INTERPOLATING SPATIAL DISTRIBUTION USING KRIGING COMBINED WITH ENVIRONMENTAL CONDITION

Ming LUO[1], Tao PEI[2], Qun DAI[3]
[1]Ph.D. Student, Department of Geography and Resource Management, The Chinese University of Hong Kong, Sha Tin, N.T., Hong Kong, China; Tel: +852-26096233;
Email: luoming@cuhk.edu.hk

[2]Prof., State Key Laboratory of Resources and Environmental Information System,
Institute of Geographical Sciences and Natural Resources Research, CAS, 11A, Datun Road, Beijing, China;
Tel: +86-10-64888960; Email: peit@lreis.ac.cn

[3]Ph.D. Student, School of Architecture, The Chinese University of Hong Kong,
Sha Tin, N.T., Hong Kong, China; Tel: +852-26961224;
Email: daiqun@cuhk.edu.hk

KEY WORDS: Spatial interpolation, Ordinary Kriging, Geostatistics, Digital soil mapping, Soil organic matter.

ABSTRACT:

Environmental condition usually impacts the spatial distribution of geographical features, and this environmental condition information can be utilized to help improve interpolating the spatial distribution of geographical feature. This paper proposes a new spatial interpolation approach combining Kriging with environmental condition. The proposed approach estimates the values at unknown locations by a weighted summation of observations at sampled neighborhood locations. The weights of estimation consist of two parts: Kriging weights and environmental similarity. Ordinary Kriging interpolation was used to solve the Kriging weights related to spatial autocorrelation, and similarity-based fuzzy set approach was used to solve the environmental similarity related to environmental condition. Then two parts of weights were integrated by introducing an adjustment coefficient under the condition of minimizing error of sample learning. An experiment, which was performed in a study area located in Northeast China to map soil organic matter content in the topsoil, showed that, compared with several other conventional approaches (e.g., *Inverse Distance Weighting, Ordinary Kriging, Ordinary CoKriging*, etc), the proposed approach improved mapping accuracy distinctly in terms of the smallest *Root Mean Squared Error* and *Mean Absolute Error*.

1. INTRODUCTION

Spatial interpolation methods aim to estimate the variables at unobserved locations in geographic space based on the values at observed locations. Conventional spatial interpolation methods, such as *Ordinary Kriging*, *Inverse Distance Weighting* provide a useful tool for mapping spatial distribution of geographical features (Cressie 1993; Deutsch and Journel 1998; Goovaerts, 1997). However, because of the high sampling cost, conventional univariant interpolation methods do not provide the required quality of detailed mapping, and a lot of methods combining exhaustive secondary information have been proposed to undertake the task. These methods include *CoKriging*, *Kriging with External Drift*, *Regression Kriging*, etc. (Knotters et al, 1995; Goovaerts, 1997). Nevertheless, these methods are limited to combine complex, nonlinear, correlation between secondary variables and target variable (the variable needed to be interpolated), and the secondary information cannot be utilized sufficiently, even though these secondary information could become one of the most important factors that influence the spatial distribution of the target variable. As a result, nonlinear correlation with more detailed secondary information (e.g. high resolution digital elevation model, vegetation cover images), with improved accuracy, reduced sampling cost, is a key issue in spatial interpolation research. It means to develop a spatial interpolation method to predict the detailed distribution information for geographical target features by utilizing the number-limited samples and the complex.

Aimed to overcome the above shortcomings of the conventional spatial prediction methods, this paper proposes a new spatial prediction scheme by integrating both Kriging and environmental condition. The details on the proposed method will be described in the next following section. The study area and dataset will be briefly introduced in Section 3, and they are followed by the interpolation results compared with several conventional methods in the same section. The discussion concludes the paper in the last section.

