

## Accepted Manuscript

Effects of different sampling densities on geographically weighted regression kriging for predicting soil organic carbon

Huichun Ye, Wenjiang Huang, Shanyu Huang, Yuanfang Huang, Shiwen Zhang, Yingying Dong, Pengfei Chen

PII: S2211-6753(17)30040-4

DOI: <http://dx.doi.org/10.1016/j.spasta.2017.02.001>

Reference: SPASTA 206

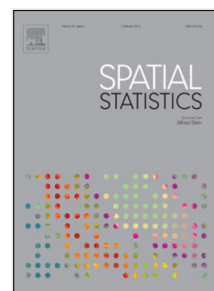
To appear in: *Spatial Statistics*

Received date: 20 April 2016

Accepted date: 8 February 2017

Please cite this article as: Ye, H., Huang, W., Huang, S., Huang, Y., Zhang, S., Dong, Y., Chen, P., Effects of different sampling densities on geographically weighted regression kriging for predicting soil organic carbon. *Spatial Statistics* (2017), <http://dx.doi.org/10.1016/j.spasta.2017.02.001>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



1        1        Effects of Different Sampling Densities on Geographically Weighted  
2  
3  
4        2        Regression Kriging for Predicting Soil Organic Carbon  
5  
6  
7  
8

9        3        Huichun Ye <sup>a,b</sup>, Wenjiang Huang <sup>a,b,\*</sup>, Shanyu Huang <sup>c</sup>, Yuanfang Huang <sup>d</sup>, Shiwen Zhang <sup>e</sup>,

10  
11        4        Yingying Dong <sup>a</sup>, Pengfei Chen <sup>f</sup>  
12  
13  
14

15  
16        5        <sup>a</sup> *Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese*  
17  
18        6        *Academy of Sciences, Beijing 100094, P.R.China*  
19  
20

21        7        <sup>b</sup> *Key Laboratory of Earth Observation, Hainan Province, Sanya 572029, P.R.China*  
22  
23

24        8        <sup>c</sup> *Institute of Geography, University of Cologne, Köln 50923, Germany*  
25  
26

27        9        <sup>d</sup> *College of Resources and Environment, China Agricultural University, Beijing 100193,*  
28  
29        10        *P.R.China*  
30  
31

32        11        <sup>e</sup> *College of Earth and Environment, Anhui University of Science and Technology, Huainan*  
33  
34        12        *232001, P.R.China*  
35  
36

37  
38        13        <sup>f</sup> *State Key Laboratory of Resources and Environment Information System, Institute of Geographic*  
39  
40        14        *Science and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101,*  
41  
42        15        *P.R.China*  
43  
44  
45  
46

47        16  
48        17  
49        18        Huichun Ye, E-mail: [yehc@radi.ac.cn](mailto:yehc@radi.ac.cn), Tel: 86-10-82178178, Fax: 86-10-82178177  
50  
51

52        20        \* Communicating author: Wenjiang Huang, E-mail: [huangwj@radi.ac.cn](mailto:huangwj@radi.ac.cn), Tel:  
53        21        86-10-82178169, Fax: 86-10-82178177  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## Effects of Different Sampling Densities on Geographically Weighted Regression Kriging for Predicting Soil Organic Carbon

### Abstract

Geographically weighted regression kriging (GWRK) is a popular interpolation method, considering not only spatial parametric non-stationarity and relationship between target and explanatory variables, but also spatial autocorrelation of residuals. However, little attention has been paid to the effects of different sampling densities on GWRK technique for estimating soil properties. Objectives of this study were: (i) comparing the GWRK predictions with those obtained from multiple linear regression kriging (MLRK) and ordinary kriging (OK), and (ii) examining how different sampling densities affect the performance of GWRK for predicting soil organic carbon (SOC). Soil samples were simulated with four sampling densities, including 0.010, 0.020, 0.041, and 0.082 sites/km<sup>2</sup>. The results showed that GWRK made less prediction errors and outperformed MLRK and OK in the case of a high sampling density, with the root mean squared errors of GWRK < MLRK < OK and coefficient of determination of GWRK > MLRK > OK. However, in the case of a low sampling density, GWRK generated larger prediction errors, exhibiting a poorer performance than MLRK and OK. Accordingly, we conclude that GWRK can be considered as the best approach for predicting SOC in these three approaches with sufficient data points, but it has a poorer performance than the other methods with sparse data points.

## Keywords

42

43 Sampling density

44 Geographically weighted regression

45 Geographically weighted regression kriging

46 Soil organic carbon

47 Spatial variation

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## 1. Introduction

Soil is the final product of the interaction of the relief (altitude, aspect, and slope of a terrain) with organisms, parent material and climate over time. Soil properties exhibit spatial heterogeneity on different spatial scales, but also spatial dependence within a certain spatial range (Jackson and Caldwell, 1993; Ettema and Wardle, 2002; Dale and Fortin, 2014; Miller et al., 2016). Sampling design is the main method for obtaining the spatial distribution of soil characteristics, and directly affects the accuracy of soil mapping (Brus and Noij, 2008; Heim et al., 2009; Arrouays et al., 2012; Zhu et al., 2015). However, collecting a large number of samples at a regional scale is unrealistic as it would consume large amounts of labor, material and financial resources. Fortunately, some soil properties are spatially related to environmental factors (Minasny et al., 2006; Chai et al., 2008; Kumar et al., 2012; Zhang et al., 2012; Ye et al., 2015, 2016). With the increasing power of geographical information system (GIS), global positioning system (GPS) and remote sensing (RS) in recent years, more and more high resolution maps, such as remote sensing images, digital elevation models (DEMs) and their derivative terrain attributes, can be readily obtained and used as ancillary data to predict spatial distribution of soil properties (McBratney et al., 2003). Thus, many studies have determined the correlative relationships between soil characteristics and environmental covariates, and integrated these relationships with geostatistical techniques for further soil mapping or for the design of optimal sampling strategies (Odeh et al., 1995; McKenzie and Ryan, 1999; Rawlins et al.,

69 [2009; Gasch et al., 2015; Sun et al., 2015; Miller et al., 2016; Somarathna et al., 2016;](#)  
70 [Song et al., 2016](#)). In these studies, the most widely used approaches were regression  
71 kriging (RK) method, which due to the fact that they combine trends fitted by global  
72 regression with kriged residuals, have shown to make a better use of the available data  
73 and thereby improved the accuracy of the predictions compared to single regression  
74 methods and kriging methods ([Bishop and McBratney, 2001; Chai et al., 2007;](#)  
75 [Kumar et al., 2012](#)). For instance, [Chai et al. \(2008\)](#) predicted the soil organic matter  
76 (SOM) content at a county scale using the multiple linear regression kriging (MLRK)  
77 method after multiple regression analysis of SOM and its different environmental  
78 variables (elevation, slope and topographic wetness index), and then suggested that  
79 combining MLRK with topographical data could further improve the accuracy of  
80 SOM spatial predictions. Moreover, [Zhang et al. \(2012\)](#) successfully introduced  
81 categorical variables (land use types, soil texture and soil genetic types) as auxiliary  
82 variables for MLRK to predict the SOM content. However, one major limitation of  
83 these approaches is that the relationships between target soil variables and  
84 environmental covariates are assumed to be stationary across space. This assumption  
85 cannot be adapted to fit data locally using neighborhoods ([Walter et al., 2001; Kumar](#)  
86 [et al., 2015](#)). Therefore, some type of non-stationarity model that is able to capture the  
87 relationship variability with absolute locations across space is needed ([Lloyd, 2010;](#)  
88 [Lark, 2012](#)).

89 Geographically weighted regression (GWR) is an extension of the traditional  
90 regression framework and, in contrast to global models, allows for spatially varying

1 91 regression coefficients of the environmental covariates at different locations  
2  
3 92 ([Brunsdon et al., 1996](#)). However, it does not directly consider spatial dependence or  
4  
5  
6 93 spatial autocorrelation of regionalized variables during the process of model  
7  
8  
9 94 development ([Kumar et al., 2012](#)). The geographically weighted regression kriging  
10  
11  
12 95 (GWRK) technique is an extension of the GWR method coupled with the kriging  
13  
14  
15 96 approach. This technique derives the benefits of GWR by adequately considering  
16  
17  
18 97 spatial parametric non-stationarity and the local relationships between the target  
19  
20  
21 98 variable and environmental covariates to fit the deterministic component (trend) of a  
22  
23  
24 99 target variable. The remaining stochastic component (residuals) can be kriged to the  
25  
26  
27 100 estimation map with the semivariogram model parameters and then added to the fitted  
28  
29  
30 101 trend ([Kumar and Lal, 2011](#)). [Kumar et al. \(2012\)](#) used a GWRK approach to examine  
31  
32  
33 102 the relationships between environmental variables and the SOC stock for the state of  
34  
35  
36 103 Pennsylvania, USA, and compared the GWRK results with those obtained by RK.  
37  
38  
39 104 More recently, [Liu et al. \(2015\)](#) evaluated the performance of GWRK in predicting  
40  
41  
42 105 the spatial distribution of SOC density compared with other common methods, such  
43  
44  
45 106 as RK, GWR, MLRK and ordinary kriging (OK), on the Jiangnan Plain, China. [Harris](#)  
46  
47  
48 107 [et al. \(2010\)](#) demonstrated that the GWRK approach can be used for estimating the  
49  
50  
51 108 residuals to predict the trend more efficiently and provide a worthy alternative when  
52  
53  
54 109 predicting with non-stationary relationships.

53 110 However, little attention has been paid to the effects of different sampling  
54  
55  
56 111 densities on the GWRK approach for estimating soil properties. In this study, the  
57  
58  
59 112 Beijing (China) region was used as the research area and the SOC content as the

113 target variable. Three spatial prediction methods, namely OK, MLRK and GWRK,  
114 were implemented with four different sampling densities (0.010, 0.020, 0.041 and  
115 0.082 sites/km<sup>2</sup>). Elevation, slope and topographic wetness index were used as  
116 auxiliary variables. The objectives of this study were as follows: (i) to compare  
117 GWRK predictions with those obtained by the MLRK and OK methods, and (ii) to  
118 examine how different sampling densities affect the performance of the GWRK  
119 method for SOC estimation. The results of this study have important implications for  
120 soil sampling strategies and soil mapping.

