ISEIS

Journal of Environmental Informatics 26(2) 91-105 (2015)

Journal of Environmental Informatics

www.iseis.org/jei

Improving Environmental Prediction by Assimilating Auxiliary Information

Y. Yang[1,2], C. T. Zhang[1,2], R. X. Zhang[1,2], and G. Christakos[3,*]

[1]*Key Laboratory of Arable Land Conservation (Middle & Lower Reaches of Yangtse River), Ministry of Agriculture, Wuhan 430070, China*
[2]*College of Resources & Environment, Huazhong, Agricultural University, Wuhan 430070, China*
[3]*Ocean College, Zhejiang University, Hangzhou 310058, China*

Received 27 September 2014; revised 15 January 2015; accepted 29 January 2015; published online October 23, 2015

ABSTRACT. The concern of this work is the systematic synthesis of site-specific samples and auxiliary information (including continuous and categorical variables) aiming at improving spatial prediction of natural attributes (soil properties, contaminant processes etc.). Bayesian Maximum Entropy (BME) is the theoretical support of the proposed synthesis. The significance of the synthesis is that it can increase the prediction accuracy of natural attributes in a physically meaningful and technically efficient manner. The spatial prediction approach is applied in a real world case study that combines soil organic matter (SOM) content samples with auxiliary information (terrain indices, soil types, and soil texture) to generate predictive maps. Prediction was affected by soil type and soil texture (prediction accuracy increased when categorical variables were included). In the same case study, the BME-based approach was compared with mainstream spatial statistics techniques, like Regression Kriging (RK) with auxiliary information, and hard data-driven Ordinary Kriging (OK). The numerical results demonstrated the superiority of the BME-based approach over the Kriging-based techniques, whereas it was found that some key BME parameters (counts of soft data, predicted variables categories, and continuous auxiliary variable categories) can have different effect on SOM prediction accuracy. The success of BME-based prediction relied heavily on finding adequate auxiliary information about the study situation.

Keywords: spatial statistics, prediction, BME, Kriging, auxiliary information, soil, nitrogen

1. Introduction

Many real world case studies are characterized by considerable amounts of auxiliary yet valuable information (categorical data, soft measurements etc.), a fact that highlights the need for sound techniques capable of assimilating efficiently diverse sources of information and generating improved soil maps (Hengl et al., 2007; Cao et al., 2014). There are various studies supporting the above considerations. If terrain attributes are used as co-variables, soil prediction can be improved considerably compared to purely data-driven Ordinary Kriging (Herbst et al., 2006). The incorporation of historical climate records, including drought indices, temperature, and precipitation is shown to enhance the quality of short-term forecast of drought indices (Liu and Hwang, 2014). Soil fuzzy membership values can increase soil prediction accuracy in areas characterized by complicated soil-terrain relationships (Zhu et al., 2010). Correlated auxiliary information improves considerably soil mapping accuracy (Lamsal et al., 2006; Li, 2010; Pei et al., 2010). Auxiliary information used for prediction include digital elevation model (DEM) and its derived terrain attributes, soil and land use mapping units, spectral readings and vegetation indices of remote sensing images, as well as other related natural attributes in a sample form (Messier et al., 2012; Yu and Wang, 2013).

Several Kriging-based techniques of classical geostatistics have used in environmental prediction, like Regression Kriging, Stratified Kriging, Co-Kriging, and Residual Kriging (Odeh et al., 1994; Brus et al., 1996; Odeh et al., 1996; Yang et al., 2004; Eldrandaly and Abu-Zaid, 2011). Classical geostatistics assumes linear prediction and Gaussian probability distributions, and it cannot not make full use of auxiliary information (Orton and Lark, 2007; Lee et al., 2008). The Bayesian Maximum Entropy theory (BME; Christakos, 1990, 1992) of modern geostatistics differs from the classical geostatistics theory in a number of key aspects: it provides the theoretical support to consider non-linear prediction and non-Gaussian probability distributions, in general, and enables the integration of auxiliary information, physical laws and different kinds of empirical relationships in a physically meaningful and mathematically rigorous manner (Christakos, 2000; Bogaert and D'Or, 2002; Yu et al., 2007b, 2013). BME applications can be found in earth and environmental sciences, ecology, public health and epidemiology (Serre et al., 2003; Bogaert and Wibrin, 2004; Douaik et al., 2004; Gesink Law et al., 2006; Wibrin et al., 2006; Orton and Lark, 2007; Yu et al., 2007a; Lee et al., 2008; Kolovos et al., 2010).

The present work acknowledges that a key element of

* Corresponding author. Tel.: +86 571 88981701.
E-mail address: gchristakos@zju.edu.cn (G. Christakos)

ISSN: 1726-2135 print/1684-8799 online
 doi: 10.3808/jei.201500316

spatial prediction is the distinction between the original study variables and newly introduced auxiliary ones, which must be properly identified and rigorously formulated. The work explores the information content of auxiliary variables, both continuous and categorical, transforming them into BME theory terms, and applying the resulting process in the study of soil organic matter (SOM) content in a region of China. It is investigated whether and to what extent the integration of predicted and auxiliary variables can affect spatial prediction leading to more accurate SOM maps. Relationships between prediction accuracy and certain BME parameters (counts of soft data, predicted variable categories, continuous auxiliary variable categories) are examined. The BME prediction results are subsequently compared to those obtained using Kriging-based techniques.

2. BME Methodology

Consider a spatially varying natural attribute, such as the SOM content of the Qingshan dataset considered in the case study below. SOM is one of the key soil quality indicators, and, accordingly, accurate information about the SOM spatial variation is important in the sustainable soil utilization and management. The SOM content can be mathematically represented by the spatial random field (Christakos, 1992):

$$Z(\mathbf{s}) \sim f_Z \tag{1}$$

where the vector $\mathbf{s} = (s_1, s_2) \in R^2$ defines the coordinates of the location of the SOM content value, and the "~" denotes that the spatial random field is defined by its probability density function (pdf) f_z and does not have a single value at each location $\boldsymbol{s}$ (this pdf represents the local randomness combined with the spatial structure of the SOM content distribution). Let the SOM dataset z_{data} consist of hard (accurate) data z_{hard} and soft (uncertain) data z_{soft}. SOM content predictions are usually sought at unsampled locations across space. Different kinds of z_{soft} include, e.g., intervals of SOM content estimated from old soil maps based on polygon's color and legends, and probabilistic functions of secondary information originated from historical attribute data or from fuzzy data obtained by means of other methods (Gesink Law et al., 2006; Heywood et al., 2006; Jiang and Woodbury, 2006; Yu et al., 2007b, 2010).

Since utilizing spatially correlated auxiliary information to improve the prediction accuracy of soil properties is widely recognized, next we discuss a technique by means of which auxiliary information can be transformed into pdf to be used in soil characterization and spatial prediction. Given that its observed range and width are $[z_{min}, z_{max}]$ and $\delta z = z_{max} - z_{min}$, respectively, the attribute $Z_{(s)}$ can be divided into n categories so that the corresponding value range of the k-th category is:

$$Z_k \in [z_{\min} + \frac{k-1}{n}\delta z, z_{\min} + \frac{k}{n}\delta z] \tag{2}$$

with the obvious mid-value $\overline{Z_k} = z_{\min} + (k - \frac{1}{2}) / n \cdot \delta z$, and sample points divided into n groups. If the auxiliary variable A is a continuous variable, the range of its values is divided into n_A categories (A_1, ..., A_n). If A is a categorical variable, n_A is the count of A's type, and (A_1, ..., A_n) are the various types. By traversing each Z group's sampling locations and recording the A category at every location, the quantitative relationship between Z and A_i (i = 1, ..., n_A) is:

$$R(Z, A_i) = \left((\frac{Count(1)_i}{Count_i}, Z_1), ..., (\frac{Count(n)_i}{Count_i}, Z_n) \right) \tag{3}$$

where $count_i$ is the count of sampling points belonging to category A_i, and $count(k)_i$, k = 1, ..., n, is the count of sampling points simultaneously belonging to categories Z_k and A_i.

If at a soft data location the observed A-value belongs to category A_i, then the probability distribution of the predicted attribute at the specified location is $P(Z,Ai) = R(Z,Ai)$. If we have other auxiliary variables, B, C etc, using the above method we can get the probability distribution of the corresponding predicted variables, $P(Z,B) = R(Z,B)$ etc. The underlying premises are that (*a*) these auxiliary variables can be observed at the predicted locations, and (*b*) their values are not beyond the range obtained at the sampling points. Considering the different correlations between Z and the auxiliary variables, the Spearman correlation coefficients between Z and the categorical auxiliary variables, and the Pearson correlation coefficients between Z and the continuous auxiliary variables, denoted as r_A, r_B, r_C…, are introduced as weights. Before they are used in the calculations, the correlation coefficients are normalized, e.g., $r_A = |r_A| / (r_A| + |r_B| + |r_C| + ...)$, so that their sum equals 1. Finally, in light of the relationships between Z and r_A, r_B, r_C…, the pdf $f_S(z_{soft})$ of Z at the soft data locations can be obtained. For example, if the auxiliary variable B is observed at the soft data location, and its value belongs to category B_3, one finds:

$$\begin{aligned} f_S &= R(Z,(A,B,...)) = \textstyle\sum_{J=A,B,...} P(Z,J) r_J \\ &= \left(\left((\frac{Count(1)_1}{Count_1})_A r_A + (\frac{Count(1)_3}{Count_3})_B r_B + ..., Z_1 \right), ..., \right. \\ &\quad \left. (\frac{Count(n)_1}{Count_1})_A r_A + (\frac{Count(n)_3}{Count_3})_B r_B + ..., Z_n \right) \end{aligned} \tag{4}$$

Otherwise said, Equation (4) is the soft datum at the specified location according to the relationship between predicted and auxiliary variables. Note that the case of collinearity among continuous auxiliary variables must be examined. If collinearity is confirmed (i.e., some of these variables experience strong linear relationships), principal component analysis (PCA) is performed to combine these variables into independent ones, before generating soft data.

