Hyperspectral estimation of petroleum hydrocarbon content in soil using ensemble learning method and LASSO feature extraction

Menghong Wu, Nan Lin, Genjun Li, Hanlin Liu and Delin Li

College of Resources and Environment, Jilin Agricultural University, Changchun, Jilin, China; College of Surveying and Exploration Engineering, Jilin Jianzhu University, Changchun, Jilin, China; Geological Survey Institute, Xining, Qinghai, China; Key Laboratory of Geological Processes and Mineral Resources of the Northern Qinghai-Tibet Plateau, Xining, Qinghai, China

ABSTRACT
Rapid and accurate estimation of soil petroleum hydrocarbon content is crucial for analyzing the degree of soil pollution and evaluating pollution status. Surface soil samples and hyperspectral measurements in the reservoir area in Lenghu Town, Qinghai Province, China, were viewed as research objects, and the correlation between different spectral forms of original data and petroleum hydrocarbon content in soil was analyzed. To improve the estimation accuracy, we proposed a solution that introduces least absolute shrinkage and selection operator (LASSO) combined with extremely randomized trees (ERT) and gradient boosting decision tree (GBDT) ensemble learning for constructing hyperspectral estimation model. The results show: LASSO algorithm can not only solve the spectral multicollinearity problem effectively but also reduce the number and calculation complexity of soil hyperspectral variables to a great extent. Compared with traditional machine learning, ERT and GBDT perform superior. In particular, the estimation accuracy of the LASSO-GBDT model is the highest.

1. Introduction

Soil is a key component of living environment [1]. From a production perspective, soil can provide fertility for green plants; from the perspective of environmental protection, it is a receptor of pollutants. With the characteristics of concealment, latency, irreversibility and chronicity, soil pollution can further cause and promote pollution of water, atmosphere, and biology. It is a vital part of environmental pollution affecting human health [2,3]. Petroleum may leak into the soil during mining, transportation, use, and storage [4]. Large quantities of petroleum hydrocarbons lead to changes in physicochemical properties of soil, fertility reduction, salinization, and desertification [5]. In addition, they endanger life and health through plants and the food chain. More importantly, the structure and function of the ecosystem can be damaged by pollution. Rapid and accurate estimation of petroleum hydrocarbons in soil is not only beneficial for analyzing and providing information on soil pollution status but also has application value and practical significance for protecting and managing the ecological environment of soil [6,7].

Traditional detection of petroleum hydrocarbon pollution in soil requires field collection of soil samples and is conducted in the laboratory [8,9]. Rapid and large-scale monitoring is not available due to the high cost of detection and the lack of timeliness and portability [10]. In recent years, hyperspectral remote sensing has been applied to soil component content estimation on account of its rapidity and accuracy [11,12]. Hyperspectral remote sensing records information on soil reflectance spectra through dozens of wavelengths arranged closely [13–15]. By utilizing characteristics of reflectance spectra to estimate physicochemical properties quantitatively, hyperspectral remote sensing has incomparable advantages over traditional detection techniques, especially for the efficient acquisition of information on soil pollution over large areas [16–18]. For the past few years, the feasibility of hyperspectral remote sensing technology in monitoring petroleum hydrocarbon pollution in soil has been widely demonstrated [7,19]. Hauser et al. established a partial least squares regression (PLSR) model to predict the TPH content of samples using the near-infrared spectroscopy (NIRS) method with spectral data collected from the training set, and the correlation coefficient between validation set data of prediction model and the measured laboratory data was relatively high, indicating that the NIRS method can be applied to predict soil TPH content effectively [20]. Scafutto et al. designed experiments to immerse oil with different concentrations into the soil and established estimation models of petroleum hydrocarbon content in soil using ground hyperspectral data.
The results of the comparison between the predicted results and measured data proved that hyperspectral remote sensing technology enables effective monitoring even at low oil concentrations [21]. Yin et al. demonstrated that the prediction result of the stepwise multiple linear regression (SMLR) model established by hyperspectral data for estimating TPH content was superior to that of the univariate regression model, which depicted that the SMLR model can be a satisfactory solution for rapid quantitative assay of TPH in soil content [22]. Chakraborty et al. established a multiple linear regression (MLR) model with wavelet transform based on hyperspectral data for predicting petroleum hydrocarbons in the case of the complex composition of soil pollutants because the spectral information of various petroleum hydrocarbon components overlaps each other [23].

With the rapid development of machine learning theory and technology, the estimation models between hyperspectral features and component content of soil based on machine learning algorithms are studied widely [24–27]. The technique can adaptively simulate the relationship between input data and output data and continuously mine soil spectral information, showing an excellent estimation effect [28]. At present, artificial neural network (ANN), classification and regression tree (CART), support vector machine (SVM), and extreme learning machine (ELM) are applied widely for the estimation of soil component content [29–32]. The process of machine learning solution can be considered the process of searching for a learning model with strong generalization ability and high robustness in hypothesis space [28,33,34]. Since the hypothesis space is artificial, the actual target hypothesis is not in the hypothesis space in most applications of machine learning. The process presents some problems such as strong sensitivity to training samples, high calculation complexity, and over-fitting. Ensemble learning is a hot topic in machine learning [35]. By constructing a learning model that combines multiple learning models with the same or different species, an ensemble learning model usually achieves better performance than a single learning model [36,37]. In terms of learner combination rules, the commonly used ensemble learning methods include bagging parallel and boosting serial ensemble learning [38]. In recent years, various ensemble learning algorithms, such as adaptive boosting (Adaboost), random forest (RF), and Stacking, have been applied widely to study the spectral quantitative estimation of soil component content [39,40]. Compared with the traditional methods, the machine learning constructed by the integrated framework performs better due to its strong leaning ability, high stability, and strong generalization ability.

