limited region. The semivariance function formula is defined as follows.

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \quad (1)$$

where $\gamma(h)$ is the semivariogram value, h is the lag between the two locations, $N(h)$ is the total number of point pairs when the interval is equal to h, and $Z(x_i)$ and $Z(x_i + h)$ are the measured values of the regional variables at x_i and $x_i + h$, respectively.

Multiple Linear Regression (MLR) & Regression Kriging (RK) RK combines MLR and OK interpolation [33]. The RK result consists of two parts: the trend item and the residual item. The trend item was obtained using the MLR model. The regression method used in this study is the input method. The MLR formula was expressed as follows:

$$y_i = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + \varepsilon_i \qquad (2)$$

where y_i is the predicted value of the model, β_0 is the intercept, β_i is the regression coefficient, x_i is the independent variable, and ε_i is the model residual.

The residual generated by the prediction of the multiple linear regression model is estimated using the OK method, and the residual item is obtained. The trend and residual items were added to obtain the prediction result of the RK model. The RK formula was calculated as follows:

$$\mathrm{Z} = \mathrm{m}(x_i) + \mathrm{e}(x_i) \qquad (3)$$

where Z is the predicted value of the RK model, $\mathrm{m}(x_i)$ is the predicted item obtained by the MLR model, and $\mathrm{e}(x_i)$ is the residual item composed of the MLR model residual estimated by OK.

Geographically Weighted Regression (GWR) & Geographically Weighted Regression Kriging (GWRK). GWRK is a combination of the GWR model and the OK method, which is similar to RK [34], and the trend item in the RK method is obtained by a regression item. The trend item in the GWRK method was simulated by the GWR model. The RK model is a global regression model,

while GWRK is a local regression model that is considered an improvement of the former. The relationship between variables is not considered in the modeling of MLR models, which are affected by spatial location, whereas GWR deals with the problem of spatial non-stationary characteristics by introducing the spatial location of sampling points and establishing local regression equations in the spatial range [35]. The regression coefficients of the GWR model change according to different spatial locations.

If there are n observation points in total, the position of each observation point is (u_i, v_i), and there are m variables involved in the modeling, then the GWR model is expressed as follows:

$$y_i = \sum_{i=0}^{n} \sum_{j=0}^{m} \beta_j(u_i, v_i) x_{ij} + \varepsilon_i \qquad (4)$$

where x_{ij} is the j th variable of observation point i, $\beta_j(u_i, v_i)$ is the regression coefficient of the j th independent variable at the position of (u_i, v_i), ε_i is the random error term, and y_i is the dependent variable.

In GWR analysis, the Gaussian kernel and bisquare kernel functions are generally used to determine the spatial weight [36]. In our model analysis, different types of kernel functions were used according to the distribution of different observation points. When the distribution of the observation points is uniform, a fixed kernel function is used. When the distribution of observation points was uneven, the effect of the adaptive kernel function was better. Owing to the uneven distribution of sampling points in this study, an adaptive Gaussian kernel function was used for modeling. In addition, there is another

important indicator in the GWR model: effective bandwidth. An optimal bandwidth is vital for the accurate description of spatial features. The effective bandwidth is a smooth parameter: if the effective bandwidth is too wide, the parameters of the entire study area may become too similar and cover up the local details of the variables; if the effective bandwidth is too narrow, it will cause too many local spatial variations, and it is difficult to identify the spatial laws between variables. The determination methods of effective bandwidth include Akaike information criterion (AIC), corrected Akaike information criterion (AICc), Bayesian information criterion (BIC), and others. GWR analysis is sensitive to the bandwidth selection of specific weight function. Therefore, this study used the AICc to determine the model bandwidth. The model simulation is based on the AICc minimization principle.

Multi-scale Geographically Weighted Regression (MGWR) & Multi-scale Geographically Weighted Regression Kriging (MGWRK). MGWRK is a combination of MGWR models and the OK interpolation method. MGWRK is an improvement of the GWRK method, in which all independent variables use the same effective bandwidth and the same spatial weight matrix. In the MGWR model, each independent variable has a different bandwidth. It considers the different spatial scales of spatial non-stationary characteristics between different independent variables, that is, multiple scales. The MGWR model is described as follows [27]:

$$y_i = \sum_{i=0}^{n}\sum_{j=0}^{m}\beta_{bwj}(u_i, v_i)x_{ij} + \varepsilon_i \quad (5)$$

where β_{bwj} is the regression coefficient corrected by the effective bandwidth of the j th independent variable.

Model Evaluation. The mean error (ME), mean absolute error (MAE), root mean square error (RMSE), and model prediction accuracy (ACC) were used to evaluate the model accuracy [37, 38], which were calculated as follows:

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}\left[Z(x_i) - \hat{Z}(x_i)\right] \quad (6)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left[\left|Z(x_i) - \hat{Z}(x_i)\right|\right] \quad (7)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[Z(x_i) - \hat{Z}(x_i)\right]^2} \quad (8)$$

$$\mathrm{ACC} = \left[1 - \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{Z}(x_i) - Z(x_i)}{Z(x_i)}\right|\right] \quad (9)$$

where $\boldsymbol{n}$ is the number of observation points used for verification, $\boldsymbol{Z(x_i)}$ is the measured value, and $\boldsymbol{\hat{Z}(x_i)}$ is the predicted value of the model. ME, MAE, and RMSE values nearer to zero and higher ACC values indicate a better prediction effect.

Results

Selection of Environmental Covariates. Many environmental factors affect SOM. In this study, ten indicators, including slope, aspect, TRI, PC, TWI, elevation, TEM, PRE, NPP, and NDVI were selected to study the impact of

each index on SOM. The spatial distribution of environmental covariates is shown in Figure 2, and the descriptive statistics of the environmental covariates are shown in Table 2.

The coefficient of variation (CV) is the ratio of the standard deviation to the mean, which can be used to reflect the spatial variation level of each index. When CV < 10%, there is a weak variation; when 10% < CV < 100%, it shows moderate variation; and when CV > 100%, there is a strong variation [39]. Table 2 shows that the coefficient of variation in some indicators belongs to weak variability, including elevation, TEM, PRE, and NDVI; among the indicators that belong to moderate variability, there are slope, aspect, TRI, and TWI; likewise, NPP, SOM, and PC are strongly variable.