In the computational procedure above, there are several parameters that need to be defined. The first one is the number n of groups in Equation (2). In previous studies, the n was determined by empirical formulas like (Sturges, 1926), $n = 1 + 3.32 \log N_{sample}$, N_{sample} = number of sampling points. The procedure used in the present study, however, allows valuable

flexibility in the choice of the n-value, see following section, so that the effect of varying n-values on prediction accuracy can be determined. The second parameter is the count of continuous auxiliary variable categories n_A. If n_A is too small, the types of soft data will be limited, whereas if it is too large, the number of sampling points belonging to each auxiliary variable category will be low. In this work, in order to assess the effect of varying n_A on prediction accuracy, we let n_A = 5, 10, 15 (the last value is the number of soft data points); and to compare prediction accuracy for different soft data densities, we set $N_{soft} = lN_{sample}$ (l = 1, …, 5).

The objective of the SOM content study was spatial prediction based on the Bayesian Maximum Entropy (BME) theory. The basic set of BME equations are (Christakos, 2000):

$$\left.\begin{aligned} \int dz_{map}(\boldsymbol{g}-\bar{\boldsymbol{g}})\, e^{\boldsymbol{\mu}\cdot\boldsymbol{g}} = 0 \\ \int d\boldsymbol{\xi}_S\, e^{\boldsymbol{\mu}\cdot\boldsymbol{g}} - A f_K(z_k) = 0 \end{aligned}\right\} \quad (5a\text{-}b)$$

where $\boldsymbol{g}$ is a vector of functions expressing mathematically the available core (or general) knowledge base (G-KB), including spatial dependence models, physical laws, and scientific theories; $\bar{\boldsymbol{g}}$ denotes the mean value of $\boldsymbol{g}$; $\boldsymbol{\xi}_S$ represents the available site-specific knowledge base (S-KB), including samples and auxiliary information (Bogaert, 2002, 2004); $\boldsymbol{\mu}$ is a vector of coefficients representing the relative importance of each $\boldsymbol{g}$-function; and A is a normalization parameter. Equations (5a-b) can be solved with respect to the pdf $f_K(z_k)$ of the unsampled SOM values z_k at all map locations of interest (i.e., locations at which predictions are sought across space). Software libraries have been developed dealing with the solution of Equations (5a-b) in real world conditions, including BMElib, SEKS-GUI, Quantum BME, and StarBME (Christakos et al., 2002; Yu et al., 2007b, 2013). BME techniques can account for the influence of categorical variables (soil texture and soil genetic types, land use types, etc.) and can systematically analyze the relationship between auxiliary variables and the soil attributes to be predicted.

As usually happens in similar cases, the proposed BME-based prediction technique was compared to other spatial prediction techniques: the hard data-driven Ordinary Kriging (OK), and the Regression Kriging (RK) with the same auxiliary information were implemented to predict SOM contents at a set of test sample locations s_j (j = 1, …, n). Since both the OK and the RK techniques are described in detail in the literature, no further technical details are given here, and the readers are referred to the relevant geostatistics literature (e.g., Odeh et al 1994; Olea, 1999).

Let $Z(s_j)$ be the SOM content at location s_j, and $\hat{Z}_T(s_j)$ the corresponding prediction obtained by the technique T = BME, OK, RK. Three quantitative measures of accuracy were computed for each spatial prediction technique: the Pearson correlation coefficient r_T measuring the strength of the linear relation between $\hat{Z}_T(s_j)$ and $Z(s_j)$ (the closer to 1 the r_T is, the more accurate the prediction), the mean error ME_T of $\hat{Z}_T(s_j)$ vs. $Z(s_j)$ over the test set of sample points (should be as close to 0 as possible), and the root mean squared error $RMSE_T$ over the same set of points (should be minimized). Moreover, the % improvement on prediction precision of a technique relative to the reference technique (in this case, OK) is measured by the relative indicator (Sumfleth and Duttmann, 2008):

$$RI_T = \frac{RMSE_{OK} - RMSE_T}{RMSE_{OK}} \times 100\% \quad (6)$$

where T = BME, RK. $RI_T > 0$ implies that the T technique is more accurate than OK (the higher the RI, the higher T's accuracy); and $RI_T < 0$ implies that T is less accurate than OK.

3. The Qingshan Dataset

The Shayang dataset used in this study consists of 288 soil samples (surface soils, at 0 ~ 20 cm depth), selected during the autumn of 2007 throughout the Shayang County situated in the central region of Hubei province of China (Figure 1). The locations of the sampling sites were recorded using a global positioning system over an area of 2044 km^2 characterized by a northwestern elevation and a low southeastern terrain (lowest and highest elevations are, respectively, 20 and 143 m). The SOM content distribution, which is the focus of the present study, is related to environmental parameters, like topography, soil type, and soil texture (Simbahan et al., 2006; Zhang et al., 2012). SOM played the role of the study predicted attribute, and the SOM content was determined using the pot assium dichromate-wet combustion procedure (NSS, 1995). The available dataset was divided into two sub-sets (Figure 1): (a) The first one included 88 randomly selected samples and served as the validation (test) sub-set assessing the performance of the different prediction techniques. (b) The second one included the remaining 200 samples and served as the training sub-set. Note that soil texture and soil type were considered simultaneously. The spatial prediction techni-

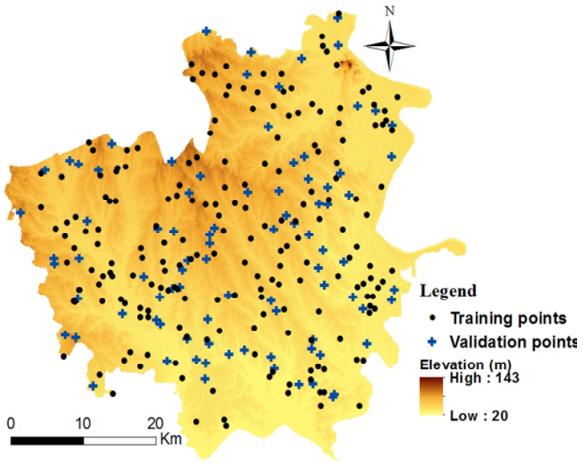


Figure 1. Map of soil sampling sites and associated elevation (Shayang County, China).

ques used were the BME-based technique and the RK technique. The SOM content distribution and summary statistics (training sub-set) are displayed in Figure 2. The calculated Kolmogorov-Smirnov (K-S) test value was 0.019 ($P < 0.005$) and, hence, it was concluded that the SOM content values at the training points did not follow a normal distribution.

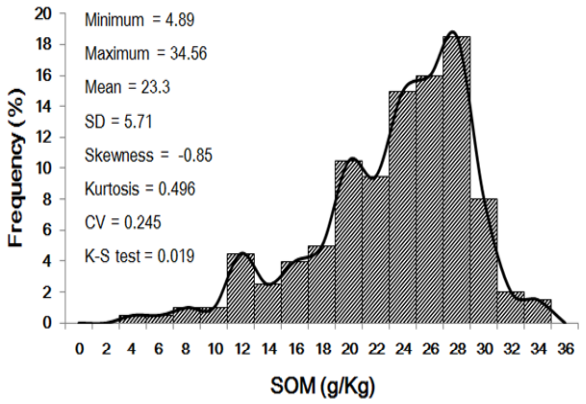


Figure 2. Statistical SOM features of the training sub-2et and significance level of Kolmogorov-Smirnov test.

A DEM was used for terrain analysis based on a 50 m grid, whereas primary and secondary terrain attributes were derived from DEM using the terrain analyzing module of the ArcMap 10.2 software. Terrain factors are closely associated with water transportation and substance migration affecting soil nutrient content. The terrain variables included: elevation from sea-level *h* (in *m*); slope (*degrees*, 0 ~ 90º); aspect angle (*degrees*, counter-clockwise from East, 90° to the North, 180º to the West, 270º to the South, and 360° to the East); topographic relief amplitude *TRA* (the difference "maximum-minimum" elevation around a grid); slope of slope *SOS* (slope change rate); length-slope factor *LS* of the universal soil loss equation (*USLE*; Wischmeier and Smith, 1978) suitable for identifying erosion processes; wetness index *WTI*, a well-studied indicator of soil property and soil moisture distribution at different landscape position (Beven and Kirkby, 1979; Pei et al., 2010); stream power index *SPI*, indicative of runoff erosion potential (Moore et al., 1993); surface roughness *M*, establishing the ratio of surface area to projected area; river kinetic energy index *RKEI*, indicative of surface runoff capability. Relationships between SOM and terrain variables were calculated in terms of Pearson correlation analysis (Table 1).

Categorical variables have been generally used as auxiliary variables to improve prediction accuracy of soil attributes. In the present study, the categorical variables included soil texture and soil type obtained from the 1:50000 soil map of the Shayang County. The main soil type was paddy soil, corresponding to approximately 72.4% of the total area, the remaining area covered mainly by moisture and yellow brown soils. Soil texture was classified according to Kaczynski's classification system (Huang, 1999). Soil texture types included sandy loam, light loam, medium loam, and heavy loam, with heavy loam covering approximately 63% of the region. Since significant soil type and texture differences may result in SOM content variability, correlations between SOM and these categorical variables were investigated using analysis of variance (ANOVA) and Spearman correlation analysis.

