ABSTRACT

Hyperspectral data offers a powerful tool for predicting soil heavy metal contamination due to its high spectral resolution and many continuous bands. However, band selection is the prerequisite to accurately invert and predict soil heavy metal concentration by hyperspectral data. In this paper, 181 soil samples were collected from the suburb of Nanjing City, and their reflectance spectra and soil lead concentrations were measured in the laboratory. Based on these dataset, we compare Least Angle Regression, which is a modest forward choose method, and least squares regression and partial least squares regression based on genetic algorithm. As a result, regression with band selection has better accuracy than those without band selection. Although both Least Angle Regression and partial least squares regression with genetic algorithm can reach 70% training accuracy, the latter based on genetic algorithm is better, because it can reach a larger solution space. At last, we conclude that partial least squares regression is a good choice for the soil lead content retrieval by hyperspectral remote sensing data, and genetic algorithm can improve the retrieval by band selection promisingly. Bands centered around 838nm, 1930nm and 2148nm are sensitive for soil lead content.

Keywords: heavy metal inversion, partial least square regression, least square regression, genetic algorithm, band selection

1. INTRODUCTION

With the rapid industrialization and urbanization, the continuous emission and accumulation of waste water and solid waste which containing heavy metal have caused increasingly environmental problems. Once soil is polluted by the heavy metal, it will cause serious harm to environmental and food security. Meanwhile, its cleaning process is extremely expensive. Therefore, heavy metal pollution has become one of the serious potential threat to environment and public health.

Geochemical remote sensing technologies provide a fast, macro way to access to the Earth's surface chemical information, thereby has been widely used in many fields such as environmental geochemistry and soil sciences. As a powerful tool for monitoring vegetation stress, hyperspectral remote sensing technique is increasingly being used, directly or indirectly to monitor the status of heavy metal pollution[1,2,3]. The mechanism for hyperspectral remote sensing of soil heavy metal lies in two facts. Hyperspectral remote sensing with its acquisition of continuous, fine

* zx@irsa.ac.cn
reflectance spectrum allows itself to capture the spectral absorption characteristics of heavy metals. In addition, soil heavy metals are often associated with some of the soil spectrum active ingredients (such as iron oxides, organic matter, etc.). With the aid of internal relevance of this soil component, it is possible to predict/retrieve soil heavy metals by reflectance spectroscopy.

However, due to the numerous and high-correlated bands of hyperspectral data, band selection or band transformation is the prerequisite for heavy metal retrieval by hyperspectral data. The typical band selection methods include forward selection, backward elimination and stepwise regression. Lasso and Forward Stagewise are two more popular methods for the present, while the Least Angle Regression (LARS) is one way to implement them[4]. In this study, Lasso implemented by LARS is selected as one of the band selection methods.

Partial Least Square regression (PLS) has been widely used in soil spectrum analysis[1,5,6] and has begun its application in assessment of contaminant elements in soils[3]. As PLS fit to deal with serious collinearity, it’s suitable for the case that sample number is less than the variable number. Meanwhile, it takes into account both the independent variable and the dependent variable during the band selection process, so PLS is often effective than PCA. However, PLS is not able to remove the noise band or the Irrelevant variables at the variable projection stage. Thus, the useless information may affect model to reveal a more reasonable relationship. Therefore, it will be helpful to perform band selection while PLS regression. Although LARS considers combination of variables, but only considers the combinations of the new selected variable and the previously selected variables. Therefore, the scope of the solution space is still reduced. The possible way to expand the scope of the solution space search is a random search, in which genetic algorithm is an effective method[7,8]. Researches have shown that genetic algorithm can aid partial regression analysis in the effective band selection [9].

This study focuses on the comparison of LARS algorithm, genetic algorithm based least square algorithm, genetic algorithm based PLS algorithm, and assesses their application in soil lead content retrieval by using hyperspectral remote sensing data. In section 2, we expatiates the specific parameters setting of LARS, genetic algorithm and PLS regression model for soil heavy metal content retrieval by hyperspectral remote sensing data. In section 3, the models are applied in soil lead retrieval in Jiangning District of Nanjing Suburb, and finally the retrieval results of different algorithms are compared and analyzed.