## 121 2. Materials and methods

### 122 2.1. Study area

123 The study area is located in the Beijing region, China (39°24' – 40°00'N,  
124 115°20' – 117°30'E), and covers an area of  $1.64 \times 10^4$  km<sup>2</sup>. The terrain is higher in  
125 the northwest and lower in the southeast with the altitude ranging from 2.5 to 2301.3  
126 m as shown in Fig. 1(a). It is bounded by mountains to the west, north and northeast,  
127 whereas the center and southeast of the region are an alluvial plain which is a part of  
128 the North China Plain. The mountainous area is mainly covered by Leptic Luvisols  
129 and Chromic Luvisols, and the flat area is composed of Eutric Cambisols, according  
130 to [IUSS Working Group WRB \(2006\)](#). Specifically, the Leptic Luvisols, Chromic  
131 Luvisols and Eutric Cambisols account for 10, 65 and 25% of the total area,  
132 respectively. The primary type of land use is agricultural land and croplands,  
133 orchards, forestland and grassland occupied 13, 8, 45 and 5% of total area,

1 134 respectively ([Beijing Statistical Bureau, 2015](#)). The Beijing region has a semi-humid  
2  
3 135 and semi-arid monsoon climate. The mean annual temperature is 8 – 12 °C. The  
4  
5  
6 136 mean annual precipitation is 600 – 700 mm, and the northwest and north mountain  
7  
8  
9 137 area has rainfall levels below 500 mm. The seasonal distribution of precipitation is  
10  
11  
12 138 uneven, with 70% of the precipitation concentrated from July to September.

## 139 *2.2. Soil sample data and environmental variables*

140 Soil samples were taken in 2010 in the autumn after harvesting to avoid the  
141 effect of fertilization. A 2-km grid was overlain on the agricultural land map across  
142 the Beijing region and a total of 1458 topsoil (0 – 20 cm) samples were selected by  
143 taking the land use types and soil types into account. A GPS receiver was used to  
144 locate the latitude and longitude for each sampling location. Five soil samples were  
145 collected from within a 10 m diameter surrounding a GPS location and then mixed.  
146 About 1 – 1.5 kg of soil per sampling site was extracted by the quarter method and  
147 then air-dried and sieved through 2 mm openings. The SOC content was determined  
148 using the potassium dichromate digestion method ([Nelson and Sommers, 1982](#)).

149 In this study, the observation dataset was randomly divided into 1338 sites as a  
150 calibration subset and 120 sites as a validation subset. To investigate the effects of  
151 different sampling densities on the prediction of the SOC content by the GWRK  
152 method, we randomly selected 50% of the sampling points from the calibration  
153 dataset and used it as a new calibration subset, while a further 50% of the randomly  
154 selected sampling points from the new subset was used as another new calibration

155 subset, and then the process was repeated. Finally, four different calibration subsets,  
 156 including 167, 334, 669, and 1338 sites, were established and treated as four different  
 157 sampling densities of the 0.010 sites/km<sup>2</sup> (Density 1), 0.020 sites/km<sup>2</sup> (Density 2),  
 158 0.041 sites/km<sup>2</sup> (Density 3), 0.082 sites/km<sup>2</sup> (Density 4), respectively, shown in Fig.  
 159 1.

160 Topographic features, including slope gradient, elevation, and topographic  
 161 wetness index (TWI), were used as environmental covariates in this study. The TWI is  
 162 described as follows:

$$163 \quad TWI = \ln \left( \frac{\alpha}{\tan \beta} \right) \quad (1)$$

164 where  $\alpha$  is the potential flow accumulation area per unit contour length (m<sup>2</sup>/m) and  $\beta$   
 165 is the local slope gradient (°) (Beven and Kirkby, 1979; Quinn et al., 1995). In this  
 166 study,  $\alpha$  was calculated using a multiple-flow-direction algorithm as proposed by  
 167 Quinn et al. (1991). The terrain topographies included the slope, elevation, and TWI  
 168 and were derived from the digital elevation models (DEM) with a 25 m spatial  
 169 resolution. The DEM data were obtained from the Beijing Digital Soil System  
 170 (BDSS).

### 171 2.3. Prediction methods

172 In geostatistics, a target variable  $Z$  at a location can be modeled as the sum of the  
 173 trend component (deterministic part)  $m$  and the residuals component (stochastic part)  
 174  $\varepsilon$ . Both components can be fitted separately and then summed to obtain the spatial  
 175 predictions (Odeh et al., 1994). The MLRK method is based on the idea that the trend

176 component of the target variable is explained by a regression model (Bishop and  
 177 McBratney, 2001), and the residuals are spatially correlated and interpolated with  
 178 kriging (Chai et al., 2008; Zhang et al., 2012; Kumar et al., 2012). The equation of the  
 179 MLRK method coupled with that of OK method is as follows:

$$180 \quad \hat{z}_{MLRK}(x_0) = \hat{m}_{MLR}(x_0) + \hat{\varepsilon}_{OK}(x_0) \quad (2)$$

$$181 \quad \begin{cases} \hat{m}_{MLR}(x_0) = \sum_{k=0}^p \hat{\beta}_k \cdot q_k(x_0) \\ q_0(x_0) = 1 \end{cases} \quad (3)$$

$$182 \quad \hat{\varepsilon}_{OK}(x_0) = \sum_{i=1}^n \lambda_i \cdot \varepsilon(x_i) \quad (4)$$

183 Where  $\hat{z}_{MLRK}(x_0)$  is the estimated value of the target variable at location  $x_0$ ,  
 184  $\hat{m}_{MLR}(x_0)$  is the fitted trend,  $\hat{\varepsilon}_{OK}(x_0)$  represents the interpolated residuals using an  
 185 OK approach,  $q_k(x_0)$  denote the environmental variables,  $k$  is the number of  
 186 predictors or environmental variables,  $\hat{\beta}_k$  represent the estimated regression  
 187 coefficients,  $\lambda_i$  denote the kriging weights determined by the spatial dependence  
 188 structure of the residuals, and  $\varepsilon(x_i)$  represent the residuals at location  $x_i$ . However,  
 189 The MLRK method is an extension of the global regression approach where no  
 190 geographical location information is considered in the estimation of the regression  
 191 model parameters and the coefficients are assumed to be spatially invariant (Li et al.,  
 192 2010; Kumar et al., 2012).

193 The GWRK method is an extension of the GWR approach, which extends  
 194 traditional regression (e.g., multiple linear regression) by allowing local rather than  
 195 global parameters to be estimated and takes the spatial locations of data points into

196 consideration (Kumar et al., 2012; Wang et al., 2013). The GWRK model used in the  
 197 present study is as follows:

$$198 \quad \hat{z}_{GWRK}(x_0) = \hat{m}_{GWR}(x_0) + \hat{\varepsilon}_{OK}(x_0) \quad (5)$$

$$199 \quad \hat{m}_{GWR}(x_0) = \sum_{k=0}^p \hat{\beta}_k(x_0) \cdot q_k(x_0) \quad (6)$$

200 where  $\hat{\beta}_k(x_0)$  are the unknown regression coefficients that are spatially variant.

201 For both the GWRK and MLRK methods, the residuals were kriged to the  
 202 prediction grid and added to the fitted regression trend to generate the final prediction  
 203 maps. The same environmental covariates were used for the estimation of the SOC by  
 204 GWRK and MLRK. The normality of the datasets was assessed using the  
 205 Kolmogorov-Smirnov (K-S) test, and the MLR analysis was performed using SPSS  
 206 16.0 (SPSS Inc., Chicago, IL, USA). GWR analysis, kriging and mapping were  
 207 performed using ArcGIS 10.0 (ESRI, Inc., Redlands, CA, USA). In the application of  
 208 GWR in ArcGIS, a fixed spatial kernel was selected to solve each local regression  
 209 analysis and the Akaike Information Criterion (AICc) was used to determine the  
 210 optimal bandwidth. Spatial kernels with a small bandwidth have a steeper  
 211 distance-decay weighting function and produce rougher surfaces than spatial kernels  
 212 with a large bandwidth.

#### 213 *2.4. Methods evaluation*

214 To evaluate the performances of the various spatial prediction methods, a total of  
 215 120 sites were randomly selected as validation set from the measured dataset. The  
 216 following indices were used to assess the performances of MLRK and GWRK

217 methods by comparing predictions of the SOC content of the validation set with the  
 218 known measured values.

219 -Both the root mean squared error (RMSE) and the mean absolute error (MAE) were  
 220 taken as indices to measure the accuracy of the predictions:

$$221 \quad MAE = \frac{1}{n} \sum_{i=1}^n |z(x_i) - \hat{z}(x_i)| \quad (7)$$

$$222 \quad RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n [z(x_i) - \hat{z}(x_i)]^2} \quad (8)$$

223 where  $n$  is the number of the validation sites,  $z(x_i)$  is the observed SOC content at  
 224 location  $x_i$ ,  $\hat{z}(x_i)$  is estimated SOC content at location  $x_i$ .

225 -The coefficient of determination ( $R^2$ ) was taken as an index to show the amount of  
 226 variance explained by the model (Sun et al., 2015):

$$227 \quad R^2 = 1 - \left\{ \frac{\sum_{i=1}^n [z(x_i) - \hat{z}(x_i)]^2}{\sum_{i=1}^n [z(x_i) - \bar{z}(x_i)]^2} \right\} \quad (9)$$

228 where  $\bar{z}(x_i)$  is the mean value of the observed SOC content at the validation sites.

229 -The median of the standardized squared error (SSE) was taken as index to evaluate  
 230 the uncertainty of the predictions (Lark, 2000):

$$231 \quad SSE(x_i) = \frac{[z(x_i) - \hat{z}(x_i)]^2}{\sigma_{OK, x_i}^2} \quad (10)$$

232 where  $\sigma_{OK, x_i}^2$  is the kriging variance of the residuals at location  $x_i$ .

233 The small values of the RMSE and MAE indicate more accurate predictions on a  
 234 point by point basis (Schloeder et al., 2001), and an  $R^2$  value equal to 1.0 indicates a  
 235 perfect prediction (Kumar et al., 2015). If the variogram is correct, the median of the

236 SSE is 0.45. A median significantly less than 0.45 suggests that kriging overestimates  
237 the variance, possibly because of the effects of outliers on the variogram estimator.  
238 On the other hand, a median significantly greater than 0.45 suggests that kriging  
239 underestimates the variance, possibly because of the effect of non-normality on a  
240 robust estimator (Lark, 2000).

### 241 3. Results

#### 242 3.1. Descriptive statistics

243 Table 1 shows the descriptive statistics of the SOC content and environmental  
244 covariates under different sampling densities. The mean values of the SOC content for  
245 Density 1, 2, 3 and 4 were 10.3, 10.4, 10.4 and 10.5 g kg<sup>-1</sup>, respectively. The  
246 skewness values and Kolmogorov-Smirnov (K-S) normal distribution tests indicated  
247 that the SOC content belonged to a positive skew distribution ( $p < 0.05$ ). The  
248 coefficients of variation (CV) of the SOC content at the four sampling densities  
249 ranged from 52 to 57%, corresponding to a middle degree of variability. Furthermore,  
250 the mean values of the elevation, slope, and TWI ranged from 221.6 to 228.6 m, 4.5 to  
251 5.3°, and 7.3 to 8.8, respectively, for the four sampling densities. The high CV of the  
252 elevation (109 to 114%), slope (167 to 189%), and TWI (101 to 121%) under the  
253 different sampling densities indicated a strong degree of variability. Overall, the  
254 observation datasets of the four different sampling densities exhibited quite small  
255 differences in the statistical characteristics, making the application of the spatial  
256 prediction methods with different sampling densities comparable.