Hyperspectral remote sensing data are large and informative because the soil spectral information is recorded by continuous and tight bands. What’s more, the number of band is so much that lead to the phenomenon of collinearity and superfluous information affecting the modeling prediction result [41]. The LASSO algorithm is widely used because it can extract key variables from high-dimensional variables efficiently, solve the multicollinearity problem of variables, and improve the accuracy of the model [42]. The algorithm is a compression estimation method constructed through reserving subsets. The refined model built by constructing the penalty function and adjusting the coefficients of compression variables by setting the values of penalty parameters. The variable coefficient keeps converging to 0 as the penalty parameter increases. When the penalty parameter is large enough, some of the variable coefficients will be compressed to 0, and the model achieves the goal of reducing the original variable [43–45].

The surface soil samples collected from reservoir areas in Lenghu Town, Haixi Prefecture, Qinghai Province, China, were used as experimental data to analyze spectral reflectance and its change characteristics. The LASSO regression algorithm used for selecting the feature band of hyperspectral data and the ensemble learning algorithm, including ERT and GBDT, were combined to build a hyperspectral estimation model of petroleum hydrocarbons in soil. The prediction performance of the models was evaluated by coefficient of determination (R²), root-mean-square error (RMSE) and residual prediction bias (RPD). The main purposes of this study were to: 1) analyzing the correlation between different transformation forms of soil spectrum and petroleum hydrocarbon content and confirming the sensitive band range; 2) constructing estimation model of petroleum hydrocarbon in soil with ensemble leaning method based on LASSO algorithm, which can improve the performance of the ensemble leaning model in estimating. The study can provide an efficient new method for estimating petroleum hydrocarbon content in soil.

2. Materials and methods

2.1. Study area and sample collection

The study area is an important petroleum industry base located in Lenghu Town, Mangya City, Haixi Mongolian, and Tibetan Autonomous Prefecture, Qinghai Province, distributed on the north margin of Qaidam Basin. The terrain is high in the northwest and low in the southeast, and the area is an extremely dry area in China with large areas of Yadan landforms. The soil sampling area is located on the west side of S210 provincial road in the southeast of Lenghu Town. First, 45 sampling sites were selected around the oil well, and the distance between sampling and the highway was 150 m at least. Then, sampling grid sized 50 × 50 m was set at every sampling site, and the five-point
sampling method was applied to collect five samples with a surface thickness of 5–15 cm in the grid. Finally, the five samples were mixed evenly and put into a sealed plastic bag. After encapsulating the soil sample, the sampling point was numbered, and the GPS positioning technology was used to record the longitude and latitude coordinates of sampling points at the center point. Thus, a total of 45 soil samples were collected (Figure 1).

2.2. Petroleum hydrocarbon content and spectral measurement

After removing straw, gravel, and other impurities, samples with grain sizes less than 0.25 mm were screened through a 0.25 mm mesh sieve after air drying (Figure 2). The qualified samples were divided into two parts: one for the determination of petroleum hydrocarbons in soil and the other for indoor high-spectral measurements. After extraction, purification, concentration, and constant volume, petroleum hydrocarbon (C10-C14) content was measured using the gas chromatographic method. The results of the statistical analysis of 45 petroleum hydrocarbon samples are illustrated in Table 1. The maximum and mean values of petroleum hydrocarbon content were 15.41 and 3.75 g/kg, respectively. According to China National Environmental Quality Standards for Soils, the standard value in soil pollution evaluation of petroleum hydrocarbon is 5 g/kg in industrial land. The pollution level of some soil samples in the study area exceeds the standard value of industrial land. In order to ensure the balance of sample content of training and validation set, 45 samples were divided into 15 sample groups according to their content level. Selecting one sample from each group randomly as the validation samples, the remaining were training samples.

The ASD FieldSpec4 spectroradiometer was used to perform spectral measurements on the collected soil samples in a dark room to obtain continuous spectral curves within 350–2500 nm. Then, 45 samples were put into a black glass container, and a 50-watt halogen lamp was placed vertically over the soil sample as an indoor light source. In order to improve the accuracy of spectral measurement data, the whiteboard was calibrated before each measurement, and the mean of 10 spectral measurements was taken as the reflectance spectral data of the soil sample (Figure 3). The spectral reflectance data distributed in the bands of 350–399 nm and 2451–2500 nm were eliminated because of the low signal-to-noise ratio. The instrument detection elements of

Figure 1. The distribution of the soil sampling points.

Figure 2. Petroleum hydrocarbon samples.
spectroradiometer are different considering different spectral band range. 1050 nm is the interface band channel of visible and short-wave infrared sensor, 1800 nm is the interface band channel of short-wave infrared and long-wave infrared sensor, and therefore, the reflectance curve of the spectroradiometer appears breakpoint jump phenomenon at 1050 nm and 1800 nm, the breakpoint of spectral data were corrected by instrument software in the study. Since the sample interval of the ASD FieldSpec4 spectroradiometer was 1 nm, the spectral resolution was high, and the number of bands was large, information overlap may exist between adjacent bands. Thus, the spectral data were resampled, and the sampling interval of wavelength was set to 5 nm. Based on denoising and resampling, the original spectral reflectance was transformed by first-order differential transformation, reciprocal logarithmic and continuum removal, etc. Different forms of spectral transformation help to find the peak value accurately and quickly, thus picking out the sensitive bands.

