A STUDY OF SOIL ORGANIC MATTER CONTENT CHANGE IN MINSK REGION BASED ON GRADIENT BOOSTED TREE CLASSIFIERS

B. Zhao

Belarusian State University, Nezavisimosti Av., 4,
220030, Minsk, Belarus, e-mail: geozhao@outlook.com

Soil organic matter content change in the Minsk region of Belarus was inverted using Sentinel-2A multispectral remote sensing images in conjunction with information on measured soil organic matter content classes. After image preprocessing, characteristic bands were selected by correlation analysis and a multispectral model was constructed using a gradient boosting classifier for the inversion of soil organic matter content in the study area. The study can provide technical support and reference for dynamic monitoring of soil organic matter content.

Keywords: remote sensing; soil; organic matter content, gradient boosting classifier.

Introduction. Soil organic matter (SOM) is an important parameter of soil quality, which can provide various nutrients for agricultural crops and has an important influence on elemental epigenetic geochemical characteristics. As a precious soil resource, the organic matter content of black soil is an important parameter reflecting soil quality [1-2]. In recent years, with the gradual degradation of black soil, the organic matter content of soil has decreased significantly, so it is important to estimate the organic matter content of black soil to reverse the declining trend of the content, which is an important measure for the protection of black soil [3]. The traditional monitoring of soil organic matter is mainly through the collection of a large number of field soil samples and indoor chemical experiments in the monitoring area for inversion, this method is long, time-consuming and labour-intensive, and the accuracy is controlled by the density of the samples, which is difficult to meet the needs of the rapid development of modern agriculture [4]. With the increasing maturity of remote sensing technology, the determination of soil organic matter content through the spectral difference of organic matter content has become an effective means.

The remote sensing inversion of soil organic matter content mainly consists of two research directions: the processing and selection of spectral information and the construction of inversion models. Spectral processing methods such as inverse, logarithmic, and de-complex line transformation are often used, but the selected organic matter characteristic bands vary according to the image data sources. Qu Ran et al. [5] selected Landsat TM images to invert the organic matter content in Fuchuan County, Guangxi Zhuang

Autonomous Region, and found that the organic matter content of the soil had the highest correlation with the DN values of Landsat TM bands 5 and 7. Chen Debao et al.[6] used Landsat 8 remote sensing images to model the organic matter in the black soil area of Nong'an County, and showed that the short-wave infrared B6 band reflectance was the best fitted model. Chen Siming et al. [7] reconstructed the Landsat 7 soil spectra with linear spectral separation, and concluded that the reconstructed spectra could significantly enhance the correlation with soil organic matter content and improve the accuracy of soil organic matter inversion. In previous studies, linear regression and partial least squares regression (PLSR) models have been used for soil organic matter inversion. Dhawale et al. [8] used the PLSR model with the organic matter content of soil samples and the corresponding soil reflectance, and the root-mean-square error (RMSE) of the model did not exceed 2.24%. Ma Chi [9] compared the multiple regression models of different band combinations of Sentinel-2A remote sensing imagery, and the R2 was greater than 0.7. Currently, correlation analysis is mainly used for the selection of organic matter sensitive bands, and most of the inverse models are fitted linearly. In the present study, we used Sentinel-2A remote sensing images, combined with the measured soil organic matter content in Minsk Region, to study the relationship between the organic matter content of the soil surface and the remote sensing images through Sperman rank correlation analysis, and modelled by Gradient Boosting Classifier, so as to achieve the high-precision and rapid inversion of the soil organic matter on the ground.

Data acquisition and processing. The Minsk Region, located in central Belarus, spans approximately 40,800 square kilometers and includes agricultural land, forests, grasslands, and urban areas. In this paper, based on the data of soil types in Minsk region of Belarus, 500 points were randomly selected as the study area using Python language, where half of the data were used in the training set and the other half in the validation set for accuracy evaluation. The specific study area and sample training set are shown in Figure 1. Sentinel-2A imagery was selected from the study area in March 2024 during a period of bare soil and no snow, with 0 % cloud cover. The image preprocessing includes geometry correction, atmospheric correction, image inlay and image cropping. In order to improve the correlation between soil organic matter and spectral reflectance (R), the remote sensing images were processed by inverse (1/R), logarithmic (lgR), power function (Ra), first-order differentiation (FDR), second-order differentiation (SDR) and inverse logarithmic first-order differentiation (FDLR).