In the process of modeling, selecting multiple variables may lead to the problem of multicollinearity between variables; that is, there is a high correlation between different variables, resulting in information redundancy, reducing model accuracy, and increasing calculation time. Therefore, the necessary condition for modeling each regression model is that there is no multicollinearity between the independent variables. Generally, the variance inflation factor (VIF) and tolerance are used to diagnose whether the data are multicollinear. Tolerance is the reciprocal of the variance inflation factor. When VIF$\geqslant$10 and tolerance$\leqslant$0.1, there is a serious collinearity problem between the variables. In this study, VIF and tolerance were used to diagnose the collinearity problem of these models, and the results are shown in Table 3. The

VIF of each variable was less than 10, and the tolerance was greater than 0.1, therefore, there were no serious collinearity problems in the model.

The final fitting model of MLR is:

SOM=16.654+0.023*Slope+0.002*Aspect-0.025*TRI+1.667*PC+0.049*TWI-0.007*Elevation-0.571*TEM-0.001*PRE+0.005*NPP+2.180*NDVI (10)

where SOM is soil organic matter, TRI is topographic relief index, PC is plane curvature, TWI is topographic wetness index, TEM is annual mean temperature, PRE is annual mean precipitation, NPP is the annual NPP, and NDVI is the annual NDVI. The variance analysis of the MLR model shows that f = 2.936, P = 0.001, the model is statistically significant [39], and the regression equation is effective. GWR and MGWR also use the same variables to build the model.

Comparison of MLR, GWR, and MGWR Models in Terms of Accuracy. Table 4 compares the effects of the MLR, GWR, and MGWR models. RSS, AICc, and R2 are typically used to evaluate the performance of the model. The smaller the RSS and AICc, and the larger the R2 value, the higher the model accuracy. Combining various accuracy indicators, we concluded that the MGWR model was the most accurate, the GWR model was the second most accurate, and the MLR model was the least accurate.

OK, RK, GWRK, and MGWRK Modeling. The residuals of each model and

SOM were fitted with a spatial semivariogram, and the model parameters and optimal models are shown in Table 5. The best-fitting model for MGWR residuals is the spherical model, whereas the fitting models of other types are exponential models. The nugget value reflects the spatial heterogeneity caused by random errors in the sampling scale. The sill value represents the total variation intensity of the variables in the study area. A greater sill value indicates higher total heterogeneity in the system. The model range reflects the maximum distance of the spatial autocorrelation. The range of each model is between 240 and 640 m, the range of the MLR model residual is the smallest, and the range of the residual MGWR model is the largest. Nugget/sill is the ratio of the nugget value to the sill value, which is close to 1, indicating that the variation in these variables within the study area is constant. Generally, a nugget/sill value > 75% indicates that the spatial variation is mainly affected by random factors. A value between 25% and 75% indicates that the variable has a moderate spatial autocorrelation. The variable is affected by the regional structure factors and random factors. A value of < 25% indicates that the variables have strong spatial autocorrelation, mainly structural variation [32]. The nugget/sill of the OK, MLR, GWR, and MGWR residuals were 14.87%, 13.56%, 13.98%, and 7.72%, respectively, all of which were less than 25%, indicating strong spatial autocorrelations. The residuals of each model were interpolated using their own semivariogram to obtain the spatial distribution map of the residuals (Figure 3).

316 The spatial distribution of MLR residuals was different from that of the
317 other two models. The MLR residuals are generally low in the west, southwest,
318 and southeast, and the high and low values clearly cluster. The spatial
319 distribution of the GWR model residuals is a staggered distribution of high and
320 low values. The staggered distribution of the MGWR model residuals is more
321 obvious than that of the GWR model, and the phenomenon of “fragmentation”
322 is more serious than that of the GWR model. The aggregation degree of the
323 high and low values of residual spatial distribution in the MGWR model is
324 weaker than that in the MLR and GWR models. From the residual distribution
325 characteristics of these three models (Figure 3), the residual range is also
326 different. The MLR residual range was the largest, while the MGWR residual
327 range was the smallest. That is, the overall error caused by the MLR model
328 was the largest, and the overall error caused by the MGWR model was the
329 smallest. Most of the residual value ranges of the MGWR model are stably
330 distributed between -2.23 and 1.25. From the perspective of residuals, the
331 MGWR model performed the best, followed by GWR and MLR, respectively.
332
333 **Spatial Prediction of Soil Organic Matter Based on Different Methods**
334 Based on the spatial prediction of organic matter by different methods (Figure
335 4), the spatial distribution trends of SOM predicted by different methods are
336 basically the same. The spatial pattern of SOM in the study area was low in the
337 southwest and southeast and high in the middle and east. The prediction of

SOM using the MLR model has the smallest value range. It is difficult to reflect the spatial distribution detail of SOM, and the MLR model has the most information distortion, followed by the GWR model. the “fragmentation” phenomenon of OK model and MLR model are also obvious. The mapping effects of the RK, GWRK, MGWR, and MGWRK models are smoother than those of the OK and MLR models, and the MGWR model performs well in a single model. Prediction and mapping effects of GWRK and MGWRK are more refined, which provides more details than other models.