2.3. LASSO algorithm

The LASSO algorithm, first proposed by Robert Tibshirani in 1996 [46], is essentially a compression estimation method that preserves subset contraction and is a way to deal with biased estimators of complex collinear data. Based on linear regression, the absolute value of the regression coefficient is constrained to a certain threshold by adding a normal function, and the variables with a correlation less than the threshold value are compressed to 0 and eliminated by optimizing the objective function. Therefore, all remaining variables are the required feature variables [42,47].

For a data set $D = (X', y_j), j = 1, 2, L \cdots n$, $X' = (X_{j1}, \cdots, X_{jm})$ represents the explanatory variable, $y_j$ represents the explained variable, and the regression coefficient is $\hat{\beta} = (\hat{\beta}_1, L, \hat{\beta}_p)^T$. Given a simple linear regression model, the optimized objective function is expressed as:

$$\hat{\beta} = \arg\min\left\{ \sum_{j=1}^{n} (y_j - \sum_{i=1}^{m} \hat{\beta}_i x_{ij})^2 \right\}$$

(1)
When the number of independent variables is large or there is a multicollinearity problem between variables, the coefficient matrix is unstable and cumbersome to solve, and the model is more complicated and easy to produce over-fitting. In order to solve the problems, a penalty function item is added to the LASSO algorithm based on the loss function to constrain the complexity of the model and achieve the purpose of selecting models with low empirical risk and model complexity. The optimization objective function after adding the penalty function is:

$$
\hat{\beta} = \arg\min \left\{ \sum_{j=1}^{n} \left( y_j - \sum_{i=1}^{m} \hat{\beta}_i x_{ij} \right)^2 + \lambda \sum_{i=1}^{m} |\beta_i| \right\}
$$

(2)

The second term represents the penalty to the coefficient, and $\hat{\lambda}$ is the adjustment coefficient that controls the compression degree of each variable. The variable is adjusted by the changes in $\hat{\lambda}$, which can compress the coefficient of unimportant variables to 0. Smaller $\hat{\lambda}$ represents smaller punishment and more reserved variables, while the larger $\hat{\lambda}$ means greater punishment, and fewer reserved variables. Therefore, the algorithm can extract feature variables effectively.

### 2.4. ERT model

The ERT method belongs to the ensemble learning method under the parallel integration framework proposed by Pierre Geurts and other scholars in 2006 after many experimental studies [48]. The method improves the similarity problem of decision trees in random forest [49]. In the extreme random tree, each base learner is trained based on the whole data set, which ensures the utilization of training samples and reduces the final prediction bias to a certain extent [50]. In order to ensure the structural difference between each decision tree, the extreme random tree introduces greater randomness in node division. Additionally, to strengthen randomness, when splitting, the split feature points of each code are randomly selected. Thus, the variance of the decision tree is reduced, and the generalization performance is improved. All base learning prediction results are recorded, and the final result is produced by a voting decision [51].

According to the ‘error-ambiguity decomposition’ theory, the generalization error of the base learner is assumed to be $Ei$. Therefore, the weighted generalization error of the learner is:

$$
\bar{E} = \sum_{i=1}^{r} \omega_i E_i
$$

(3)

assuming that the bifurcation value of the base learner is $Ai$, the weighted bifurcation value of the learner is:

$$
\bar{A} = \sum_{i=1}^{r} \omega_i A_i
$$

(4)

the generalization error of the extreme random tree is expressed as:

$$
E = \bar{E} - \bar{A}
$$

(5)

improving the diversity of base learners and the prediction accuracy of learners can further promote the prediction accuracy of the extreme random tree model.

### 2.5. GBDT model

The GBDT is an algorithm that combines the decision tree and boosting thought proposed by Professor Friedman of Stanford University in 2001 [52]. It is an additive model of a weighted combination of multiple weak learners into strong learners and can be expressed as [53]:

$$
F(x) = \sum_{t=1}^{T} a_t h_t(x)
$$

(6)

where $F(x)$ is the model objective function; $T$ is the number of trees to be constructed in the decision tree for degree improvement; $h_t(x)$ is the decision tree of weak learners; $a_t$ is the weight of the $t$th tree.

The type of decision tree determines what kind of problems can be solved by the gradient boosting tree. When it is a classification problem, the decision tree is a binary classification tree; when it is a regression problem, the decision tree is a binary regression tree.

The GBDT adopts a forward distribution algorithm [54], and $F_0(x)$ is the initial value of the model that is always constant. When running at $m$, the model is expressed as:

$$
F_m(x) = F_{m-1}(x) + a_m h_m(x)
$$

(7)

where $F_{m-1}(x)$ is the current model, and the new classification regression tree $h_m(x)$ is obtained by minimizing the loss function:

$$
\hat{h}_m = \arg\min_h \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + h(x_i))
$$

(8)

where $N$ is the number of samples. The gradient boosting decision tree solves optimal model by gradient descent method, and the descent direction is defined as the negative gradient value of $F_{m-1}(x)$.

$$
F_m(x) = F_{m-1}(x) - a_m \sum_{i=1}^{N} \nabla_{x} L(y, F_{m-1}(x_i))
$$

(9)

$a_m$ is acquired by line search:
\[ a_m = \arg \min_h \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i)) - \alpha \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \]  

(10)

regularization of decision trees is promoted by gradient, which can be conducted by setting the learning rate:

\[ F_m(x) = F_{m-1}(x) + \nu a_m h_m(x) \]  

(11)

where \( \nu \) represents the learning rate. A lower learning rate means that more decision trees are needed, and the final error will be smaller, but it will also increase the training time.