257 Table 1

258 Descriptive statistics of the SOC content and environmental variables under different  
 259 sampling densities.

Variables	Min.	Max.	Mean	SD	CV (%)	Skewness	K-S $\rho$ value
Density 1 (0.010 sites/km <sup>2</sup> )							
SOC (g kg <sup>-1</sup> )	1.1	41.0	10.3	5.4	52	2.1	0.000
Elevation (m)	11.8	1775.6	224.5	252.5	112	1.6	0.000
Slope (°)	0.0	50.0	4.5	8.5	189	3.0	0.000
TWI	-11.1	29.8	7.8	9.5	121	-0.5	0.000
Density 2 (0.020 sites/km <sup>2</sup> )							
SOC (g kg <sup>-1</sup> )	0.9	50.1	10.4	5.9	57	2.6	0.000
Elevation (m)	9.3	1775.6	224.9	244.7	109	1.5	0.000
Slope (°)	0.0	50.0	4.8	8.2	170	2.4	0.000
TWI	-11.1	29.8	8.3	9.3	112	-0.4	0.000
Density 3 (0.041 sites/km <sup>2</sup> )							
SOC (g kg <sup>-1</sup> )	0.9	50.1	10.4	5.8	56	2.6	0.000
Elevation (m)	10.3	1812.6	228.6	260.2	114	1.5	0.000
Slope (°)	0.0	52.6	5.3	8.9	170	2.2	0.000
TWI	-11.8	34.3	8.8	9.0	101	-0.4	0.000
Density 4 (0.082 sites/km <sup>2</sup> )							
SOC (g kg <sup>-1</sup> )	0.9	50.1	10.5	5.9	56	2.5	0.000
Elevation (m)	9.3	1812.6	221.6	251.5	114	1.4	0.000
Slope (°)	0.0	52.6	5.2	8.6	167	2.2	0.000
TWI	-11.8	34.3	8.7	9.2	106	-0.3	0.000
Verification sites (n=120)							
SOC (g kg <sup>-1</sup> )	1.1	48.2	10.2	5.3	52	1.5	0.000
Elevation (m)	10.9	1128.7	222.6	248.3	112	1.5	0.000
Slope (°)	0.0	38.4	4.5	7.4	171	2.2	0.000
TWI	-10.8	26.6	8.3	10.3	125	-0.2	0.000

260 SD = standard deviation; CV = coefficient of variation; K-S = Kolmogorov-Smirnov test.

### 261 3.2. Trend analysis

262 In geostatistics, a target variable is composed of two components, deterministic  
 263 (trend) and stochastic (residuals), which are modeled separately. If the trend (such as  
 264 global trend or local trend) of the target variable can be accurately identified,  
 265 quantified and removed in the process of variogram analysis and kriging, a

266 short-range random variation of the residuals can be accurately simulated (Tang and  
267 Yang, 2006).

268 As shown in Fig. 2, the SOC contents of samples under all the four sampling  
269 densities exhibited a second-order global trend in the east-west direction and an  
270 approximately first-order linear global trend in the north-south direction. These trends  
271 are consistent with the basic topographical distribution of the region. Several studies  
272 have also shown a good correlative relationship between the soil properties and  
273 environmental information (Chai et al., 2008; Zhang et al., 2012; Somarathna et al.,  
274 2016; Song et al., 2016). In order to determine the topographic features (including  
275 elevation, slope gradient and TWI) dominating the spatial variation of the SOC  
276 content, a stepwise multiple regression analysis was performed to obtain the  
277 regression equations. The results indicated that all the topographic features were  
278 selected as explanatory variables. Then, both the MLR and GWR methods were  
279 applied to fit the global and local SOC trend models, respectively. Table 2 lists the  
280 summary statistics of the regression coefficients and parameters. The regression  
281 coefficients for the intercept, elevation, slope, and TWI determined by MLR were  
282 spatially invariant. For instance, the coefficients of intercept, elevation, slope, and  
283 TWI gained by MLR under Density 4 were 8.603, 0.004, 0.185, and 0.004,  
284 respectively. However, the regression coefficients determined by GWR were spatially  
285 variant. For example, the coefficients of intercept, elevation, slope, and TWI for  
286 GWR under Density 4 ranged from -9.392 to 26.706, -0.411 to 0.440, -3.812 to  
287 11.730, and -0.237 to 0.393, respectively. Fig. 3 shows the distribution of the

1 288 regression coefficients by GWR in the case of Density 4. It can be seen that the high  
2  
3  
4 289 coefficients of elevation and slope were mainly distributed in the southeast region,  
5  
6 290 whereas those of intercept and TWI were mainly distributed in the west region.  
7  
8  
9 291 Moreover, the goodness of fit (adjusted  $R^2$ ) for the MLR approach in the case of low a  
10  
11 292 sampling density (i.e. 0.439 for Density 1 and 0.485 for Density 2) was larger than  
12  
13  
14 293 that in the case of a high sampling density (i.e. 0.376 for Density 3 and 0.393 for  
15  
16  
17 294 Density 4). However, the GWR approach exhibited the opposite results with the  
18  
19  
20 295 adjusted  $R^2$  under the condition of small samples (i.e. 0.450 for Density 1 and 0.406  
21  
22  
23 296 for Density 2) less than that under the condition of large samples (i.e. 0.451 for  
24  
25  
26 297 Density 3 and 0.501 for Density 4). Within a same sampling density (except for  
27  
28  
29 298 Density 1), the adjusted  $R^2$  for GWR method was larger than that for the MLR method.  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

299 Table 2

300 Summary of the regression coefficients and parameters using the MLR and GWR  
 301 models under different sampling densities.

Variables	MLR		GWR					
	Coefficients	Adjusted $R^2$	Range of coefficients	Median of coefficients	Adjust $R^2$	Effective number of parameters	Bandwidth (m)	AICc
Density 1 (0.010 sites/km <sup>2</sup> )								
Intercept	8.753	0.439	3.838–16.354	8.757	0.450	33.8	27718	894.9
Elevation	0.004		–0.067–0.020	0.002				
Slope	0.174		–0.095–1.133	0.202				
TWI	–0.015		–0.241–0.245	–0.013				
Density 2 (0.020 sites/km <sup>2</sup> )								
Intercept	7.991	0.485	3.679–12.811	9.269	0.406	56.5	21429	1822.4
Elevation	0.007		–0.153–0.045	0.006				
Slope	0.202		–0.306–3.907	0.147				
TWI	0.011		–0.177–0.119	0.006				
Density 3 (0.041 sites/km <sup>2</sup> )								
Intercept	8.767	0.376	–3.379–23.949	9.167	0.451	133.1	12802	3630.2
Elevation	0.005		–0.221–0.163	0.003				
Slope	0.115		–28.646–8.480	0.104				
TWI	–0.015		–0.328–0.444	–0.014				
Density 4 (0.082 sites/km <sup>2</sup> )								
Intercept	8.603	0.393	–9.392–26.706	8.868	0.501	202.7	10455	7125.1
Elevation	0.004		–0.411–0.440	0.003				
Slope	0.185		–3.812–11.730	0.130				
TWI	0.004		–0.237–0.393	–0.004				

### 302 3.3. Semivariogram analysis

303 In this study, we only use the exponential model for semivariogram to compare  
 304 GWRK with MLRK under different sampling densities, because exponential model is  
 305 stable in nonlinear least squares fitting and was found to represent most soil properties  
 306 (Minasny and McBratney, 2005; Chai et al., 2007). For all the four sampling densities,  
 307 the K-S normal distribution tests indicated that the SOC contents were log-normally

1 308 distributed, whereas the MLR and GWR residuals were normally distributed, as  
2  
3 309 shown in [Table 3](#). Accordingly, we performed firstly logarithm transform over the  
4  
5  
6 310 original SOC data using Geostatistical Analyst in ArcGIS before interpolating with  
7  
8  
9 311 single OK method. After making predictions on the transformed scale, the software  
10  
11 312 automatically transformed the predictions back to the original scale. The  
12  
13 313 semivariogram parameters of LnC, MLR residuals and GWR residuals are listed in  
14  
15  
16  
17 314 [Table 3](#), and the experimental and estimated semivariograms are shown in [Fig. 4](#).

18  
19  
20 315 Overall, with the increase of the sampling density, the experimental  
21  
22 316 semvariograms (dots) for the LnC, MLR residuals and GWR residuals gradually  
23  
24  
25 317 became smooth and steady with better fitting effects on the estimated semvariograms  
26  
27  
28 318 (solid lines). The nugget, sill and nugget/sill ratio also increased as the sampling  
29  
30  
31 319 density increased, indicating that the data with a small sampling interval presents the  
32  
33  
34 320 large random variation and its proportion in total variation. Within a same sampling  
35  
36  
37 321 density, the GWR residuals had a lower range and a larger nugget/sill ratio compared  
38  
39  
40 322 to the MLR residuals and LnC. For instance, in the case of Density 4, the nugget/sill  
41  
42  
43 323 ratios for the LnC, MLR residuals and GWR residuals were 39.13, 35.02 and 40.21%,  
44  
45  
46 324 respectively. The ranges for the LnC, MLR residuals and GWR residuals were 25.1,  
47  
48  
49 325 37.8 and 11.7 km, respectively. The comparatively small range and large nugget/sill  
50  
51  
52 326 ratio for the GWR residuals indicated that much of the structured variation in the SOC  
53  
54  
55 327 content is explained by its local relationship to environmental variables. Many studies  
56  
57  
58 328 have also reported a shorter range for GWR residuals for estimating target variables  
59  
60  
61 329 ([Lloyd, 2010](#); [Kumar et al., 2012](#)). Furthermore, the range was reduced by 1.4 km

330 from 11.5 km of MLR residuals to 10.1 km of GWR residuals in the case of Density 1.  
 331 Similarly, by comparing the MLR residuals with the GWR residuals, the range was  
 332 also reduced by 2.4 km from 7.9 to 5.5 km in the case of Density 2, by 17.1 km from  
 333 28.5 to 11.4 km in the case of Density 3, and by 26.1 km from 37.8 to 11.7 km in the  
 334 case of Density 4, respectively. The large variation in range under Density 3 and 4  
 335 indicated that the GWR approach is much more suitable than the MLR method for  
 336 modeling the deterministic component of the SOC content in the condition with high  
 337 sampling densities.

