Comparative Analysis for Grey Relation Estimation Models of Soil Organic Matter based on Hyperspectral Data

LI Xinhao\textsuperscript{1a}, LI Jiangong\textsuperscript{2b}

\textsuperscript{1}College of Information Science and Engineering, Shandong Agricultural University, Taian, Shandong, China
\textsuperscript{2}College of Landscape Architecture and Art Design, Hunan Agricultural University, Changsha, Hunan, China

\textsuperscript{a}lxh17866703382@163.com \textsuperscript{b}li769890843@163.com

Abstract: Rapid and accurate acquisition of farmland soil organic matter content is of great significance for the rapid monitoring soil fertility and the development of precision agriculture. Based on the hyperspectral data and organic matter content data of 74 soil samples in Zhangqiu District, Jinan City, Shandong Province, first transform soil spectral reflectance by the logarithm of the first-order differential, square, reciprocal, logarithm and square root, the estimation factors were selected according to the principle of maximum relation. Second, the hyperspectral estimation of soil organic matter was carried out by using six models such as grey close relation degree, and the estimation results of different methods were combined and analyzed. The results show that the combined grey relation model can effectively improve the estimation accuracy, and the average relative error of 16 test samples is 8.931%. Studies have shown that using the combined model of grey relation degree to estimate soil organic matter content with hyperspectral data is valid.

1. Introduction

Soil organic matter is an important indicator of soil fertility. Rapid and accurate acquisition of farmland soil organic matter content is of great significance for the rapid monitoring soil fertility and the development of precision agriculture. In recent years, the development and application of hyperspectral technology has provided technical support for rapid monitoring of soil fertility and development of fine agriculture\cite{1}. Since the 1960s, international attempts have been made to carry out quantitative inversion studies of soil organic matter reflectivity spectra using spectral detection instruments\cite{2-4}. Ben-Dor conducted controlled experiments on the decomposition process of organic matter in soils and pointed out that all traditional estimates of soil organic matter from albedo are biased by the stage of decomposition of organic matter in the soil\cite{5}. Galvão confirmed, through an indoor study, that the soil reflection spectrum at the absorption peak of 550~700nm is mainly caused by organic matter in the soil and was determined using AVIRIS image hyperspectral data was used to analyze the relationship between organic matter and soil reflection spectra\cite{6}. At present, the method of building hyperspectral inversion models of soil organic matter mainly adopts one-dimensional or multivariate linear statistical regression analysis, and the general linear analytical statistical model inversion accuracy is not satisfactory due to the more complex nonlinear relationship between soil organic matter and its hyperspectral characteristics\cite{7}. Soil organic matter content and reflectivity data are affected by numerous factors, making the obtained experimental data inevitably...
subject to random errors and grey in their connotation[8] and ambiguity in their delineation[9-10], resulting in poor correlation between soil organic matter and spectral characteristics. Therefore, in order to improve the accuracy of soil organic matter spectral estimation, genetic analysis, statistical analysis and grey correlation analysis were used[11] to establish a soil organic matter spectral estimation model based on 74 soil sample data from Zhangqiu District, Jinan City, and to compare and analyze the effects of different grey correlation identification methods and their fusion models on the estimation results, and more satisfactory results were obtained.

2. Data acquisition and processing

2.1 Overview of the study area
The study area was selected in Zhangqiu District, Jinan City (latitude 36°25′-37°09′ and longitude 117°10′-117°35′), which is located in Jinan city in the east. Zhangqiu District contains Mountain District, hills and Plains. The high south and low north, the Yellow River flows through the northern border. The higher the temperature, the greater the precipitation. There are 4 soil classes, 11 subclasses, 20 soil genera and 87 soil species in Zhangqiu District. The distribution of sampling sites within the study area is shown in Figure 1.

![Figure 1 Distribution area of sampling sites in the study area](image)

2.2 Data acquisition and pre-processing
2017 Year Mar. 74 soil samples were collected in the study area in March. In the experimental area, the sampling area selection criteria were gentle terrain and uncovered surface area, sampling method was plum sampling method, sampling principle was uniformity and representativeness, collecting surface soil within 20 cm of the surface was collected as samples, and the organic matter content was determined using the ferrous sulfate titration method, and the numerical characteristics of the samples are shown in Table 1.

| Characteristic index     | characteristic value |
|-------------------------|----------------------|
| Minimum / mg            | 9.626                |
| Maximum / mg            | 29.387               |
| Mean / mg               | 20.760               |
| Variance / mg           | 25.813               |
| Coefficient of variation /% | 24.472             |