Model Accuracy Evaluation The accuracy of the seven prediction models was evaluated by calculating five indices: mean error (ME), mean absolute error (MAE), root mean square error (RMSE), R^2, and evaluation accuracy (ACC). ME is an index to measure the deviation degree of model predictions, the closer it is to 0, the better the prediction effect. According to ME, the models ranked accordingly from best prediction effect to worst: MGWRK> GWRK>RK > OK > GWR > MLR. MAE and RMSE are indicators for measuring the prediction accuracy of the model. RMSE is sensitive to particularly large or small errors in the prediction results. The smaller the MAE and RMSE values, the better the prediction effect of the model was. The MAE values from small to large are GWRK> OK > RK > MGWRK > MGWR> GWR > MLR; and the RMSE values from small to large are GWRK> RK> OK > MGWRK> MGWR> GWR > MLR. R^2 indicates the degree of coincidence

between the predicted value and the real value of the model. The order of R^2 values from large to small are GWRK> OK > RK> MGWRK> MGWR> GWR > MLR. In addition to the above four indicators, ACC is also introduced to evaluate model accuracy. ACC values closer to 1 indicate better model accuracies. According to the ACC index, the accuracy of the models rank accordingly from best to worst: GWRK > RK > OK > MGWRK> MGWR> GWR > MLR. Except for the unstable performance of individual models in the evaluation indicators, the performance trends of the models are generally the same, with the best performance of GWRK and the worst performance of MLR.

Spatial Distribution of Regression Coefficients of Soil Organic Matter Influencing Factors Based on MGWR Model. According to the MLR equation of SOM, SOM is positively correlated with slope, aspect, PC, TWI, NPP, and NDVI, and negatively correlated with TRI, elevation, TEM, and PRE, and there was only one regression coefficient on the global scale. In fact, the influential extent of influencing factors on SOM cannot be constant in space, and the relationship between dependent variables and independent variables will change with geographical location, that is, with spatial non-stationarity [40]. The MGWR model performed better than the GWR model. Therefore, the MGWR model was used to study the spatial distribution of the standardized regression coefficients of different influencing factors. As shown in Figure 5, the regression coefficients and positive or negative correlations between

382 different influencing factors and SOM are different at different spatial locations.
383 The MGWR model considers the nonlinear relationship between SOM and
384 influencing factors.
385

386 ## Discussion

387 When exploring the accuracy of each regression model, the MLR model,
388 as a global model, only considers the numerical relationship between variables
389 and does not consider that the relationship between variables will also change
390 owing to the change in spatial location. In the MLR model, although the
391 regression equation obtained is statistically significant, the regression equation
392 obtained does not have an effective guiding significance. MLR models have a
393 low model fitting degree and ignore the spatial non-stationary characteristics
394 between SOM and its environmental auxiliary variables at the sampling sites;
395 therefore, its accuracy proved the worst, compared to the other models
396 examined. As local models, GWR and MGWR models can be regarded as an
397 improvement of the MLR model. They consider the non-stationary
398 characteristics between variables. The AICc values of the MLR, GWR, and
399 MGWR models are 15 016.906, 14 973.07, and 14 872.466, respectively. R^2
400 increased slightly from 0.101 (GWR) to 0.162 (MGWR). Although the fitting
401 degree is still low, in large-scale research areas, the accuracy is greatly
402 improved when compared with the MLR global model (0.013). The MGWR
403 model has a higher precision than the GWR model owing to its use of multiple

scales; the variables in the MGWR model have different bandwidths, but each variable of the GWR model shares one bandwidth. The MGWR model can better explain the relationship between the independent and dependent variables than the GWR model.

In the residual distribution range among models, that of the MGWR model is the smallest overall, and the interleaving of high and low values is not as serious as that of the MLR and GWR models, indicating that the multi-scale function of the MGWR model plays an obvious role in reducing the spatial residual. Because the effective bandwidth between various variables in the GWR model is the same, the phenomenon of fragmentation is more serious than in the MGWR model. As the global model, MLR has the largest residual range, which shows that the local regression model can better reduce the residual item of the regression model than the global model.

In terms of the visualization effect of the spatial prediction of SOM, the mapping of the RK model is smoother than that of the other hybrid models. The reason may be that the MLR model only retains the spatial trend of large-scale SOM in the prediction process, and its prediction effect is poor in narrow space and small scale. The MLR model cannot reflect variations in local details and excludes too many local features, and the mapping of the MLR model has high information distortion. Comparing the GWR and MGWR models, the latter has a better mapping effect. These two models are superior to the MLR model because they consider the spatial non-stationary characteristics of SOM. The

MGWR, GWRK, and MGWRK models are based on the improvement of the local regression model, and the model details perform well. The MGWR and MGWRK models also consider the scale problem based on local theory. MGWRK is the next step, and it shows more details in the spatial prediction distribution of the model than the other six models.

Seven models were evaluated and discussed. Comparing the index evaluation accuracy, ACC, of each model, the GWRK method is the best. As a common spatial prediction method, OK also achieves good accuracy, which is flush with the RK model. Although the MGWR model is superior to the GWR model, its accuracy is inferior to that of the GWRK model. Combined with the analysis of the residual distribution map, although the local regression model better reduces the residual term of the regression model, multi-scale is both an advantage and an interference in spatial prediction. The prediction accuracy of MGWRK reached 47.72% for the OK method, 47.72% for the RK method, and 40.08% for the GWRK method. Many studies show that the GWRK model achieves better prediction accuracy in most cases because the GWR model is more mature [38, 41]. As a new improved model, although the accuracy of MGWR model needs to be improved, it still provides a new idea for spatial prediction and has a great exploration value.

In terms of the spatial distribution of regression coefficients, the regression coefficients of most environmental variables have the same correlation in the overall environment of the study area; only the strength of the regression

coefficients is different in space. It is inferred that TRI, TEM, and PRE may have a strong effect on the change in SOM according to the size of the regression coefficients. The regression coefficients of TRI, TWI, and NPP are mixed with positive and negative values. The spatial positive and negative correlations between SOM and these three indicators were different. The regularity of this situation is not obvious, which may be related to the complex landforms and vegetation distribution in Lingshi County. Slope, aspect, and plane curvature are positively correlated with SOM, which may restrict human activities and reduce the interference of human factors and SOM [42]. Elevation and PRE are negatively correlated with SOM, which may be because the lower the altitude, convenient transportation, flat terrain, and the application of chemical fertilizer and organic fertilizer provided good conditions to the SOM, so the content of SOM is high [7, 43]; due to the influence of leaching, the rainfall is greater, and more SOM will be lost [44].

