Using Geographically Weighted Regression Kriging for Soil Organic Matter Maping in Red Soil Region of Southern China

Siming Chen[1,*], Ning Wang[2]

[1]Ocean College, Mingjiang University, Fuzhou, China

[2]College of Forestry, Fujian Agriculture and Forestry University, Fuzhou, China

[*]Corresponding Author

Keywords: Soil organic matter, Geographically weighted regression kriging, Auxiliary variables, Spatial prediction

Abstract: In order to accurately predict the spatial distribution of soil organic matter (SOM), a case study was conducted in the south-western Fujian Province, south-east China. A total of 24 environmental factors were extracted by ArcGIS Geostatistical analyst and remote sensing image analysis technique. Then, the stepwise regression model was used to select the best combination of environmental variables. Finally, a hybrid model of geographically weighted regression kriging (GWRK) was adapted to predict the spatial distribution of SOM, and to compare with regression Kriging (RK). The results show that:(1) B5, B6, NDVI, CI, ELE, SPI, TRI and TMEAN as the most important factors affecting the spatial variability of SOM. The determination coefficients(R2) of stepwise regression model between these variables and SOM is 0.34, and the significance probability value show that $P < 0.0001$; (2) The SOM spatial distribution patterns derive with the RK and GWRK models are quite similar, showing a spatial pattern of “high in the middle, low in the north and south “. The GWRK model is the highest in prediction accuracy, and the prediction results are more consistent with the actual situation, reflecting the detailed information about spatial distribution of SOM. This method can provide a methodological support for the study of spatial distribution of soil organic matter in the same region.

1. Introduction

Soil organic matter (SOM) is one of the most important indicators of soil quality, which directly impacts soil physical and chemical properties such as soil nutrients and soil texture. There is a growing demand for accurate spatial information of SOM in local agricultural development and environmental management. Conventional field soil sampling can provide high accurate SOM content information, but is difficult, costly, and time consuming. Hence a more effective way to acquire the dynamic changes in SOM content is needed.

As a widely applicable method, kriging interpolation use spatial correlation and variation functions to estimate soil properties at unsampled location, but do not easy explain the driving mechanism between the soil dependent variable and environmental parameters, and thus oversimplify the reality. In recent years, more studies have suggested hybrid geostatistical procedures, which combine a linear or non-linear algorithm and spatial interpolation [1]. Regression kriging (RK) are one of more widely used hybrid techniques, which simultaneously analyses the approximate ability between soil properties and correlated environmental variables, and the spatial autocorrelation of the soil properties [2]. In fact, it is difficult for RK model to capture local variation of soil property on complex environment, because it assumes that the relationship between target data and environmental variables is globally constant across space [3].

Geographically weighted regression kriging (GWRK) is the summation of Geographical weighted regression and kriging interpolation, which can represent the local variation obscured by spatial non-stationary and the spatial autocorrelation of target variable [4]. However, little attention has been paid to the effect of environmental variables in estimating SOM by GWRK. Therefore, the main objective is to investigate the relationship SOM content and environmental covariates, and

  DOI: 10.25236/ic3ec.2020.020

compare the ability of RK and GWRK models for SOM content mapping in the red soil region.

2. Materials and Method

2.1 Research Area

The study area is situated in the south-western Fujian Province, south-east China, between 25° 23′ to 25° 48' N and116°16′ to 116° 31′ E. The total area covers approximately 591.71 km^2, with an altitude elevation ranging from 164 to 812 m above sea level. Red soil in this region, which accounts for approximately 80 % of the total, developed on coarse-grained granite parent material with a high sediment concentration and deep weathering crust. This area is a traditional agricultural and protected forested zone, which comprise forest land, agricultural land, grass land, economic planting areas, water bodies and other land use area.