Using a wavelength range of 350-2500nm of ASD Field Spec Pro FR portable field spectrometer. Spectral data acquisition with a spectral bandwidth of 3nm(350-1000nm),10nm (1000-2500nm) and then the interpolation points are automatically selected for interpolation, and the spectral re-sampling interval is 1nm. The original spectral profile is shown in Figure 2.
Due to the influence of various condition factors, the spectral data of some soil samples may be abnormal in the spectral measurements, so it is necessary to remove the data of samples with abnormal spectral reflectivity from the collected soil. Based on the negative correlation between soil spectral reflectivity and organic matter content, the reflectivity spectra were ranked according to the level of organic matter content, and the trend was analyzed comprehensively to remove the spectral curve samples with obvious anomalies. The remaining 71 samples with sample numbers 32, 33, and 57 are normal samples.

2.3 Hyperspectral Characterization of Soil Organic Matter Content
According to the organic matter content, the 71 soil samples were divided into three groups, and the spectral reflectivity of all samples in each group was averaged, and the spectral characteristic grouping curves are shown in Fig. 3.

As shown in Fig. 2, the trend of the characteristic curves is consistent, and the organic matter content is negatively correlated with the hyperspectral reflectivity which is consistent with the confirmed findings.

Airborne moisture has a strong influence on the band, making the spectral noise in 1400nm and 1850nm near the band very strong and therefore not considered. Also the soil spectral reflectivity is at 350-500nm at the same time as 1300nm increases slowly. The growth was faster at 500-1300nm locations.

2.4 Spectral feature extraction
Respectively, using the first-order differentiation of the logarithm of the reflectivity, the square, inverse, first-order differentiation, logarithm and square root transformations. The original spectral reflectivity was transformed using six transformation methods to improve the correlation between organic matter
The processing results showed that the soil spectral data correlation with organic matter was significantly enhanced only after the first-order differential transformation of the logarithm, with little improvement in the correlation for other changes. Therefore, only feature indicators are selected from the logarithmically first-order differentially transformed spectral data.

2.5 Grey correlation estimation models and combination methods

The basic idea of grey correlation is to determine the proximity and similarity of curve geometries of different data series. Only the grey proximity correlation model is presented below[11].

Let the known patterns be used as the reference series and the samples to be identified as the comparison series. With data sequences

\[ X_i = (x_i(1), x_i(2), \cdots, x_i(n)) \]  

where \( i \) denotes the serial number, is said to be

\[ X_i(t) = \{x_i(k) + (t-k)(x_i(k+1) - x_i(k))\} | k = 1, 2, \cdots, n-1, t \in [k, k+1] \]

is the fold corresponding to the data sequence \( X_i \), then

\[ s_i(k) = \int_{k}^{k+1} X_i(t) dt \]

If there is a data sequence \( X_j, X_j \), among \( k = 1, 2, \cdots, n-1 \), then

\[ s_i(k) - s_j(k) = \int_{k}^{k+1} (X_i(t) - X_j(t)) dt \]

where \( X_i(t), X_j(t) \) denotes the first known pattern of the sample to be identified \( i \) and the first known pattern of \( j \), respectively indicator \( k \) value.

The grey proximity correlation is

\[ \rho_{ij} = \frac{1}{1 + |s_i - s_j|} \]

where \( \rho_{ij} \) denotes the grey correlation of the sample to be identified \( i \) with the first \( j \) known pattern.

From equation (5), the grey correlation sequence between the to-be-identified and known patterns can be calculated, and according to the maximum correlation principle, the organic matter content of the known pattern closest to the to-be-identified sample is taken as the estimated value of the to-be-identified sample.

In addition to modeling with the above grey proximity correlation, this study also used Dunn's grey correlation, combined Euclidean distance and grey correlation analysis, combined Hemming distance and grey correlation analysis, combined Euclidean, Hemming distance and grey correlation analysis, similar correlation, and combined similar correlation and proximity correlation analysis, for a total of seven modeling methods. The predictions of the grey proximity correlation model were combined with the predictions of the other six grey correlation models to determine the optimal prediction based on the mean relative error of the test sample.

3. Analysis of results

3.1 Characterization of indicators

Five bands were discrete selected throughout the band and avoided in 1350nm and 1850nm nearby bands were selected for the characteristic indicators. The selected bands and correlation coefficients are shown in Table 2.
Table 2 characteristic bands of organic matter content

| Characteristic index (nm) | 544  | 1469 | 1656 | 2059 | 2318 |
|--------------------------|------|------|------|------|------|
| Correlation coefficients | -0.685 | -0.677 | -0.715 | -0.701 | 0.725 |

3.2 Estimation results based on grey proximity correlation

Based on the organic matter content, after excluding abnormal samples, randomly selected samples from 71 samples were selected from 16 samples as samples to be identified (test samples), and another 55 samples were used as known models for grey correlation estimation and comparative analysis.