Post-hoc (*a posteriori*) tests may expand considerably the range and capability of exploratory research techniques (these tests, e.g., limit the probability that significant effects will seem to have been detected between soil types when none actually exist). The present study is concerned with the significant differences in SOM content among different soil types. The relevant ANOVA results shown in Table 2 indicate that the average SOM content was highest in yellow brown soil (24.64 g/kg) and lowest in moisture soil (12.9 g/kg). The average SOM content is ranked as follows: moisture soil < paddy soil < yellow brown soil.

In particular, the SOM content in moisture soil was much lower than in paddy and yellow brown soils, whereas there was no significant difference between paddy and yellow brown soils (Spearman correlation coefficient = 0.276). It is worth-noticing that there are 21 training points without soil type information. Soil texture can affect several soil properties and processes. For illustration, the relevant ANOVA results for varying soil texture characteristics are listed in Table 3 (*post hoc* analysis). The average SOM content is ranked as follows: sandy loam < light loam < medium loam < heavy loam.

Specifically, the heavier the soil texture, the higher the SOM content. These findings are in agreement with those of some earlier studies (Zhang et al., 2012). The calculated Spearman correlation coefficient in Table 3 is 0.276 (i.e., same value as in Table 2), confirming the positive effect of soil texture on SOM content (again, there are 21 training points lacking soil texture information).

4. Soft Data Generation

Soft data were generated from the relationships between SOM content and auxiliary variables. Note that, as was mentioned earlier, according to the empirical formula $n = 1 + 3.32 \log N_{sample}$ the *n*-value should be 7 (and the sample points would be divided into 7 groups). However, in the present study a more realistic quantitatively analysis was considered letting $n = 5, 7, 9$ (Table 4). In this way, the effect on prediction accuracy produced by different *n*-values can be determined.

Table 1. Pearson Correlation Analysis of Terrain Variables and SOM

Variable	*h*	*β*	*α*	*TRA*	*SOS*	*LS*	*WTI*	*SPI*	*M*	*RKEI*
SOM	0.324**	0.212**	0.055	0.262**	0.23**	0.141*	0.005	0.233**	0.005	0.083

* Correlation is significant at 0.05 level (2-tailed).
** Correlation is significant at 0.01 level (2-tailed).

Table 2. Results of Post Hoc Tests in ANOVA with Mean SOM Values for Different Soil Types

Soil types	Number of samples	Mean and significance test	Rank
Moisture soil	19	12.9a	1
Paddy soil	150	24.62b	2
Yellow brown soil	10	24.64b	3
Spearman correlation coefficient 0.276**			

* Values in each row with the same letter are not significant ($p < 0.05$).
** $p < 0.01$ (2-tailed).

Table 3. Post Hoc Tests in ANOVA with Mean SOM Values for Different Soil Texture Characteristics

Soil texture	Number of samples	Mean and significance test	Rank
Sandy loam	7	12.76a	1
Light loam	41	20.69b	2
Medium loam	22	24.25c	3
Heavy loam	109	24.89c	4
Spearman correlation coefficient 0. 276**			

* Values in each row with the same letter are not significant ($p < 0.05$).
** $p < 0.01$ (2-tailed).

Table 4. Value Ranges and Counts of Sampling Points of Groups According to Different n

n	Value range, and count of sampling points for each value range
5	4.89-10.82(6), 10.82-16.76(19), 16.76-22.69(51), 22.69-28.63(97), 28.63-34.56(27)
7	4.89-9.13(4), 9.13-13.37(14), 13.37-17.61(13), 17.61-21.84(39), 21.84-26.08(58), 26.08-30.32(61), 30.32-34.56(11)
9	4.89-8.19(4), 8.19-11.48(6), 11.48-14.78(10), 14.78-18.08(14), 18.08-21.37(33), 21.37-24.67(37), 24.67-27.97(54), 27.97-31.26(36), 31.26-34.56(6)

The probability distribution patterns of SOM content according to soil type and soil texture (the two categorical auxiliary variables considered) and different number of SOM categories are shown in Figure 3 (left column) and Figure 3 (right column), respectively. For moisture soil and sandy loam soil, the probability distribution shapes are left-skewed, without high SOM content, indicating that the SOM content in moisture soil and sandy loam soil was comparatively low. For paddy soil, yellow brown soil type, heavy loam, and medium loam soil texture, however, the probability distribution shapes were right-skewed, implying that the SOM content in paddy soil, yellow brown soil, heavy soil, and medium loam was high by comparison. Furthermore, the probability distribution patterns were considerably different in paddy soil vs. yellow brown soil, and medium loam texture vs. heavy loam texture, although the average means of SOM content in those two kinds of soil type and the two kinds of soil texture were similar. For light loam soil, the probability shape was close to the normal distribution. In conclusion, the distribution patterns of SOM content vary significantly among different soil types and soil textures. On the other hand, the probability distribution patterns of SOM categories for different *n*-values were generally similar. One may also notice that bigger *n*-values may result in more detailed (informative) probability distributions.

Terrain factors that exhibited significant correlations with SOM were considered as auxiliary variables to be involved in subsequent calculations. In particular, *TRA*, *SOS*, *SPI*, *h*, β, and *LS* were selected as continuous auxiliary variables. Considering the relationships between these six auxiliary variables, before generating soft data, the collinearity diagnostics were performed to identify the collinearity relationship among the six auxiliary variables. As is shown in Table 5, some of the continuous auxiliary variables have strong linear relationships due to the corresponding eigenvalues being close to 0, with some of the condition indices being bigger than 10. PCA and Pearson correlation analysis were implemented to quantify in a systematic manner associations between SOM and the continuous auxiliary variables above. Two principal components were obtained, accounting for 72% of the total variance. Factor 1 is dominated by *SOS*, *TRA*, *h* and β, accounting for 46.12% of the total variance. Factor 2, is dominated by *LS* and *SPI*, accounting for 25.88% of the total variance. The regression equations of PCA are as follows:

$$F1 = 0.32LS + 0.431SPI + 0.728SOS + 0.898TRA + 0.636h + 0.859\beta \quad (7)$$

$$F2 = 0.895LS + 0.8SPI - 0.069SOS - 0.221RTA - 0.225h - 0.266\beta \quad (8)$$

The Pearson correlation coefficients between $F1$ and SOM, and between $F2$ and SOM, were 0.368 and 0.05, respectively. Hence, the first factor ($F1$), replacing original terrain factors, was utilized to generate soft data. In this case, the $F1$ of the training points ranged from 22.8 to 82.51. $F1$ was divided in to 5, 10, and 15 categories (marked as n_A) for

Table 5. Results of Collinearity Diagnostics

Model	Dimension	Eigenvalue	Condition index	Variance Proportions						
				Constant	*LS*	*SPI*	*SOS*	*h*	*β*	*TRA*
1	1	5.583	1	0	0	0	0.01	0	0	0
	2	0.857	2.552	0	0.53	0	0	0	0	0
	3	0.287	4.412	0.02	0.01	0.02	0.3	0.02	0.04	0
	4	0.175	5.65	0	0.01	0	0.65	0	0.16	0.02
	5	0.055	10.108	0.05	0.06	0.09	0	0.92	0.02	0
	6	0.028	14.045	0	0	0.03	0.03	0.03	0.74	0.94
	7	0.016	18.707	0.93	0.38	0.86	0.01	0.03	0.04	0.03

different category counts, leading to different prediction accuracies. Table 6 shows the probabilities of SOM categories according to different $F1$ intervals conditioned to $n = 7$ and $n_A = 10$. As is shown in Table 6, $F1$ was divided into 10 categories, and the corresponding probabilities were significantly different between each category and SOM. The results indicated that $F1$ could influence the pattern of SOM content.

We will use a numerical example to illustrate how to generate soft data at any location of the study area. Suppose we can observe 2 of the 3 auxiliary variables, soil type (= paddy soil) and $F1$ (= 50) at an unmeasured location. If, say, $n = 7$ and $n_A = 10$, the corresponding probabilities (%) of each category of the predicted SOM variable are shown in Table 7: 0.0, 0.86, 5.51, 14.7, 48.35, 25.68, and 4.9. Hence, in this example, although the soil texture is not available, we still obtained soft data based on the remaining auxiliary variables. Note that if all auxiliary variables were not available at the prediction location, the soft data were represented by the probability distribution at all training points.