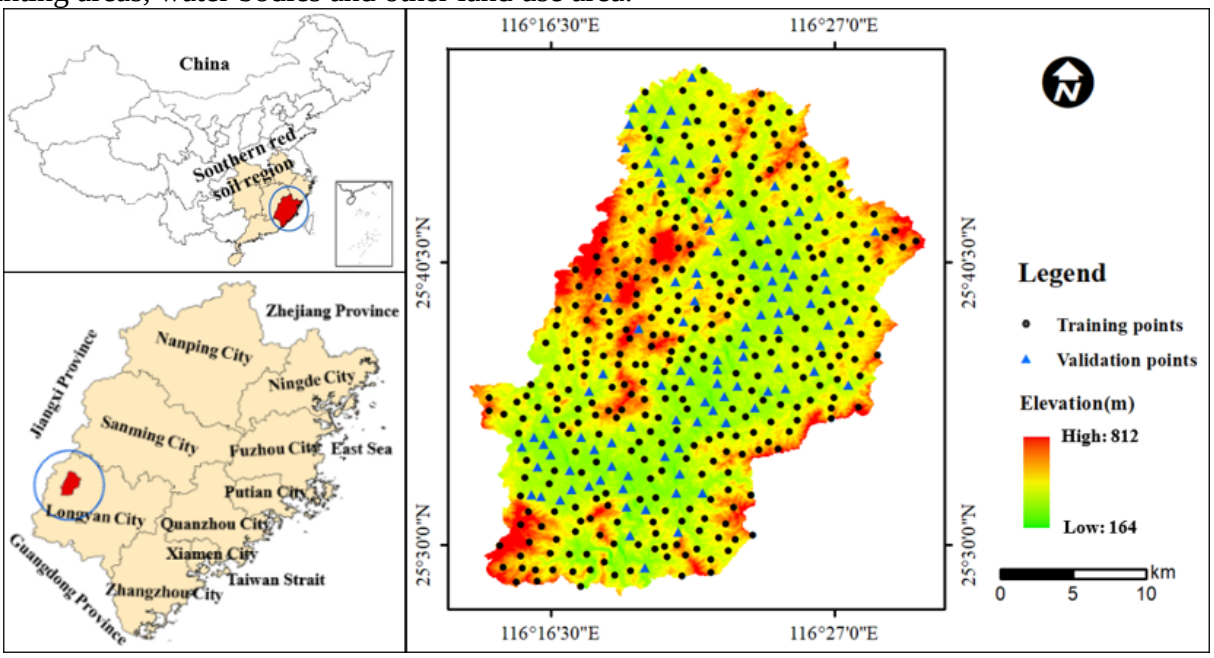

Fig.1 Position of Study Area and Distribution of the Sample Sites.

2.2 Soil Sampling and Analysis

Soil samples were taken in December 2016 after crops have been harvested, to reduce the effect of soil moisture, vegetation growth and fertilization for SOM content. A total of 390 topsoil (0-30 cm) samples (Figure 1) were collected from the field using a uniform random sampling technique, and each sampling site were selected with around 1.5 km intervals in between by taking the land type, landscape protection and topography into consideration. Global position system (GPS) was used to provide the exact longitude and latitude of each sample site. All soil samples were air-dried indoors for 5 days and then sieved through a 2-mm mesh after removing stone and plant residues. The SOM content was determined with the potassium dichromate oxidation-outer heating method.

2.3 Auxiliary Variables Data

The Landsat 8 OLI image acquired on 16 December 2016 was obtained from the United States Geological Survey because it was close to date of soil sampling and had the least cloud cover. This image was converted from DN value to the apparent reflectance (0-1) by radiometric calibration and atmospheric correction in ESRI ENVI 5.2. Then, blue band 2 (B2), green band 3 (B3), red band 4 (B4), NIR band 5 (B5,), SWIR band 6 (B6) and SWIR band 7 (B7) were selected as the soil spectral data. In addition, Normalized Difference Vegetation Index (NDVI), Ratio Vegetation Index (RVI), Difference Vegetation Index (DVI) were used as the quantitative estimation of vegetation. Clay Index (CI) and Grain Size Index (GSI) were extracted to quantify the clay content and the fine sand content of soil with the following equation.

A digital elevation model (DEM) with a spatial resolution of 30 m was obtained from Computer Network Information Centre, Chinese Academy of Sciences. From this DEM dataset, eight topographic parameters including Elevation (ELE, m), Slope (SLO, °), Plan Curvature (PLANC), Profile Curvature (PROFC), Valley Depth (VD, m), Terrain Ruggedness Index (TRI), Roughness(R), Topographic Wetness Index (TWI), Stream Power Index (SPI) were derived using ArcGIS 10.2 software.

The Climate data with a 1 km spatial resolution were obtained from the Geospatial Data Cloud of China and National Meteorological Information Centre, China. From these data, Average Monthly Precipitation (PREC, mm) and Average Monthly Temperature (TMEAN, C°) were selected as climatic variables, and resampled to 30 m using bilinear interpolation in ArcGIS 10.2 software.

2.4 Som Modelling and Prediction

2.4.1 Exploratory Data Analysis

In this study, we recorded the descriptive statistics of SOM content and used K-S test to verify normality of the data. Thereafter, the stepwise linear regression (SLR) was conducted to solve the linear relationships between the auxiliary variables. SLR is applied as an effective method to determine the contribution rate of the environmental variables, which can be used to eliminate weak significant variables step by step, and to find the optimal combination by F test and P test.

