ESTIMATING SOIL PARAMETERS FROM HYPERSPECTRAL IMAGES USING MACHINE LEARNING

Julius Maina

The author is a research student in data analytics at KCA University in Kenya

October 28th, 2023

Abstract *– This paper presents a machine learning approach for soil parameter estimation, utilizing LightGBM and HistGradientBoosting. The research aims to automate the estimation of crucial soil parameters—potassium (K), phosphorus pentoxide (P2O5), magnesium (Mg), and pH. The adoption of LightGBM and HistGradientBoosting reflects the cutting-edge methodology employed to advance soil analysis and contribute to sustainable agriculture.*

Keywords – Singular Value Decomposition, Wavelet Transform, Gradient Calculation, Fast Fourier Transform, Spectral Curve Filtering, Data Augmentation, In-Orbit Processing, Hyperspectral Imaging.

1. INTRODUCTION

ESA Φ-lab, KP Labs, and QZ Solutions join forces for an innovative farming revolution, blending Earth observation and AI in an ambitious challenge. Our mission: enhance farm sustainability by leveraging recent advances in space tech and AI. Modern agriculture grapples with the dual challenge of affordable food production and eco-friendly practices. Key to this is swift access to soil data for optimized fertilization—a hurdle our challenge tackles head-on.

Traditionally, soil analysis is laborious and limited, relying on in-situ methods that are neither scalable nor efficient. Enter hyperspectral imaging tech from satellites, offering a game-changing alternative. Our mission focused on advancing soil parameter retrieval from airborne hyperspectral data, setting the stage for the groundbreaking Intuition-1 satellite mission. This 6U-class satellite from KP Labs boasts a hyperspectral instrument and AI processing unit—the world's first.

In this challenge, we dived into automatic estimation of crucial soil parameters—potassium (K), phosphorus pentoxide (P2O5), magnesium (Mg), and pH. The synergy of cutting-edge imaging and AI, coupled with in-orbit processing, aims not just for sustainable agriculture but a transformative future for our planet.

2. DATASET

Agricultural datasets often present unique challenges due to their multidimensional nature, requiring specialized preprocessing techniques for effective feature extraction. In this section, we provide an overview of the dataset used for our study and detail the preprocessing steps, including feature engineering, applied to prepare the data for regression tasks related to soil property prediction.

The dataset comprises 3D cubes representing individual agricultural fields, with each cube consisting of two main components: the raw data and a corresponding binary mask. The raw data encodes information about the agricultural field, while the mask indicates regions with meaningful data. Additionally, ground truth labels for soil properties—phosphorus (P), potassium (K), magnesium (Mg), and soil acidity (pH)—are provided for the training set.

The dataset comprised 2886 patches in total (2 m GSD), of which 1732 patches for training and 1154 patches for testing. The patch size varied (depending on agricultural parcels) and was 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.

A training set of 1732 training examples was provided. The examples were hyperspectral image patches with the corresponding ground-truth information. Each masked patch corresponds to a field of interest, as presented in Fig. 1. Ground truth are the soil parameters obtained for the soil samples collected for each field of interest in the process of laboratory analysis, and is represented by a 4-value vector.

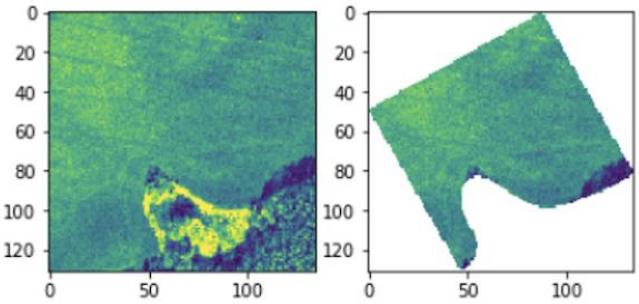

Fig. 1: The masked hyperspectral image patches corresponding to the fields of interest (one masked patch is a single field), and should estimate the four soil parameters based on the available image data.