5. Spatial Prediction

An objective of the present case study was to establish predictive models incorporating auxiliary information for spatial SOM content prediction purposes. In order to compare BME prediction accuracy assuming different spatial densities of soft data, we considered 5 groups of soft data points with corresponding counts n = 200, 400, 600, 800, and 1000. The coordinates of the points in every one of these group were random. Based on the procedure described earlier, we can obtain probability distributions at every soft data location. Figure 4 presents the spatial distribution of the fifth group of so-

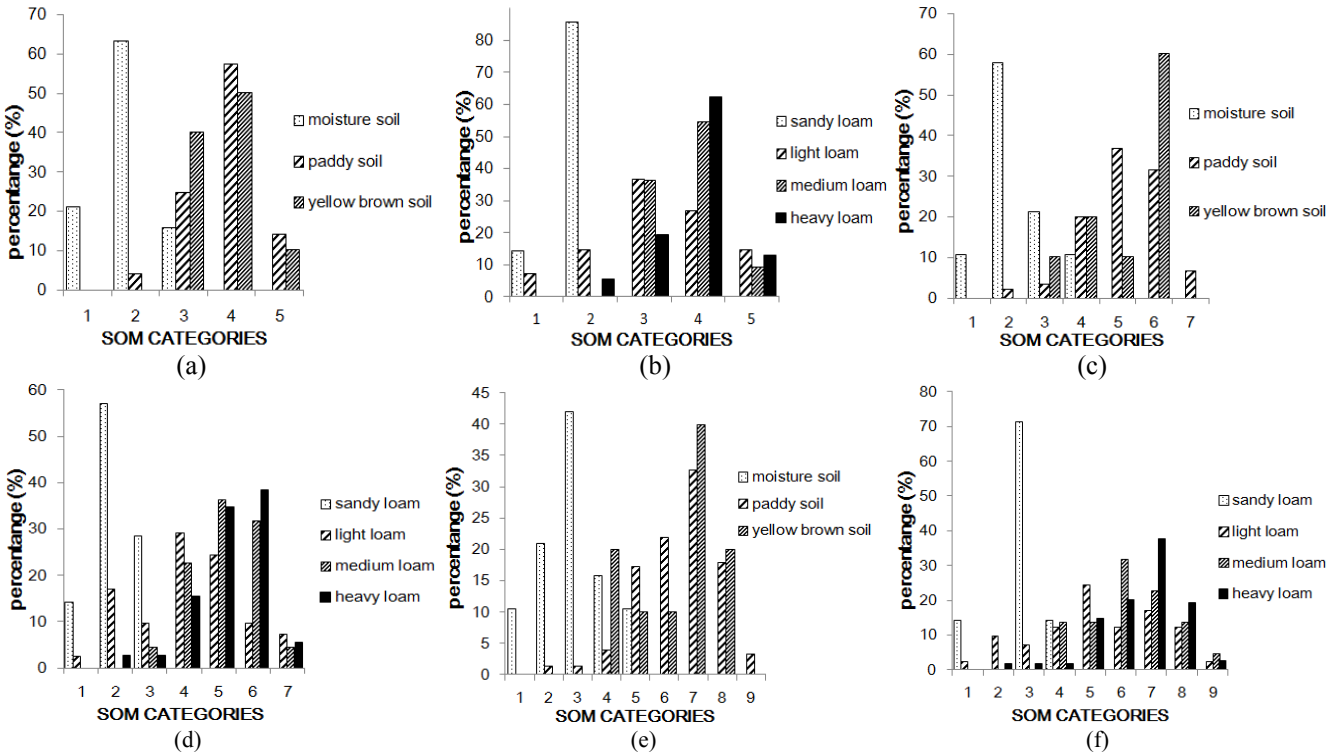


Figure 3. Probability distribution patterns of SOM according to soil types (left), soil texture (right) with n = 5 (a, b), 7 (c, d), and 9 (e, f).

Table 6. Probabilities of SOM Categories for Different *F*1 Intervals Conditioned to $n = 7$, $n_A = 10$

Ordinal	Value range (m)	Probability of each predicted SOM category (%)						
		1	2	3	4	5	6	7
1	22.80-28.77	14.28	33.33	9.52	19.05	9.52	9.52	4.76
2	28.77-34.74	2.44	12.19	9.76	21.95	21.95	29.27	2.44
3	34.74-40.71	0	4.17	12.5	29.17	25	25	4.17
4	40.71-46.68	0	3.13	3.13	25	21.87	40.62	6.25
5	46.68-52.66	0	0	7.14	10.71	57.14	21.43	3.57
6	52.66-58.63	0	0	0	16.67	27.78	50	5.55
7	58.63-64.60	0	0	0	5.26	36.84	42.11	15.79
8	64.60-70.57	0	0	14.29	28.57	28.57	28.57	0
9	70.57-76.54	0	0	0	14.29	28.57	42.86	14.29
10	76.54-82.51	0	0	0	50	50	0	0

Table 7. An Example of Calculation Soft Data at one Location

Variable	Correlation Coefficient	Weight	Value	Probability of each predicted SOM category (%)						
				1	2	3	4	5	6	7
Soil texture	0.255	0	Not available	--	--	--	--	--	--	--
Soil type	0.276	0.429	paddy soil	0	2	3.33	20	36.67	31.33	6.67
*F*1	0.368	0.571	50	0	0	7.14	10.71	57.14	21.43	3.57
Result				0	0.86	5.51	14.7	48.35	25.68	4.9

Table 8. Variogram Model Parameters of Square-transformed SOM Contents and SOM Residuals

Types	Model form	Range	Nugget (C_0)	Sill ($C_0 + C$)	C_0/Sill
Square(SOM)	Exponential	24000	17936	56817	0.316
Residuals	Exponential	24000	12.512	18.2	0.687

ft points ($n = 1000$), and three examples of probability distribution at the corresponding soft locations. It was noticed that the results of the K-S test ($P = 0.019 < 0.05$) showed that the SOM contents of the 200 training samples were not normally distributed. Yet, the distribution of the square transform of the SOM values was consistent with a normal distribution ($P = 0.478 > 0.05$).

Numerical SOM content predictions were generated in the Shayang County (China) by implementing first the OK technique of mainstream geostatistics (Olea, 1999). Specifically, by using the geostatistical analyst extension of ArcGIS10.2 (ESRI, 2013), the variogram models of the square (SOM) were obtained, see Table 8 and Figure 5(a). The predicted map of spatial SOM content distribution obtained using OK is displayed in Figure 6(a). One easily notices the rather smooth spatial pattern of the SOM map. The next two prediction techniques aimed at examining whether inclusion of categorical variables can improve the numerical accuracy of SOM prediction based on a systematic variability analysis. Specifically, Regression Kriging (RK; Odeh et al., 1994) is a popular technique used to combine sampling points and auxiliary variables characterized by a mesh distribution. According to Li (2010), RK's implementation usually involves three steps: performing multiple linear regression between the target variable and the auxiliary variables or environmental correlation, detrending local means of the regression, and calculating residuals of the regression in terms of the variograms and OK. In the present case study, in order to utilize categorical variables (including soil types and soil texture) in the RK context, the categorical variables were converted into dummy variables, using the procedure described next.

Soil types included three values: paddy soil, moisture soil, and yellow brown soil. We set two dummy variables for soil types, paddy and moisture, with the rules:

IF *paddy* = 1 and *moisture* = 0, THEN soilType = *paddy soil.*

IF *paddy* = 0 and *moisture* = 1, THEN soilType = *moisture soil.*

IF *paddy* = 0 and *moisture* = 0, THEN soilType = *yellow brown soil.*

In the same way, we set three dummy variables for soil texture: heavy, medium, and light, with the following rules:

IF *heavy* = 1, *medium* = 0, and *light* = 0 THEN soil texture = *heavy loam.*

IF *heavy* = 0, *medium* = 1, and *light* = 0 THEN soil texture = *medium loam.*

IF *heavy* = 0, *medium* = 0, and *light* = 1 THEN soil texture = *light loam.*

IF *heavy* = 0, *medium* = 0, and *light* = 0 THEN soil texture = *sandy loam.*

Usually, multiple linear stepwise regression (MLSR) is used to predict the deterministic component of the target vari-

able in RK. In the present study, the regression equation of the SOM model was expressed as:

$$SOM = 16.148 + 0.474SPI + 0.064h - 9.097Moisture + 1.544Heavy \quad (9)$$

where R^2 = 0.416, $p < 0.001$. The variogram models of the SOM residuals were obtained by using the geostatistical analyst extension of ArcGIS 10.2 (ESRI, 2013), see Figure 5(b) and Table 8. The resulting map of SOM content (50 m × 50 m grid size) obtained by RK is shown in Figure 6(b).

Lastly, the BME-based prediction technique integrated hard data (training points) and soft data derived from the physical relationships between SOM and auxiliary variables to predict the spatial distribution of SOM content. In order to compare BME accuracy for different densities of soft data, we considered 5 groups of soft data points. The count n_s of soft data points for these groups was set to 200, 400, 600, 800, and 1000. Also, different values were assumed for the count n_A of categories of continuous auxiliary variables, and the count n of categories of the predicted variable. The computational analysis was done using the BMElib toolbox (Christakos et al., 2002) written for Matlab. The generated BME maps are plotted in Figure 7 below.

As is shown in Figures 6 and 7, the general spatial trends of the SOM prediction maps produced by the three different techniques turned out to be rather similar (e.g., the SOM content was higher in the western-central part and lower in the eastern part of the study area). There were, however, certain important differences. Overall, the SOM prediction ranges obtained by the BME technique were closer to the sample ranges. The OK predictions have the narrowest SOM content range, due to OK's smoothing effect. The RK prediction range was narrower than that of BME and wider than that of OK. The RK polygons were more fragmentized than those of OK and BME, due to the linear regression equation of the SOM model being composed of four auxiliary variables, thus, resulting in higher spatial variability.