2. METHODOLOGY PROPOSED

This paper proposed an optimal spatial interpolation method combining Kriging with environmental condition. Similar to *Ordinary Kriging* interpolation (Journel and Huijbregts, 1978; Goovaerts, 1997), the proposed approach estimates the values at unknown locations by a linear combination of weighted observations at sampled neighborhood locations (Formula 1). The estimating weights of proposed interpolation methods consist of two parts, one related to Kriging weights and other related to environmental condition.

$$z^*(\mathrm{u}) = \sum_{\alpha=1}^{n(\mathrm{u})} w_\alpha z(\mathrm{u}_\alpha) = \sum_{\alpha=1}^{n(\mathrm{u})} \frac{(1-\kappa)\lambda_\alpha + \kappa \cdot s_\alpha}{(1-\kappa)\cdot\sum_{\beta=1}^{n(\mathrm{u})}\lambda_\beta + \kappa\cdot\sum_{\beta=1}^{n(\mathrm{u})} s_\beta} z(\mathrm{u}_\alpha) \tag{1}$$

where $z^*(\mathrm{u})$ is the unknown value at unobserved location u; $z(\mathrm{u}_\alpha), \alpha = 1,2,\ldots,n(\mathrm{u})$ is the observations at location u_α, $n(\mathrm{u})$ is the number of observations in the neighborhood of u; w_α is the weight of observation at location u_α, it consists of two parts, Kriging weights λ_α and environmental similarity s_α, where λ_α and s_α are the Kriging weight and the environmental similarity respectively between observed location u_α and unobserved location u; κ is the adjustment coefficient, which is used to adjust the ratio of λ_α and s_α.

In Formula (1), there are three parameters, λ_α, s_α, and κ, They need to be calculated to estimate the unobserved value $z^*(\mathrm{u})$. In this paper, λ_α is estimated by solving *Ordinary Kriging* equations (Deutsch and Journel 1998): $\lambda = 1/d(\mathrm{u}_\alpha, \mathrm{u})^2$; s_α is calculated by Fuzzy Membership Function (Zhu, 2008) which can be used to characterize the degree of environmental similarity between location u_α and u. The adjustment coefficient κ is estimated under the condition given the minimum the Mean Squared Error (MSE) (see Formula (2)) when λ_α and s_α are normalized for each observed location:

$$\text{Mean Squared Error } = \frac{1}{N}\sum_{\alpha=1}^{N}\left(z_\alpha^* - z_\alpha\right)^2 \to \min \tag{2}$$

The value of the adjustment coefficient κ ranges from 0 to 1. As we can see from Formula (1), when κ equals 0, the weight of the proposed approach equals *Ordinary Kriging*. When κ equals 1, the proposed approach only considers only correlation with secondary variables. So, the values of κ have substantial meaning: the greater is the coefficient the stronger is the influence of secondary variable; the smaller is the coefficient the stronger is the influence of Kriging weights.

3. A CASE STUDY

3.1 Study area and environmental data

To test the proposed interpolation approach, an experiment has been studied in a area located in Heshan Farm, Nenjiang County, Heilongjiang Province, Northeast China (Figure 1.a), and the proposed spatial prediction approach was applied to map soil organic matter (SOM) content in the topsoil throughout the study area. The area is 64 km^2 with elevation ranging from 278m to 361m and average slope gradient almost 2°. The soils in the area are formed on deposits of silt loam loess and have a thick A-horizon with high organic matter content (Zhu et al, 2010). The average annual temperature at the site is 12.2 °C, and the average annual precipitation is between 400 and 600 mm (Pei et al, 2010).

