Predicting Soil Elements Using Hyperspectral Images

Sarita Bha

Abstract

Soil elements are important indicators of soil quality and fertility. However, traditional methods of measuring soil elements are costly, time-consuming, and destructive. Hyperspectral imaging is a promising technique that can provide rapid and non-invasive assessment of soil elements using spectral reflectance. In this project, we aim to predict four soil elements: phosphorus pentoxide (P2O5), potassium (K), magnesium (Mg), and pH using hyperspectral images. We use various statistical features, vegetation and soil indices, and band ratios derived from the hyperspectral images as input variables. We compare four machine learning models: random forest regression, extra trees regression, light gradient boosting machine regression, and k-nearest neighbor regression. We use the out-of-fold predictions of the training data to find the optimal weights for combining the predictions of the four models. We evaluate the performance of the models on the test data using root mean squared error (RMSE). We find that some models perform better on individual soil elements, and that using band ratios and indices improves the models slightly. We also find that using Savitzky-Golay filter to smooth the spectral data.

Introduction

Soil elements are essential for plant growth and soil health. They affect various soil properties such as nutrient availability, water retention, and pH. However, measuring soil elements using conventional methods is expensive, laborious, and destructive. Therefore, there is a need for alternative methods that can provide fast and non-destructive assessment of soil elements.

Hyperspectral imaging is a technique that captures the spectral reflectance of an object in hundreds of narrow and contiguous bands. Hyperspectral imaging can provide detailed information about the chemical and physical characteristics of an object based on its spectral signature. Hyperspectral imaging has been widely used for various applications such as mineral exploration, crop monitoring, and food quality assessment.

In recent years, hyperspectral imaging has also been applied for soil analysis. Hyperspectral imaging can capture the spectral variations caused by different soil elements and their interactions. Several studies have shown that hyperspectral imaging can be used to estimate soil elements such as organic matter, nitrogen, phosphorus, potassium, calcium, magnesium, iron, and pH .

However, hyperspectral imaging also poses some challenges for soil analysis. First, hyperspectral images have high dimensionality and redundancy, which require appropriate feature selection and extraction methods. Second, hyperspectral images are affected by external factors such as illumination, moisture, and noise, which require appropriate preprocessing and calibration methods. Third, hyperspectral images have complex and nonlinear relationships with soil elements, which require appropriate modeling and prediction methods.

In this project, we aim to address these challenges and predict four soil elements: phosphorus pentoxide (P2O5), potassium (K), magnesium (Mg), and pH using hyperspectral images. We use various statistical features, vegetation and soil indices, and band ratios derived from the hyperspectral images as input variables. We compare four machine learning models: random forest regression, extra trees regression, light gradient boosting machine regression, and k-nearest neighbor regression. We use the out-of-fold predictions of the training data to find the optimal weights for combining the predictions of the four models. We evaluate the performance of the models on the test data using root mean squared error (RMSE).

Data

The dataset comprises 2886 patches in total (2 m GSD), of which 1732 patches for training and 1154 patches for testing. The patch size varies (depending on agricultural parcels) and is on average around 60x60 pixels. Each patch contains 150 contiguous hyperspectral bands (462-942) nm, with a spectral resolution of 3.2 nm), which reflects the spectral range of the hyperspectral imaging sensor deployed on-board Intuition-1.

The training dataset is further split in ratio 7:3 for training and validation.

Data Preprocessing

The hyperspectral images were preprocessed using the following steps:

- **Statistical features**: For each soil sample, we calculated the mean, standard deviation, skewness, and kurtosis of the spectral reflectance across all bands. These features capture the dispersion, asymmetry, and peakedness of the spectral data.
- **Vegetation and soil indices**: We calculated several vegetation and soil indices using different combinations of bands. These indices are designed to enhance the spectral contrast between vegetation and soil, and to highlight the spectral features related to soil elements. The indices used in this project are:
    - Normalized Difference Vegetation Index (NDVI): NDVI is a widely used index to measure the greenness and biomass of vegetation.
    - Soil Adjusted Vegetation Index (SAVI): SAVI is a modified version of NDVI that reduces the influence of soil background.
    - Enhanced Vegetation Index (EVI): EVI is a measure of vegetation greenness that is more sensitive to high biomass areas and less affected by atmospheric and soil conditions than the normalized difference vegetation index (NDVI).
- **Band ratios**: We calculated the ratios of different bands using the formula Ri/Rj, where Ri and Rj are the reflectance mean values at band i and band j, respectively (bands being, B, G, R and N-IR).
- **Savitzky-Golay filter**: We applied a Savitzky-Golay filter to smooth the spectral data and reduce the noise. A Savitzky-Golay filter is a polynomial smoothing filter that preserves the shape and features of the spectral data. We used a window size of 3 and a polynomial order of 1 for the filter. This filter is applied only to the mean of the reflectance values.