338

339 Table 3

340 Semivariogram parameters of LnC and the MLR residuals and GWR residuals under different  
 341 sampling densities.

Variables	K-S $\rho$ value	Nugget	Sill	Nugget/sill ratio (%)	Range (km)	$R^2$	RSS
LnC							
Density 1	0.141	0.03	0.22	13.64	24.0	0.75	0.00
Density 2	0.107	0.09	0.23	39.13	15.7	0.76	0.00
Density 3	0.079	0.11	0.23	47.83	25.2	0.82	0.00
Density 4	0.064	0.09	0.23	39.13	25.1	0.99	0.00
MLR residuals							
Density 1	0.252	0.01	21.61	0.05	11.5	0.71	151.00
Density 2	0.091	0.01	19.27	0.05	7.9	0.64	77.40
Density 3	0.159	6.11	25.66	23.81	28.5	0.93	29.00
Density 4	0.128	9.30	26.56	35.02	37.8	0.99	2.05
GWR residuals							
Density 1	0.132	0.01	14.03	0.07	10.1	0.70	48.60
Density 2	0.086	0.01	13.64	0.07	5.5	0.56	53.60
Density 3	0.324	3.23	14.29	22.60	11.4	0.83	13.50
Density 4	0.237	5.42	13.48	40.21	11.7	0.85	3.78

342 RSS = residual sum of squares.

### 343 3.4. Validation of predictions

344 Three approaches (OK, MLRK and GWRK) with four sampling densities of data  
345 points were used to predict the SOC contents at the validation sites. The validation  
346 results are shown in Table 4. As the sampling density increased, the MAE and RMSE  
347 values decreased and the  $R^2$  increased, indicating that the accuracy of the predictions  
348 under a high sampling density is better than that under a low sampling density. Within  
349 a same sampling density, the  $R^2$  values for MLRK were larger than those for OK (i.e.  
350 0.53 vs. 0.51 for Density 1, 0.59 vs. 0.57 for Density 2, 0.67 vs. 0.66 for Density 3,  
351 and 0.71 vs. 0.69 for Density 4), and the former MAE and RMSE values were smaller  
352 than those of the latter (as RMSE, 3.05 vs. 3.15 g kg<sup>-1</sup> for Density 1, 2.87 vs. 2.93 g  
353 kg<sup>-1</sup> for Density 2, 2.56 vs. 2.61 g kg<sup>-1</sup> for Density 3, and 2.46 vs. 2.47 g kg<sup>-1</sup> for  
354 Density 4). These suggested that the MLRK method combined with the environmental  
355 covariates can further enhance the precision of the SOC content estimations compared  
356 to the OK method. In the case of Density 1 and 2, the  $R^2$  values of GWRK (0.16 and  
357 0.49, respectively) were also lower than those of OK (0.51 and 0.57, respectively) and  
358 MLRK (0.53 and 0.59, respectively). Additionally, the RMSE values of GWRK (4.10  
359 and 3.19 g kg<sup>-1</sup>, respectively) were greater than those of OK (3.15 and 2.93 g kg<sup>-1</sup>,  
360 respectively) and MLRK (3.05 and 2.87 g kg<sup>-1</sup>, respectively). In the case of Density 3  
361 and 4, however, the  $R^2$  values of GWRK (0.75 and 0.70, respectively) were higher  
362 than those of OK (0.67 and 0.71, respectively) and MLRK (0.66 and 0.69,  
363 respectively), and the RMSE values of GWRK (2.43 and 2.24 g kg<sup>-1</sup>, respectively)  
364 were also lower than those of OK (2.61 and 2.47 g kg<sup>-1</sup>, respectively) and MLRK

365 (2.56 and 2.46 g kg<sup>-1</sup>, respectively). Thus, high sampling densities (Density 3 and 4)  
 366 resulted in the prediction performance in the order GWRK > MLRK > OK, whereas  
 367 low sampling densities led to the performance in the order MLRK > OK > GWRK.  
 368 This revealed that the GWRK method gives the best prediction of the SOC content in  
 369 the three approaches when the data points are sufficient, whereas it exhibits a worse  
 370 performance than the other two methods when the data points are sparse.

371  
 372 Table 4

373 Results of the validation for the OK, MLRK, and GWRK methods under different  
 374 sampling densities.

Methods	MAE (g kg <sup>-1</sup> )	Median SEE	RMSE (g kg <sup>-1</sup> )	$R^2$
OK				
Density 1	2.27	0.18	3.15	0.51
Density 2	2.10	0.21	2.93	0.57
Density 3	2.00	0.24	2.61	0.66
Density 4	1.93	0.75	2.47	0.69
MLRK				
Density 1	2.29	0.24	3.05	0.53
Density 2	2.14	0.77	2.87	0.59
Density 3	1.95	0.77	2.56	0.67
Density 4	1.90	1.05	2.46	0.71
GWRK				
Density 1	2.73	0.19	4.10	0.16
Density 2	2.31	0.85	3.19	0.49
Density 3	1.91	0.74	2.43	0.70
Density 4	1.75	0.65	2.24	0.75

375 MAE = mean absolute error; RMSE = root mean squared error; SEE = standardized squared error.

### 376 3.5. Uncertainty of predictions

377 The median of the standardized squared error (SEE) at the validation sites for the

1 378 different densities are listed in [Table 4](#). The median of the SEE for the GWRK  
2  
3 379 method under Density 4 is 0.19, which is less than 0.45. This suggested that the  
4  
5  
6 380 variances were overestimated by kriging of residuals ([Lark, et al., 2000](#)). While the  
7  
8  
9 381 medians of the SEE for the GWRK method under Density 2, 3 and 4 were 0.85, 0.74  
10  
11 382 and 0.65, respectively, which were all larger than 0.45. These results suggested the  
12  
13  
14 383 underestimation of the variance ([Lark, et al., 2000](#)). Meanwhile, the result using  
15  
16  
17 384 GWRK under Density 4 seemed less biased than those under Density 2 and 3.  
18  
19  
20 385 Similarly, the median of the SEE for the MLRK under Density 1 was also less than  
21  
22 386 0.45 and those under the Density 2, 3 and 4 were all greater than 0.45. However, the  
23  
24  
25 387 median of the SEE for the MLRK under Density 4 (1.05) was obviously greater than  
26  
27  
28 388 those under Density 2 and 3 (both were 0.77). This indicated that the MLRK could  
29  
30  
31 389 result in a large uncertainty of the SOC predictions in the condition with high  
32  
33  
34 390 sampling density, perhaps because of the strong non-stationary relationships between  
35  
36  
37 391 the SOC content and the environmental covariates. And the spatial non-stationary  
38  
39  
40 392 relationships cannot be well fitted by the MLRK method. Unlike the GWRK and  
41  
42 393 MLRK methods, the OK method made the medians of the SEE under Density 2 and 3  
43  
44  
45 394 less than 0.45. A comparison of the different prediction approaches revealed that the  
46  
47  
48 395 medians of the SEE for both the MLRK and GWRK methods were greater than those  
49  
50  
51 396 for the OK method in the case of Density 2 and 3. This suggested that the introduction  
52  
53  
54 397 of environmental information into the prediction model contributed to solve the  
55  
56 398 problem of inadequate sample sites, but the uncertainty of the predictions still  
57  
58  
59 399 remained. Besides, in the case of the Density 2, 3 and 4, the GWRK method gave  
60  
61  
62  
63  
64  
65

1 400 lower medians of the SEE than the MLRK method, with the GWRK method  
2  
3 401 achieving a smaller uncertainty than the MLRK method.  
4  
5  
6

### 7 402 *3.6. Spatial distribution of the SOC content*

10  
11 403 [Fig. 5](#) shows the spatial distribution of the SOC mapped using the OK, MLRK  
12  
13  
14 404 and GWRK methods with different sampling densities. It can be seen that all the maps  
15  
16  
17 405 displayed the same overall patterns of the SOC content, with the high SOC contents  
18  
19  
20 406 (above 25 g kg<sup>-1</sup>) mainly distributing in the western region and the low SOC contents  
21  
22  
23 407 (below 10 g kg<sup>-1</sup>) primarily distributing in the central and southeastern region. This  
24  
25  
26 408 distribution pattern may be related to the topography of the region, where the west  
27  
28  
29 409 mainly consisted of mountains and the central and southeast regions were covered by  
30  
31 410 an alluvial plain.  
32

33  
34 411 In some local regions, however, there are a few differences in the spatial  
35  
36 412 distribution of the SOC content among the three prediction methods. For instance, as  
37  
38  
39 413 the western region with an SOC value above 25 g kg<sup>-1</sup> in the case of Density 4, the  
40  
41  
42 414 OK approach yielded a smoother prediction map with more integrated polygons in its  
43  
44  
45 415 smoothing effect, as shown in [Fig. 5\(d\)](#). Whereas the MLRK and GWRK methods  
46  
47  
48 416 with environmental covariates obtained much more realistic and fragmented  
49  
50  
51 417 prediction maps, as shown in [Fig. 5\(h\)–\(l\)](#). This was especially true for the GWRK  
52  
53  
54 418 approach, which can better expose the local details than MLRK. Similar results have  
55  
56 419 also been reported in previous studies ([Kumar et al., 2012](#); [Liu et al., 2015](#)).  
57

58 420 Moreover, there are great differences in the spatial pattern of the SOC content  
59  
60  
61  
62  
63  
64  
65

1 421 among the different sampling densities. For the GWRK method, as shown in Fig.  
2  
3 422 5(i)–(l), the highest SOC content regions (above  $25 \text{ g kg}^{-1}$ ) mainly distributed in the  
4  
5  
6 423 northwest under Density 1. While in the case of Density 2, 3 and 4, the highest SOC  
7  
8  
9 424 content regions were in the southwest, and the area in the case of Density 2 was larger  
10  
11  
12 425 than those in Density 3 and 4. Moreover, the difference of the distribution patterns  
13  
14  
15 426 between Density 3 and Density 4 was slight.

#### 18 427 4. Discussion

22 428 Any relationship between the soil properties and environmental covariates that is  
23  
24 429 often non-stationary over space is not well represented by a global statistic and,  
25  
26 430 indeed, this global value may be very misleading locally (Fotheringham et al., 2002).  
27  
28  
29 431 The GWRK is a local spatial interpolation model, which not only considers the spatial  
30  
31  
32 432 heterogeneity and dependency of the target variable, but also the different spatial  
33  
34 433 weights of environmental variables to the target variable (Kumar et al., 2012; Liu et  
35  
36 434 al., 2015). In this study, a GWRK approach was used to predict the spatial distribution  
37  
38  
39 435 of the SOC content in the Beijing region, China, and compare it with that predicted by  
40  
41  
42 436 the OK and MLRK approaches. Although the local relationships between the SOC  
43  
44 437 content and its environmental covariates were still fitted by a multiple linear  
45  
46 438 regression analysis for the GWRK method, the varying regression coefficients for the  
47  
48  
49 439 environmental covariates at different locations also indicated the non-stationarity of  
50  
51  
52 440 the relationships. As expected, the results showed that the application of the GWRK  
53  
54 441 approach for predicting SOC content outperformed those of the MLRK and OK  
55  
56 442 approaches in the case of an appropriate sampling density, such as in the cases of  
57  
58  
59 443 Density 3 and 4 in this study. Similar conclusions were also reported by Kumar et al.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
444 (2012), who compared the GWRK and RK approaches to examine the relationships  
445 between environmental variables and the SOC stocks, and their results showed that  
446 GWRK was more accurate in representing the heterogeneity of the SOC stocks. Liu et  
447 al. (2015) compared the GWRK method with other common methods (RK, GWR,  
448 MLR and OK) for estimating the spatial distribution of the SOC density by using the  
449 correlated environmental variables, namely terrain factors, distance factors, land cover  
450 type and spectral indices. Their results indicated that the GWRK method was a more  
451 suitable spatial interpolation model than other approaches for predictive mapping of  
452 the SOC density when multiple covariables were available. Harris et al. (2010)  
453 demonstrated that the GWRK approach can be used for estimating the residuals to  
454 predict the trend more efficiently and provide a worthy alternative when predicting  
455 with non-stationary relationships. However, these studies still lacked concrete  
456 experimental designs supporting the assessment of the effects of different sampling  
457 densities for SOC content estimation by GWRK approach.