Using equation (4) to calculate separately 16 samples to be identified with the area difference between the 16 specimens to be identified and the 55 known patterns was calculated, and the grey proximity correlation series was calculated according to equation (5), and the closest known pattern to the specimen to be identified was judged based on the principle of maximum correlation, and its organic matter content was taken as the estimated value of the sample to be identified. 16 The estimated results for the test sample are shown in Table 3.

| Sample number | Measured value | Predictive value | Relative error | Average relative error % |
|---------------|----------------|-----------------|---------------|--------------------------|
| 1             | 20.635         | 17.2628         | -16.3441      |                          |
| 6             | 25.7723        | 18.3757         | -28.6997      |                          |
| 8             | 24.4649        | 26.5891         | 8.6828        |                          |
| 14            | 17.6664        | 20.5191         | 16.0802       |                          |
| 18            | 26.4139        | 25.5445         | -3.2915       |                          |
| 21            | 25.9854        | 26.2421         | 0.9880        |                          |
| 23            | 15.3785        | 11.4882         | -25.2971      |                          |
| 25            | 22.2062        | 25.4246         | 14.4977       | 13.4893                  |
| 31            | 21.7325        | 18.9955         | -12.594       |                          |
| 36            | 23.7329        | 25.4398         | 7.192         |                          |
| 38            | 15.4669        | 17.7636         | 14.8492       |                          |
| 40            | 21.2764        | 16.7874         | -21.0966      |                          |
| 45            | 15.9661        | 12.5805         | -21.205       |                          |
| 55            | 19.1453        | 21.6795         | 13.2368       |                          |
| 59            | 27.2175        | 26.5891         | -2.3085       |                          |
| 71            | 28.7554        | 26.2421         | -8.7403       |                          |

As seen in Table 3, the maximum relative error in estimating soil organic matter content using grey proximity correlation was -28.7%, the minimum relative error was 0.988% and the average relative error was 13.49%.

3.3 Comparative analysis of the estimation results of different grey correlation models

In addition to estimation using grey proximity correlation, this study then used Dunn's correlation (E1), Euclidean distance correlation (E2), Hemming distance correlation (E3), Euclidean Hemming correlation (E4), similar correlation (E5) and similar proximity correlation (E6) for estimation. The six grey correlation models were used to calculate separately 16 samples to be identified with 55 grey correlation sequences of the known patterns. Based on the principle of maximum correlation, the closest known pattern organic matter content to the sample to be identified was used as an estimate for the sample to be identified, and the results are shown in 4.

| Sample number | Grey relational degree model |
|---------------|-----------------------------|
|               | E1  | E2  | E3  | E4  | E5  | E6  |
| 1             | 36.125 | -44.328 | -44.328 | -44.328 | -44.328 | -18.648 |
| 6             | -2.322 | -1.290 | -1.290 | -1.290 | -2.276 | -15.880 |
| 8             | 8.683 | 8.683 | 8.683 | 8.683 | -17.869 | 8.683 |
| 14            | -28.383 | 1.410 | -28.383 | 1.410 | 48.794 | 1.898 |
As seen in Table 4, similar proximity correlation (E6) has the smallest mean error with a mean error of 10.3%, while similar grey correlation (E5) has the largest mean error with a mean error of 34.3%.

3.4 Comparative analysis of the estimation results of different combinations of grey correlation models

To improve the accuracy of the grey correlation estimation results, the estimation results of the grey proximity correlation model were fused with the estimation results of Dunn's correlation (E1), Euclidean distance correlation (E2), Hemming distance correlation (E3), Euclidean Hemming correlation (E4), similar correlation (E5) and similar proximity correlation (E6), respectively. The fused models are denoted by E01, E02, E03, E04, E05, E06 respectively, and the results are shown in Table 5.