Although the new MGWRK method improves the prediction accuracy of the model in part, this improvement is not obvious and fails to achieve the expected effect. The operation cost of the MGWRK model may be higher than that of OK; when the study area is large and the sampling point density is high, the OK method can also fully provide good support for spatial prediction, and the process of mapping is simpler and quicker, but in front of the manpower and material resources saved by accurately predicting SOM, the operation cost of the model can be ignored. Composite models may have higher

accuracy than unitary models in theory; they are not as mature as common models and do not have as widespread usage. However, the MGWR model has the advantage of revealing spatial non-stationary characteristics that other models do not. Although the application effect of MGWRK model in soil spatial prediction is unsatisfactory, it can still be used as a new progress in the application of regression method in spatial prediction, but it needs to explore more suitable application scenarios.

The prediction accuracy of the regression model was highly sensitive to the correct selection of auxiliary variables. There are many factors influencing SOM through a complex mechanism. It is crucial to scientifically and reasonably select the influencing factors of SOM to improve the accuracy of the model.

The study only selected some environmental variables as auxiliary variables for modeling and did not consider human factors. The spatial variation in SOM is also affected by human activities and other factors. In a follow-up study, the cultivation management measures should be quantified and included in the scope of model measurements to further improve the influence mechanism of SOM, to improve the accuracy of simulation and prediction.

After the standardization of the MGWR model, this study only makes a visual study on the regression coefficient, and it does not involve an in-depth examination of the local R^2 and other indicators. Subsequent studies can

deeply excavate other indicators of the standardized MGWR model.

Conclusion

In summary, this study compared the performance of seven different methods (OK, MLR, RK, GWR, GWRK, MGWR, and MGWRK) in the spatial prediction of SOM in Lingshi County, Jinzhong City, Shanxi Province, and applied the MGWR model to explore the factors influencing cultivated SOM.

In terms of the spatial prediction of SOM, the general trend of these models is the same, only differing in the granular degree of the mapping effect. The overall spatial pattern of SOM in the study area was low in the southwest and southeast, and high in the middle and east. The prediction accuracy of the GWRK model was the highest, and that of the MLR model was the worst. The accuracy of the RK model was greatly improved after the residual of the MLR model was estimated using the OK method. In general, the hybrid models (RK, GWRK, and MGWRK) were better than the single models (MLR, GWR, and MGWR).

The MGWR model achieves a different spatial distribution of regression coefficients, breaks the limitation of global correlation coefficients, uses different optimal bandwidths for each variable, which is more in line with the actual situation, and more accurately describes the spatial nonstationary characteristics of each variable. It also shows a complex nonlinear relationship between SOM and the influencing factors.

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

References

1. Sun, Y., Wang ,W., Li, G. Spatial distribution of forest carbon storage in Maoershan region，Northeast China based on geographically weighted regression kriging model. *Chin. J. Appl. Ecol.* **30,** 1642-1650 (2019).
2. Wang, S., Zhou, C., Li, K., Zhu, S., Huang, F. Analysis on Spatial Distribution Characteristics of Soil Organic Carbon Reservoir in China. *Acta Geogr. Sin.* 533-544 (in Chinese) (2000).
3. Simbahan, G.C., Dobermann, A., Goovaerts, P., Ping, J., Haddix, M.L. Fine-resolution mapping of soil organic carbon based on multivariate secondary data. *Geoderma* **132,** 471-489 (2006).
4. Wu, T., Schoenau, J.J., Li, F., Qian, P., Zhang, S., Malhi, S.S., Wang, F. Concepts and relative analytical techniques of soil organic matter. *Chin. J. Appl. Ecol.* 717-722 (in Chinese) (2004).
5. Zamulina, I.V., Gorovtsov, A.V., Minkina, T.M., Mandzhieva, S.S., Bauer, T.V. Soil organic matter and biological activity under long-term contamination with copper. *Environ. Geochem. Health.* (2021).
6. Zhang, J.; Sayer, E.J.; Zhou, J.; Li, Y.W.; Li, Y.X.; Li, Z.; Wang, F. Long-term fertilization modifies the mineralization of soil organic matter in response to added substrate. *Sci. Total Environ.* 798 (2021).
7. Yang, J.; Han, X.; Huang, J.; Pan, Q. The dynamics of soi lorganic matter in cropland responding to agricultural practices. *Acta Ecol. Sin.* 787-796 (in Chinese) (2003).
8. Hanli, Q.; Xiubo, S.; Huazhou, C.; Jingyi, L.; Shaoyong, H. Effective prediction of soil

organic matter by deep SVD concatenation using FT-NIR spectroscopy. *Soil Tillage Res.* **215,**(2022): doi:10.1016/J.STILL.2021.105223.

9. Zhao, M., Liu, B., LV, H., Li, D., Zhang, G. Spatial modeling of soil organic matter over low relief areas based on geographically weighted regression. *Trans. Chin. Soc. Agric. Eng.* **35,** 102-110 (in Chinese) (2019).
10. Eldeiry, A.A., Garcia, L.A. Comparison of Ordinary Kriging, Regression Kriging, and Cokriging Techniques to Estimate Soil Salinity Using LANDSAT Images. *ASCE J. Irrig. Drain. Div.* **136,** 355-364 (2010).
11. Mcbratney, A.; Santos, M.; Ma, B.J.G. On Digital Soil Mapping. *Geoderma.* **117,** 3-52 (2003).
12. Chen, F., Qin, F., Li, X., Peng, G. Inversion for spatial distribution of soil organic matter content based on multivariate geostatistics. *Trans. Chin. Soc. Agric. Eng.* **28,** 188-194+297 (2012).
13. Li, Q., et.al. Spatially distributed modeling of soil organic matter across China: An application of artificial neural network approach. *Catena.* **104,** 210-218 (2013). doi:10.1016/j.catena.2012.11.012.
14. Li, Q., et.al. Prediction of distribution of soil organic matter based on qualitative and quantitative auxiliary variables: a case study in Santai County in Sichuan Province. *Prog.Geogr.* **33,** 259-269 (2014).
15. Lian, G., Guo, X., Fu, B., Hu, C. spatial variability and prediction of soil nutrients on a county scale on the loess plateau—a case study of heng shan county,Shaanxi province. *Acta Pedol. Sin.* 577-584 (in Chinese) (2008).

