Prediction of Soil Carbon and Nitrogen Content Using Hyperspectral Image with A New Feature Selection Algorithm

Xueying Li[1, 2]*, Zongmin Li*[3]*, Pingping Fan*[2]*, Huimin Qiu*[2]*, Guangli Hou*[2]

[1]School of Geosciences, China University of Petroleum (East China), Qingdao, 266580, China
[2]Institute of oceanographic Instrumentation, Qilu University of Technology (Shandong Academy of Sciences), Qingdao, 266061, China
[3]College of computer science and technology, China University of Petroleum (East China), Qingdao, 266590, China

ABSTRACT

The real-time and rapid detection of soil carbon and nitrogen content is of great significance to promote ecosystem carbon balance, ensure the healthy growth of crops, sustainable land use, prediction and early warning of soil pollution. Hyperspectral image is a promising alternative to predict soil carbon and nitrogen elements. Considering some noise and interference information in hyperspectral image, a new feature selection algorithm, extend successive projections algorithm (ESPA), was proposed. Compared with the prediction using full spectrum, successive projection algorithm (SPA), uninformative variable elimination (UVE), genetic algorithm (GA) and manual selection of feature spectra, the prediction was improved by using ESPA selecting the spectral region. The result indicates that the selected spectral region using ESPA is effective in prediction of soil total carbon (TC) and total nitrogen (TN) content, providing an alternative to predict carbon and nitrogen content in soil using hyperspectral image.

Index Terms—Hyperspectral image, soil, carbon, nitrogen, feature selection

1. INTRODUCTION

Carbon and nitrogen are the main nutrient elements in soil nutrient [1]. Appropriate content of carbon and nitrogen has a great impact on promoting plant growth. Soil nitrogen plays an important role in promoting the growth of crop roots, stems and leaves. Insufficient nitrogen content in the soil will affect the growth of crops. And excessive fertilization will lead to excess nutrients, which will cause environmental pollution and waste of resources. When the soil carbon content is low, crops will not be able to absorb organic carbon from the soil, resulting in carbon deficiency in roots. Carbon deficiency will lead to plant root weakness, premature senescence of crops, yellowing of leaves and decline of disease resistance. Meanwhile, as an important component of terrestrial ecosystem, soil is the main exchange and storage of carbon [2]. Compared with other types of soil, the carbon pool of farmland soil is more active, and the change of carbon content is more obvious. Through the monitoring of soil carbon and nitrogen content, the real-time and rapid detection of soil carbon and nitrogen content is of great significance to promote ecosystem carbon balance, ensure the healthy growth of crops, sustainable land use, prediction and early warning of soil pollution.

Hyperspectral image can be regarded as a cube containing two-dimensional spatial information and the third dimensional spectral information, which is rich in spectral information and spatial information [3]. The spectral information contained in hyperspectral image usually consists of visible and near-infrared spectra. Visible and near infrared spectroscopy technology is a fast, non-destructive measurement method. It has been widely used in medicine, agriculture, oil and other field [4, 5]. Hyperspectral camera has the advantages of visible and near-infrared spectrum, and it is more suitable for in-situ and field data acquisition, and the collected spectral information and image information are more comprehensive, which is more conducive to subsequent analysis.

The traditional chemical method for determination of soil carbon and nitrogen has the problems of high cost, long time and complex operation. And the measurement methods of different nutrient elements are different, so each nutrient content needs to be determined separately. Based on the spectral detection of soil carbon and nitrogen content, some scholars have studied the relevant basic research [6, 7]. The content of soil organic carbon and nitrogen determined by Hyperspectral technique reached certain progress [8-11]. In [12-13] compared with the AgriSpec portable spectrometers and other spectrometers, the better prediction results determined by Hyperspectral technique were obtained. Because the soil composition is very complex, the obtained spectral information also contains a lot of noise and interference information. Therefore, it is particularly important to extract the characteristic wavelength, eliminate redundant information and improve the accuracy of the model. In [14], two feature selection algorithms, which were continuous wavelet transform (CWT) and competitive adaptive reweighted sampling (CARS), was applied to optimize the performance of the prediction models. In [3], the successive projections algorithm (SPA) method was utilized to select effective wavelengths from the full spectrum, the better correct classification rate was achieved.

In this paper, considering some noise and interference information in hyperspectral image, a new feature selection algorithm, extend successive projections algorithm (ESPA), is proposed. The partial least squares (PLS) method was used to establish the quantitative analysis model to realize the rapid detection of soil total carbon (TC) and total nitrogen (TN).