3. Results

3.1. Analysis of soil sample spectral characteristic

Spectral curves of indoor soil samples after treatment are shown in Figure 4. It can be seen from Figure 4(a) that the original reflectance of the collected soil samples is between 0 and 0.7, and the spectral curve fluctuations of each sample are similar. In the visible band, the reflectance increases sharply with the increase in wavelength. The spectral reflectance of soil in the near-infrared region is generally higher than that in the visible region. Two obvious troughs distributed around 1400 nm and 1900 nm are mainly caused by the absorption of residual water in the soil and water vapor in the air. There is a slight depression at 2200 nm, which is mainly affected by clay minerals in

![Figure 4: Original and transformed spectral reflectance curves of soil: (a) Original spectral reflectance (R); (b) Spectral first-order differential transformation (FD-R); (c) Spectral reciprocal logarithmic transformation (lg1/R); (d) Spectral reciprocal logarithmic first-order differential transformation (FD-(lg1/R)); (e) Spectral continuum removal transformation (CR); (f) Spectral continuum removal first-order differential transformation (FD-(CR)).](image-url)
the soil. **Figures 4** (b)–(f) show the spectral curves of the original reflectance transformed by the first-order differential transformation (FD-R), reciprocal logarithmic transformation (lg1/R), reciprocal logarithmic first-order differential transformation (FD- lg1/R)), continuum removal transformation (CR), and continuum removal first-order differential transformation (FD-CR), respectively. According to the change results, the various transformations of spectral can amplify the original spectral change significantly, which is consistent with previous research result [55].

### 3.2. Spectral correlation coefficient analysis

The Pearson correlation coefficient between petroleum hydrocarbon content in soil and soil spectral reflectance was calculated, and a correlation graph was drawn. According to the calculation result of the correlation coefficient of the original spectrum in **Figure 5(a)**, the petroleum hydrocarbon content in soil is negatively correlated with the spectral reflectance. Compared with the original spectral reflectance, the spectral data after characteristic transformation show an improved correlation with petroleum hydrocarbon content (**Figure 5** (a)–(f)). Among them, the first-order differential transformation has a significant

---

**Figure 5.** Soil spectral correlation coefficient curves: (a) Original spectral correlation coefficient curves(R); (b) Spectral first-order differential transformation correlation coefficient curves(FD-R); (c) Spectral absorbance transformation correlation coefficient curves(lg1/R); (d) Spectral absorbance first-order differential transformation correlation coefficient curves(FD- lg1/R)); (e) Spectral continuum removal transformation correlation coefficient curves(CR); (f) Spectral continuum removal first-order differential transformation correlation coefficient curves(FD-CR).
effect. After the first-order differential transformation, the correlation coefficient between petroleum hydrocarbon content in soil and spectral reflectance fluctuates positively and negatively and the number of peaks was increased. In addition, the peak of correlation coefficient increases significantly, which is conducive to improving the prediction accuracy of the spectral estimation model.

A total of 410 spectral bands were obtained after spectral denoising and resampling at 5 nm resolution. However, the bands are still very dense. Therefore, the bands with high correlation were selected as the training data of the ensemble learning model according to the calculation result of correlation coefficients. In this study, the statistical test of significance was used to judge the degree of correlation between petroleum hydrocarbon content and spectral reflectance of samples, with a corresponding critical value of 0.38. The statistical result shows that, the correlation between spectral reflectance and petroleum hydrocarbon content is most significant after first order differential transformation, and 112 bands pass the significance test (Table 2). Compared with other transformations, the correlation coefficient of the FD-R curve has the largest peak value at −0.724, the corresponding wavelength range is around 995 nm.

3.3. Feature bands extraction based on LASSO algorithm

Compared with 410 bands after denoising and resampling, the number of bands reduces to 112 to a certain extent after the correlation coefficient significance test.

Table 2. Maximum correlation coefficients and sensitive bands.

| Evaluation index | FD−R | Lg(1/R) | FD−Lg(1/R) | CR | FD−CR |
|------------------|------|---------|------------|----|-------|
| Pearson Maximum  | −0.724 | 0.414  | −0.585 | −0.678 | −0.601 |
| Bands            | 995  | 2025    | 2215 | 595 | 985   |
| Number           | 112  | 103     | 78   | 76  | 107   |

Figure 6. The calculation results of RMSE with the different values of penalty coefficient $\lambda$ (Unit: g/kg).

Nevertheless, due to the continuity and high dimension of hyperspectral measurement data, data redundancy may exist among these bands. Therefore, the LASSO algorithm is used to carry out dimensional reduction calculation of spectral band information and extract feature bands as training sample data for the ensemble learning model. Penalty coefficient $\lambda$, an important parameter of the model, reduces the complexity of the model by controlling the size of the penalty to regulate the overall convergence degree of the model, improving the performance of the model ultimately. In this study, 112 sensitive bands after first-order differential transformation were taken as input samples by the grid search method, the minimum RMSE was set as the fitness value of the optimization objective function, and the optimal value of $\lambda$ was searched by the 10-fold cross-validation method. As shown in the Figure 6, when the value of $\lambda$ is 0.09, the RMSE achieves the minimum value of 1.063, and the number of feature bands reduces from 112 to 14, effectively compressing the number of feature bands.

Figure 7. Reflectance spectrum characteristics of petroleum hydrocarbon based on the LASSO selection results.
As shown in Figure 7, 14 feature bands extracted by LASSO algorithm mainly distributed in near-infrared and shortwave-infrared spectral range.