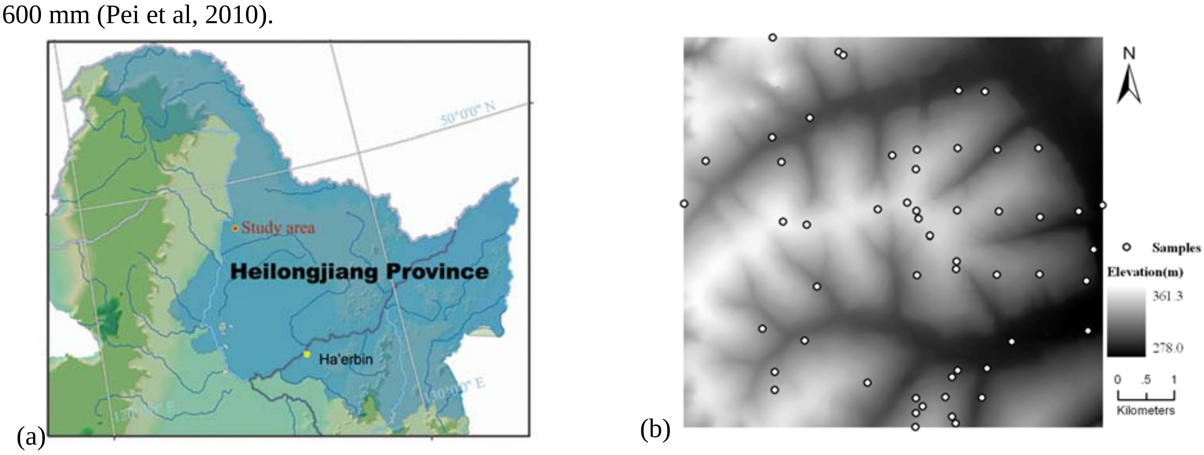

Figure 1. Study area and locations of samples.

(a) Location of the study area; (b) locations of samples and the DEM of the study area.

A 10-m resolution DEM was derived from a 1:10,000 scale topographic map (published by Chinese Bureau of Surveying and Mapping) and 54 soil samples at different locations were observed in July 2005 (Figure 1.b), 1.2m-depth profiles at each location were dug and samples were collected from the A-horizon. The samples were analyzed for Soil Organic Matter (SOM), which is the target variable to be interpolated in this experiment. Terrain variables have been used widely in soil organic matter (SOM) content mapping as they can be incorporated into geostatistical methods and used as secondary variables (Bell et al, 2000; Mueller and Pierce, 2003). Hence the following four topographic variables, elevation, slope gradient, two topographic wetness indices (TWI_{SFD}, TWI_{MFD}) based on the single-flow-direction algorithm and the multiple-flow-direction algorithm (Quinn et al, 1991; Holmgren, 1994), were used in this study to characterize the environmental similarities.

3.2 Correlation analysis

The Pearson's correlation coefficients (Rodgers and Nicewander, 1998) between the topographical variables derived from the DEM and processed SOM measurements are listed in Table 1. Correlation coefficients for TWI_{MFD} and elevation are the strongest and considerably higher than slope, whose correlation is non-significant. Correlations between the TWI_{MFD} and SOM were stronger than those between TWI_{SFD} and SOM, which is likely due to the TWI_{MFD}'s better representation of soil moisture. As a result, elevation and TWI_{MFD}, with highest correlation coefficients with SOM, were selected as the secondary environmental variables to map SOM.

Table 1. Correlation cofficients between SOM and different topographical variables.

	Slope	**Elevation**	TWI_{SFD}	**TWI_{MFD}**
SOM	0.021	**-0.591***	0.321*	**0.487***

* Significant at the 0.05 level.

3.3 Regionalization

Then interpolation results of proposed method compared with several other conventional approaches, including *Inverse Distance Weighting* (IDW), *Ordinary Kriging* (OK), and *Ordinary CoKriging* (OCK). Before mapping, the experimental semi-variograms have been calculated for solving Kriging weights. Due to the limited number of samples only the omnidirectional semi-variogram was computed. In previous SOM mapping studies, the spherical, Gaussian and exponential have been the most widely used models in fitting semi-variograms. According to the performance of experiment semi-variograms of these variables, we use the spherical model to fit the experimental semi-variogram of SOM, elevation and TWI_{MFD}. In the same way, the cross semi-variograms of SOM and elevation, SOM and TWI_{MFD} are fitted. The parameters of fitted semi-variogram models are listed in Table 2.

Table 2. Parameters of semi-variogram and cross semi-variogram models.