2. STUDY AREA AND DATA

2.1 Study area and experimental materials

The study area was in the Fushan Mountain foothills and Licun River, Qingdao City, Shandong Province, China. The soil of Fushan Mountain foothills was sandy loam, and the soil of Licun River belonged to silt loam. 60 Mountain foothills and 60 River topsoil (0 – 20 cm) samples were collected. 5 soil samples were removed because of man-made cause anomalies.

2.2 Chemical analysis

The soil samples were dried to constant weight at 50 ℃ and passed through 0.45mm nylon sieve. 5-10g soil was taken out from soil samples, TC and TN content for soil samples were determined by elemental analyzer. Statistics of TC and TN contents were given in Table 1.

Table 1. Statistics of TC and TN contents for soil samples

	Min	Max	Mean	SD
TC(mg/kg)	1.904	13.400	7.041	4.019
TN(mg/kg)	0.209	1.736	0.921	0.533

2.3 spectral measurement

The portable hyperspectral camera GaiaField-V10 was used to obtain hyperspectral images of soil samples. The spectral range was 400-1100 nm. The sampling interval was 3.2 nm. Soil samples were gently flattened in a rectangular box. The hyperspectral camera was placed on a tripod and soil samples were taken vertically, as Figure 1 shown. Three soil samples were simultaneously collected from each hyperspectral image.

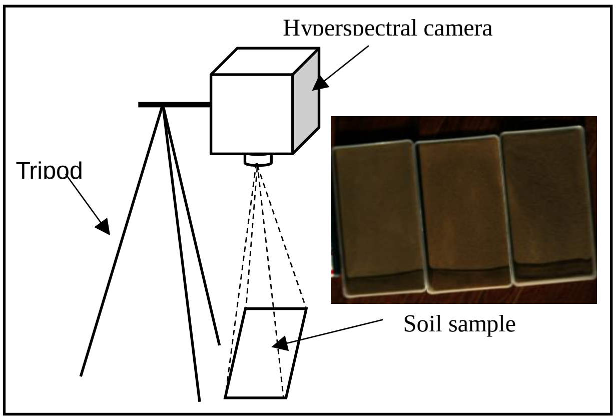

Figure 1. The schematic diagram of soil hyperspectral data measurements.

The region of interest (ROI) of the hyperspectral image was extracted by a rectangular figure of 100*100 pixels size. The average spectral value of each point in the ROI region was obtained, as Figure 2 shown. They was used to build soil TC and TN contents models.

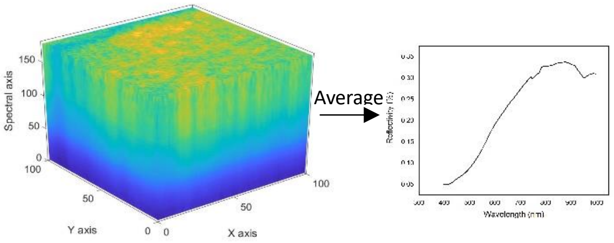

Figure 2. The acquisition process of average spectrum.

3. METHOD

3.1 Spectral Band Selection

The spectral information obtained from hyperspectral image has the characteristics of multiple bands and multiple noises. Extracting spectral characteristic wavelength can eliminate redundant information to a certain extent and improve the accuracy of the model. Common feature variable extraction methods include successive projection algorithm (SPA) [15], uninformative variable elimination (UVE) [16] and genetic algorithm (GA) [17]. These algorithms all select the spectral band data that can be used for modeling from a large number of original spectral data.

Based on SPA, we propose a new feature selection method, extend SPA (ESPA). SPA fully finds the variable group with the minimum redundant information from the spectral information, so that the collinearity between variables can be minimized. The redundant information in the original spectral matrix is eliminated by extracting several characteristic wavelengths of the whole band. ESPA is an extension of the spectral wavelength selected by spa. If n characteristic wavelength points are selected by SPA, m wavelength points around n wavelength points and n characteristic wavelength points are selected as the final characteristic wavelength. There is a certain correlation between adjacent spectra, and the selected m spectral band can contain more useful information. The selection of m is also very important for the prediction accuracy. If m value is too large, redundant information will be introduced, and if m value is too small, it can not fully contain useful information.

3.2 Model calibration and validation

Partial least squares regression (PLSR) is a method to realize the prediction of unknown samples, based on the known spectral data and the chemical value data [18].