### 3.4. Ensemble learning model of petroleum hydrocarbon content in soil

The training sample contains input and output variables of models, with the 14 feature spectral bands selected by LASSO algorithm as the input variables and the petroleum hydrocarbon content as the output variable. The ERT and GBDT, which belong to the ensemble learning model, are used for the hyperspectral estimation experiment of petroleum hydrocarbon in soil. Ensemble learning is a method to form strong learners by training several weak learners and referring to various combination strategies. From the aspect of the learner combination rule, the method of ensemble learning includes bagging parallel and boosting serial ensemble learning, and ERT and GBDT in this study are typical representatives. The base learner of the two patterns is called the regression tree, which performs outstandingly in solving regression estimation problems. When using the ensemble learning model for regression prediction, the prediction accuracy and generalization performance of the model can be improved effectively by determining the key internal parameters of the model. In the ERT model, both the depth and the number of regression tree are the key to the model, and the learning rate and the number of regression tree are the important parameters of the GBDT model. Grid search-cross validation method is applied for parameter optimization calculation in the two models, and mean absolute percentage error (MAPE) is viewed as the fitness value of optimizing objective function, which can be obtained by iterative loop calculation. The results show that: 1) MAPE is the minimum when the depth of regression tree is 5, and the number of regression tree is 20 in the ERT model; 2) when the learning rate is 0.8, and the number of regression tree is 3 in the GBDT model (Figure 8).

The ERT and GBDT models were used for hyperspectral estimation of petroleum hydrocarbon contents in soil by referring to the parameter optimization calculation results. The prediction results of the ensemble learning model without LASSO algorithm dimension reduction and the traditional machine learning SVM model were calculated (Table 3). The evaluation index of training and validation from various models were also calculated. The indexes include $R^2$, RMSE, and RPD. Among them, $R^2$ presents the stability of the model. The larger the $R^2$, the closer it is to 1, and the more stable the model is. A close of $R^2$ to 0 indicates that the predicted value is not correlated with the measured value.

As shown in the table, the modeling and prediction accuracy of both the SVM machine learning and ensemble learning models, including ERT and GBDT, are improved after feature extraction by the LASSO algorithm. The SVM model shows the greatest improvement, with $R^2$ increasing from 0.557 to 0.649 for the training data set and PRD increasing from 1.051 to 1.450 for the validation data set. According to the result of indexes, the accuracy of ERT and GBDT ensemble learning models is better than that of traditional machine learning models both in training and validation sets, and the evaluation indexes of the LASSO-GBDT model are superior than those of other models. In order to further compare and analyze the fitting and prediction effects of different models, the fitting effect diagram was drawn by using the predicted values and the measured values (Figure 9). In the figure, the horizontal axis is the measured value of soil sample, the vertical axis is the predicted value of the model, the pink point is the training sample, and the blue point is the verification sample. The closer the sample point in the figure is to line 1:1, the closer the measured value is to the predicted value. Figure 9 shows that the evaluation indexes of ERT and GBDT models were higher than those of the SVM model, and the samples of the training set and validation set were closer to the 1:1 line. That is, the ensemble learning model is superior to the traditional machine learning model in terms of fitting ability and stability, and the hyperspectral estimation model of petroleum hydrocarbons based on the ensemble theory is more outstanding.

![Figure 8](image_url)

*Figure 8. Parameter optimization of ERT and GBDT models: (a) Optimization results of the ERT parameter; (b) Optimization results of the GBDT parameter.*
4. Discussion

4.1. Spectral feature summary

Hyperspectral technology can provide rich spectral information and is often used to capture subtle changes in the content of substances in soil. However, studies have shown that when the detected element content is relatively low in the soil composition, the difference of the original spectral reflectance caused by the change of the element content is small [56]. Through mathematical transformation of the spectral data, it is found that the absorption and reflection characteristics of the spectral curve are obviously enhanced, and the correlation coefficient between the spectral reflectance and the petroleum hydrocarbon content in soil is also improved. Among them, the method of first-order differential transformation performed best, which is consistent with the research results of Ou et al. [57]. Therefore, it is necessary to improve the response degree of spectral characteristics through mathematical transformation before constructing the estimation model of soil petroleum hydrocarbon content, so as to determine the best form of spectral transformation. In addition, the spectral band information obtained from soil hyperspectral measurement is redundant to a certain extent. Some studies have shown that the introduction of high-dimensional variables complicates the calculation, reduces the effect of model fitting, and affects the calculation speed [58]. Thus, it is crucial to select appropriate spectral band for constructing estimation model of petroleum

| Ensemble learning model | Training $R^2$ | RMSEC | Validation $R^2$ | RMSEP | RPD |
|-------------------------|----------------|-------|----------------|-------|-----|
| SVM                     | 0.557          | 0.631 | 0.461          | 0.758 | 1.051 |
| LASSO-SVM               | 0.649          | 0.599 | 0.505          | 0.712 | 1.450 |
| ERT                     | 0.631          | 0.615 | 0.512          | 0.733 | 1.390 |
| LASSO-ERT               | 0.708          | 0.538 | 0.563          | 0.694 | 1.472 |
| GBDT                    | 0.716          | 0.348 | 0.607          | 0.609 | 1.403 |
| LASSO-GBDT              | 0.782          | 0.393 | 0.679          | 0.554 | 1.762 |