2.1 Data Loading and Augmentation

We loaded the raw data from the provided 3D cubes using a custom loading function as provided in the starter notebook. However, the function was adjusted for training set to introduce data augmentation by randomly cropping and adding noise to the larger agricultural fields. This augmentation aimed to diversify the training set and improve the model's robustness.

2.2 Data Preprocessing

Singular Value Decomposition (SVD)

We performed SVD on the 3D cubes to extract features capturing spatial relationships within the fields. The singular values served as indicators of the importance of different modes of variability within the data.

Wavelet Transform

We utilized both Symlet and Daubechies wavelets to transform the 3D data. The resulting coefficients provide information about local variations at different scales.

Gradient Calculation

Gradient information was computed to capture changes in data values, providing insights into field topography.

Fast Fourier Transform (FFT)

The FFT was applied to both the original data and the singular values, capturing frequency domain characteristics.

Feature Concatenation

Different feature combinations were created based on the regression task—Random Forest (RF). Features included raw data, wavelet coefficients, gradients, singular values, and FFT results.

Spectral Curve Filtering

To create a spectral curve, we apply a custom filtering function to aggregate pixel values within one band, providing a compact representation of the spectral characteristics of each field.

Data Reshaping and Merging

The next steps involve reshaping the data for better compatibility with downstream tasks. The data was melted, creating a 'Target' column for each soil property, and then merged with the original data frame.

These preprocessing steps ensured that the data is in a suitable format for training machine learning models, allowing for accurate soil property predictions based on the extracted features. The resulting CSV files ('train.csv' and 'test.csv') were utilized in the subsequent models' training and evaluation stages. The shape of the two datasets were 13856 rows × 2407 columns for train and 4616 rows × 2402 columns for test.

3. MODELLING

In this section, we detail the process of developing predictive models for soil property estimation, focusing on parameters such as Phosphorous (P), Potassium (K), Magnesium (Mg), and pH. The modeling approach encompasses feature selection, model training, and evaluation, aiming to provide accurate predictions for each soil property.

3.1 Dataset Preparation

The soil dataset is divided into distinct subsets for each soil property: P, K, Mg, and pH. This partitioning facilitates tailored modeling for individual properties, acknowledging their unique characteristics.

3.2 Feature Selection

To enhance model performance, a customized feature selection process is employed. The Univariate Feature Selection class, leveraging scikit-learn's feature selection methods, is defined and utilized. It allows for the selection of a specific percentage or a fixed number of the most influential features. This adaptive approach ensures that each model is trained on the most relevant information for a given soil property.

3.3 Model Training and Evaluation

A robust K-Fold cross-validation strategy is adopted to train and evaluate the models. With k set to 5, the dataset is partitioned into five subsets, with each serving as a testing set while the remaining four are

utilized for training. This process is repeated, ensuring comprehensive evaluation.

3.4 LightGBM Regression

LightGBM regressor and HistGradientBoosting regressor models are chosen for their efficiency and effectiveness in handling large datasets. Only default parameters are used and there is no any kind of tuning done for simplicity and efficiency. These boosting algorithms provide accurate predictions while maintaining computational efficiency, making them suitable for soil property estimation.

3.5 Evaluation Metric: Root Mean Squared Error (RMSE)

The performance of the models is assessed using the Root Mean Squared Error (RMSE) metric. This metric quantifies the deviation between predicted and actual values, providing a comprehensive measure of the model's predictive accuracy.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2}$$

3.6 Prediction and Submission

The finalized models are applied to predict soil properties on the Zindi test set. Predictions are normalized, and a submission file is generated for evaluation on the Zindi platform.

4. RESULTS

Results for each soil property, including the selected models and corresponding RMSE for each cross-validation fold, are presented.