To evaluate the prediction, four parameters including the determination coefficient of calibration and test sets (R_c^2, R_c^2), root mean square error of prediction (RMSEP), and ratio of prediction to deviation (RPD) were adopted [19]. When RPD < 1.5, the prediction of the model is poor, it can be used for qualitative prediction and quantitative prediction; when 1.5 < RPD < 2.0, it is a good model, which can make rough quantitative prediction; when RPD > 2.0, it is a very good model, which can make accurate quantitative prediction.

4. RESULTS AND DISCUSSION

4.1 Prediction of TC and TN content in soil

The 115 soil samples were divided into calibration and test sets according to sampling sequence. Every other two samples were grouped as test samples, total of 38 soil samples. Others were grouped as calibration samples, total of 77 soil samples.

PLSR was run to predict TC and TN content using the full spectrum (none), three feature variable extraction methods (SPA, UVE and GA), and manual selection of feature spectra. Manual selection was the best spectral band by testing a large number of band prediction results. Manual selection band of TC and TN were 700-820nm and 420-630nm. The prediction accuracies and BN (the number of spectral bands) of TC and TN content in soil were given in Table 2 and 3.

Table 2. Prediction accuracies and BN of TC content using full spectrum, three feature variable extraction methods and manual selection.

Method	BN	R_c^2	R_t^2	RMSEP	RPD
None	176	0.923	0.862	1.461	2.704
SPA	17	0.932	0.874	1.441	2.741
UVE	70	0.880	0.819	1.696	2.328
GA	44	0.919	0.794	1.927	2.050
Manual	35	0.912	0.859	1.472	2.683

Table 3. Prediction accuracies and BN of TN content using full spectrum, three feature variable extraction methods and manual selection.

Method	BN	R_c^2	R_c^2	RMSEP	RPD
None	176	0.930	0.868	0.192	2.757
SPA	30	0.931	0.871	0.189	2.793
UVE	77	0.917	0.852	0.202	2.609
GA	37	0.929	0.842	0.270	1.956
Manual	64	0.919	0.861	0.195	2.715

The method with the highest prediction accuracy was SPA, the R_c^2, R_t^2, RMSEP and RPD values of TC content were 0.923, 0.862, 1.461 and 2.704. And next were using full spectrum, manual selection, UVE and GA. The prediction accuracy TN also showed a consistent results. Among the five methods, the minimum number of spectral bands in TC and TN content was selected using SPA, which were 17 and 30. Therefore, by using feature selection method SPA could improve the prediction accuracy with less spectral bands, which could make accurate quantitative prediction of soil TC and TN content.

4.2 Prediction of TC and TN content with ESPA

To further explore the effectiveness of predicting TC and TN content in soil using SPA. ESPA was proposed as a new feature selection method, because of the correlation between adjacent spectra. Based on the characteristic wavelength points of TC and TN content selected by SPA, around characteristic wavelength points m were 1, 2, 3, 4, 5, 6 and 7 respectively. The prediction accuracies and BN of TC and TN content in soil using ESPA were given in Table 4 and 5.

Table 4. Prediction accuracies and BN of TC content using ESPA.

Method	BN	R_c^2	R_t^2	RMSEP	RPD
SPA±1	42	0.929	0.871	1.408	2.805
SPA±2	53	0.924	0.871	1.407	2.807
SPA±3	73	0.925	0.873	1.403	2.815
SPA±4	83	0.916	0.834	1.588	2.487
SPA±5	92	0.917	0.838	1.570	2.516
SPA±6	98	0.918	0.842	1.554	2.542
SPA±7	104	0.923	0.871	1.414	2.794

Table 5. Prediction accuracies and BN of TN content using ESPA.

Method	BN	R_c^2	R_t^2	RMSEP	RPD
SPA±1	75	0.932	0.878	0.184	2.875
SPA±2	107	0.932	0.881	0.182	2.910
SPA±3	128	0.932	0.885	0.179	2.949
SPA±4	145	0.932	0.887	0.178	2.966
SPA±5	158	0.932	0.888	0.177	2.984
SPA±6	165	0.932	0.887	0.178	2.976
SPA±7	171	0.930	0.869	0.191	2.767

In the prediction accuracies of TC content, the RPD values increased first then decreased, and then increased as m increased. When m were 1, 2 and 3, the RPD values were on the rise, and higher than 2.8. When m was 3, the RPD value had the max value. When m were 4, 5, 6 and 7, the RPD values were also on the rise and were between 2.4 and 2.7. The R_c^2 and R_t^2 also showed a consistent trend. When m was 3, the R_c^2 and R_t^2 had the maximum, the RMSEP had the minimum. The prediction accuracies using ESPA were better than that using SPA, when m were 1, 2 3 and 7. ESPA could effectively improve the prediction accuracies of TC content. The optimal prediction (m was 3) of TC content using ESPA was illustrated with scatter plot in Figure 3.