Figure 9. A comparison between the measured values and predicted values of the different regression models: (a) SVM model; (b) LASSO-SVM model; (c) GBDT model; (d) LASSO-GBDT model; (e) ERT model; (f) LASSO-ERT model.
hydrocarbon in soil. Pabon et al. discovered that the hydrocarbon diagnostic features are mainly seen in two spectral regions: 1700–1780 nm and 2290–2367 nm in the laboratory experiment [17]. Chen et al. stated that the absorption valleys are found near 1200–1220, 1700–1770, 2309–2311, and 2346–2350 nm of the spectral curve [59]. The feature bands extracted in the paper are consistent with the above findings. However, instead of using successive projections algorithm (SPA), stable competitive adaptative reweighted sampling (CARS), kernel principle component analysis (KPCA) and other algorithms for extracting feature spectral directly [60–62], Pearson correlation coefficient signification threshold was combined with LASSO regression algorithm to extract feature bands in the study. By adding the constraint condition that the sum of the absolute values of the coefficients of all independent variables is less than a threshold value, the LASSO algorithm makes the coefficient of the independent variables weakly correlated with the estimation target approach zero when solving the multiple regression model, which improves the representativeness of the selected variables and ensures the explanatory and concise of the model. Compared with using merely band extraction algorithm for hyperspectral data, the combination of Pearson correlation coefficient threshold with LASSO algorithm ensured the reliability of extraction results and greatly decreased computation.

4.2. Ensemble learning modelling summary

The traditional solution process of machine learning can be viewed as the process of searching for an appropriate learning model with strong generalization ability and high robustness in the hypothesis space. However, the search process is difficult. Many studies have shown that ensemble learning is a combinatorial optimization learning method, which can combine multiple simple models to get better performance combinatorial models. Liu et al. found that the estimation accuracy of extreme gradient boosting (XGBoost) is superior to other methods in predicting soil properties quantitatively [63]. Clingensmith et al. proved that the RF is more accurate than PLSR in prediction Soil organic carbon [64]. This is consistent with the results of this study, which shows that the ensemble learning model is superior to the traditional machine learning model in fitting ability and stability, and the petroleum hydrocarbon hyperspectral estimation model based on ensemble theory is more prominent. According to the prediction results of ERT and GBDT models, the various prediction indexes of the GBDT model built on the boosting serial learning principle are higher than those of the ERT model built on the bagging parallel learning principle. According to the data of soil sample content (Table 1), the large standard deviation of the petroleum hydrocarbon content value indicates that the dispersion degree is high and the data is affected by the noise information. The boosting ensemble learning model represented by the GBDT model can make each round of basic learners pay more attention to the samples of the previous round of learning errors in the training process in some way. Moreover, the boosting ensemble learning model has a relatively good fault tolerance to noise information. Therefore, the model generally performs better than the bagging ensemble learning model in handling small sample data set problems. In addition, it can be seen from the fitting effect diagram that the fitting effect of the training set is better than that of the validation set, and the distribution of sample points of the predicted validation set is relatively scattered, indicating that both the traditional machine learning model and the ensemble machine learning model have high requirements on the reliability and validity of the training sample data.

5. Conclusions

This study attempts to explore the method and feasibility of estimating the petroleum hydrocarbons in soil from hyperspectral data. Taking the surface soil samples and soil hyperspectral measurements in Lenghu Town, Qinghai Province, China, the correlation between soil spectral and petroleum hydrocarbon content was analyzed, and a method combining LASSO algorithm and ensemble learning was proposed to construct the estimation model. The results show that: 1) Pre-processing of hyperspectral measurement data, such as spectral denoising and resampling, can effectively reduce the influence of measurement environment and noise. Based on spectral data preprocessing, the original spectral reflectance is characterized by first-order differential transformation, which effectively highlight peaks and valleys of correlation curves; 2) LASSO algorithm effectively solves the problem of information redundancy existing hyperspectral band variables and improves the representativeness and integrity of characteristic data. The convergence rate of the model is speed up, and the problem of model overfitting is avoided, thus ensuring the interpretability and simplicity of the model; 3) The established ensemble learning model has higher stability, better learning effect, and more outstanding generalization performance than the traditional machine learning model. For sample data with high dispersion, the GBDT model built by the boosting serial learning principle has higher prediction accuracy and has stronger fault tolerance than the ERT model built by the bagging parallel learning principle.

Disclosure statement

No potential conflict of interest was reported by the author(s).


**Funding**

This work was supported by the National Natural Science Foundation of China (41702357 and 52178042); the Scientific and Technological Transformative Special Project of Qinghai Province (2020-SF-150); and the Science and Technology Development Project of Jilin Province (2021020301065F).

**References**

[1] Blum WEH. Functions of soil for society and the environment. Rev Environ Sci Biotechnol. 2005;4:75–79.

[2] Lu YL, Song S, Wang RS, et al. Impacts of soil and water pollution on food safety and health risks in China. Environ Int. 2015;77:5–15.

[3] Zhou D, Song X, Zhao FJ, et al. Soil environment and pollution remediation preface. Pedosphere. 2017;27:387–388.

[4] Li ZY, Ma ZW, van der Kuijp TJ, et al. A review of soil heavy metal pollution from mines in China: pollution and health risk assessment. Sci Total Environ. 2014;468:843–853.

[5] Trellu C, Mousset E, Pechaud Y, et al. Removal of hydrophobic organic pollutants from soil washing/flushing solutions: a critical review. J Hazard Mater. 2016;306:149–174.

[6] Guo P, Li HN, Li F, et al. Review of petroleum hydrocarbon pollution and microorganism remediation. Biotechnol Bull. 2017;33:18–25.