2.4.2 Regression Kriging

Regression kriging (RK) is a hybrid geostatistical method, which combines regression values with kriging values of the regression residuals [5]. In this study, the predictive trend was obtained using the multiple linear regression (MLR) method, which represent the approximate ability between the soil properties and auxiliary variables. Then, the regression residuals were interpolated by OK, which reflect the spatial autocorrelation of the soil properties. Finally, the interpolation results of RK were performed by summing the predicted trend and residuals. This process can be expressed as:

$$Z_{RK}(x_0) = \sum_{k=0}^{p} \beta_k q_k(x_0) + \sum_{i=1}^{n} \lambda_i e(x_i) \quad (1)$$

Where β_k corresponds to the estimated deterministic model coefficients, $q_k(x_0)$ represents the predictive variables at location x_0, λ_i is the kriging weight determined by the spatial autocorrelation structure of the residual, e (x_i) is the residual of the regression model at site x_i. The MLR analysis was performed using SPSS 16.0 software.

2.4.3 Geographically Weighted Regression Kriging

Geographically weighted regression kriging (GWRK) is a spatial interpolation technique, in which the global regression in RK is replaced by the local regression of Geographically weighted regression (GWR) [6]. In the process of GWRK, the regression coefficient of soil properties at x_0 for the auxiliary variables is an estimated coefficient obtained of the local weighted regression using adjacent observations, and the predicted values of the residuals were interpolated using OK. The equation for the GWRK model is:

$$Z_{GWRK}(x_0) = \beta_0(x_0) + \sum_{k=1}^{p} \beta_k(x_0) q_k(x_0) + \sum_{j=1}^{n} \lambda_j e(x_j) \quad (2)$$

Where $\beta_0(x_0)$ is the intercept at location x_0, $\beta_k(x_0)$ is the estimated local coefficient of the independent variable x_0, $q_k(x_0)$ is the explanatory variables, $e(x_j)$ is the residual of the regression model at site x_j. The GWR model was tested in GWR 4.0 software. An adaptive spatial kernel and optimal band width determined by Akaike information criterion (AICc) method.

2.4.4 Model Assessment

The correlation coefficient (R^2), root mean square error (RMSE), and the mean absolute error (MAE) were computed to determine which models had more precision in the estimation of SOM content mapping. MAE and RMSE denote the precision, stability, and performances of the models, whereas R^2 reflects relationship between observed and predicted values. The calculated equations are as follows:

3. Results

3.1 Descriptive Statistics

The SOM content for all the sample sites showed a range of 1.2 to 38.5 $g \cdot kg^{-1}$, with a mean of 19.33 $g \cdot kg^{-1}$ and standard deviation of 8.59 $g \cdot kg^{-1}$. The coefficients of skewness and kurtosis were 0.26 and -0.41 in the measured SOM respectively, indicating that the target data was normally distributed. The coefficient of variation of SOM (CV=44.48 %) exhibited a moderate variability. Among auxiliary variables data, PROFC had the largest variability (134.46 %), whereas PREC and TMEAN showed slightly.

3.2 Slr Analysis of Auxiliary Variables

Eight auxiliary variables were screen from all the original data using the SLR model, which included B5, B6, NDVI, CI, ELE, SPI, TRI and TMEAN. The determination coefficients (R^2) of the SLR between the selected variables and SOM content were 0.34, and the significance probability value showed that $p < 0.0001$, which can explain most of the information of the total variance.

3.3 Spatial Prediction of Som by Rk and Gwrk

3.3.1 Variogram of Som

To accurately predict the spatial distributions of SOM, the data set was divided into two independent groups (training data set included 273 samples and testing data set included 117 samples). The SOM content was used as the target variable, and B5, B6, NDVI, CI, ELE, SPI, TRI and TMEAN used as explanatory variables for the MLR and GWR. The MLR model for the test data showed an adjusted R-square of 0.34, whereas GWR model had larger R^2 and lower RMSE values, which were 0.47 and 8.56.

As shown in Table 1, SOM, MLR residuals and GWR residuals were determined by geo-statistical analysis using GS 1.0 software. The best-fit variogram models selected were spherical and exponential models. The nugget/sill ratio was used to represents the degree of spatial dependence and random variation for the SOM content. The nugget/sill ratio of SOM, MLR residuals and GWR residuals were 0.40, 0.33 and 0.29 respectively, indicating that the sampled spatial dependence was moderate.

Table 1 Semi-Variogram Model Parameters for Som Content and Residuals of Mr and Gwr.

Variogram	Model	Range (km)	Nugget	Still	Nugget/Still	R^2
SOM	Spherical	6.1	0.17	0.25	0.40	0.90
Residuals of MLR	exponential	5.3	48.45	97.75	0.33	0.89
Residuals of GWR	exponential	5.6	37.78	90.62	0.29	0.91

3.3.2 Spatial Distribution of Som

For the SOM mapping, the values of both ME (4.16) and RMSE (7.32) produced with the GWRK approach were lower than those generated with the RK method. More importantly, the correlation coefficient (R^2) of the GWRK model was higher, which exhibited an efficient analytic ability between the correlative factors and SOM content. As can be shown in Figure 2, the higher SOM were distributed in the middle of the study area, and lower ones in the south and north part. The varied range of SOM content is from 1.18 $g \cdot kg^{-1}$ to 38.06 $g \cdot kg^{-1}$, which were closer to the measured value and had a deeper nonlinear analytic ability for SOM. Meanwhile, transitions

among the spatial heterogeneity of local SOM had better gradients in the prediction maps produced by GWRK than by the other methods.