Algorithmic principle. Spearman's rank correlation is a non-parametric statistical method used to measure the strength and direction of a monotonic relationship between two variables. Unlike Pearson's correlation, Spearman's method does not assume linearity or normal distribution of the data, making it

suitable for non-linear relationships. The core idea is to rank the raw values of the two variables, X and Y, transforming them into ranks R_i and S_i. The Spearman correlation coefficient, denoted as ρ, is calculated as:

$$\rho = 1 - \frac{6\Sigma(R_i - S_i)^2}{n(n^2-1)} \quad (1)$$

where n is the number of data points, and $(R_i - S_i)^2$ is the squared difference in ranks for each observation.

ρ=1: Perfect positive monotonic correlation.

ρ=−1: Perfect negative monotonic correlation.

ρ=0: No monotonic correlation.

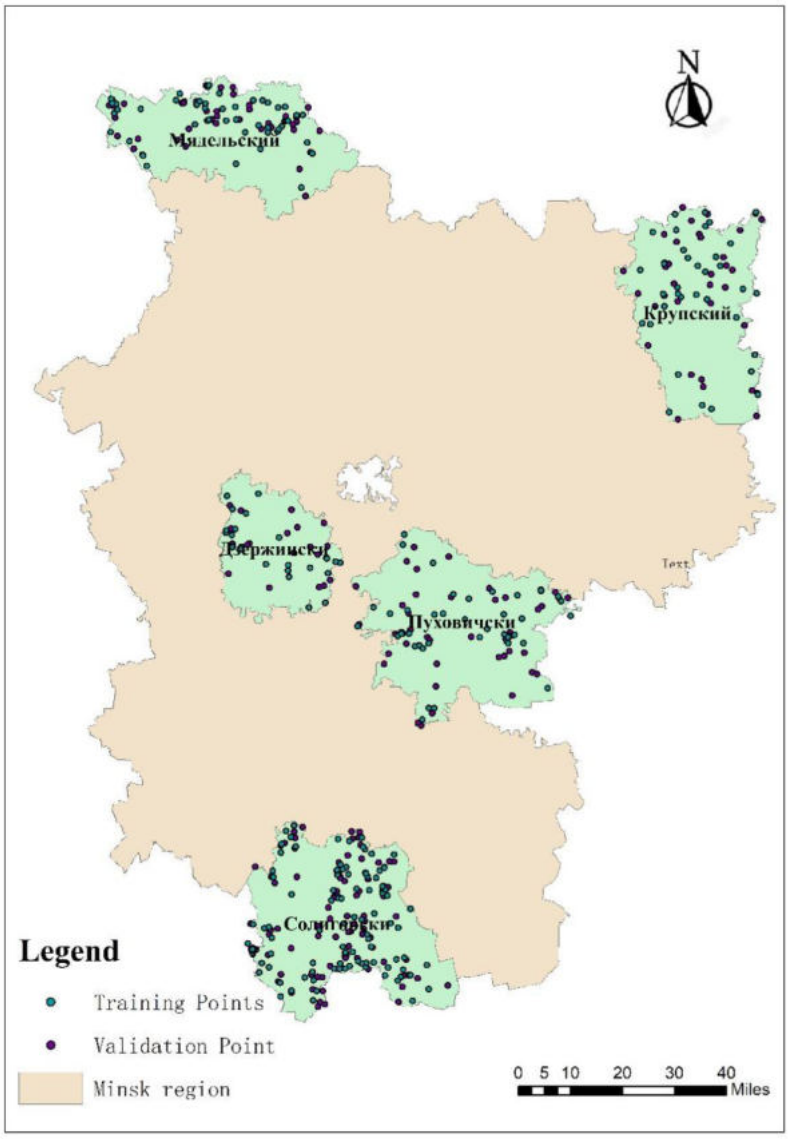

Fig. 1. Study area and sample points

Spearman's rank correlation is robust against outliers since it uses rank values instead of raw data, and it excels in identifying relationships that may not be linear but are monotonic in nature.