[7] Chen ZL, Yin WQ, Liu HT, et al. Review of monitoring petroleum-hydrocarbon contaminated soils with visible and Near-Infrared spectroscopy. Spectrosc Spectr Anal. 2017;37:1723–1727.

[8] Yang ZY, Wang ZD, Yang C, et al. Evaluation of total petroleum hydrocarbons (TPH) measurement methods for assessing oil contamination in soil. Environ Forensics. 2013;14:193–203.

[9] Saari E, Peramaki P, Jalonen J. Measurement uncertainty in the determination of total petroleum hydrocarbons (TPH) in soil by GC-FID. Chemometrics Intell Lab Syst. 2008;92:3–12.

[10] Ping LF, Luo YM, Zhang HB, et al. Distribution of polycyclic aromatic hydrocarbons in thirty typical soil profiles in the Yangtze River Delta region, east China. Environ Pollut. 2007;147:358–365.

[11] Shen HY, Zhou PC, Pan BL. Research on Visible and Near Infrared Spectral-polarimetric properties of soil polluted by crude oil. In: Proceedings of the Annual Conference of the Chinese-Society-for-Optical-Engineering (CSOE) on Applied Optics and Photonics China (AOPC) - Optical Spectroscopy and Imaging, Beijing, China, 2017 Jun 04-06.

[12] Schwartz G, Ben-Dor E, Eshel G. Quantitative analysis of total petroleum hydrocarbons in soils: comparison between reflectance spectroscopy and solvent extraction by 3 certified laboratories. Appl Environ Soil Sci. 2012;1–11. doi:10.1155/2012/751956.

[13] Dematte JAM, Alves MR, Gallo BC, et al. Hyperspectral remote sensing as an alternative to estimate soil attributes. Rev Cienc Agron. 2015;46:223–232.

[14] Ewing J, Oommen T, Jayakumar P, et al. Utilizing hyperspectral remote sensing for soil gradation. Remote Sens. 2020;12:1–13.

[15] Srivastava PK, Gupta M, Singh U, et al. Sensitivity analysis of artificial neural network for chlorophyll prediction using hyperspectral data. Environ Dev Sustain. 2021;23:5504–5519.

[16] Pelta R, Ben-Dor E. An exploratory study on the effect of petroleum hydrocarbon on soils using hyperspectral Longwave Infrared Imagery. Remote Sens. 2019;11:15.

[17] Pabon REC, de Souza CR. Spectroscopic characterization of red latosols contaminated by petroleum-hydrocarbon and empirical model to estimate pollutant content and type. Remote Sens Environ. 2016;175:323–336.

[18] Sun WC, Liu S, Zhang X, et al. Performance of hyperspectral data in predicting and mapping zinc concentration in soil. Sci Total Environ. 2022;824:1–14.

[19] Webster GT, Soriano-Disla JM, Kirk J, et al. Rapid prediction of total petroleum hydrocarbons in soil using a hand-held mid-infrared field instrument. Talanta. 2016;160:410–416.

[20] Hauser A, Ali F, Al-Dosari B, et al. Solvent-free determination of TPH in soil by near-infrared reflectance spectroscopy. Int J Sus Dev Plann. 2013;8:413–421.

[21] Scafutto RDM, de Souza CR, de Oliveira WJ. Hyperspectral remote sensing detection of petroleum hydrocarbons in mixtures with mineral substrates: implications for onshore exploration and monitoring. J Photogramm Remote Sens. 2017;128:146–157.

[22] Yin WQ, Chen ZL, Jiao YW, et al. Research on visible-near infrared spectral characterization of purified soil contaminated with petroleum hydrocarbon and estimation of pollutant content. Spectrosc Spectr Anal. 2017;37:3924–3931.

[23] Chakraborty S, Weindorf DC, Zhu YD, et al. Spectral reflectance variability from soil physicochemical properties in oil contaminated soils. Geoderma. 2012;177:80–89.

[24] Hamidisepehr A, Sama MP, Hamidisepehr A. Moisture content classification of soil and stalk residue samples from spectral data using machine learning. Trans ASABE. 2019;62:1–8.

[25] Gruszczynski S, Gruszczynski W. Supporting soil and land assessment with machine learning models using the Vis-NIR spectral response. Geoderma. 2022;405:1–17.

[26] Gholizadeh A, Saberion M, Ben-Dor E, et al. Monitoring of selected soil contaminants using proximal and remote sensing techniques: background, state-of-the-art and future perspectives. Crit Rev Environ Sci Technol. 2018;48:243–278.

[27] Peng YP, Zhao L, Hu YM, et al. Prediction of soil nutrient contents using visible and Near-Infrared reflectance spectroscopy. ISPRS Int J Geo. 2019;8:437.

[28] Bi QF, Goodman KE, Kaminsky J, et al. What is machine learning? A primer for the epidemiologist. Am J Epidemiol. 2019;188:2222–2239.

[29] Wang S, Guan KY, Zhang CH, et al. Using soil library hyperspectral reflectance and machine learning to predict soil organic carbon: assessing potential of airborne and spaceborne optical soil sensing. Remote Sens Environ. 2022;271:1–15.

[30] Song YQ, Zhao X, Su HY, et al. Predicting spatial variations in soil nutrients with hyperspectral remote sensing at regional scale. Sensors. 2018;18:1–18.

[31] Xue Y, Zou B, Wen YM, et al. Hyperspectral inversion of chromium content in soil using support vector machine combined with lab and field spectra. Sustainability. 2020;12:1–16.
[32] Wei LF, Yuan ZR, Zhong YF, et al. An improved gradient boosting regression tree estimation model for soil heavy metal (Arsenic) pollution monitoring using hyperspectral remote sensing. Appl Sci. 2019;9:1–16.