Modeling

We used four machine learning models to predict the soil elements using the preprocessed hyperspectral data as input variables. The models used in this project are:

a) **Random forest regression**: A random forest is an ensemble of decision trees that are trained on bootstrap samples of the data and use random subsets of features at each split. A random forest regression model predicts the output by averaging the predictions of the individual trees. A random forest regression model can handle high-dimensional and nonlinear data, and can reduce the variance and overfitting of a single decision tree. We used the scikit-learn implementation of random forest regression with the tuned parameters.

b) **Extra trees regression**: An extra trees regression model is similar to a random forest regression model, except that it uses extremely randomized trees that split the nodes randomly rather than using the best split. An extra trees regression model can reduce the bias and variance of a random forest regression model, and can be faster and more robust to noise. We used the scikit-learn implementation of extra trees regression with the tuned parameters.

c) **Light gradient boosting machine regression**: A light gradient boosting machine (LightGBM) regression model is a gradient boosting framework that uses tree-based learning algorithms. A gradient boosting model builds an ensemble of weak learners (usually decision trees) in a sequential manner, where each learner tries to correct the errors of the previous learners. A LightGBM regression model can handle high-dimensional and sparse data, and can be faster and more efficient than other gradient boosting models. We used the LightGBM implementation of gradient boosting regression with the tuned parameters.

d) **K-nearest neighbor regression**: A k-nearest neighbor (KNN) regression model is a non-parametric and lazy learning algorithm that predicts the output based on the similarity of the input to the training data. A KNN regression model finds the k closest neighbors of the input in the feature space, and predicts the output by averaging the outputs of the neighbors. A KNN regression model can handle nonlinear and complex data, but it can be slow and sensitive to noise and outliers. We used the scikit-learn implementation of KNN regression with the tuned parameters.

We used 4-fold cross-validation to train and validate the models on the training data. We used root mean squared error (RMSE) as the evaluation metric. We used the out-of-fold predictions of the training data to find the optimal weights for combining the predictions of the four models using a linear regression model. We then re-trained the four models on the whole training data and made predictions on the test data using the weighted average of the four models. We compared the performance of the individual models and the combined model on the test data using RMSE.

Findings

The RMSE values of the individual models and the combined model on the test data are shown in the table below:

Model	**P2O5**	**K**	**Mg**	**pH**
Random forest	24.306	55.107	37.89	0.242
Extra trees	24.446	54.687	38.042	0.239
LightGBM	24.345	54.864	37.742	0.236
KNN	24.692	57.725	38.982	0.249
Combined	24.143	54.503	37.800	0.239
Baseline(mean)	25.066	59.119	39.252	0.257

The table shows that the combined model has the lowest RMSE for all the soil elements, indicating that it can improve the prediction accuracy by leveraging the strengths of the individual models. The table also shows that some models perform better on individual soil elements. For example, LightGBM has the lowest RMSE for pH, Extra trees has the lowest RMSE for K, and Random forest has the lowest RMSE for Mg. This suggests that different models can capture different spectral features and relationships related to the soil elements.

Summary and Conclusion

In this project, we predicted four soil elements: phosphorus pentoxide (P2O5), potassium (K), magnesium (Mg), and pH using hyperspectral images. We used various statistical features, vegetation and soil indices, and band ratios derived from the hyperspectral images as input variables. We compared four machine learning models: random forest regression, extra trees regression, light gradient boosting machine regression, and k-nearest neighbor regression. We used the out-of-fold predictions of the training data to find the optimal weights for combining the predictions of the four models. We evaluated the performance of the models on the test data using root mean squared error (RMSE).

We found that the combined model has the lowest RMSE for all the soil elements, indicating that it can improve the prediction accuracy by leveraging the strengths of the individual models. We also found that some models perform better on individual soil elements, and that using band ratios and indices improves the models slightly. Most important step in data-processing is smoothing the signal using savgol-filter, using raw reflectance values gives results close and sometimes worse than baseline(mean) model.

Credits

This project is based on the following sources:

[1] A S, Akg A, Bsd B, Nr B. Application of VIS-NIR spectroscopy for estimation of soil organic carbon using different spectral preprocessing techniques and multivariate methods in the middle Indo-Gangetic plains of India—ScienceDirect. Geoderma Regional. 2020. | https://doi.org/10.1016/j.geodrs.2020.e00349

[2] Yuanyuan Shi, Junyu Zhao, Xianchong Song, Zuoyu Qin, Lichao Wu, Huili Wang, Jian Tang. Hyperspectral band selection and modeling of soil organic matter content in a forest using the Ranger algorithm. Remote Sensing of Environment. 2021. | https://doi.org/10.1371/journal.pone.0253385