34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
458 Fotheringham et al. (2002) have pointed out that all types of spatial analysis and  
459 spatial modelling are affected by scale to some degree. For instance, local modelling  
460 by GWRK or GWR is particularly sensitive to the representativeness of the sample  
461 data (Pasculli et al., 2014). In this study, four different sampling densities (0.010,  
462 0.020, 0.041, and 0.082 sites/km<sup>2</sup>) were used to examine how different sampling  
463 densities affect the performance of the GWRK method for predicting the SOC content,  
464 and the results showed that the performance of the GWRK method was affected by  
465 sampling density. The more important concern is that the GWRK approach performed  
466 better at predicting the spatial distribution of the SOC content in the case of a high  
467 sampling density, but performed poorer in the case of a low sampling density,  
468 compared with the MLRK and OK approaches. A possible reason is that, on the one

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

469 hand, the GWRK method uses many more parameters than the other models so that  
470 the risk of overfitting is highest in the application of the GWRK method in the case of  
471 few observations combined with many parameters. On the other hand, the application  
472 of the GWRK method with fixed spatial kernels on very few data points might give  
473 rise to parameter estimates with large variances and be increasingly unreliable  
474 (Fotheringham et al., 2002). In addition, the kernels in the GWRK method have larger  
475 bandwidths when the data are sparse (Table 2 shows the bandwidths in the cases of  
476 Density 1, 2, 3 and 4 were 27718, 21429, 12802 and 10455 m, respectively), which to  
477 some extent reflects that the correlations between topographic factors and the SOC  
478 content are more representative of global trends than local trends. However, the  
479 application of the GWRK approach is inferior to that of the MLRK approach in fitting  
480 the global trends.

481 As can be seen from Table 4, although the MLRK and GWRK with the  
482 covariates performed better than the single OK method, the precisions of interpolating  
483 were still improved slightly. One possible reason is the limitation and uneven  
484 distribution of the soil samples (82.7% of the soil samples were collected in the areas  
485 with altitude below 500 m), which may not well reflect the impacts of topographic  
486 factors on SOC. Besides, human activity in low altitude areas would increase the  
487 spatial variability of soil, which may reduce the influence of topographic factors on  
488 SOC.

## 489 5. Conclusions

490 In this study, a GWRK approach was used to predict the spatial distribution of  
491 the SOC content under four different sampling densities (0.010, 0.020, 0.041, and  
492 0.082 sites/km<sup>2</sup>). Two other geostatistical approaches (OK and MLRK) were also

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

493 studied to compare their performance with GWRK. Results show that the  
494 performance of the GWRK method was affected by sampling density. As the sampling  
495 density increased, GWRK produced a lower uncertainty and higher accuracy of SOC  
496 predictions. Compared to the MLRK and OK, GWRK yielded smaller prediction  
497 errors, and outperformed MLRK and OK in the case of high sampling densities, with  
498 the root mean squared errors of  $GWRK < MLRK < OK$  and coefficient of  
499 determination ( $R^2$ ) of  $GWRK > MLRK > OK$ . This is because the GWRK approach  
500 considers the spatial non-stationarity of the relationships coupled with spatial  
501 autocorrelation of the residuals. In the case of low sampling densities, however,  
502 GWRK generated larger prediction errors, exhibiting a poorer performance than  
503 MLRK and OK. Accordingly, we conclude that GWRK can be considered as the best  
504 approach for predicting SOC in the three approaches when the data points are  
505 sufficient, but it has a poorer performance than the other two methods when data  
506 points are sparse.

## 507 Acknowledgements

508 This research was supported by the Science & Technology Basic Research  
509 Program of China (2014FY210100), the National Natural Science Fund of China  
510 (41501468, 41471186), the Natural Science Foundation of Hainan Province, China  
511 (20154177, 2016CXTD015).

## 512 References

513 Arrouays, D., Marchant, B.P., Saby, N.P.A., Meersmans, J., Orton, T.G., Martin, M.P.,  
514 Bellamy, P.H., Lark, R.M., Kibblewhite, M., 2012. Generic issues on

- 1 515 broad-scale soil monitoring schemes: a review. *Pedosphere* 22(4), 456-469.  
2  
3 516 [http://dx.doi.org/10.1016/S1002-0160\(12\)60031-9](http://dx.doi.org/10.1016/S1002-0160(12)60031-9).  
4  
5  
6 517 Beijing Statistical Bureau, 2015. Beijing statistical yearbook 2015. Beijing: Chinese  
7  
8 518 Statistics Press.  
9  
10  
11 519 Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model  
12  
13 520 of basin hydrology. *Hydrolog. Sci. J.* 24, 43-69.  
14  
15 521 <http://dx.doi.org/10.1080/02626667909491834>.  
16  
17  
18 522 Bishop, T.F.A., McBratney, A.B., 2001. A comparison of prediction methods for the  
19  
20 523 creation of field-extent soil property maps. *Geoderma* 103, 149-160.  
21  
22 524 [http://dx.doi.org/10.1016/S0016-7061\(01\)00074-X](http://dx.doi.org/10.1016/S0016-7061(01)00074-X).  
23  
24  
25 525 Brunsdon, C., Fotheringham, A.S., Charlton, M., 1996. Geographically weighted  
26  
27 526 regression: a method for exploring spatial non-stationarity. *Geogr. Anal.* 28,  
28  
29 527 281–298. <http://dx.doi.org/10.1111/j.1538-4632.1996.tb00936.x>.  
30  
31  
32 528 Brus, D.J., Noij, I.G.A.M., 2008. Designing sampling schemes for effect monitoring  
33  
34 529 of nutrient leaching from agricultural soils. *Eur. J. Soil Sci.* 59(2), 292-303.  
35  
36 530 <http://dx.doi.org/10.1111/j.1365-2389.2007.00996.x>.  
37  
38  
39 531 Chai, X., Shen, C., Yuan, X., Huang, Y., 2008, Spatial prediction of soil organic  
40  
41 532 matter in the presence of different external trends with REML-EBLUP.  
42  
43 533 *Geoderma* 148, 159-166. <http://dx.doi.org/10.1016/j.geoderma.2008.09.018>.  
44  
45  
46 534 Dale, M.R., Fortin, M.J., 2014. Spatial analysis: a guide for ecologists. Cambridge  
47  
48 535 University Press, Cambridge.  
49  
50  
51 536 Ettema, C.H., Wardle, D.A., 2002. Spatial soil ecology. *Trends Ecol. Evol.* 17,  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- 1 537 177-183.  
2  
3  
4 538 Fotheringham, A.S., Brunsdon, C., Charlton, M., 2002. Geographically weighted  
5  
6 539 regression: the analysis of spatially varying relationships. John Wiley & Sons  
7  
8  
9 540 Ltd., Chichester, UK.  
10  
11 541 Gasch, C. K., Hengl, T., Gräler, B., Meyer, H., Magney, T. S., Brown, D. J., 2015.  
12  
13  
14 542 Spatio-temporal interpolation of soil water, temperature, and electrical  
15  
16  
17 543 conductivity in 3D+ T: The Cook Agronomy Farm data set. *Spat. Stat.* 14, 70-90.  
18  
19  
20 544 <http://dx.doi.org/10.1016/j.spasta.2015.04.001>.  
21  
22  
23 545 Harris, P., Fotheringham, A.S., Crespo, R., Charlton, M., 2010. The use of  
24  
25 546 geographically weighted regression for spatial prediction: an evaluation of  
26  
27  
28 547 models using simulated data sets. *Mathematical Geosciences*, 42(6), 657-680.  
29  
30  
31 548 <http://dx.doi.org/10.1007/s11004-010-9284-7>.  
32  
33  
34 549 Heim, A., Wehrli, L., Eugster, W., Schmidt, M.W.I., 2009. Effects of sampling design  
35  
36 550 on the probability to detect soil carbon stock changes at the Swiss CarboEurope  
37  
38  
39 551 site Lägeren. *Geoderma* 149(3), 347-354.  
40  
41  
42 552 <http://dx.doi.org/10.1016/j.geoderma.2008.12.018>.  
43  
44  
45 553 IUSS Working Group WRB, 2006. World reference base for soil resources 2006.  
46  
47 554 World Soil Resources Reports No. 103. FAO, Rome.  
48  
49  
50 555 Jackson, R.B., Caldwell, M.M., 1993. Geostatistical patterns of soil heterogeneity  
51  
52  
53 556 around individual perennial plants. *J. Ecol.* 81, 683-692.  
54  
55  
56 557 <http://dx.doi.org/10.2307/2261666>.  
57  
58  
59 558 Kumar, S, Lal, R., Liu, D., 2012. A geographically weighted regression kriging  
60  
61  
62  
63  
64  
65