Semi-variogram	C_0	Range (m)	C
$\gamma_{\mathrm{SOM}}(h)$	0.03	2000	0.85
$\gamma_{\mathrm{Elev}}(h)$	-0.10	2000	1.20
$\gamma_{\mathrm{TWI_{MFD}}}(h)$	0.38	2000	0.67
$\gamma_{\mathrm{SOM\text{-}Elev}}(h)$	-0.20	2000	0.78
$\gamma_{\mathrm{SOM\text{-}TWI_{MFD}}}(h)$	-0.04	2000	0.53

3.4 Mapping results

After semi-variogram modeling, SOM maps with four mapping approaches (e.g., IDW, OK, OCK and proposed approach) were generated. The mapping results are displayed in Figure 2. All maps show the same two high value areas in the east and northwest. Apart from this, the maps differ significantly. Compared with the methods which did not use secondary variable, e.g., IDW, OK, which show smooth SOM surface, the other two methods, OCK and proposed method prediction maps are influenced significantly by the secondary variables and reveal more detail (Figure 2c–f). These can be validated through consulting the TWI and elevation distributions (Figure 1b). The impact of elevation is obvious in maps generated by OCK and the proposed approach using elevation (Figure 2c and 2e) while the impact of TWI_{MFD} is more pronounced in Figure 2d and 2f (the map of TWI_{MFD} is now showed).

3.4 Accuracy assessment

K-fold cross validation was used to assess the accuracy of interpolation, and the number of folds, K=10, more details about the K-Fold cross validation method should be referred to (Kohavi, 1995; Boyce et al, 2002; McLachlan et al, 2004). Two different accuracy indicators, *Root Mean Squared Error* (RMSE) and *Mean Absolute Error* (MSE) were taken to measure the accuracies of different methods and the results are shown in Table 3.