16. Qi, Y., Wang, Y., Chen, Y., Liu, J., Zhang, L. Soil organic matter prediction based on remote sensing data and random forest model in Shaanxi province. *J. Nat. Resour.* ***32*****,** 1074-1086 (in Chinese) (2017).
17. Zhang, S., et.al. Prediction of spatial distribution of soil nutrients using terrain attributes andremote sensing data. *Trans. Chin. Soc. Agric. Eng.* **26,** 188-194 (In Chinese) (2010).
18. López-Granados, F., Jurado-Expósito, M., Using geostatistical and remote sensing approaches for mapping soil properties. *Eur. J. Agron.* **23,** 279-289 (2005).
19. Duan, L., Guo, L, Zhang, H., Ju, Q. Prediction of spatial distribution of soil organic matter based on improved OK models: A case study of Honghuatao Town in Yidu City. *Chin. J. Eco-Agric.* **27,** 131-141 (in Chinese) (2019).
20. Hengl, T., Heuvelink, G., Stein, A. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma.* ***120*****,** 75-93 (2004).
21. Niu, L., et.al. Research on China's surface urban heat island drivers and its spatial heterogeneity. *China Environ. Sci.***2021**, 1-11.
22. Yu, f., Liu, J., Xia, L., Xu, Z., Long, X. Spatial Prediction Method of Farmland Soil Organic Matter in Weibei Dryland of Shaanxi Province. *Chin. J. Environ. Sci.* 1-13(in Chinese) (2021).
23. Ding, Y., et.al. Spatial distribution characteristics and influencing factors of soil organic carbonin Yihe River Basin based on GWR model *Acta Ecol. Sin.***41,** 4876-4885 (in Chinese) (2021).
24. Yang, S.-H., et.al. Mapping topsoil electrical conductivity by a mixed geographically

602 weighted regression kriging: A case study in the Heihe River Basin, northwest China.
603 *Ecol. Indic.* ***102,*** 252-264(2019). doi:10.1016/j.ecolind.2019.02.038.
604 25. Liu, Y., Guo, L., Jiang, Q., Zhang, H., Chen, Y. Comparing geospatial techniques to
605 predict SOC stocks. *Soil Tillage Res.****148,***46-58(2015).doi:10.1016/j.still.2014.12.002.
606 26. Ye, H.; Huang, W.; Huang, S.; Huang, Y.; Zhang, S.; Dong, Y.; Chen, P. Effects of
607 different sampling densities on geographically weighted regression kriging for
608 predicting soil organic carbon. *Spatial Stat.* ***20,*** 76-91 (2017).
609 doi:10.1016/j.spasta.2017.02.001.
610 27. Fotheringham, A.S., Yang, W., Kang, W. Multiscale Geographically Weighted
611 Regression (MGWR). *Ann. Am. Assoc. Geogr.****107,***1247-1265 (2017).
612 doi:10.1080/24694452.2017.1352480.
613 28. Qiu, L.; Yang, C.; Lin, F; Yang, N.; Zheng, X.; Xu, H.; Wang, K. Spatial pattern of soil
614 fertility in Bashan tea garden: A prediction based on environmental auxiliary variables.
615 *Chin. J. Appl. Ecol.* ***21,*** 3099-3104 (in Chinese) (2010).
616 29. Jiang, Y.; Li, Q.; Zhang, X.; Liang,W. Kriging prediction of soil zinc in contaminated
617 field by using an auxiliary variable. *Chin. J. Appl. Ecol.* 97-101 (in Chinese) (2006).
618 30. Zeng, C.; Yang, L.; Zhu, A.X.; Rossiter, D.G.; Liu, J.; Liu, J.; Qin, C.; Wang, D.
619 Mapping soil organic matter concentration at different scales using a mixed
620 geographically weighted regression method. *Geoderma.* ***281,*** 69-82 (2016).
621 doi:10.1016/j.geoderma.2016.06.033.
622 31. Bao, S. Soil agrochemical analysis *.3 rd ed.*; China Agriculture Press: Beijing, China,
623 25-38 (in Chinese) (2000).

624 32. Liu, A.; Wang,P.; Ding, Y. *Introduction to Geostatistics*; Science Press: Beijing,
625 China, 82 (in Chinese) (2012).
626 33. Odeh, I.; Mcbratney, A.B.; Chittleborough, D.J. Further results on prediction of soil
627 properties from terrain attributes: heterotopic cokriging and regression-kriging.
628 *Geoderma* ***67*****,** 215-226 (1995).
629 34. Yang, S.; Zhang, H.; Guo, L.; Ren, Y. Spatial interpolation of soil organic matter using
630 regression Kriging and geographically weighted regression Kriging. Chinese. *Chin. J.*
631 *Appl. Ecol.* ***26*****,** 1649-1656 (in Chinese) (2015).
632 35. Griffith, D. Spatial-filtering-based contributions to a critique of geographically
633 weighted regression (GWR). *Environment Planning A* **40,** 2751-2769 (2008).
634 36. Fotheringham; A.; S.; Charlton; M. Geographically weighted regression: a natural
635 evolution of the expansion method for spatial data. *Environment Planning A.* **30,**
636 1905-1927 (1998).
637 37. Yang, S., et.al. Predicting soil organic matter content in a plain-to-hill transition belt
638 using geographically weighted regression with stratification. *Arch. Agron. Soil Sci.* **65,**
639 1745-1757 (2019). doi:10.1080/03650340.2019.1576171.
640 38. Gao, L., et.al. Comparative Study on Spatial Digital Mapping Methods of Soil
641 Nutrients Based on Different Geospatial Technologies. *Sustainability* ***13*****,** (2021)
642 doi:10.3390/su13063270.
643 39. Liu, H.; Mao, W.; Modern statistics. 2nd ed.; Sun Yat-sen University Press:
644 Guangzhou, China (in Chinese) (2002).
645 40. Wang, X, et.al. comparing Interpolation Methods to Predict the Spatial Distribution of

Soil Available Phosphorus in the Mollisol Watershed of Northeast China. *Res. Soil Water Conserv.* ***28*****,** 33-40 (2021).