- 1 559 approach for mapping soil organic carbon stock. *Geoderma* 189-190, 627-634.  
2  
3 560 <http://dx.doi.org/10.1016/j.geoderma.2012.05.022>.  
4  
5  
6 561 Kumar, S., 2015. Estimating spatial distribution of soil organic carbon for the  
7  
8 562 Midwestern United States using historical database. *Chemosphere* 127: 49-57.  
9  
10 563 <http://dx.doi.org/10.1016/j.chemosphere.2014.12.027>.  
11  
12  
13 564 Kumar, S., Lal, R., 2011. Mapping the organic carbon stocks of surface soils using  
14  
15 565 local spatial interpolator. *J. Environ. Monit.* 13, 3128-3135.  
16  
17 566 <http://dx.doi.org/10.1039/C1EM10520E>.  
18  
19  
20 567 Lark, R.M., 2000. A comparison of some robust estimators of the variogram for use in  
21  
22 568 soil survey. *Eur. J. Soil Sci.* 51(1), 137–157.  
23  
24 569 <http://dx.doi.org/10.1046/j.1365-2389.2000.00280.x>.  
25  
26  
27 570 Lark, R. M., 2012. Towards soil geostatistics. *Spat. Stat.* 1, 92-99.  
28  
29 571 <http://dx.doi.org/10.1016/j.spasta.2012.02.001>.  
30  
31  
32 572 Li, S., Zhao, Z., Xie, M., Wang, Y., 2010. Investigating spatial non-stationary and  
33  
34 573 scaledependent relationships between urban surface temperature and  
35  
36 574 environmental factors using geographically weighted regression. *Environ.*  
37  
38 575 *Modell. Softw.* 25, 1789-18. <http://dx.doi.org/10.1016/j.envsoft.2010.06.011>.  
39  
40  
41 576 Liu, Y., Guo, L., Jiang, Q., Zhang, H., Chen, Y., 2015. Comparing geospatial  
42  
43 577 techniques to predict SOC stocks. *Soil Till. Res.* 148, 46-58.  
44  
45 578 <http://dx.doi.org/10.1016/j.still.2014.12.002>.  
46  
47  
48 579 Lloyd, C.D., 2010. Nonstationary models for exploring and mapping monthly  
49  
50 580 precipitation in the United Kingdom. *Int. J. Climatology* 30, 390-405.  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- 1 581 <http://dx.doi.org/10.1002/joc.1892>.
- 2
- 3 582 McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping.
- 4
- 5
- 6 583 Geoderma 117(1), 3-52. [http://dx.doi.org/10.1016/S0016-7061\(03\)00223-4](http://dx.doi.org/10.1016/S0016-7061(03)00223-4).
- 7
- 8
- 9 584 McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using
- 10
- 11 environmental correlation. Geoderma 89, 67-94.
- 12 585
- 13
- 14 586 [http://dx.doi.org/10.1016/S0016-7061\(98\)00137-2](http://dx.doi.org/10.1016/S0016-7061(98)00137-2).
- 15
- 16
- 17 587 Miller, B.A., Koszinski, S., Hierold, W., Rogasik, H., Schröder, B., Van Oost, K.,
- 18
- 19 Wehrhana, M., Sommer, M., 2016. Towards mapping soil carbon landscapes:
- 20 588
- 21 Issues of sampling scale and transferability. Soil Till. Res. 156, 194-208.
- 22 589
- 23
- 24
- 25 590 <http://dx.doi.org/10.1016/j.still.2015.07.004>.
- 26
- 27
- 28 591 Minasny, B., McBratney, A.B., 2005. The Matérn function as a general model for soil
- 29
- 30 variograms. Geoderma 128, 192-207.
- 31 592
- 32
- 33 593 <http://dx.doi.org/10.1016/j.geoderma.2005.04.003>.
- 34
- 35
- 36 594 Minasny, B., McBratney, A.B., Mendonca-Santos, M.L., Odeh, I.O.A., Guyon, B.,
- 37
- 38 2006. Prediction and digital mapping of soil carbon storage in the Lower Namoi
- 39 595
- 40 Valley. Soil Res. 44, 223-244. <http://dx.doi.org/10.1071/SR05136>.
- 41 596
- 42
- 43
- 44 597 Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute
- 45
- 46 prediction using terrain analysis. Soil Sci. Soc. Am. J. 57, 443-452.
- 47 598
- 48
- 49 599 <http://dx.doi.org/10.2136/sssaj1993.03615995005700020026x>.
- 50
- 51
- 52
- 53 600 Nelson, D.W., Sommers, L.E., 1982. Total carbon, organic carbon, and organic Matter,
- 54
- 55 in: Page, A.L., Miller, R.H., Keeney, D.R. (Eds), Methods of soil analysis, Part
- 56 601
- 57
- 58 602 2. Chemical and microbiological properties. American Society of Agronomy,
- 59
- 60
- 61
- 62
- 63
- 64
- 65

- 1 603 Soil Science Society of America, Madison, Wisconsin, USA, pp. 539-579.  
2  
3  
4 604 Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1994. Spatial prediction of soil  
5  
6 605 properties from landform attributes derived from a digital elevation model.  
7  
8  
9 606 Geoderma 63, 197-214. [http://dx.doi.org/10.1016/0016-7061\(94\)90063-9](http://dx.doi.org/10.1016/0016-7061(94)90063-9).  
10  
11 607 Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1995. Further results on  
12  
13 608 prediction of soil properties from terrain attributes: heterotopic cokriging and  
14  
15 609 regressionkriging. Geoderma 67, 215–226.  
16  
17  
18  
19  
20 610 [http://dx.doi.org/10.1016/0016-7061\(95\)00007-B](http://dx.doi.org/10.1016/0016-7061(95)00007-B).  
21  
22 611 Pasculli, A., Palermi, S., Sarra, A., Piacentini, T., Miccadei, E., 2014. A modelling  
23  
24 612 methodology for the analysis of radon potential based on environmental  
25  
26  
27 613 geology and geographically weighted regression. Environ. Modell. Softw. 54,  
28  
29 614 165-181. <http://dx.doi.org/10.1016/j.envsoft.2014.01.006>.  
30  
31  
32  
33 615 Quinn, P., Beven, K., Chevallier, P., Planchon, O., 1991. Prediction of hillslope flow  
34  
35 616 paths for distributed hydrological modelling using digital terrain models. Hydrol.  
36  
37  
38 617 Process. 5, 59-79. <http://dx.doi.org/10.1002/hyp.3360050106>.  
39  
40  
41 618 Quinn, P.F., Beven, K.J., Lamb, R., 1995. The  $\ln(a/\tan\beta)$  index: How to calculate it  
42  
43 619 and how to use it within the Topmodel framework. Hydrol. Process. 9, 161-182.  
44  
45  
46 620 <http://dx.doi.org/10.1002/hyp.3360090204>.  
47  
48  
49 621 Rawlins, B.G., Marchant, B.P., Smyth, D., Scheib, C., Lark, R.M., Jordan, C., 2009.  
50  
51 622 Airborne radiometric survey data and a DTM as covariates for regional scale  
52  
53 623 mapping of soil organic carbon across Northern Ireland. Eur. J. Soil Sci. 60,  
54  
55  
56 624 44-54. <http://dx.doi.org/10.1111/j.1365-2389.2008.01092.x>.  
57  
58  
59  
60  
61  
62  
63  
64  
65

- 1 625 Schloeder, C.A., Zimmerman, N.E., Jacobs, M.J., 2001. Comparison of methods for  
2  
3 626 interpolating soil properties using limited data. *Soil Sci. Soc. Am. J.* 65,  
4  
5  
6 627 470-479. <http://dx.doi.org/10.2136/sssaj2001.652470x>.  
7  
8  
9 628 Somarathna, P.D.S.N., Malone, B.P., Minasny, B., 2016. Mapping soil organic carbon  
10  
11 629 content over New South Wales, Australia using local regression kriging.  
12  
13  
14 630 *Geoderma Regional* 7(1), 38-48.  
15  
16  
17 631 <http://dx.doi.org/10.1016/j.geodrs.2015.12.002>.  
18  
19  
20 632 Song, X.D., Brus, D.J., Liu, F., Li, D.C., Zhao, Y.G., Yang, J.L., Zhang, G.L., 2016.  
21  
22 633 Mapping soil organic carbon content by geographically weighted regression: A  
23  
24  
25 634 case study in the Heihe River Basin, China. *Geoderma* 261, 11-22.  
26  
27  
28 635 <http://dx.doi.org/10.1016/j.geoderma.2015.06.024>.  
29  
30  
31 636 Sun, W., Minasny, B., McBratney, A., 2012. Analysis and prediction of soil  
32  
33 637 properties using local regression-kriging. *Geoderma* 171-172, 16-23.  
34  
35  
36 638 <http://dx.doi.org/10.1016/j.geoderma.2011.02.010>.  
37  
38  
39 639 Sun, W., Zhu, Y., Huang, S., Guo, C., 2015. Mapping the mean annual precipitation  
40  
41  
42 640 of China using local interpolation techniques. *Theoretical and Applied*  
43  
44  
45 641 *Climatology*, 2015. 119(1): p. 171–180.  
46  
47  
48 642 <http://dx.doi.org/10.1007/s00704-014-1105-3>.  
49  
50  
51 643 Tang, G., Yang, X., 2006. Experimental course of ArcGIS spatial analysis. Science  
52  
53 644 press, Beijing, China.  
54  
55  
56 645 Walter, C., McBratney, A.B., Donuaoui, A., Minasny, B., 2001. Spatial prediction of  
57  
58  
59 646 topsoil salinity in the Chelif valley, Algeria, using local ordinary kriging with  
60  
61  
62  
63  
64  
65

- 1 647 local variograms versus whole-area variogram. *Soil Res.* 39, 259-272.  
2  
3 648 <http://dx.doi.org/10.1071/SR99114>.  
4  
5  
6 649 Wang, K., Zhang, C., Li, W., 2013. Predictive mapping of soil total nitrogen at a  
7  
8 650 regional scale-A comparison between geographically weighted regression and  
9  
10 651 cokriging. *Appl. Geogr.* 42, 73-85.  
11  
12 652 <http://dx.doi.org/10.1016/j.apgeog.2013.04.002>.  
13  
14  
15 653 Wang, K., Zhang, C., Li, W., 2012. Comparison of geographically weighted  
16  
17 654 regression and regression kriging for estimating the spatial distribution of soil  
18  
19 655 organic matter. *GISci. Remote Sens.* 49, 915-932.  
20  
21 656 <http://dx.doi.org/10.2747/1548-1603.49.6.915>.  
22  
23  
24 657 Wang, S., Huang, M., Shao, X., Mickler, R. A., Li, K., & Ji, J. (2004). Vertical  
25  
26 658 distribution of soil organic carbon in China. *Environ. Manage.* 33(1),  
27  
28 659 S200-S209. <http://dx.doi.org/10.1007/s00267-003-9130-5>.  
29  
30  
31 660 Wu, J., Jones, B., Li, H., Loucks, O. L., 2006. *Scaling and uncertainty analysis in*  
32  
33 661 *ecology: methods and applications*. Springer, Dordrecht, the Netherlands.  
34  
35  
36 662 Ye, H., Huang, Y., Chen, P., Huang, W., Zhang, S., Huang, S., Hou, S., 2016. Effects  
37  
38 663 of land use change on the spatiotemporal variability of soil organic carbon in an  
39  
40 664 urban-rural ecotone of Beijing. *J. Integr. Agr.* 15, 918-928.  
41  
42 665 [http://dx.doi.org/10.1016/S2095-3119\(15\)61066-8](http://dx.doi.org/10.1016/S2095-3119(15)61066-8).  
43  
44  
45 666 Ye, H., Shen, C., Huang, Y., Huang, W., Zhang, S., Jia, X., 2015. Spatial variability  
46  
47 667 of available soil microelements in an ecological functional zone of Beijing.  
48  
49 668 *Environ. Monit. Assess.* 187(13), 1-12.  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- 1 669 <http://dx.doi.org/10.1007/s10661-014-4230-7>.
- 2  
3 670 Zhang, C., Tang, Y., Xu, X., Kiely, G., 2011. Towards spatial geochemical modelling:  
4  
5 671 use of geographically weighted regression for mapping soil organic carbon  
6  
7 672 contents in Ireland. *Appl. Geochem.* 26, 1239-1248.  
8  
9  
10 673 <http://dx.doi.org/10.1016/j.apgeochem.2011.04.014>.
- 11  
12  
13 674 Zhang, S., Huang, Y., Shen, C., Ye, H., Du, Y., 2012. Spatial prediction of soil organic  
14  
15 675 matter using terrain indices and categorical variables as auxiliary information.  
16  
17  
18 676 *Geoderma* 171, 35-43. <http://dx.doi.org/10.1016/j.geoderma.2011.07.012>.
- 19  
20  
21 677 Zhu, A.X., Liu, J., Du, F., Zhang, S.J., Qin, C.Z., Burt, J., Behrens T., Scholten, T.,  
22  
23  
24 678 2015. Predictive soil mapping with limited sample data. *Eur. J. Soil Sci.* 66(3),  
25  
26  
27 679 535-547. <http://dx.doi.org/10.1111/ejss.12244>.