For accuracy assessment purposes, the outcomes of the four validation criteria (ME, RMSE, r and RI) for the three SOM prediction techniques discussed above are listed in Table 9. Clearly, the OK prediction performance is the poorest; it has the largest bias (ME), the largest RMSE, and the lowest r. The RK results show better SOM content predictions than those of OK, and poorer predictions than those of BME (the outcomes of the RK's four validation criteria are inbetween those of BME and OK). Overall, the BME technique demonstrated the best prediction accuracy (it has the smallest ME and RMSE, as well as the largest r). Furthermore, the choice of BME parameters can directly affect prediction accuracy. In this study, in particular, different values for three parameters were selected (number of soft points n_s), number of categories of predicted variable n. Concerning the n_s effect, it was found

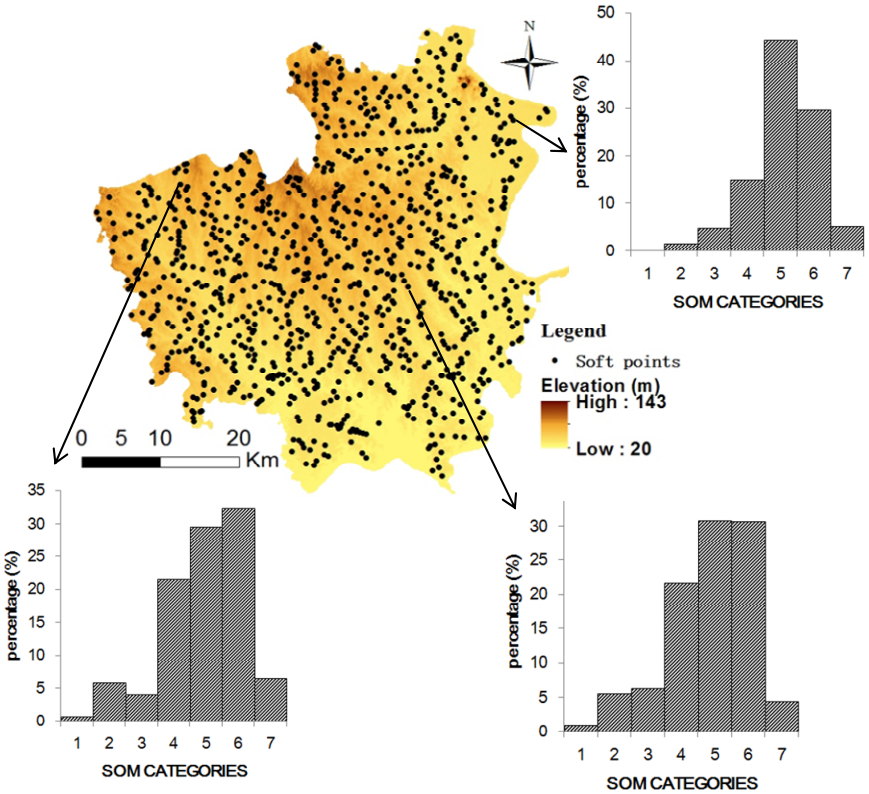
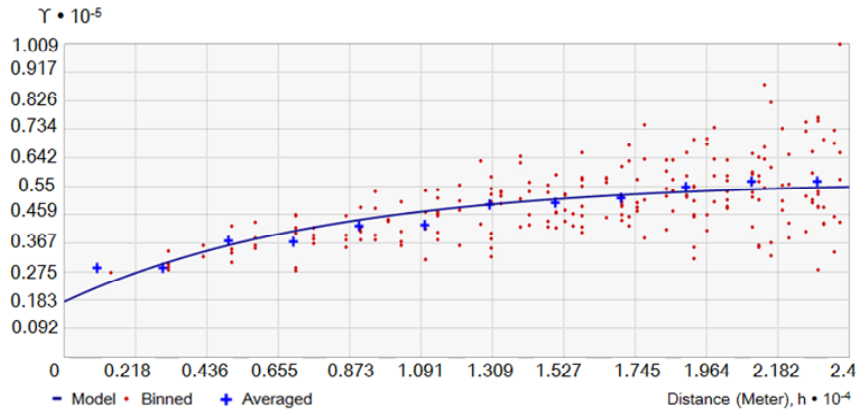
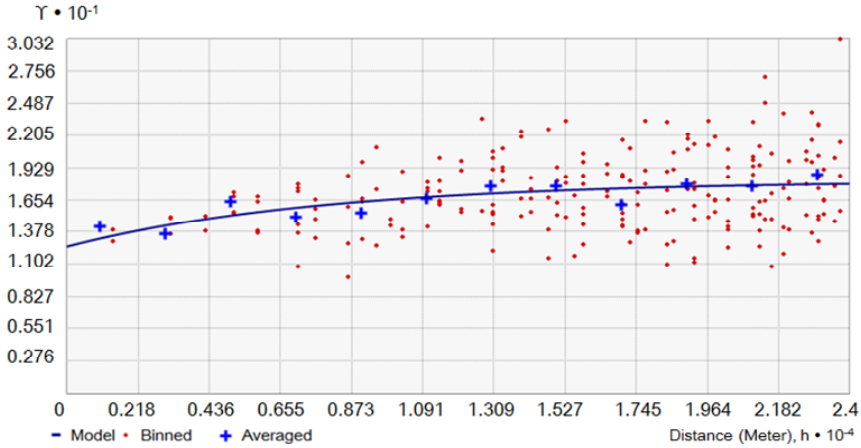


Figure 4. Spatial distribution of soft points (n = 1000), and three examples of probability distributions at soft data points.



(a)



(b)

Figure 5. Variogram models for (a) SOM content transformed by squares and (b) SOM residual.

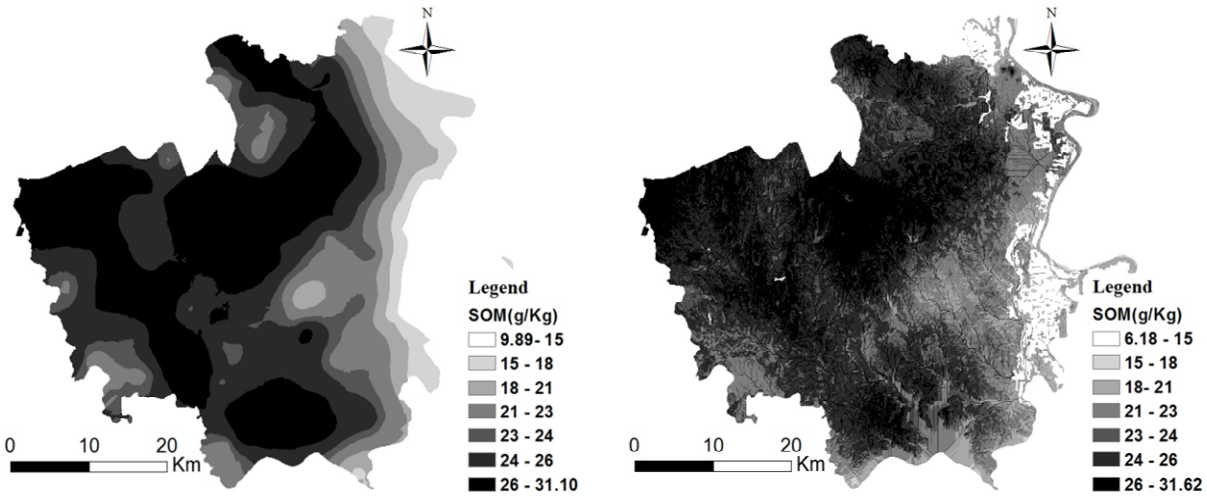


Figure 6. Maps of spatial distribution of SOM using OK (a) and RK (b).

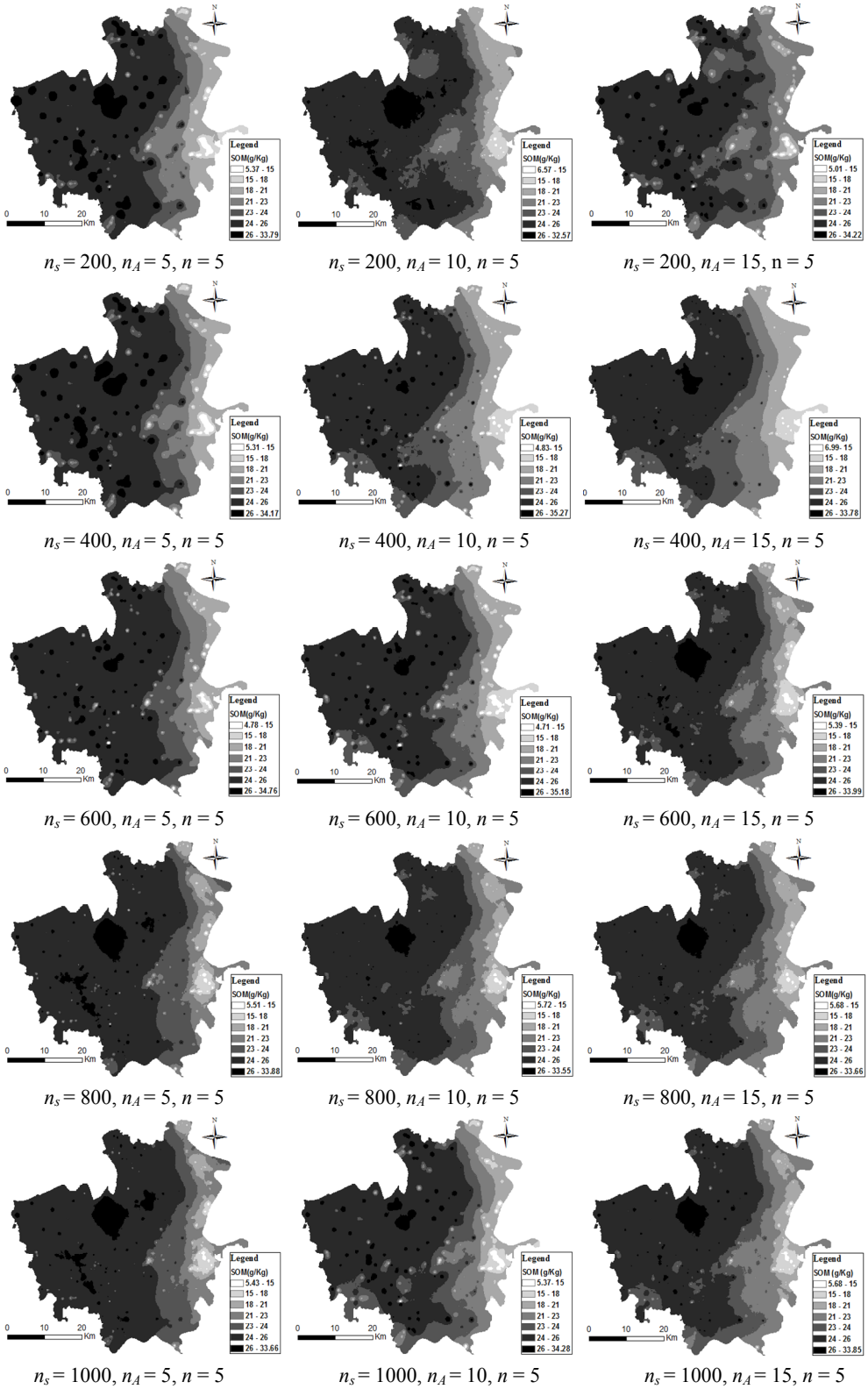


Figure 7 (Part 1). Maps of spatial SOM estimates using BME with different parameters (n_s: count of soft points; n_A: count of categories of continuous auxiliary variable; n: count of categories of predicted variable.