In the prediction accuracies of TN content, the RPD values increased first then decreased as m increased. When m was 5, the RPD value, R_c^2 and R_t^2 had the maximum, the RMSEP had the minimum. The number of spectral bands was 158. Due to the more wavelength points of TN content using SPA, there were more spectral bands in the TN content compared with the TC content. ESPA could also effectively improve the prediction accuracies of TN content. The optimal prediction (m was 5) of TN content using ESPA was illustrated with scatter plot in Figure 4.

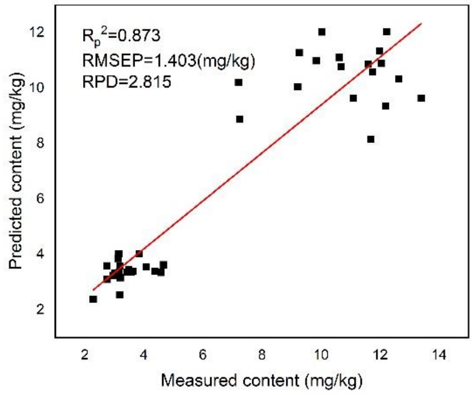

Figure 3. Scatter plot of the measured against predicted TC content using ESPA.

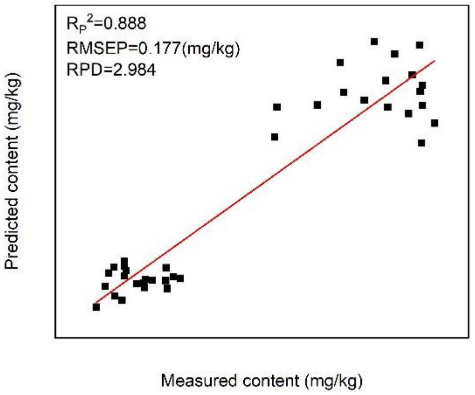

Figure 4. Scatter plot of the measured against predicted TN content using ESPA.

5. CONCLUSION

This article adopts hyperspectral image to provide an effective prediction of TC and TN content in soil using a new feature selection algorithm, extend successive projections algorithm (ESPA). This method considers some noise and interference information, and extracts spectral feature of TC and TN content in soil. Compared with the prediction using full spectrum, SPA, UVE, GA and manual selection of feature spectra, the prediction was improved by using ESPA selecting the spectral region. The R_c^2, R_t^2, RMSEP and RPD values of TC content were 0.925, 0.873, 1.403 and 2.815, and these of TN content were 0.932, 0.888, 0.177 and 2.984. . The number of spectral bands of TC and TN content were 73 and 158. The result indicates that the selected spectral region using ESPA is effective in prediction of soil TC and TN content, providing an alternative to predict carbon and nitrogen content in soil using hyperspectral image.

6. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (31700447), and the Natural Science Foundation of Shandong Province, China (ZR2018LD007, ZR2017BB037).

7 REFERENCES

[1] S. D. Veresoglou, B. Chen, and M. C. Rillig, "Arbuscular mycorrhiza and soil nitrogen cycling," *Soil Biology & Biochemistry*, vol. 46, pp. 53-62, 2012.

[2] M. Roberto, S. Marinari, P. Brunetti, E. Radicetti, and E. Campiglia, "Organic mulching, irrigation and fertilization affect soil CO2 emission and C storage in tomato crop in the Mediterranean environment," *Soil & Tillage Research* , vol. 152, pp. 39-51, 2012.

[3] S. Jia, H. Li, Y. Wang, R. Tong, and Q. Li, "Hyperspectral Imaging Analysis for the Classification of Soil Types and the Determination of Soil Total Nitrogen," *Sensors*, vol. 17, pp. 2252, 2017.

[4] W. Ng, B. P. Malone, and B. Minasny, "Rapid assessment of petroleum-contaminated soils with infrared spectroscopy," *Geoderma*, vol. 289, pp. 150-160, 2017.