28  
29 680  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## 681 Figures

682

683 Fig. 1. Distribution map of the elevation, validation sites, and calibration sites  
684 under different sampling densities.

685

686 Fig. 2. Trend analysis of the SOC content under different sampling densities.

687

688 Fig. 3. Maps of the regression coefficients obtained by the GWR model in the  
689 case of Density 4.

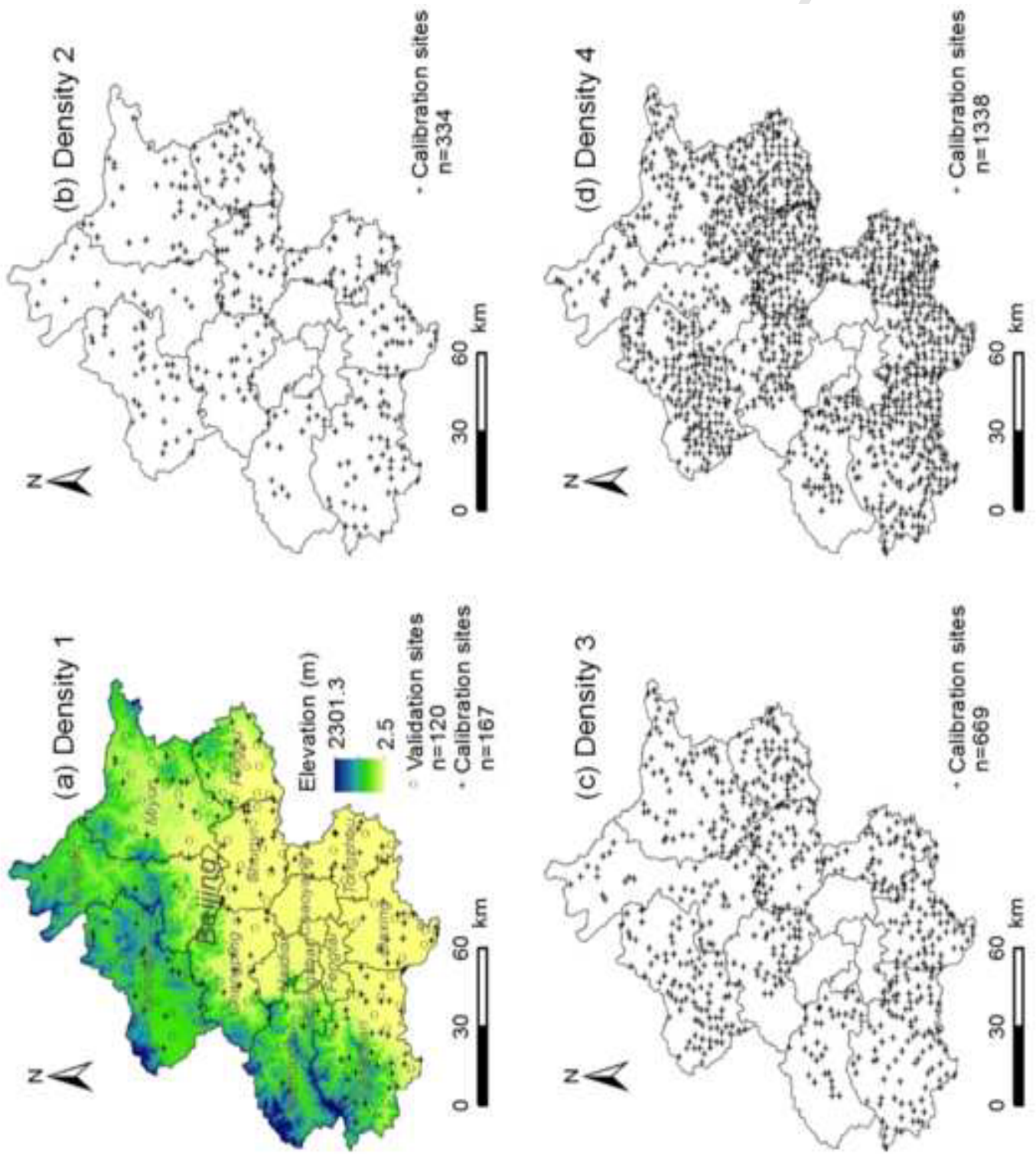
690

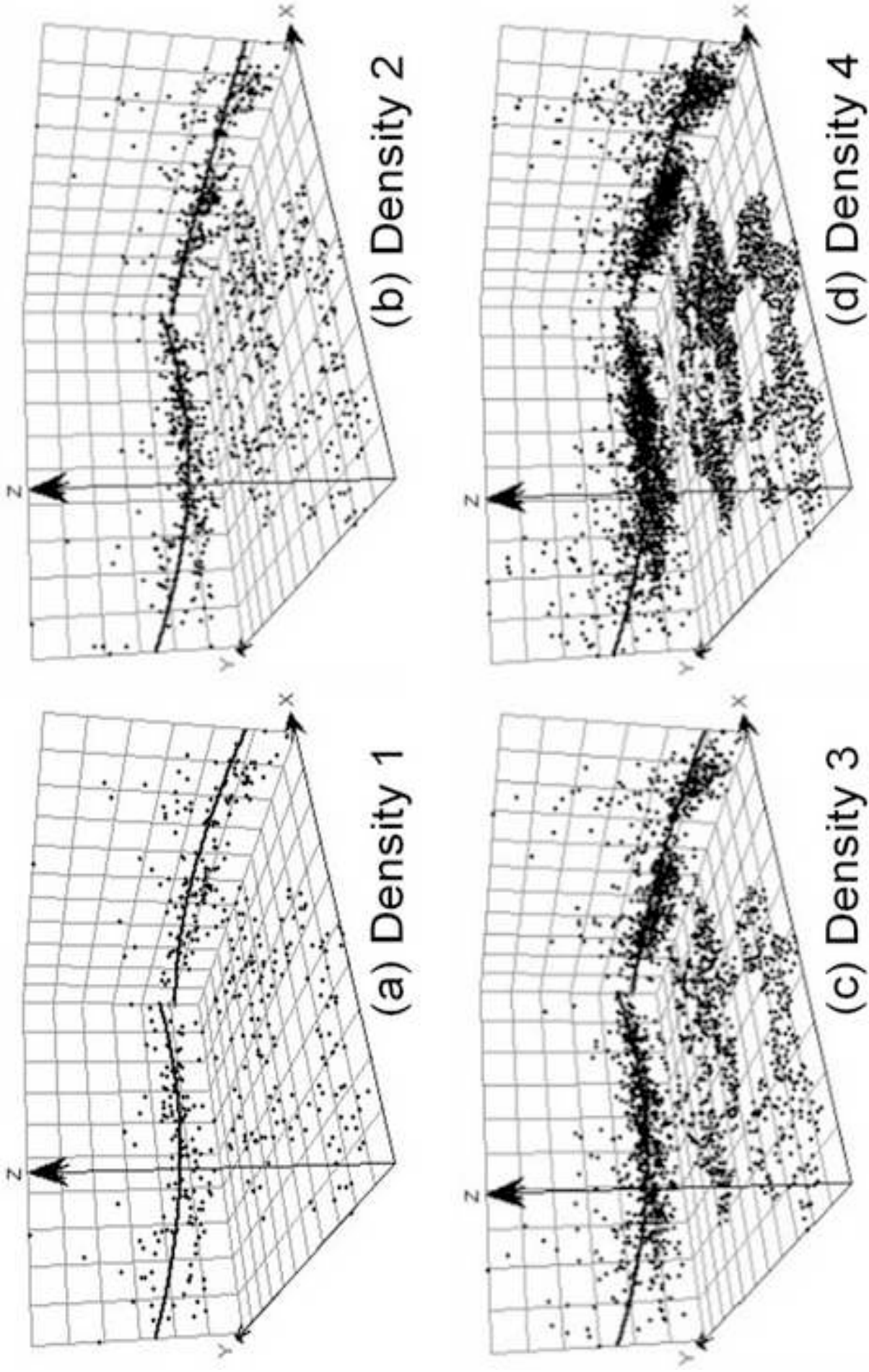
691 Fig. 4. Semivariogram models of the LnC, MLR residuals and GWR residuals  
692 under different sampling densities.

693

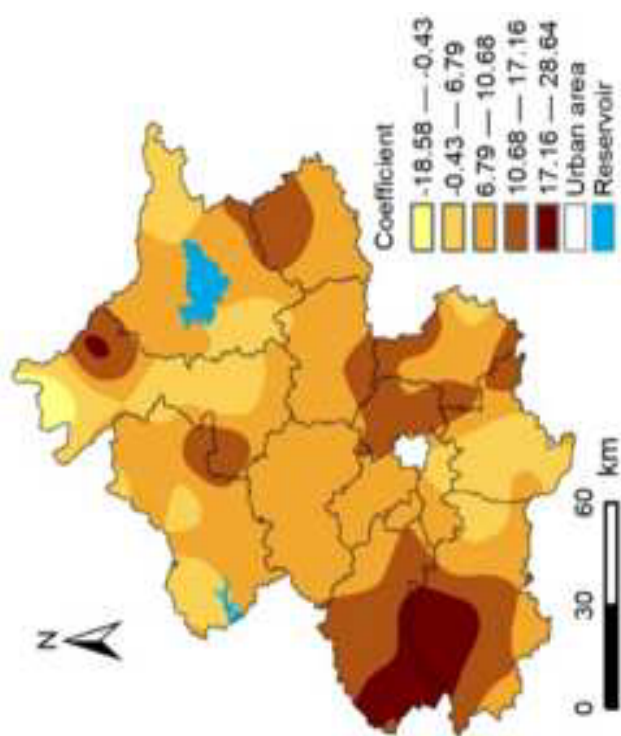
694 Fig. 5. Maps of the spatial distribution of the SOC content generated by the OK,  
695 MLRK and GWRK methods under different sampling densities.