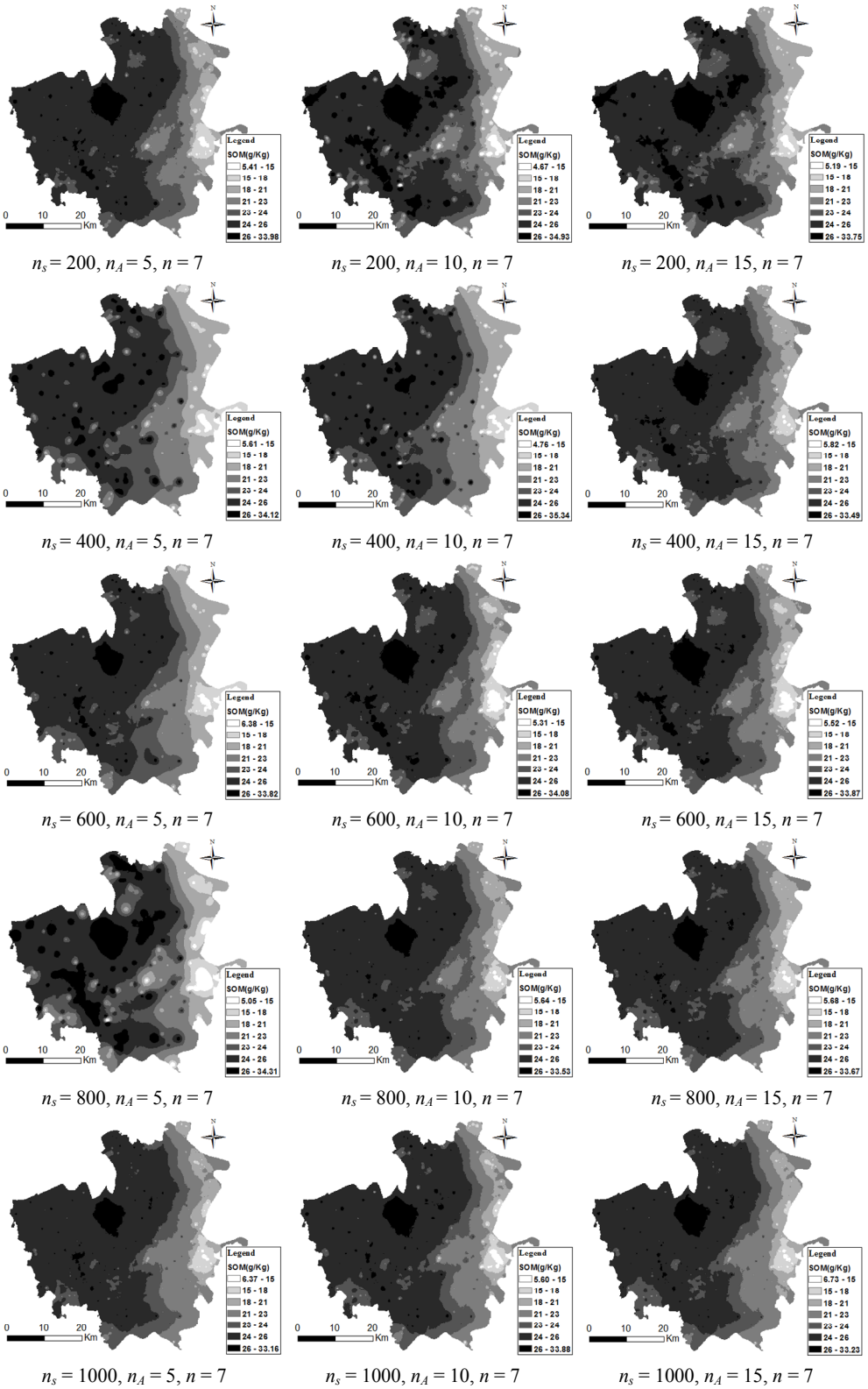


Figure 7 (Part 2).

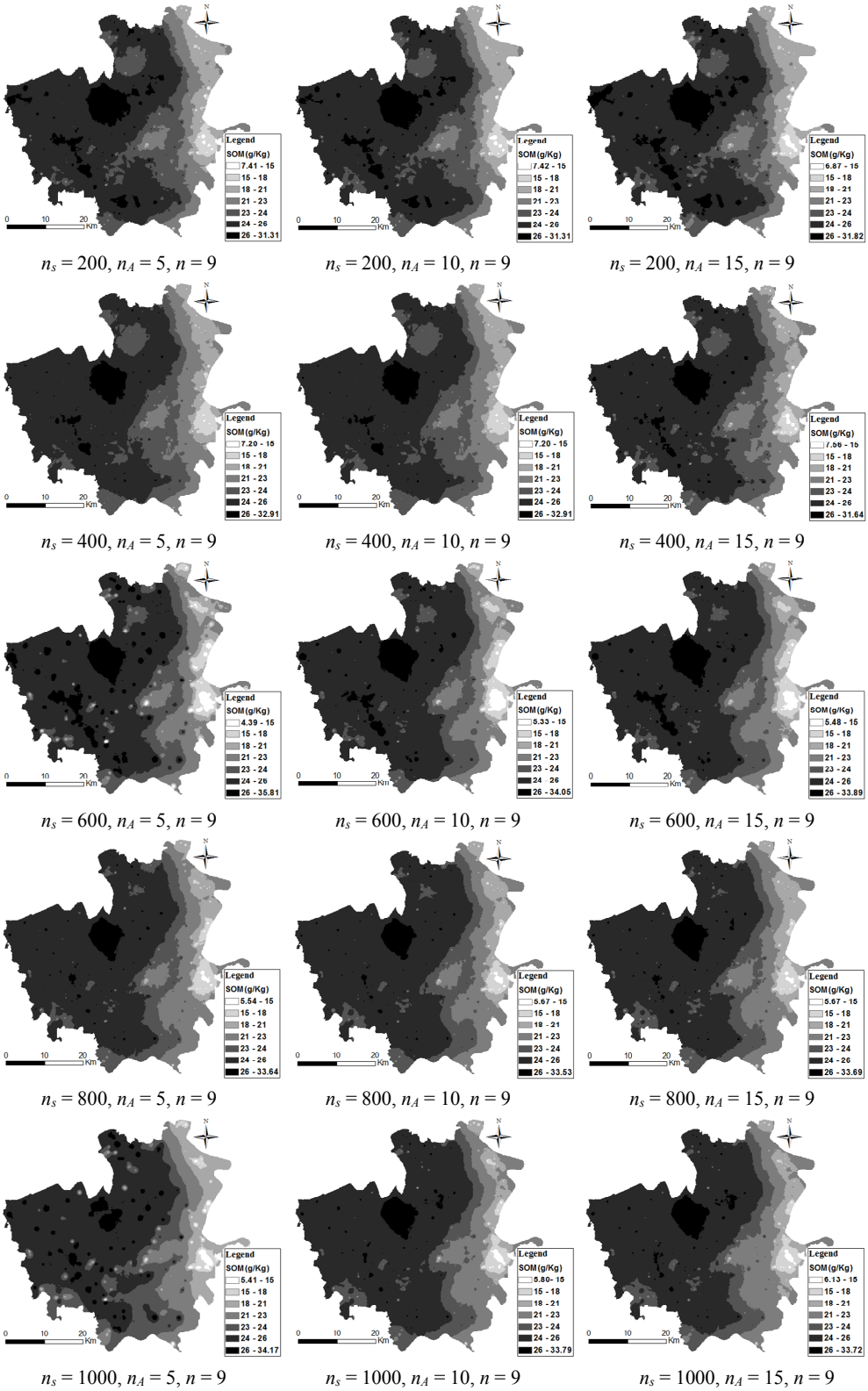


Figure 7 (Part 3).

that, after a certain number, when BME was used with more soft data points, the prediction accuracy was not significantly improved (compared to BME with fewer soft data points). In this case, we got the highest average prediction accuracy when using 600 soft data in BME. Adding more soft data did not lead to higher prediction accuracy, potentially because (a) not a sufficient number of additional soft data were implemented when predicting SOM content at these unmeasured positions, (b) the number of 600 soft data constituted the saturation level in the specific case, and (c) choosing soft data locations randomly may not result in high prediction accuracy even when their number increases. In view of these findings, future work will investigate the determination of soft data locations in a way that the selected number will increase in areas where there is a hard data deficit and decrease in areas where there is a plethora of hard data. Notice that for n_A = 5, 10, and 15, the corresponding average (RMSE values were 4.00, 3.95, and 3.98, respectively, indicating that the number of continuous auxiliary variable categories has little effect on SOM prediction accuracy. On the other hand, for n = 5, 7, and 9, the average RMSE values were 4.09, 3.91, and 3.94, respectively, indicating that the number of predicted variable categories has a significant effect on prediction, and that a smaller n-value, say 5, may result in poor prediction accuracy.