2. MODEL AND PARAMETER SETTINGS

2.1 LARS

LARS is a forward selection model, but it’s different from traditional forward selection models. The traditional forward selection models select the most relevant independent ones in all the independent variables to make the regression, so they belong to excessive greedy mode. Lasso is essentially a less greedy regression model, which retains part of correlation to improve the overall fitting results. It adopts small-step fitting instead of the largest step fitting. As a fast algorithm of small-step fitting, LARS can be employed to achieve Lasso regression.
To start LARS, set all the coefficients as 0, then select the highest-correlation independent variable $X_j$ to fit. Suppose the fitting step is $\gamma$, then the correlation between the current residual $y - \beta X_j$ and $X_j$ is equal to the correlation of the independent variable to be selected at the next step and the residual. This step determination method makes the next step fitting go along the direction which has equal angle with each selected variable. The concrete step is: Set initial fitting vector $\mu_0 = 0$, followed by $m$ step LARS fitting. Assume the current fitting vector is $\hat{\mu}_s$, then $\hat{C} = X(y - \hat{\mu} A)$ is vector composed by the correlation coefficients between residual and independent variable $X$. Choose the highest-correlation independent variables and calculate the equal-angle vector $\mu_s$ and the step $\hat{\gamma}$, then the fitting vector of next step will be

$$\hat{\mu}_{s+1} = \hat{\mu}_s + \hat{\gamma} \mu_s.$$  

2.2 Genetic algorithm based multiple linear regression

Genetic Algorithm is a random search method which references the biological evolution law. In the genetic algorithm, the solution of optimization problem is called as the individual, which is expressed as a list of parameters and called chromosome or gene sequence. In this study, the length of chromosome is number of the selected spectral bands. Firstly, select randomly part of all the bands to generate a certain number of individuals which is also known as population. The population size is required to some extent larger than the number of selected bands. However, in order to guarantee the convergence rate, the population size can not be too large. Here the population size is set to 100. Secondly, adaptive evaluation is conducted for each individual. Here the least square fitting is employed on the selected bands, then CV value calculated by Leave One Out Cross Validation is used as the adaptability evaluation indicator. Thirdly, retain two optimal individuals. The remaining 80% of the individuals exchange. In exchange process, select half bands respectively from the two parent individuals to generate a new generation of individual. The last remaining individuals mutate. Set the basic mutation step to 100(bands). If the mutation individuals have good adaptability, then lessen the step. Otherwise, increase the step. So diversity exists in the last group, but the search scope is ensured large enough which is more beneficial to search for the best offspring.

2.3 Genetic algorithm based PLS fitting

Compared with the traditional multiple linear regression models, PLS can conduct regression modeling in terms of serious multiple correlation existing in independent variables. In addition, it allows the number of sample points less than the number of variables. For the heavy metal content retrieval, there is only one dependent variable, denoted by $y$. The hyperspectral band independent variable is denoted by $X=(x_1, X_2, \ldots, x_p)$. Components of $t$ and $u$ is produced by PLR algorithm from the independent variable metrix $X$ and dependent variable $y$ both composed by $n$ observations. The components $t$ and $u$ should contain as much as possible the Variation of respective data, and reach the maximum correlation. After the first component is extracted, PLS are conducted on the regression of $X$ on $t$ and regression of $Y$ on $t$. If the regression equation reaches its satisfactory accuracy, then the algorithm terminates. Otherwise, the new round of component extraction is conducted on the remained information after $X$ is explained by $t$ as well as the remained
information after y is explained by t. This loop is executed until a more satisfactory accuracy is gained.

Here we employ the genetic algorithm similar with the previous section, that is, same selection, exchange and mutation method, but substitute Adaptability evaluation function with PLS. Similarly use the CV value by leave one out cross-examination to determine the adaptability and search the optimal band combination.

3. MATERIALS AND VALIDATION

3.1 Data and processing
A total of 120 soil samples provided by Nanjing university were collected from the Jiangning district of Nanjing suburb. The sample spot distribution was shown in figure 1. The sampling density is 2km2 for each sample and the sampling depth is 0~20cm. On each sample location, 3 sub-samples were collected by snake-shape sampling method, and mix the sub-samples into one sample. The Soil samples were dried naturally in the laboratory and were ground and passed the 100-mesh nylon sieves. The final soil sample each were equally divided into two parts, one for chemical analysis by a typical method called ICP-AES, the other part for spectral measurements by the Lambda900 spectrometer (spectral resolution: 2nm; wavelength range: 400-2500nm).