Table 5 estimation accuracy analysis of organic matter content using grey correlation degree combination model

| Sample number | E01 | E02 | E03 | E04 | E05 | E06 |
|---------------|-----|-----|-----|-----|-----|-----|
| 1             | 9.891 | 30.336 | 30.336 | 30.336 | 30.336 | 17.496 |
| 6             | 15.511 | 14.995 | 14.995 | 14.995 | 15.488 | 22.290 |
| 8             | 8.683 | 8.683 | 8.683 | 8.683 | 4.593 | 8.683 |
| 14            | 5.787 | 9.109 | 5.787 | 9.109 | 32.802 | 9.353 |
| 18            | 1.618 | 3.518 | 4.434 | 3.518 | 30.098 | 4.434 |
| 21            | 14.148 | 25.912 | 14.148 | 14.148 | 25.912 | 1.313 |
| 23            | 21.746 | 21.746 | 21.746 | 21.746 | 19.199 | 21.746 |
| 25            | 15.880 | 15.880 | 15.880 | 15.880 | 12.534 | 0.017 |
| 31            | 4.507 | 4.507 | 4.507 | 4.507 | 30.108 | 1.085 |
| 36            | 9.275 | 9.275 | 9.275 | 9.275 | 22.423 | 7.192 |
| 38            | 15.012 | 17.372 | 15.012 | 17.372 | 5.778 | 3.190 |
| 40            | 6.629 | 6.629 | 6.629 | 6.629 | 1.120 | 17.199 |
| 45            | 4.547 | 4.547 | 4.547 | 4.547 | 21.252 | 4.547 |
| 55            | 24.626 | 24.626 | 24.626 | 24.626 | 20.745 | 13.603 |
| 59            | 4.420 | 4.420 | 4.420 | 4.420 | 4.420 | 2.309 |
As seen in Table 5, the average error of all combined models is reduced to some extent, and the average error of E01, E02, E03, E04, E06 is lower than the average error of both models used for the combination, and the estimation accuracy is effectively improved. Among them, the combined grey proximity correlation with and similar proximity correlation model (E06) has the highest estimation accuracy with an average relative error of 8.931%. This indicates that the combined model using grey correlation to estimate soil organic matter content is effective.

4. Conclusion
A hyperspectral estimation model of soil organic matter content based on grey correlation was developed to address the grey nature and uncertainty of organic matter content in soil. The results show that the estimation accuracy of a single grey correlation model is not satisfactory, while the combination of grey proximity correlation with and similar proximity correlation model can effectively improve the estimation accuracy. Compared to statistical modeling and other methods, the physical-mathematical meaning of the grey correlation estimation model is clear and simple to calculate, and the results of the study provide a new way for hyperspectral estimation of soil organic matter content.

References
[1] Application of imaging spectrum technology in land dynamic monitoring” research group. (2005) Application of imaging spectrum technology in land dynamic monitoring [M]. Geological Publishing House, Beijing.
[2] HE Ting, WANG Jing, LIN Zongjian, CHENG Ye. (2006) Spectral Features of Soil Organic Matter [J]. Geomatics and Information Science of Wuhan University, Wuhan. 31(11): 975-978.
[3] Verstraete M M, Pinty B. (1991) The Potential Contribution of Satellite Remote Sensing to the Understanding of Arid Lands Processes [R]. Vegetation, world, 59-72.
[4] Smith M O, Ustin S L, Adams J B, et al. (1990) Vegetation in Deserts: I A Regional Measure of Abundance from Multi-spectral Images [J]. Remote Sens Environ, 31: 1-26.
[5] Ben-Dor E, Banin A. (1994) Visible and Near Infrared (0.4 —1.1um) Analysis of Arid and Semi-arid Soil [J]. Remote Sens Environ, 41: 10-124.
[6] Galvão L S, Vitorello Í. (1998) Variability of laboratory measured soil lines of soils from southeastern Brazil. Remote Sensing of Environmen, 63: 166-181.
[7] Chang, C. W., Laird, D. A, et al. (2002) Near-infrared reflectance spectroscopic analysis of soil C and N [J]. Soil Sci, 167(2): 110-116.
[8] LI Mingliang, LI Xican, ZHANG Shuang. (2016) Grey Relation Estimating Pattern of Soil Water Content Based on Hyper-Spectral Data [J]. Journal of Geomatics Science and Technology, 33(02): 163-168.
[9] LI Xican, ZHAO Gengxing, CHEN Hongyan, et al. (2016) The Interval Estimation of Soil Organic Matter Content Based on Hyper-Spectral Data. Journal of Geomatics Science and Technology, 31(6): 592-597.
[10] XU You-you, LI Xi-can, SHANG Xuan, et al. (2019) Estimation Model of Soil Moisture Content Based on Hyperspectral Data and Semi-supervised Fuzzy Recognition. Geography and Geoinformation Science, 35(2): 33-37.
[11] Liu SiFeng, et al. (2017) Grey system theory and its application [M]. Science Press, Beijing.