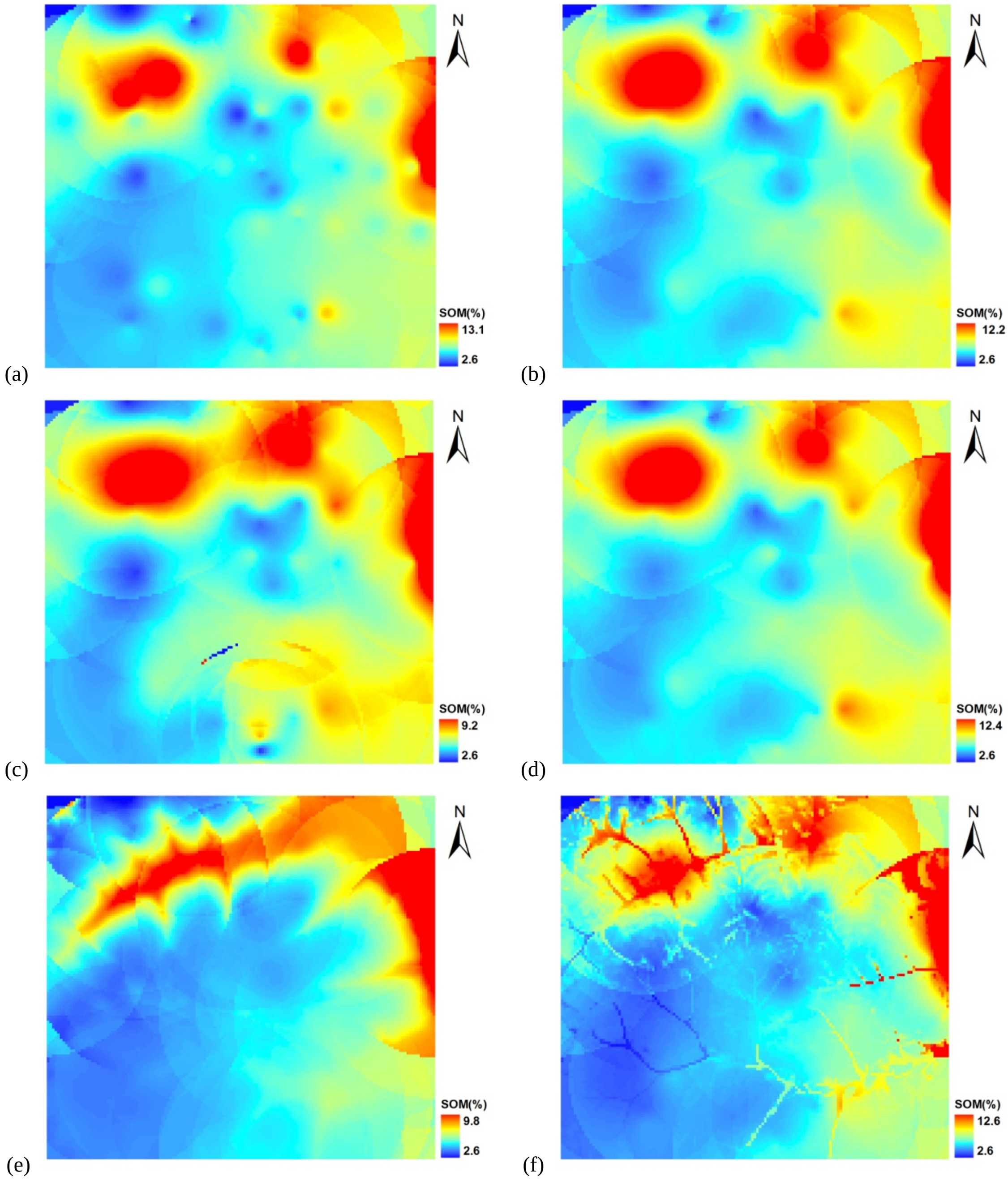

Figure 2. SOM prediction maps of different spatial prediction methods integrating elevation and TWI_{MFD} data. (a) IDW prediction; (b) OK prediction; (c) OCK prediction using elevation; (d) OCK prediction using TWI_{MFD}; (e) The proposed approach prediction using elevation; (f) The proposed approach prediction using TWI_{MFD}.

As we can see from K-fold cross validation results, it can be found that the proposed approach using elevation generated the best results with smallest RMSE and MAE. The results generated by OCK, the proposed approach (except for OCK using TWI_{MFD}) are better than IDW and OK, which utilized target variable only. This may be due to that utilizing secondary variables can help improve the interpolation accuracy. We can also find that utilizing elevation as secondary variable produced better results than using TWI_{MFD} for the same method (e.g., OCK or proposed approach) because of its higher correlation with SOM (Table 1). Compared with conventional interpolation method, both proposed methods using elevation and TWI_{MFD} performed better results since the RMSE and MAE of the proposed approach using elevation are smaller than OCK using elevation as secondary information, meanwhile, result generated from the proposed approach is better than OCK when using TWI_{MFD}. This can be explained that the proposed approach did not require a linear relationship between target variable (e.g., SOM) and environmental variables (e.g., elevation and TWI_{MFD}) while the OCK assumed that there exists a kind of linear relationship with environmental variables.

Table 3. K-fold cross validation results.

	IDW	OK	OCK (Elevation)	Proposed approach (Elevation)	OCK (TWI_{MFD})	Proposed approach (TWI_{MFD})
RMSE	1.712	1.696	1.65	1.514	1.703	1.667
MAE	1.076	1.094	1.083	0.941	1.09	1.022

In addition, the adjustment coefficients of the proposed approach using elevation and TWI_{MFD} are 0.78, 0.38 respectively. As we discussed in section 2.2, the adjustment coefficient represent the ratio between Kriging weights and environmental similarity. The values of adjustment coefficients agreed with this statement, since the higher correlation of elevation (0.591, see Table 2) generated stronger environmental similarity with a higher adjustment coefficient, 0.78, while smaller correlation of TWI_{MFD} (0.487, see Table 2) produced weaker environmental similarity and stronger Kriging weights with a smaller adjustment coefficient, 0.38. From this view, since the value of adjustment coefficient is calculated by minimizing mean squared error, the proposed approach can be considered as the automatic optimal spatial interpolation method.