[33] Kim N, Lee Y-W. Machine learning approaches to corn yield estimation using satellite images and climate data: a case of Iowa State. J Korean Soc Survey Geodesy Photograph Cartogr. 2016;34:383–390.

[34] Jia X, O’Connor D, Shi Z, et al. VIRS based detection in combination with machine learning for mapping soil pollution. Environ Pollut. 2021;268:115845.

[35] Sagi O, Rokach L. Ensemble learning: a survey. Wiley Interdiscip Rev. 2018;8:1–18.

[36] Ang KLM, Seng JKP. Big data and machine learning with hyperspectral information in agriculture. IEEE Access. 2021;9:36699–36718.

[37] Nguyen KA, Chen W, Lin B, et al. Comparison of ensemble machine learning methods for soil erosion Pin Measurements. ISPRS Int J Geo. 2021;10:42.

[38] Webb GI, Zheng ZJ. Multistrategy ensemble learning: reducing error by combining ensemble learning techniques. IEEE Trans Knowl Data Eng. 2004;16:980–991.

[39] Taghizadeh-Mehrjardi R, Schmidt K, Amirian-Chakan A, et al. Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by Stacking machine learning models and rescanning covariate space. Remote Sens. 2020;12:1–26.

[40] Wei LF, Yuan ZR, Wang ZX, et al. Hyperspectral inversion of soil organic matter content based on a combined spectral index model. Sensors. 2020;20:1–17.

[41] Rao H, Shi XZ, Rodrigue AK, et al. Feature selection based on artificial bee colony and gradient boosting decision tree. Appl Soft Comput. 2019;74:634–642.

[42] Rajaratnam B, Roberts S, Sparks D, et al. Lasso regression: estimation and shrinkage via the limit of Gibbs sampling. J R Stat Soc Ser B. 2016;78:153–174.

[43] Tibshirani RJ. The lasso problem and uniqueness. Electron J Stat. 2013;7:1456–1490.

[44] Meinshausen N, Buhlmann P. Stability selection. J R Stat Soc Ser B. 2010;72:417–473.

[45] Meier L, van de Geer S, Buhlmann P. High-dimensional additive modeling. Ann Stat. 2009;37:3779–3821.

[46] Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B. 1996;58:267–288.

[47] Tibshirani RJ, Taylor J. The solution path of the generalized LASSO. Ann Stat. 2011;39:1335–1371.

[48] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learn. 2006;63:3–42.

[49] Simm J, Magrans De Abri I, Sugiyama M. Tree-based ensemble multi-task learning method for classification and regression. IEICE Trans Inf Syst. 2014;E97D:1677–1681.

[50] Maree R, Geurts P, Piater J, et al. Random subwindows for robust image classification. In Proceedings of the Conference on Computer Vision and Pattern Recognition, San Diego, CA, 2005 Jun 20-25; pp.34–40.

[51] Paul A, Furmanchuk A, Liao WK, et al. Property prediction of organic donor molecules for photovoltaic applications using extremely randomized trees. Mol Inf. 2019;38:1–11.

[52] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189–1232.

[53] Si S, Zhang H, Keerthi SS, et al. Gradient boosted decision trees for high dimensional sparse output. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017 Aug 06-11.

[54] Wang JD, Li P, Ran R, et al. A short-term photovoltaic power prediction model based on the gradient boost decision tree. Appl Sci Basel. 2018;8:1–14.

[55] Lin N, Liu HQ, Yang JJ, et al. Hyperspectral estimation of soil composition contents based on kernel principal component analysis and machine learning model. J Appl Remote Sens. 2020;14:19.

[56] Patel AK, Ghosh JK. Soil fertility status assessment using hyperspectral remote sensing. In Proceedings of the Conference on Remote Sensing for Agriculture, Ecosystems, and Hydrology XXI held at SPIE Remote Sensing, Strasbourg, FRANCE, 2019 Sep 09-11.

[57] Ou D, Tan K, Lai J, et al. Semi-supervised DNN regression on airborne hyperspectral imagery for improved spatial soil properties prediction. Geoderma. 2021;385:114875.

[58] Howley T, Madden MG, O’Connell ML, et al. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. Knowledge Based Syst. 2006;19:363–370.

[59] Chen CQ, Jiang QQ, Zhang ZC, et al. Hyperspectral inversion of petroleum hydrocarbon contents in soil based on continuum removal and wavelet packet decomposition. Sustainability. 2020;12:13.

[60] Peng XT, Shi TZ, Song AH, et al. Estimating soil organic carbon using VIS/NIR spectroscopy with SVMR and SPA methods. Remote Sens. 2014;6:2699–2717.

[61] Uddin MP, Al Mamun MA, Hossain MA. PCA-based feature reduction for hyperspectral remote sensing image classification. IETE Tech Rev. 2021;38:377–396.

[62] Zheng KY, Li QQ, Wang JJ, et al. Stability competitive adaptive rewighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra. Chemometrics Intell Lab Syst. 2012;112:48–54.

[63] Liu LF, Ji M, Dong YY, et al. Quantitative retrieval of organic soil properties from Visible Near-Infrared Shortwave Infrared (Vis-NIR-SWIR) spectroscopy using fractional-based feature extraction. Remote Sens. 2016;8:18.

[64] Clingensmith CM, Grunwald S. Predicting soil properties and interpreting Vis-NIR models from across continental United States. Sensors. 2022;22:17.