The two spectral ranges of 840-900nm and 2300-2500nm were removed from the whole spectra due to the low signal-to-noise, hence 922 effective bands were used for further analysis. Some of the removed spectral bands were made up by the parabolic interpolation. Outliers were examined on all these data and 2 samples were removed. 118 samples remained for soil lead content modeling and analysis.

3.2 Modeling and validation
21 samples were chosen randomly as Independent validation sample, and the remaining samples were used for modeling. For LARS modeling, cross-examination was conducted by leave one out method, and select 23 spectral bands with the minimum CV values for modeling. The selected bands were shown in figure 2(c). Figure 1(a) showed the corresponding prediction results. Regarding on the number of selected bands of LARS, we directly set the band number to be selected of GA-PLS as 23. In order to demonstrate the significance of band selection, PLS were performed firstly on the overall 922 bands. The number of components extracted by PLS was also determined by leave one out cross-validation. This study used 6 components, so PLS also extracted same number of components to fit. Figure 1(c) showed the fitting result at once GA-PLS fit. To show the band significance, GA-PLS was repeated 50 times, and the frequency of selected bands was shown in figure 2(b).
Figure 1. Inversion result of different methods

(a) LARS, (b) GA-MLR, (c) GA-PLS, (d) PLS
For GA-MLR regression, we tried to select 23 bands, but found that $R^2$ can reach over 90% for training data, but less than 40% for test data. Hence we reduce the selected bands number to 6, the fitting result and frequency of selected band after 50 times repeat were shown in figure 1(b) and figure 2(c) respectively.

Figure 2 and figure 1 showed that, the accuracy of PLS regression after GA band selection was better than that of all-band based least square algorithm, which indicated the significance of band selection for PLS fitting. The result of GA band selection was better than LARS-based Lasso band selection, which was caused by the larger solution space of GA than LARS. GA-PLS was better than GA-MLR, indicating partial square regression could improve fitting accuracies. Band selection results showed that all the selected bands were located either in the absorption vale or reflectance peak of the soil spectral curve. Especially, the bands of 838nm, 1930nm and 2148nm were chosen by all the algorithms, indicating that these bands maybe sensitive to soil lead content.

4. CONCLUSION

This study compared the less greedy forward selection model based LARS algorithm and GA-based least square algorithm as well as GA-based PLS algorithm for soil lead retrieval by hyperspectral remote sensing data. Firstly, the algorithm theory and parameter setting were introduced, then the algorithms were validated by the 120set of soil spectra and their lead content. Result and analysis indicated, the GA-based PLS algorithm (GA-PLS) performed best, which gained the highest training accuracy of 73.2%. GA was superior to LARS on respect of band selection. The spectral bands of 838nm, 1930nm and 2148nm were chosen by all the algorithms. These three bands were located either located...
either in the absorption vale or reflectance peak of the soil spectral curve, hence they could be the sensitive bands with physical meaning for soil lead retrieval by hyperpectral remote sensing data.

As the models built in this study is based on the ideal laboratory measurements of soil reflectance spectra, when applying them to field measurements, airborne or spaceborne hyperspectral remote sensing image, effects of atmosphere, topography and soil roughness as well as mixed pixel should be considered. Nevertheless, the GA-based PLS algorithm can provide a new way to retrieve soil heavy metal content by hyperspectral remote sensing data.

ACKNOWLEDGEMENT

This research was supported by the National High Technology Research and Development Program “Research on identification and quantitative inversion of material composition of lunar surface” and the National Natural Science Foundation of China (40971205) and Free Exploration Basis and Technology Research of SLRSS “Rock and mineral information extraction based on mixed spectrum decomposition if CE-1 IMM image”. We greatly appreciate Dr. Wu Yunzhao of International Institute of Earth System Science, Nanjing University for providing the data in this study.

REFERENCES


Proc. of SPIE Vol. 7831  78311K-7