[5] S. Jia, H. Li, Y. Wang, R. Tong, and Q. Li, "Recursive variable selection to update near-infrared spectroscopy model for the determination of soil nitrogen and organic carbon," *Geoderma*, vol. 268, pp. 92-99, 2016.

[6] J. Wetterlind, B. Stenberg, and M. Söderström, "Increased sample point density in farm soil mapping by local calibration of visible and near infrared prediction models," *Geoderma*, vol. 156, pp. 152-160, 2010.

[7] R. A. V. Rossel and R. Webster, "Predicting soil properties from the Australian soil visible-near infrared spectroscopic database," *European Journal of Soil Science*, vol. 63, pp. 848-860, 2012.

[8] L. Guo, T. Shi, M. Linderman, Y. Chen, and P. Fu, "Exploring the influence of spatial resolution on the digital mapping of soil organic carbon by airborne hyperspectral vnir imaging," *Remote Sensing*, , vol. *11*, pp. 1032, 2019.

[9] H. Yu, B. Kong, G. Wang, R. Du, and G. Qie, "Prediction of soil properties using a hyperspectral remote sensing method," *Archives of Agronomy & Soil Science*, vol. 64, pp. 546-559, 2018.

[10] I. Tahmasbian, Z. Xu, S. Boyd, J. Zhou, R. Esmaeilani, R. Che, and S. H. Bai, "Laboratory-based hyperspectral image analysis for predicting soil carbon, nitrogen and their isotopic compositions," *Geoderma*, vol. 330, pp. 254-263, 2018.

[11] L. Lin, Z. Gao, and X. Liu, "Estimation of soil total nitrogen using the synthetic color learning machine (SCLM) method and hyperspectral data," *Geoderma*, vol. 380, pp. 114664, 2020.

[12] G. Cécile, R. A. V. Rossel, and A. B. Mcbratney, "Soil organic carbon prediction by hyperspectral remote sensing and field vis-nir spectroscopy: an australian case study," *Geoderma*, vol. *146*, pp. 403-411, 2008.

[13] J. Peon, S. Fernandez, C. Recondo, and J. F. Calleja, "Evaluation of the spectral characteristics of five hyperspectral and multispectral sensors for soil organic carbon estimation in burned areas," *International Journal of Wildland Fire*, vol. *26*, pp. 230-239, 2017.

[14] Y. S. Hong, S. C. Chen, Y. Y. Chen, M. Linderman, A.M. Mouazen, Y. L. Liu, L. Guo, L. Yu, Y. F. Liu, H. Cheng, and Y. Liu, "Comparing laboratory and airborne hyperspectral data for the estimation and mapping of topsoilorganic carbon: feature selection coupled with random forest," *Soil and Tillage Research*, vol. 199, pp. 104589, 2020.

[15] M. Insausti, A. A. Gomes, F. V. Cruz, M. F. Pistonesi, M. C. Araujo, R. K. Galvao, C. F. Pereira, and B. S. Band, "Screening analysis of biodiesel feedstock using UV-vis, NIR and synchronous fluorescence spectrometries and the successive projections algorithm," *Talanta*, , vol. 97, pp. 579–583, 2012.

[16] Z. Li, J. Wang, Y. Xiong, Z. Li, and S. Feng, "The determination of the fatty acid content of sea buckthorn seed oil using near infrared spectroscopy and variable selection methods for multivariate calibration," *Vibrational Spectroscopy*, vol. 84, pp. 24-29, 2016.

[17] A. L. Collins, L. J. Williams, Y. S. Zhang, M. Marius, J. A. J. Dungait, D. J. Smallman, E. R. Dixon, A. Stringfellow, D. A. Sear, and J. I. Jones, "Catchment source contributions to the sediment-bound organic matter degrading salmonid spawning gravels in a lowland river, southern England," *Science of the Total Environment*, vol. 456-457, pp. 181-195, 2013.

[18] S. Jia, X. Yang, J. Zhang, and G. Li, "Quantitative analysis of soil nitrogen, organic carbon, available phosphorous, and available potassium using near-infrared spectroscopy combined with variable selection," *Soil Science*, vol. 179, pp. 211-219, 2014.

[19] R. K. Douglas, S. Nawar, M. C. Alamar, A. M. Mouazen, and F. Coulon, "Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques," *Science of the Total Environment*, vol. 616, pp. 147-155, 2017.