6. Conclusions

In environmental practice one aims at improving prediction by integrating different sources of relevant information. This can, however, be a difficult task, because the limited knowledge-synthesis power of mainstream data-driven techniques often prevents the rigorous assimilation of auxiliary information. Accordingly, a goal of the present work was to efficiently utilize auxiliary variables (including continuous and categorical variables) in an efficient and physically meaningful manner to improve the accuracy of spatial prediction within the framework of the BME theory.

In order to achieve this goal, we used a technique that transforms auxiliary variables into soft data in the form of probability distributions that can be subsequently used in spatial prediction. For numerical validation purposes, soil SOM content values were used as experimental data. Six terrain indices, soil type, and soil texture were considered as auxiliary variables to generate soft data. Principal component analysis and Pearson correlation analysis assessed quantitatively the relationship between SOM and continuous auxiliary variables. ANOVA and Spearman correlation analysis determined the quantitative relationship between SOM and categorical auxiliary variables. The spatial distribution of SOM content was predicted in terms of the BME and RK techniques with the same auxiliary variables, as well as by means of the mainstream OK technique. The Pearson r, ME, RMSE, and the relative improvement value (RI) of RMSE were employed to assess the prediction accuracies of the three techniques. Based on these criteria, we concluded that the BME predictions were less biased and more accurate than those of the two Kriging techniques (OK and RK). The OK predictions were more biased and less accurate than those obtained by the RK technique. Also, the results indicated that by adequately incorporating auxiliary variables BME generated more accurate predictions than both the OK and RK techniques. The spatial distribution of SOM content was mainly affected by soil type and soil texture (the RMSE of SOM content predictions was reduced when categorical variables were included). In addition, we explored the relationships between prediction accuracy and three BME modelling parameters: counts of soft data, predicted variables categories, and continuous auxiliary variable categories. Comparative analysis showed that the first two parameters had significant effect on SOM content prediction, however, the third parameter had a poor effect on prediction accuracy. In sum, the study suggested that introduction of auxiliary information improved the prediction accuracy, and that the systematic assessment of the relationship between the predicted SOM content and auxiliary variables is important in ensuring accurate SOM content mapping.

Acknowledgements. The research was supported by National Natural Science Foundation of China (Grant No. 41171174), the Fundamental Research Funds for the Central Universities (Grant No. 266-2014PY062), and China Scholarship Council. Opinions in the paper do not constitute an endorsement or approval by the funding agencies and only reflect the personal views of the authors.

Table 9. Quantitative Criteria for the Comparison of OK, RK, and BME with Different Parameters

Parameters	r	ME	RMSE	RI (%)
n_s = 200, n_A = 5, n = 5	0.764	-0.79	4.09	10.89
n_s = 200, n_A = 10, n = 5	0.806	-0.77	3.99	13.07
n_s = 200, n_A = 15, n = 5	0.701	-0.70	4.45	3.05
n_s = 400, n_A = 5, n = 5	0.74	-0.90	4.21	8.28
n_s = 400, n_A = 10, n = 5	0.783	-0.30	4.00	12.85
n_s = 400, n_A = 15, n = 5	0.801	-0.27	3.96	13.73
n_s = 600, n_A = 5, n = 5	0.754	-0.84	4.23	7.84
n_s = 600, n_A = 10, n = 5	0.771	-0.51	4.00	12.85
n_s = 600, n_A = 15, n = 5	0.789	-0.65	3.95	13.94
n_s = 800, n_A = 5, n = 5	0.771	-1.01	4.20	8.50
n_s = 800, n_A = 10, n = 5	0.795	-0.57	4.04	11.98
n_s = 800, n_A = 15, n = 5	0.799	-0.55	4.00	12.85
n_s = 1000, n_A = 5, n = 5	0.787	-0.98	4.16	9.37
n_s = 1000, n_A = 10, n = 5	0.756	-0.49	4.05	11.76
n_s = 1000, n_A = 15, n = 5	0.792	-0.60	4.06	11.54
n_s = 200, n_A = 5, n = 9	0.801	-0.74	3.94	14.16
n_s = 200, n_A = 10, n = 9	0.805	-0.76	3.90	15.03
n_s = 200, n_A = 15, n = 9	0.797	-0.74	3.89	15.25
n_s = 400, n_A = 5, n = 9	0.792	-0.61	4.00	12.85
n_s = 400, n_A = 10, n = 9	0.794	-0.62	3.94	14.16
n_s = 400, n_A = 15, n = 9	0.804	-0.65	3.99	13.07
n_s = 600, n_A = 5, n = 9	0.775	-0.68	3.87	15.69
n_s = 600, n_A = 10, n = 9	0.790	-0.70	3.87	15.69
n_s = 600, n_A = 15, n = 9	0.796	-0.68	3.85	16.12
n_s = 800, n_A = 5, n = 9	0.795	-0.50	3.91	14.81
n_s = 800, n_A = 10, n = 9	0.794	-0.59	3.95	13.94
n_s = 800, n_A = 15, n = 9	0.801	-0.59	3.95	13.94

n_s = 1000, n_A = 5, n = 9	0.758	-0.47	4.00	12.85
n_s = 1000, n_A = 10, n = 9	0.789	-0.67	4.01	12.64
n_s = 1000, n_A = 15, n = 9	0.794	-0.63	3.98	13.29
n_s = 200, n_A = 5, n = 7	0.789	-0.65	3.95	13.94
n_s = 200, n_A = 10, n = 7	0.778	-0.71	3.87	15.69
n_s = 200, n_A = 15, n = 7	0.792	-0.72	3.86	15.90
n_s = 400, n_A = 5, n = 7	0.768	-0.27	3.92	14.60
n_s = 400, n_A = 10, n = 7	0.779	-0.27	3.93	14.38
n_s = 400, n_A = 15, n = 7	0.803	-0.65	3.99	13.07
n_s = 600, n_A = 5, n = 7	0.791	-0.40	3.89	15.25
n_s = 600, n_A = 10, n = 7	0.792	-0.65	3.85	16.12
n_s = 600, n_A = 15, n = 7	0.794	-0.64	3.87	15.69
n_s = 800, n_A = 5, n = 7	0.772	-0.28	3.61	21.35
n_s = 800, n_A = 10, n = 7	0.798	-0.57	3.98	13.29
n_s = 800, n_A = 15, n = 7	0.801	-0.56	3.93	14.38
n_s = 1000, n_A = 5, n = 7	0.796	-0.53	3.99	13.07
n_s = 1000, n_A = 10, n = 7	0.790	-0.62	3.98	13.29
n_s = 1000, n_A = 15, n = 7	0.798	-0.58	4.00	12.85
Summary	*r*	ME	RMSE	RI (%)
n_s = 200 (average)	0.781	-0.73	3.99	13.07
n_s = 400 (average)	0.785	-0.50	3.99	13.07
n_s = 600 (average)	0.783	-0.64	3.93	14.38
n_s = 800 (average)	0.792	-0.58	3.95	13.94
n_s = 1000 (average)	0.784	-0.62	4.03	12.20
n_A = 5 (average)	0.775	-0.64	4.00	12.85
n_A = 10 (average)	0.788	-0.59	3.95	13.94
n_A = 15 (average)	0.79	-0.61	3.98	13.29
n = 5 (average)	0.77	-0.66	4.09	10.89
n = 7 (average)	0.79	-0.54	3.91	14.81
n = 9 (average)	0.79	-0.64	3.94	14.16
OK	0.56	-1.33	4.59	----
RK	0.64	-1.03	4.22	8.06
BME (average)	0.79	-0.61	3.98	13.30

References

Beven, K.J., and Kirkby, M.J. (1979). A physically based variable contributing area model of basin hydrology. *Hydrol. Sci. J.,* 24(1), 46-58. http://dx.doi.org/10.1080/02626667909491834

Bogaert, P. (2002). Spatial prediction of categorical variables: the BME approach. *Stochastic Environ. Res. Risk Assess.,* 16(6), 425-448. http://dx.doi.org/10.1007/s00477-002-0114-4

Bogaert, P., and D'Or D. (2002). Estimating soil properties from thematic soil maps: the Bayesian Maximum Entropy approach. *Soil Sci. Soc. Am. J.,* 66(5), 1492-1500. http://dx.doi.org/10.2136/sssaj 2002.1492

Bogaert, P. (2004). Predicting and simulating categorical random fields: the BME approach. *In: Proc. of the 1st intern. Confer. for advances in mineral resources management & environmental geotechnology (AMIREG 2004),* 119-126, Chania, Crete, Jun 7-9 20 04.

Bogaert, P., and Wibrin, M.A. (2004). Combining categorical and continuous information within the BME paradigm. *In: Proc. of the GeoEnv V-Geostatistics for Environmental Applications,* Neuchatel, Switzerland, 13-15 Oct 2004.