41. Qiao, L., Zhang, W., Huang, M., Wang, G., Ren, J. Mapping of Soil Organic Matter and Its Driving Factors Study Based on MGWRK. *Chin. Agric. Sci.* ***053*****,** 1830-1844 (2020).
42. Liu, J., Xu, X., Simulating the effects of different management modes on the ecosystem services of agroecosystems: a case study of the Taihu Lake Basin, China. *Acta Ecol. Sin.* ***39*****,** 9314-9324 (in Chinese) (2019).
43. Hu, Z., Liu, S., Shi, Z., Liu, X., He, F. Variations of Soil Nitrogen and Microbial Biomass Carbon and Nitrogen of Quercus aquifolioides Forest at Different Attitudes in Balangshan，Sichuan. *Forest Research* ***25*****,** 261-268 (2012).
44. Pingshun, G., Mingzhu, H., Guangzu, B., Pengshan, Z., Qingxia, G. Spatial distribution characteristics of soil organic matter in Chinese herbal medicine planting area in Longxi County. *Agricultural Research in the Arid Areas.* ***38*****,** 192-197+251 (in Chinese) (2020).

Acknowledgements

This study was funded by by State Key Laboratory of Integrative Sustainable Dryland Agriculture（in preparation）, Shanxi Agricultural University (No. 202105D121008-1-6)

Author Contributions

Conceptualization, X.Z.; Methodology, X.Z., L.G. and L.Q.; Software, X.Z. and

L.G.; Validation, X.Z. and G.L.; Formal Analysis, G.W. and L.Y.; investigation, X.Z., L.G. and L.Q.; Resources, W.Z.; Data Curation, X.Z.; Writing—Original Draft Preparation, X.Z.; Writing—Review & Editing, W.Z. and X.Z.; Visualization, X.Z.; Supervision, W.Z.; Project Administration, W.Z. and M.H.; Funding Acquisition, W.Z. and M.H. All authors have read and agreed to the published version of the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: https://www.resdc.cn (acceded on 23 April 2021); http://www.gscloud.cn (acceded on 23 April 2021).

690 # Table and Fig.

Datatype	Index	Extraction methods
Terrain data	Aspect Slope	Extracted from DEM through ArcGIS surface analysis function
	Topographic relief index, TRI	TRI=E max- E min E max Maximum elevation E min Minimum elevation
	Plan curvature, PC	Extracted from DEM through ArcGIS surface analysis function
	Topographic wetness index, TWI	Extracted from DEM through SAGA-GIS terrain analysis function
	Elevation	Extracted from DEM through ArcGIS spatial analysis tool
Meteorological	Annual mean precipitation, PRE Annual mean temperature, TEM	Interpolation software is ANUSPLIN （www.resdc.cn）
Remote sensing data	The Annual Net primary productivity, NPP	Calculated based on GLO_PEM model。Unit: gC/m3（www.resdc.cn）
	The Annual Normalized differential vegetation index, NDVI	Calculated based on MVC method Through SPOT/VEGETATION NDVI remote sensing data, NDVI is the maximum value of monthly NDVI in a year.（www.resdc.cn）

691 **Table 1.** Environmental covariates and exaction method

692

693

694

695

696

Index	Unit	Min	Max	Mean	S.D.	Variance	Coefficient of Variance(%)
Soil organic matter	g/kg	0.10	48.00	4.82	6.87	47.18	142.62
Slope	°	0.00	31.48	9.84	4.96	24.59	50.39
Aspect	°	-1.00	359.21	189.81	98.19	9641.52	51.73
Topographic relief index		7.00	150.00	54.97	20.11	404.47	36.59
Plan curvature		-0.71	1.10	0.01	0.17	0.03	1240.56
Topographic wetness index		4.86	22.04	7.29	2.40	5.76	32.95
Elevation	m	674.00	1385.00	978.71	93.69	8778.63	9.57
Annual mean temperature	℃	8.98	12.38	10.74	0.49	0.24	4.55
Annual mean precipitation	mm	483.29	576.15	529.13	17.57	308.77	3.32

The Annual NPP	gC/㎡	71.20	505.40	223.90	63.81	4071.88	28.50
The Annual NDVI		0.46	0.73	0.64	0.04	0.00	6.17

697 **Table 2.** Descriptive statistics of soil organic matter and environmental
698 covariates
699
700
701

Index	Tolerance	VIF
Slope	0.369	2.711
Aspect	0.934	1.071
Topographic relief index	0.432	2.312
Plan curvature	0.866	1.154
Topographic wetness index	0.543	1.841
Elevation	0.195	5.120
Annual mean temperature	0.136	7.370
Annual mean precipitation	0.291	3.433
The Annual NPP	0.834	1.199
The Annual NDVI	0.720	1.389

702 **Table 3.** Multicollinearity test of models.
703

Model	RSS	AIC	AICc	R2	Adjust R2
MLR	104493.078	15014.766	15016.906	0.013	0.009
GWR	95172.649	14968.948	14977.07	0.101	0.063
MGWR	88684.105	14859.383	14872.466	0.162	0.116

704 **Table 4.** Accuracy diagnosis of regression models.
705

Type	Fitting-Model	Nugget(C0)	Sill(C0+C)	Range(m)	Nugget/Sill(C0/C0+C)
OK	Exponential model	7.1	47.75	250	14.87
MLR Residual	Exponential model	6.4	47.2	240	13.56
GWR Residual	Exponential model	6.1	43.63	290	13.98
MGWR Residual	Spherical model	3.1	40.13	640	7.72

706 **Table 5.** Semivariance parameters of each model.
707

Model	ME	MAE	RMSE	R2	ACC
OK	0.0364	3.8997	5.5864	0.4065	-0.8298
MLR	-0.3714	5.3039	6.8970	0.0122	-1.6960
RK	0.0229	3.9039	5.5777	0.4064	-0.8275
GWR	0.0912	4.7905	6.5767	0.1058	-1.2373
GWRK	0.0127	3.7022	5.2862	0.4840	-0.7369
MGWR	-0.1003	4.5931	6.3483	0.1630	-1.2032

MGWRK	-0.0099	4.4036	6.2235	0.1940	-1.0842

Table 6. Comparison of prediction accuracy of seven models.