Gradient Boosting Classifier(GBC) is a widely used ensemble machine learning algorithm designed to improve the accuracy and robustness of predictive models. It works by combining the outputs of multiple weak learners, typically shallow decision trees, in an iterative manner to create a strong predictive model. The process begins by fitting a simple base model, such as predicting the mean value in regression tasks or the most common class in classification tasks. In subsequent iterations, each new tree is trained to minimize the errors, or residuals, made by the current model. These residuals are treated as pseudo-responses, with the new tree learning how to correct the previous predictions. The model is updated by adding the predictions of the new tree, weighted by a learning rate, to the cumulative prediction. This iterative process is guided by the gradient of a specified loss function, such as log loss for classification or mean squared error for regression. By successively reducing the loss, the GBT classifier achieves higher accuracy with each iteration. The learning rate controls the contribution of each tree, and hyperparameters such as the number of trees, maximum tree depth, and regularization terms help balance model complexity and performance. Gradient Boosting Trees are particularly effective at handling non-linear relationships and complex interactions among features, making them versatile for a wide range of applications. Proper tuning of parameters ensures the model avoids overfitting and generalizes well to unseen data. The ability to optimize arbitrary differentiable loss functions and its inherent robustness make the GBT classifier a cornerstone of modern machine learning.

Inversion of soil organic matter content. The correlation between soil organic matter content and reflectance of Sentinel-2A remote sensing images and their transforms was calculated under SPSS 27. As shown in Fig. 2, organic matter content and spectral reflectance are negatively correlated, but the correlation is not high, and the highest values of correlation coefficients of different mathematical transforms of different bands appear in the FDLR transform, which indicates that this preprocessing method can effectively improve the correlation between reflectance and soil organic matter in Sentinel-2A. The bands that passed the significance tests of the different mathematical transforms were used as the characteristic bands for the inversion, and the transforms with the highest correlation coefficients were combined as a kind of response band reference.

After selecting the feature bands and performing the relevant mathematical transformations, the appropriate band information is selected for classification by Gradient Boosting Classifier. The classifier features are trained as follows,

number Of Trees: 200, shrinkage: 0.1, sampling Rate: 1.0, maxNodes: 16. A training sample of soil organic matter in the study area is obtained after processing in Google Earth Engine using this method. With the training sample as shown in the figure, the result of hierarchical classification of soil organic matter content in the study area has an accuracy of 68 %.

		first_derivative	log	original	power	reciprocal	reciprocal_log	second_derivative
B1	Correlation coefficient	0.0183	0.217**	0.217**	0.217**	-0.217**	-0.217**	-0.220**
	Significance p	0.7760	0.0065	0.0065	0.0065	0.0064	0.0064	0.0054
	N	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000
B2	Correlation coefficient	-0.0803	0.231**	0.231**	0.231**	-0.231**	-0.231**	-0.246**
	Significance p	0.2116	0.0028	0.0028	0.0028	0.0027	0.0028	0.0011
	N	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000
B3	Correlation coefficient	-0.130*	0.241**	0.241**	0.241**	-0.240**	-0.241**	-0.258**
	Significance p	0.0431	0.0015	0.0015	0.0015	0.0015	0.0015	0.0044
	N	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000
B4	Correlation coefficient	-0.0639	0.292**	0.292**	0.292**	-0.292**	-0.292**	-0.296**
	Significance p	0.3208	0.0031	0.0031	0.0031	0.0031	0.0031	0.0021
	N	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000
B5	Correlation coefficient	-0.0865	0.270**	0.270**	0.270**	-0.271**	-0.270**	-0.307**
	Significance p	0.1779	0.0018	0.0018	0.0018	0.0018	0.0018	0.0000
	N	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000
B6	Correlation coefficient	-0.0231	0.174**	0.174**	0.174**	-0.174**	-0.174**	-0.207**
	Significance p	0.7193	0.0651	0.0651	0.0651	0.0649	0.0651	0.0116
	N	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000
B7	Correlation coefficient	-0.0773	0.161*	0.161*	0.161*	-0.161*	-0.161*	-0.203**
	Significance p	0.2292	0.0118	0.0118	0.0118	0.0120	0.0120	0.0141
	N	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000
B8	Correlation coefficient	-0.0172	0.199**	0.199**	0.199**	-0.199**	-0.199**	-0.195**
	Significance p	0.7889	0.0178	0.0178	0.0178	0.0183	0.0183	0.0217
	N	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000
B8A	Correlation coefficient	-0.0493	0.189**	0.189**	0.189**	-0.190**	-0.189**	-0.231**
	Significance p	0.4437	0.0298	0.0298	0.0299	0.0290	0.0298	0.0028
	N	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000
B9	Correlation coefficient	-0.0731	0.242**	0.242**	0.242**	-0.242**	-0.242**	-0.236**
	Significance p	0.2550	0.0013	0.0013	0.0013	0.0013	0.0013	0.0020
	N	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000
B10	Correlation coefficient	0.0467	0.154*	0.154*	0.154*	-0.157*	-0.154*	-0.147*
	Significance p	0.4678	0.0160	0.0160	0.0160	0.0140	0.0160	0.0220
	N	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000
B11	Correlation coefficient	-0.0541	0.322**	0.322**	0.321**	-0.322**	-0.322**	-0.346**
	Significance p	0.4017	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	N	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000
B12	Correlation coefficient	-0.0937	0.304**	0.304**	0.304**	-0.304**	-0.304**	-0.333**
	Significance p	0.1445	0.0010	0.0010	0.0010	0.0010	0.0010	0.0000
	N	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000	244.0000
** Significant correlation at the 0.01 level (two-tailed).								
* Significant correlation at the 0.05 level (two-tailed).								

