GeoAI Challenge - Estimating Soil Parameters from Hyperspectral Images

5th Place
Technical Report

By

Jacob Ojumu
(Zindi Username : Enigmatic)
(B. Sc Chemical Engineering)

Introduction

ESA Φ-lab, in collaboration with KP Labs and partner QZ Solutions, has introduced an innovative initiative set to transform the landscape of farming by leveraging in-orbit processing. Addressing the pressing need for enhancing agricultural management practices and ensuring farm sustainability, the project incorporates recent advancements in Earth observation and artificial intelligence. This not only assists farmers in meeting the challenge of producing cost-effective food but also marks a significant stride towards eco-friendly agriculture.

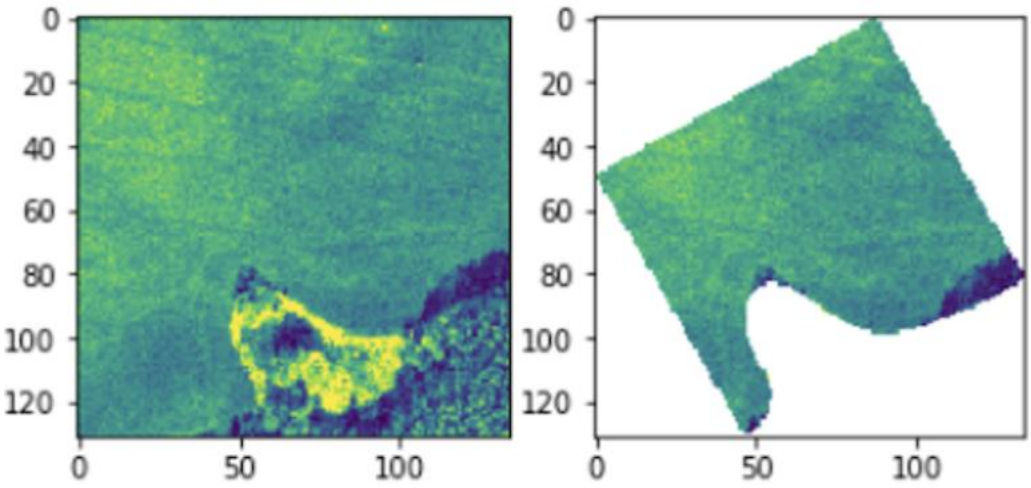
This challenge proposes the utilization of cutting-edge airborne and satellite hyperspectral imaging technology to foster sustainable agriculture, contributing to a more environmentally conscious future. The primary goal is to advance soil parameter retrieval from hyperspectral data in preparation for the upcoming Intuition-1 mission. This challenge's focus is on the automatic estimation of specific soil parameters, namely potassium (K), phosphorus pentoxide (P2O5), magnesium (Mg) and pH, heralding a significant leap in agricultural innovation.

Data

The goal of this challenge is to automate the accurate estimation of specific soil parameters—potassium (K), phosphorus pentoxide (P2O5), magnesium (Mg), and pH—by extracting information from airborne hyperspectral images taken over undisclosed agricultural locations in Poland. The precision of parameter estimation is crucial for practical application in real-world scenarios.

The dataset consists of a total of 2886 patches (2 m Ground Sampling Distance), with 1732 patches allocated for training and 1154 patches for testing. Patch sizes vary based on agricultural parcels and have an average dimension of around 60x60 pixels. Each patch encompasses 150 contiguous hyperspectral bands (462-942 nm, with a spectral resolution of 3.2 nm), mirroring the capabilities of the hyperspectral imaging sensor on Intuition-1.

Each masked patch corresponds to a specific field of interest, as illustrated in the figure 1 below. Ground truth encompasses soil parameters obtained through laboratory analysis of soil samples collected for each field of interest and is represented by a 4-value vector.



Data Processing and Feature engineering:

The data consisted of 150 unique columns which represent information at different wavelengths. Due to the closeness of the values, I created a function to calculate the average columnwise. For example, averaging of 2 implies that 1 and 2 columns were averaged then 3 and 4 were also averaged, 5 and 6 were also averaged and it continues to the last two columns. I did an averaging for two consecutive columns, three consecutive columns, five consecutive columns, seven consecutive columns and 10 consecutive columns. I stacked the resulting averaging arrays on each other and also stacked it with the original data.

Standardization of the data is also important, therefore I used the standard scaler to perform the standardization. This improved the results of the models. Also, I explored the polynomial features in sklearn to amplify the data. I used a degree of 2 and at this point, the size of the train data was about 36856 columns. This was really massive but to reduce the curse of dimensionality, I used the Principal component analysis(PCA) to reduce the dimension of the data.

Model Development

Since the elements that were considered in the competition were P, K , Mg and pH, I created 4 different files housing the preprocessing and modelling for each element. They follow similar modelling techniques but I tweaked them because each element had peculiar characteristics and generalizing the model for all the elements wouldn't have been the best approach.

In the modelling, I did a 5 fold CV to experiment with the data and made useful predictions. Within the folds, I used PCA to reduce the dimension, I tested several dimensions but I found that the best were within 60 to 120. I also employed the use of the catboost model during my modelling. Catboost outperformed all other algorithms, hence I used catboost for my final submissions.

Model Evaluation

The root mean squared error (RMSE) was calculated upon training each fold, I tested the hold out set against the ground truth. Before calculating the rmse value, the predictions were first divided by their mean in the training data. Overall, the validation rmse was very close to the predictions after the competition ended.

Insights:

Polynomial features is an excellent way of populating data when one is out of ideas on feature engineering steps, it proved very useful in this competition and it should be implement in future models.