Cao, G., Yoo, E.Y., and Wang, S. (2014). A statistical framework of data fusion for spatial prediction of categorical variables. *Stochastic Environ. Res. Risk Assess.,* 28(7), 1785-1799. http://dx.doi.org /10.1007/s00477-013-0842-7

Christakos G. (1990). A Bayesian/maximum-entropy view to the spatial estimation problem. *Math. Geol.,* 22(7): 763-777. http://dx.doi. org/10.1007/BF00890661

Christakos, G. (1992). *Random Field Models in Earth Sciences,* Academic Press, San Diego, CA.

Christakos, G. (2000). *Modern Spatiotemporal Geostatistics,* Oxford University Press, New York.

Christakos, G., Bogaert, P., Serre, M.L. (2002). *Temporal GIS*. Springer-Verlag, New York.

Douaik, A., van Meirvenne, M., Toth, T., and Serre, M.L. (2004). Space-time mapping of soil salinity using probabilistic BME. *Stochastic Environ. Res. Risk Assess.,* 18(4), 219-227. http://dx.doi. org/10.1007/s00477-004-0177-5

Eldrandaly, K.A., Abu-Zaid, M.S. (2011). Comparison of six GIS-based spatial interpolation methods for estimating air temperature in western Saudi Arabia. *J. Environ. Inform*., 18(1): 38-45.

Gesink Law, D.C., Bernstein, K.T., Serre, M.L., Schumacher, C.M., Leone, P.A., Zenilman, J.M., Miller, W.C., and Rompalo, A.M. (2006). Modeling a syphilis outbreak through space and time using the Bayesian Maximum entropy approach. *Ann. Epidemiol.,* 16 (11), 797-804. http://dx.doi.org/10.1016/j.annepidem.2006.05.003

Hengl, T, Toomanian, N, Reuter, H, and Malakouti, M. (2007) Methods to interpolate soil categorical variables from profile observations: lessons from Iran. *Geoderma,* 140(4), 417-427. http://dx.doi. org/10.1016/j.geoderma.2007.04.022

Herbst, M., Diekkruger, B., and Vereecken, H. (2006). Geostatistical co-regionalization of soil hydraulic properties in a microscale catchment using terrain attributes. *Geoderma,* 132(1-2), 206-221. http://dx.doi.org/10.1016/j.geoderma.2005.05.008

Heywood, B.G., Brierley, A.S., and Gull, S.F. (2006). A quantified Bayesian maximum entropy estimate of Antarctic krill abundance across the Scotia Sea and in small-scale management units from the CCAMLR-2000 survey. *CCAMLR Science,* 13, 97-116.

Huang, C.Y. (1999). *Soil science,* China Agriculture Press, Beijing.

Jiang, Y.F., and Woodbury, A.D. (2006). A full-Bayesian approach to the inverse problem for steady-state groundwater flow and heat transport. *Geophys. J. Int.,* 167(3):1501-1512. http://dx.doi.org/10. 1111/j.1365-246X.2006.03145.x

Kolovos, A., Skupin, A., Christakos, G., and Jerrett, M. (2010). Multi-perspective analysis and spatiotemporal mapping of air pollution sensor data. Environmental Science and Technology. http:// dx.doi.org/10.1021/es1013328

Lamsal, S., Grunwald, S., Bruland, G.L., Bliss, C.M., and Comerford, N.B. (2006). Regional hybrid geospatial modeling of soil nitrate-nitrogen in the Santa Fe River Watershed. *Geoderma,* 135, 233-247. http://dx.doi.org/10.1016/j.geoder ma.2005.12.009

Law, D.C., Bernstein, K., Serre, M.L., Schumacher, C.M., Leone, P.A., Zenilman, J.M., et al. (2006). Modeling a Syphilis out- break through space and time using the Bayesian Maximum Entropy approach. *Ann. Epidemiol.,* 16(11), 797-804. http://dx.doi.org/10. 1016/j.annepidem.2006.05.003

Lee, S.J, Balling, R., and Gober, P. (2008). Bayesian maximum entropy mapping and the soft data problem in urban climate research. *Ann. Assoc. Am. Geogr.,* 98(2), 309-322. http://dx. doi.org/10.1080 /00045600701851184

Li, Y. (2010). Can the spatial prediction of soil organic matter contents at various sampling scales be improved by using regression kriging with auxiliary information? *Geoderma,* 159 (1-2), 63-75. http://dx.doi.org/10.1016/j.geoderma.2010.06.01 7

Liu, Y., and Hwang, Y. (2014). Improving drought predictability in Arkansas using the ensemble PDSI forecast technique. *Stochastic Environ. Res. Risk Assess.,* 29(1), 79-91. http://dx.doi.org/10.1007/ s00477-014-0930-3

Messier, K.P., Akita, Y., and Serre, M.L. (2012). Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. *Environ. Sci. Technol.,* 46(5), 2772-2780. http://dx.doi.org/10.1021/es203152a

Moore, I.D., Gessler, P.E., Nielsen, G.A., and Peterson, G.A. (1993). Soil attributes prediction using terrain analysis. *Soil Sci. Soc. Am. J.,* 57(2), 443-452. http://dx.doi.org/10.2136/sssaj1993.03615995005700020026x

National Soil Survey Office (NSS), 1995. *Chinese soil Genus Records,* vol. 1-6. China Agricultural Press, Beijing.

Odeh, I.O.A., McBratney, A.B., and Chittleborough, D.J. (1994). Spatial prediction of soil properties from landform attributes derived from a digital elevation model. *Geoderma,* 63(3-4), 197-214. http://dx.doi.org/10.1016/0016-7061(94)90063-9

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J. (1996). Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma,* 67(3-4), 215-226. http://dx.doi.org/10.1016/00167061(95)00007-B

Olea, R.A. (1999). *Geostatistics for Engineers and Earth Scie-n tists,* Kluwer Acad. Publ., Boston, MA. http://dx.doi.org/10.1007/978-1-4615-5001-3

Orton, T.G., and Lark, R.M. (2007). Accounting for the uncertainty in the local mean in spatial prediction by BME. *Stochastic Environ. Res. Risk Assess.,* 21(6), 773-784. http://dx.doi.org/10.1007/s00477-006-0089-7

Pei, T., Qin, C.Z., Zhu, A.X., Yang, L., Luo, M., Li, B.L., and Zhou, C.H. (2010). Mapping soil organic matter using the topographic wetness index: a comparative study based on different flow-direction algorithms and kriging methods. *Ecol. Indicators,* 10(3), 610-619. http://dx.doi.org/10.1016/j.ecolind. 2009.10.005

Serre, M. L., Kolovos, A., Christakos, G., and Modis, K. (2003). An application of the holistochastic human exposure methodlogy to naturally occurring Arsenic in Bangladesh drinking water. *Risk Anal.,* 23(3), 515-528. http://dx.doi.org/10.1111/1539-6924.t01-1-00332

Simbahan, G.C., Dobermann, A., Goovaerts, P., Ping, J., Haddix, M.L. (2006). Fine-resolution mapping of soil organic carbon based on multivariate secondary data. *Geoderma,* 132(3-4), 471-489. http://dx.doi.org/10.1016/j.geoderma.2005.07.001

Sumfleth, K., and Duttmann, R. (2008). Prediction of soil property distribution in paddy soil landscapes using terrain data and satellite information as indicators. *Ecol. Indicators,* 8(5), 485-501. http://dx.doi.org/10.1016/j.ecolind.2007.05.005

Wibrin, M.A., Bogaert, P., and Fasbender, D. (2006). Combining categorical and continuous spatial information within the Bayesian maximum entropy paradigm. *Stochastic Environ. Res. Risk Assess.,* 20(6), 423-433. http://dx.doi.org/10.1007/s00477-006-0035-8

Yang, J.S., Wang, Y.Q., and August, P.V. (2004). Estimation of land surface temperature using spatial interpolation and satellite-derived surface emissivity. *J. Environ. Inform*., 4(1): 37-44.

Yu, H.L., Chen, J.C., Christakos, G., and Jerrett, M. (2007a). Estimating residential level ambient PM10 and ozone exposures at multiple time-scales in the Carolinas with the BME method. *Environ. Health Perspect.,* 117(4), 537-544. http://dx.doi.org/10.1289/ehp.0800089

Yu, H.L., Kolovos, A., Christakos, G., Chen, J.C., Warmerdam S., and Dev, B. (2007b). Interactive spatiotemporal modelling of health systems: the SEKS-GUI framework. *Stochastic Environ. Res. Risk Assess,* 21(5), 555-572. http://dx.doi.org/10.1007/s00477-007-0135-0

Yu, H.L., Chiang, C.T., Lin, S.D., and Chang, T.K. (2010). Spatiotemporal analysis and mapping of oral cancer risk in Changhua County (Taiwan): an application of generalized Bayesian maximum entropy method. *Ann. Epidemiol.,* 20(2), 99-107. http:// dx.doi.org/10.1016/j.annepidem.2009.10.005

Yu, H.L., and Wang, C.H. (2013). Quantile-Based Bayesian Maximum Entropy approach for spatiotemporal modeling of ambient air quality levels. *Environ. Sci. Technol.,* 47(3): 1416-1424. http://dx.doi.org/10.1021/es302539f

Zhang, S.W., Huang, Y.F., Shen, C.Y., Ye, H.C., and Du, Y.C. (2012). Spatial prediction of soil organic matter using terrain indices and categorical variables as auxiliary information. *Geoderma,* 171-172, 35-43. http://dx.doi.org/10.1016/j.geoder ma.2011.07.012

Zhu, A.X., Qi, F., Moore, A., and Burt, J.E. (2010). Prediction of soil properties using fuzzy membership. *Geoderma,* 158(2), 199-206. http://dx.doi.org/10.1016/j.geoderma.2010.05.001