Fig. 2. Sentinel 2A bands and their mathematical variant forms with soil organic matter Spearman rank correlation analysis

Conclusion. The reflectance of Sentinel-2A multispectral remote sensing image was transformed by 1/R, lgR, Ra, FDR, SDR and FDLR, and the inversion of soil organic matter was realised by combining different models, which achieved good results. The following conclusions are drawn:

1. The FDLR transformation model of reflectance fits the best when modelling by correlation analysis, and the combination of bands can effectively

improve the modelling accuracy of soil organic matter content in inversion. 2) The multispectral remote sensing image spectra of soil organic matter can be used for the inversion of soil organic matter content in different models.

2. The spectral resolution of multispectral remote sensing images is low, so the linear fitting model cannot accurately estimate the soil organic matter content, and a nonlinear model is needed to effectively fit the spectral information to the organic matter content.

3. Under the Gradient Boosting Classifier, the classification accuracy of soil organic matter content in the study area is 68%, but it shows the potential of this method for classification and extraction, and provides a new direction for the next research on soil organic matter.

References

1. *Kong M., Yang S. P.* Preliminary research into the disturbed principle of organic material to character of supergene-geochemistry in forest marsh landscape area // Geophysical and Geochemical Exploration. 2008. Vol. 32, No. 1. P. 31–32, 74.

2. *Rasmussen C., Heckman K., Wieder W. R.* Beyond clay: Towards an improved set of variables for predicting soil organic matter content // Biogeochemistry. 2018. Vol. 137, No. 5. P. 297–306.

3. *Dai H. M., Liu K., Song Y. H.* Black soil degradation and intensity in northeast China: Geochemical indication // Geology and Resources. 2020. Vol. 29, No. 6. P. 510–517.

4. *Liu H. J., Zhang M. W., Yang H. X.* Inversion of cultivated soil organic matter content combining multi-spectral remote sensing and random forest algorithm // Transactions of the Chinese Society of Agricultural Engineering. 2020. Vol. 36, No. 10. P. 134–140.

5. *Qu R., Zhang Y. Q., Nie Y. H.* Inversion of surface soil organic matter content in Fuchuan county based on multi-spectral remote sensing image // Environment and Sustainable Development. 2019. Vol. 44, No. 1. P. 154–157.

6. *Chen D. B., Chen G. F.* Inversion of soil organic matter content in black soil region based on Landsat 8 remote sensing image // Journal of Chinese Agricultural Mechanization. 2020. Vol. 41, No. 6. P. 194–198.

7. *Chen S. M., Zou S. Q., Mao Y. L.* Inversion of soil organic matter content in wetland using multispectral data based on soil spectral reconstruction // Spectroscopy and Spectral Analysis. 2018. Vol. 38, No. 3. P. 912–917.

8. *Dhawale N. M., Adamchuk V. I., Prasher S. O.* Proximal soil sensing of soil texture and organic matter with a prototype portable mid-infrared spectrometer // European Journal of Soil Science. 2015. Vol. 66, No. 4. P. 661–669.

9. *Ma C.* Inversion of soil organic matter content based on Sentinel-2A remote sensing image // Northern Horticulture. 2020. No. 2. P. 94–100.