Table 2 Accuracy Assessment Statistics for Interpolation Methods At the Test Sites.

Interpolation Methods	ME($g \cdot kg^{-1}$)	RMSE($g \cdot kg^{-1}$)	R^2
RK	6.54	8.96	0.42
GWRK	4.16	7.32	0.58

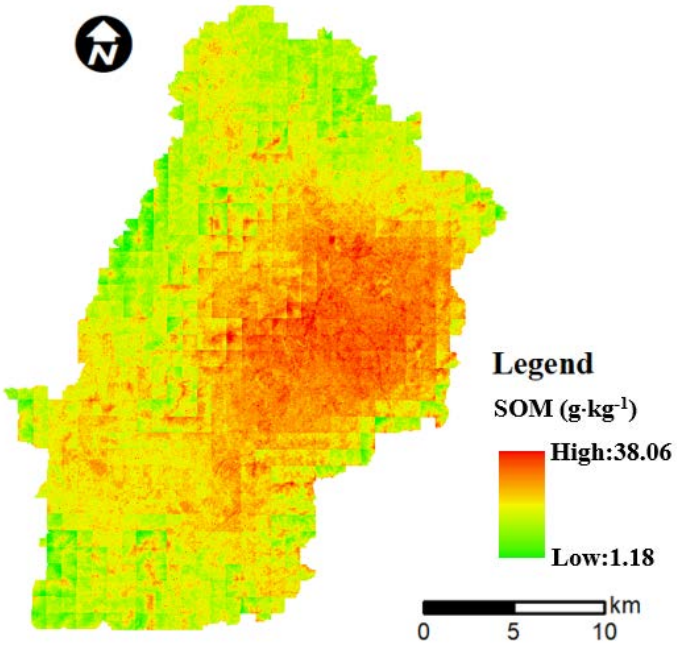

Fig.2 Predicted Som Content Maps by Gwrk.

4. Conclusion

Generally, SOM content varies significantly with climate, topography, and vegetation type, showing obvious spatial heterogeneity. In this study, a GWRK approach was used to predict the spatial distribution of SOM content, and to compare with geostatistical methods, including OK and RK. The results indicated that a driving force analysis of the auxiliary variables identified B5, B6, NDVI, CI, ELE, SPI, TRI and TMEAN as the most important factors affecting the spatial variability of SOM. The prediction accuracy of SOM content was raised by GWRK, with a lower RMSE value (7.32 $g \cdot kg^{-1}$) and higher R^2 (0.58) in comparison with the other methods. The spatial variation of SOM is mainly affected by structural factors, showing a spatial pattern of “high in the middle, low in the north and south “. This finding could prove a deeper driven significance for the nonlinear and multi-dimensional hierarchy relationship between auxiliary variables

Acknowledgment

This work is supported by projects of the Fujian Provincial Natural Science Foundation Projects.

References

[1] Mohammadi, J., Mirzaee, S., Chorbani-Dashtaki, S., et al. Spatial variability of soil organic matter using remote sensing data.Catena, vol. 145, pp. 118-127, 2016.

[2] Pham, T.G., Kappas, M., Van Huynh, C., et al. Application of Ordinary Kriging and Regression Kriging Method for Soil Properties Mapping in Hilly Region of Central Vietnam. ISPRS International Journal of Geo-information, vol.8, no.3, pp.1-17, 2019.

[3] Yang, S.H., Liu, F., Song, X.D., et al. Mapping topsoil electrical conductivity by a mixed

geographically weighted regression kriging: A case study in the Heihe River Basin, northwest China. Ecological Indicators, vol.102, pp.252-264, 2019.

[4] Pereira, O.J.R., Melfi, A. J., Montes, C.R., et al. A Downscaling of ASTER Thermal Images Based on Geographically Weighted Regression Kriging. Remote Sensing, vol.10, no.4, pp.1-20, 2018.

[5] Zhang, Y.K., Ji, W.J., Saurette, D.D., et al. Three-dimensional digital soil mapping of multiple soil properties at a field-scale using regression kriging. Geoderma, vol.366, pp.1-20, 2020.

[6] Liu, Y., Li, L.H., Chen, X., et al. Spatial distribution of snow depth based on geographically weighted regression kriging in the Bayanbulak Basin of the Tianshan Mountains, China. Journal of Mountain Science, vol.15, no. 1, pp. 33-45, 2018.