696



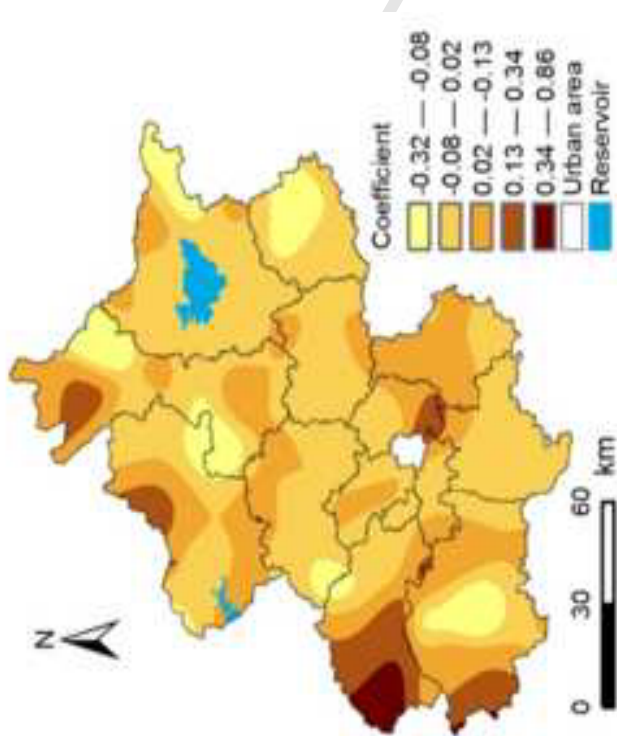


X, east direction; Y, north direction; Z: SOC content

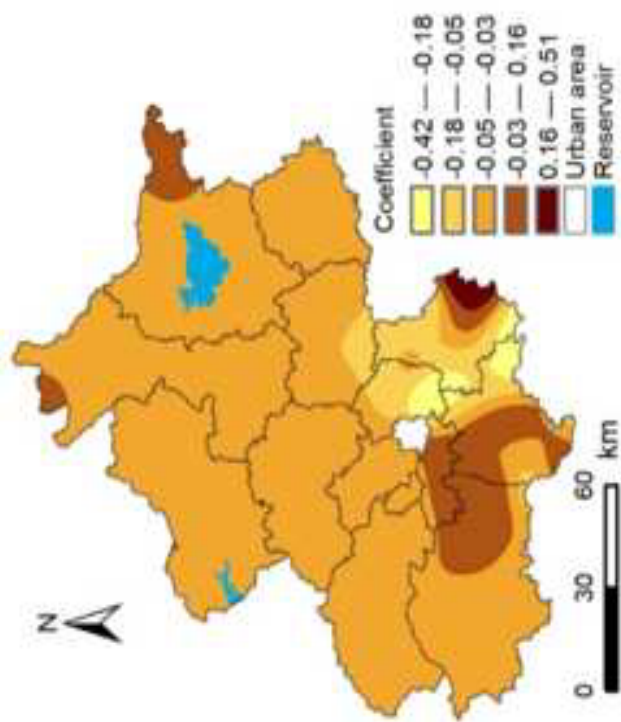


(a) Intercept

(b) Elevation

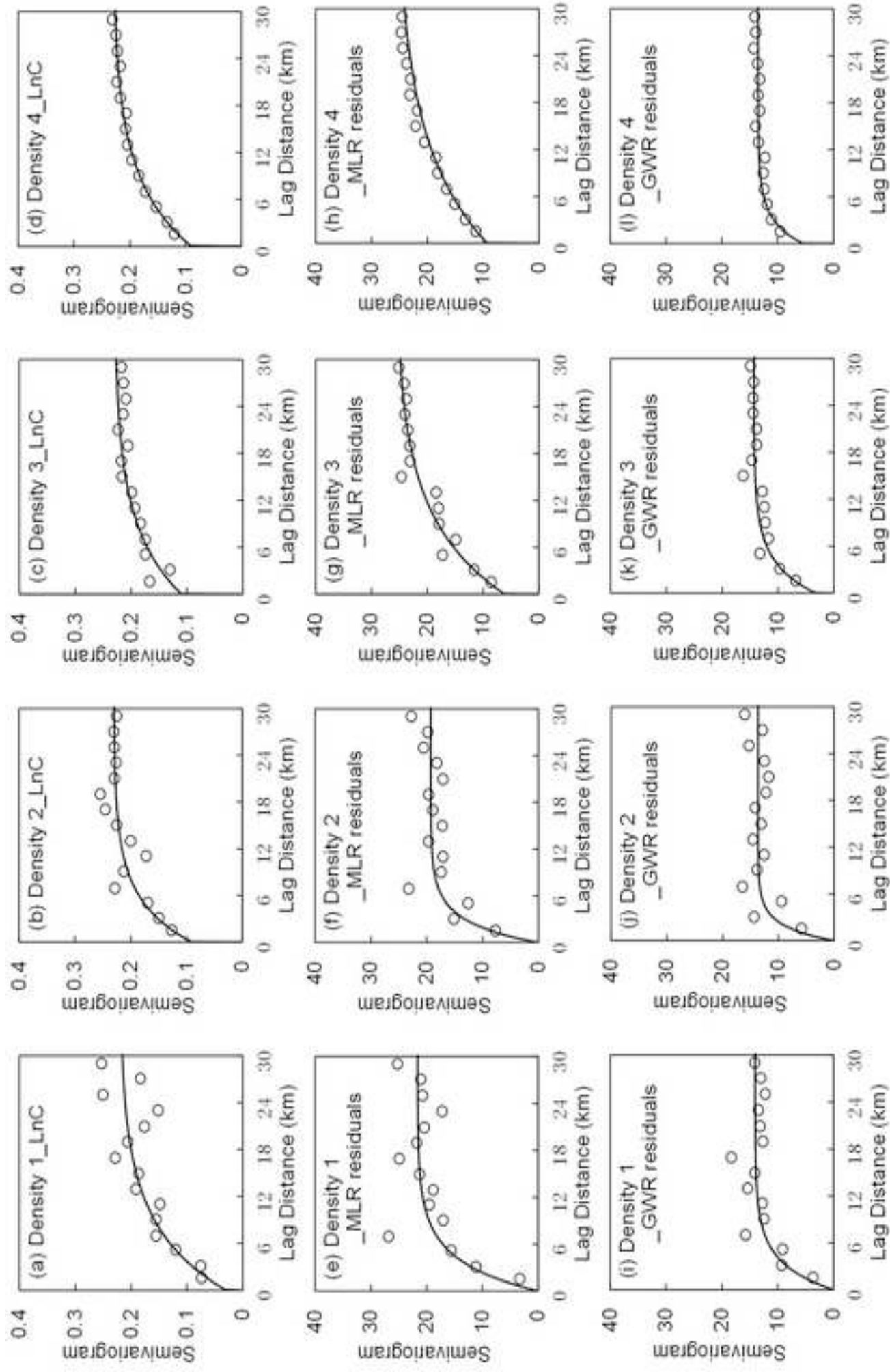


(d) Topographic wetness index (TWI)



(c) Slope

Figure 4  
[Click here to download high resolution image](#)



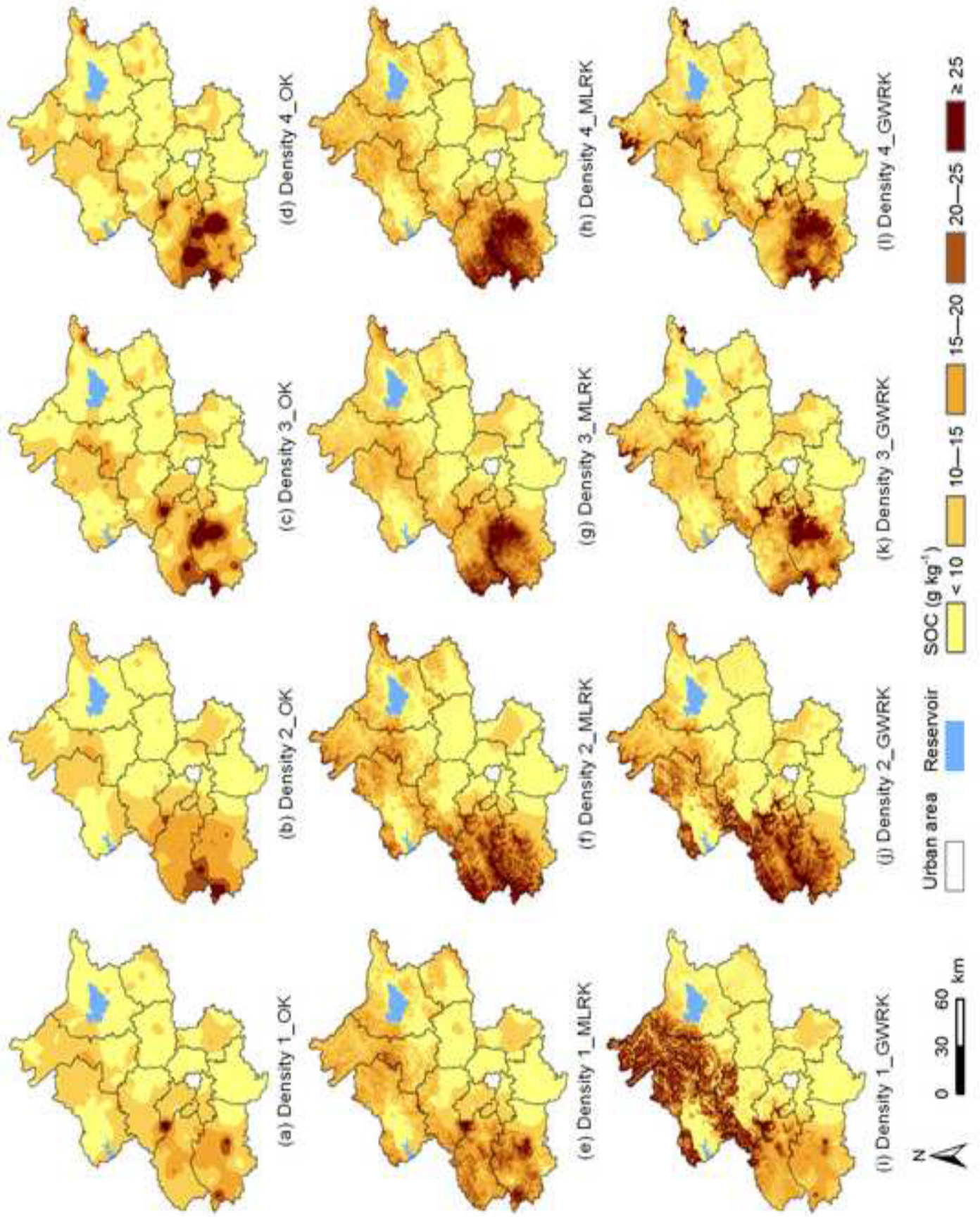


Figure 5  
[Click here to download high resolution image](#)