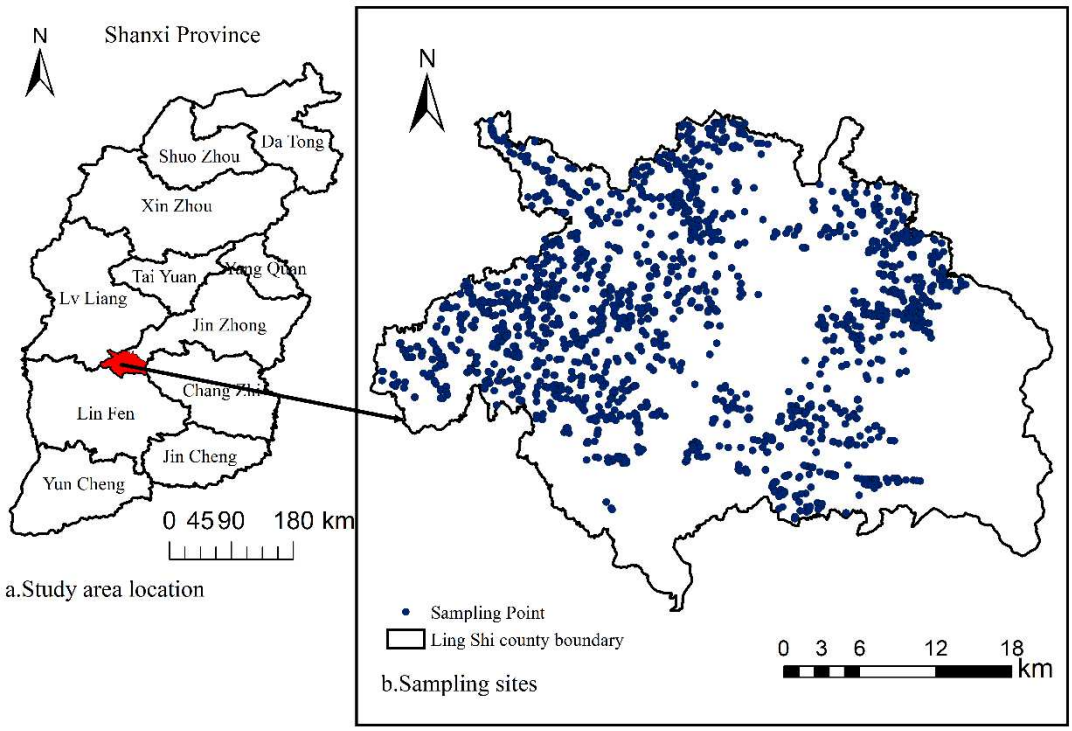

Figure 1. The study area and the sampling point distribution

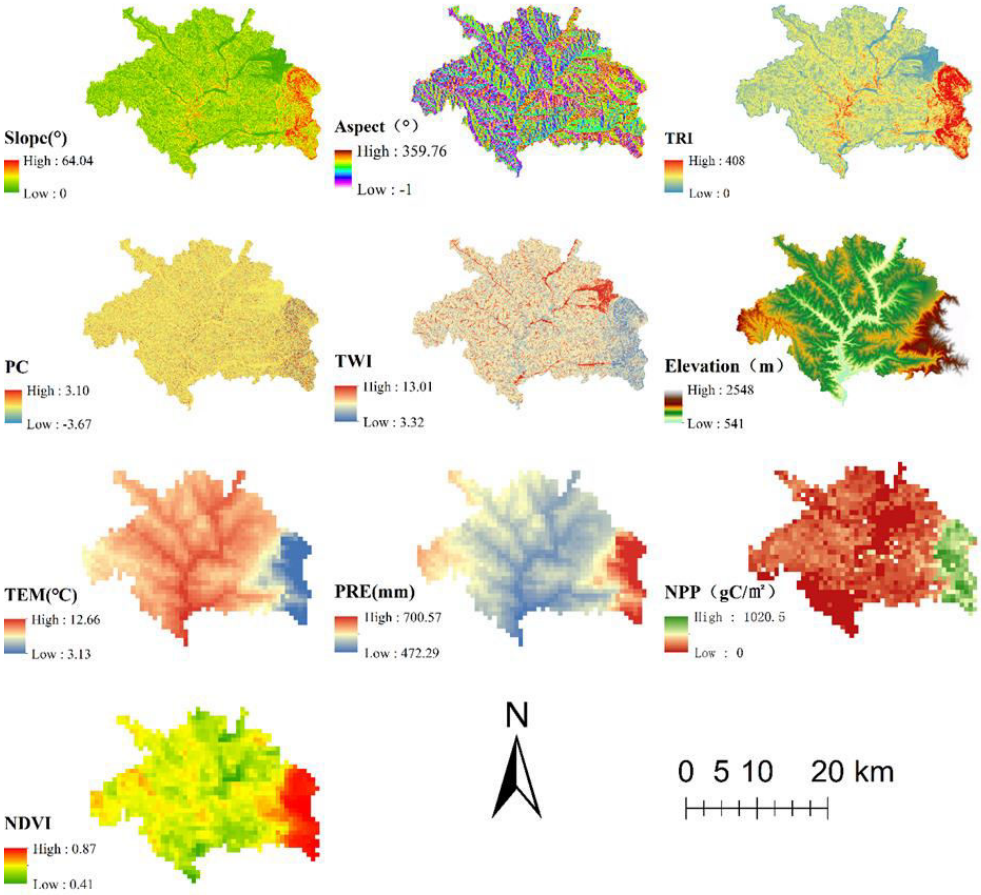

Figure 2. Spatial distribution of environmental covariates

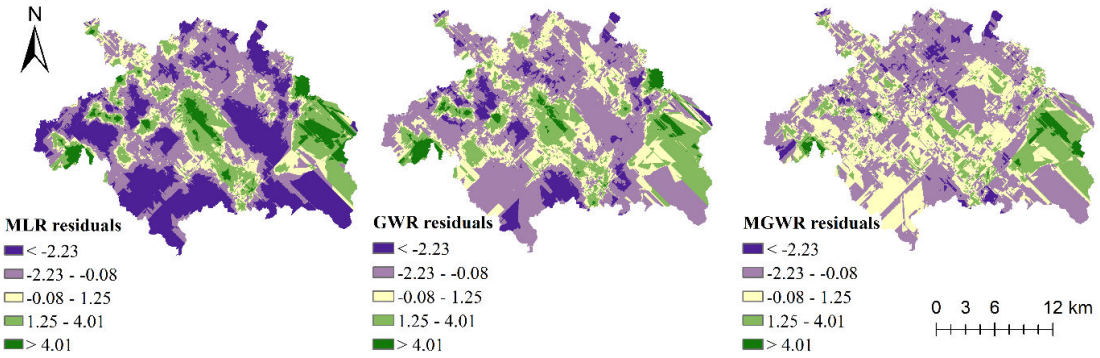

Figure 3. Spatial distribution map of MLR, GWR, and MGWR residuals