Table. 1 – Performance of the models for phosphorus pentoxide (P)

	Hist Gradient Boosting	Light-GBM
Fold 1	22.2547	20.6428
Fold 2	15.1639	14.6404
Fold 3	14.4393	14.2951
Fold 4	17.7383	17.7139
Fold 5	13.2555	13.6841
Average	**16.5703**	**16.1953**

Table. 2 – Performance of the models for potassium(K)

	Hist Gradient Boosting	Light-GBM
Fold 1	33.5658	33.5101
Fold 2	40.9432	40.6872
Fold 3	27.4212	26.485
Fold 4	39.9622	40.5385
Fold 5	35.598	36.0465
Average	**35.4981**	**35.4535**

Table. 3 – Performance of the models for magnesium (Mg)

	Hist Gradient Boosting	Light-GBM
Fold 1	29.8413	28.4277
Fold 2	14.6883	14.4926
Fold 3	22.7145	23.2754
Fold 4	18.4412	19.3031
Fold 5	15.2905	15.3049
Average	**20.1952**	**20.1607**

Table. 4 – Performance of the models for pH

	Hist Gradient Boosting	Light-GBM
Fold 1	0.16585	0.144834
Fold 2	0.142672	0.142109
Fold 3	0.131732	0.125057
Fold 4	0.141275	0.134002
Fold 5	0.135748	0.133562
Average	**0.143455**	**0.135913**

This is how the models performed on the unseen test data from Zindi.

Model	Public Leaderboard	Private Leaderboard
Hist Gradient Boosting	0.267040451	0.273326801
LightGBM	0.265251639	0.272199595
Ensemble	**0.262584911**	**0.269867471**

5. DISCUSSIONS

The results highlight the models' varying performances across different soil properties, shedding light on their strengths and weaknesses in predicting phosphorus pentoxide (P), potassium (K), magnesium (Mg), and pH.

Phosphorus Pentoxide (P) and Potassium (K): Both Hist Gradient Boosting and Light-GBM exhibit comparable performances, with average RMSE values around 16 for P and 35 for K. While the models consistently perform well across folds, the relatively wide range of RMSE values suggests some variability in predicting these soil properties.

Magnesium (Mg) and pH: The models demonstrate their adaptability in predicting different soil properties, showcasing average RMSE values around 20 for Mg and exceptionally low values around 0.14 for pH. These findings indicate a strong predictive capacity for magnesium and pH levels, particularly notable in the tight range of RMSE values across cross-validation folds.

Ensemble Performance: The ensemble model outperforms individual models on the public and private leaderboards, underscoring the effectiveness of combining the strengths of Hist Gradient Boosting and Light-GBM. This ensemble approach potentially mitigates individual model shortcomings, leading to improved overall predictive performance.

The models' performance on the unseen test data from Zindi, as indicated by the leaderboard RMSE values, aligns with their performance in the cross-validation setting. This consistency is reassuring, suggesting that the models generalize well to new data.

6. CONCLUSION

The study demonstrates the viability of employing machine learning models, particularly Hist Gradient Boosting and Light-GBM, for predicting soil properties. The models exhibit notable proficiency in predicting magnesium and pH levels. The ensemble model emerges as a robust choice, showcasing its superiority on the leaderboards. However, further exploration into the models' interpretability and potential areas of improvement could enhance the applicability of these findings in real-world agricultural scenarios. Overall, the study contributes valuable insights into leveraging machine learning for soil property prediction, with implications for optimizing agricultural practices and resource management.

REFERENCES

1. Relevant work:

   https://platform.ai4eo.eu/seeing-beyond-the-visible

2. Relevant github repositories: hyperspectral-image-classification · GitHub Topics.

3. This problem statement work can be found on this github repository:

   ITU-GeoAI-Challenge/Estimating-Soil-Parameters-from-Hyperspectral-Images-by-ITU-2023: This solution by username JuliusFx was for the challenge "Estimating-Soil-Parameters-from-Hyperspectral-Images" by ITU in 2023 and which was hosted on Zindi competition platform (github.com)