4. CONCLUSION AND DISCUSSION

The proposed approach was intended to integrate the Kriging weights with nonlinear related environmental secondary variables to predict the spatial distribution of target variable. The weights of estimation included two parts, Kriging weights and the other related to environmental similarity, and the ratio of the parts is determined by minimizing summary error. The comparison results of case study predicting SOM showed that the proposed approach is not only a practicable solution for spatial interpolation, but also an automatic optimal approach, in terms of the smallest error. However, the limitations of our study are twofold: (1) the comparison between the proposed approach and conventional method was implemented in a low-relief area only; (2) multivariate environmental similarity in the proposed approach prediction needs to be studied further.

At the same time, as an exploratory research, this research has far-reaching significance. We proposed an attempt to integrate related environmental condition into conventional interpolation method, such attempt can improve interpolation accuracy, and useful detailed information (e.g., DEM, remote sensing images) can be utilized sufficiently and more detailed maps can be generated. Such attempt should be paid more attention and will have broad application prospects.

REFERENCE

[1] Bell, J.C., Grigal, D.F., and Bates, P. C., 2000. A soil-terrain model for estimation spatial patterns of soil organic carbon. In: J.P. Wilson and J.C. Gallant, Editors, Terrain Analysis: Principles and Applications, John Wiley & Sons, Inc., NY, USA, pp. 295–309.

[2] Boyce, M. S., Vernier, P. R., Nielsen, S. E., Schmiegelow, F. K. A., 2002. Evaluating resource selection functions. Ecological Modelling, 157(2-3), pp.281-300.

[3] Cressie, N., 1993. Statistics for Spatial Data. John Wiley & Sons, New York.

[4] Deutsch, C. V., Journel, A. G., 1998. GSLIB: Geostatistical software library and user's guide. Oxford University Press, New York.

[5] Goovaerts, P., 1997. Applied geostatistics for natural resources evaluation. Oxford University Press, New York, USA.

[1] Holmgren, P., 1994. Multiple flow direction algorithms for runoff modeling in grid based elevation models - an empirical-evaluation. Hydrological Processes, 8(4), pp.327-334.

[2] Journel, A.G., Huijbregts, C.J., 1978. Mining geostatistics. Academic Press, New York.

[3] Knotters, M., Brus, D. J., Voshaar, J. H. O., 1994. A comparison of Kriging, Co-Kriging and Kriging combined with regression for spatial interpolation of horizon depth with censored observations . Geoderma, 67(3-4), pp.227-246 .

[4] Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Articial Intelligence (IJCAI), Montreal.

[6] Luo, M., Pei, T, 2009. Review on soft spatial data and its spatial interpolation methods (in Chinese). Progress in Geography, 28(5), pp. 663-672.

[7] McLachlan, G. J., Do, K. A., Ambroise, C., 2004. Analyzing microarray gene expression data. Wiley-Interscience , New Jersey.

[8] Mueller, T.G. and Pierce, F.J., 2003. Soil carbon maps: enhancing spatial estimates with simple terrain attributes at multiple scales, Soil Science Society of America Journal, 67, pp. 258–267.

[9] Pei, T., Qin, C. Z., Zhu, A. X., Yang, L., Luo, M., Li, B., Zhou, C.H., 2010. Mapping soil organic matter using the topographic wetness index: a comparative study based on different flow-direction algorithms and kriging. Ecological Indicators, 10(3), pp. 610-619.
[10] Quinn, P., Beven, K., Chevalier, P., and Planchon, O., 1991. The prediction of hillslope flow paths for distributed hydrological modeling using digital terrain models. Hydrological Processes, 5, pp. 59–79.
[11] Rodgers, J. L. and Nicewander, W.A., 1998. Thirteen ways to look at the correlation coefficient, The American Statistician, 42, pp. 59–66.
[12] Zhu, A. X., Li, B. L., Pei, T., et al, 2008. Detail digital soil survey: models and methods (in Chinese). Science Press, Beijing.
[13] Zhu, A. X., Yang, L., Li, B. L., Qin, C. Z., Pei, T., Liu, B. Y., 2010. Construction of membership functions for predictive soil mapping under fuzzy logic. Geoderma, 155(3-4), pp.164-174.