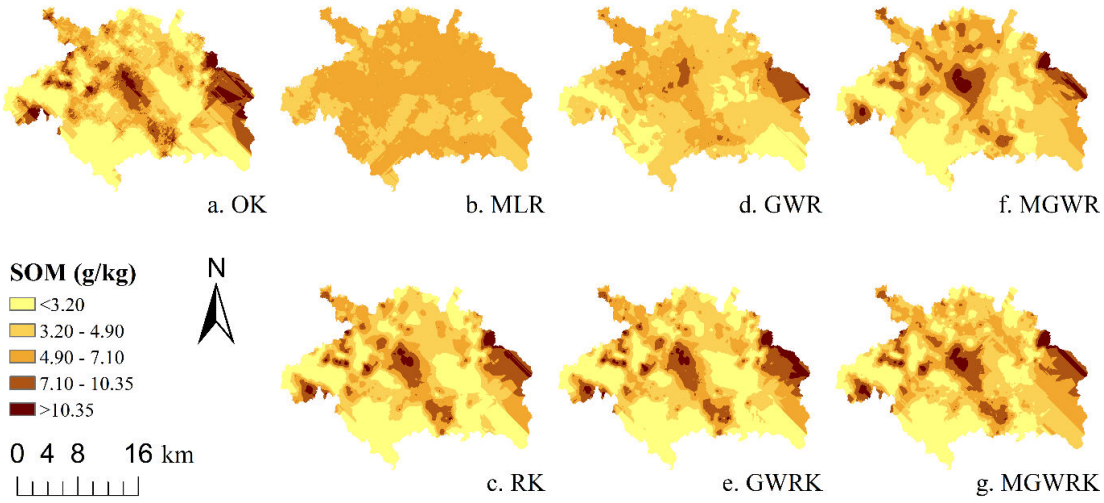

Figure 4. Prediction of spatial distribution of soil organic matter by different methods

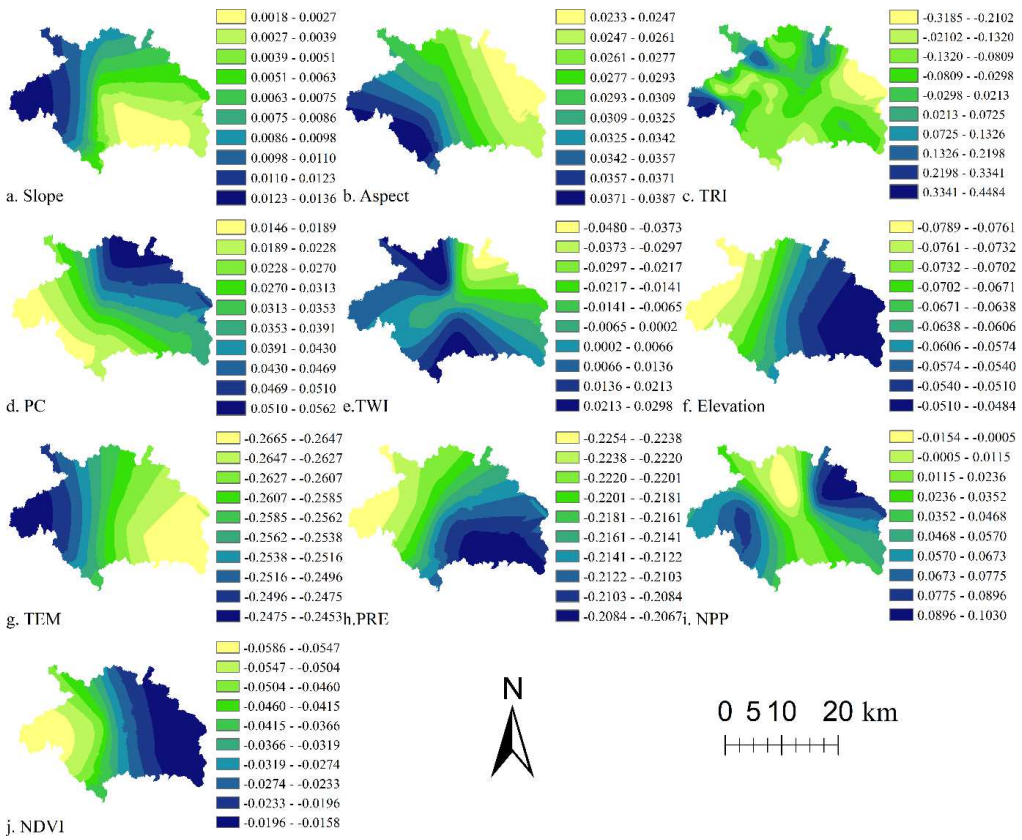

Figure 5. Spatial distribution of normalized regression coefficients of influencing factors based on MGWR