SOIL NUTRIENTS ANALYSIS USING HYPERSPECTRAL IMAGE PROCESSING AND PERFORMANCE ANALYSIS OF REGRESSION MODELS

CSD300 : PROJECT-1
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, IIIT KOTA

Submitted by

Vandan Jai Rahi | Dhanish Kumar | Abhinav Jain
2020kucp1036 | 2020kucp1040 | 2020kucp1069

Submitted to

Dr. Gyan Singh Yadav (Sir)

Date of Submission - May 24, 2023

Abstract

Soil nutrient analysis plays a crucial role in agricultural practices, environmental monitoring, and land management. Traditional soil sampling and laboratory-based testing methods are time-consuming and labor-intensive. However, recent advancements in remote sensing technologies, particularly hyperspectral imaging, offer a promising solution for efficient and non-destructive soil nutrient analysis. This study explores the integration of hyperspectral image processing techniques with Principal Component Analysis (PCA) and various regression models to estimate soil nutrient concentration. By using the spectral information captured by hyperspectral sensors, combined with machine learning algorithms, this approach provides a valuable tool for determining and monitoring soil nutrient.

Introduction

Soil testing is a procedure used to analyze the soil sample. It is prominently used to determine nutrients present in soil and their features such as pH level and chemistry. The proper soil testing can produce useful estimation of abundance or lack of these nutrients, which may or may not affect the applicability and the efficiency of land utility [1]. As a result the farmers and landowners are provided with a measure to estimate the amount of essential supplements needed in the soil which further enhance the soil utility [2]. The Traditional methods for soil testing are expensive and time consuming some agricultural practitioners cannot afford regular testing. Thus, more affordable alternatives are required and these can be achieved through the use of new technologies like remote sensing techniques [3]. There are broadly three classes of remote sensing techniques which differ from each other due to the instruments and techniques used to gather data of soil samples. Before talking about those classes, let's take a look at the three categories of image data. The digital images that are captured by our cameras consist of three channels of electromagnetic spectrum namely RGB stands for red green blue wavelengths. The MSI stands for Multispectral images that capture a broad number of electromagnetic spectrum in the form of bands that are non-continuous i.e., the bands are not continuous over the spectrum. On the other hand, HSI stands for Hyperspectral images that capture continuous bands over the electromagnetic spectrum (Figure-1). Therefore, it can be seen that from more bands more information can be gathered about the target that is being captured. In this work, the same relation is established among hyperspectral images captured by remote sensing satellites and soil survey reports that provide the concentration data of various nutrients present in the area of survey. Spectral reflectance curve, which is a curve between all the bands in the range of spectrum and reflectance value at each of

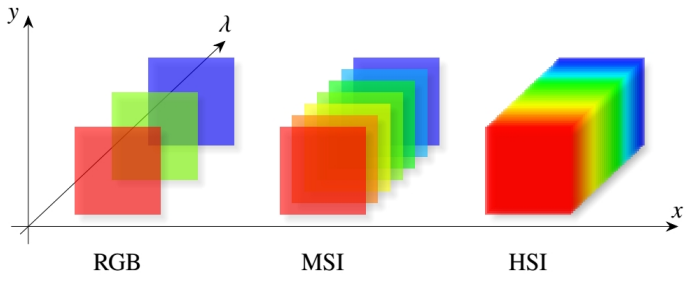


Figure-1: Various imaging techniques and difference in number of bands it captures

those bands is used as a major feature of this study's methodology. Use of various preprocessing steps to correct the hyperspectral images before feature extraction is done using image processing software. Further, the soil survey reports of the area of survey are used to infer the concentration of target nutrients in that region. Later, various regression models are applied on this dataset to do predictions on unseen data.

Literature Review

After thorough literature review of several papers, we found that there are basically three broad classes of remote sensing techniques that were used by researchers for their work. These techniques differ on the basis of the source of hyperspectral data. The first technique is to use laboratory or portable spectroradiometers to get the spectral information of the soil sample. Second technique is to use satellite hyperspectral data of the region of interest. Third and last, to use airborne vehicles attached with hyperspectral sensors to scan the area of interest and get the spectral information. Researchers have developed advanced hyperspectral imaging systems that capture high-resolution soil images across numerous spectral bands. These hyperspectral images along with the actual nutrient concentrations are fed into various machine learning models (PLSR, SVM) to accurately predict the nutrients. Laboratory techniques use spectroradiometer along with various spectral unmixing techniques for identification and quantification of different soil components and nutrients. Researchers have explored various feature selection and dimensionality reduction techniques to improve the efficiency and performance of soil nutrient analysis. These methods aim to identify the most informative spectral bands to reduce complexity and improve prediction accuracy. Overall, hyperspectral image processing techniques have demonstrated great potential in soil nutrient analysis. They offer non-destructive, rapid assessment of soil nutrient content, enabling precise nutrient management and fostering sustainable agricultural practices. Ongoing research aims to further refine these techniques and develop robust models for accurate and efficient soil nutrient estimation.

Table-1: Showing the compilation of literature reviews related to our study, features like data source, preprocessing techniques, models used for prediction, results and literature references are shown in the table below

Literature	Data Source	Techniques	Models	Results
A.S Reis [4]	- Spectroradiometer	- PCA - LDA	- PLSR	R.M.S.E = 3.44
Lixin Lin [5]	- Spectroradiometer	- FDR	- LCMCS - LCM - CS - PLSR	R.M.S.E = 0.898 R.M.S.E = 1.191 R.M.S.E = 1.147 R.M.S.E = 1.373
Yiping peng [6]	- Spectroradiometer - Flame Photometer	- LASSO - GBDT - PCC	- MLR - Ridge Regression - SVM - GABP	R.M.S.E = 0.47 R.M.S.E = 0.44 R.M.S.E = 0.41 R.M.S.E = 0.30 [only for P]
Jingjing Ma [7]	- Spectroradiometer	- Savitzky - Golay - Wavelet Domain Denoising - MSC - Spectral derivative analysis - PCA	- BP - RBF (NN) - ELM - SVR - PLSR	R.M.S.E = 0.875 R.M.S.E = 1.092 R.M.S.E = 1.041 R.M.S.E = 0.758 R.M.S.E = 1.117
Naveen J.P. Anne [8]	Hyperion Satellite	- Atmospheric Corr. - Noise reduction - Geometric Corr.	- Single band - SR - NDI - PLSR	R.M.S.E = 2.87 R.M.S.E = 2.5 R.M.S.E = 2.53 R.M.S.E = 2.58
Hengliang Guo [9]	ZY1-O2D Satellite	- CC - LASSO	- MLR - PLSR - RF	R.M.S.E = 1.925 R.M.S.E = 1.704 R.M.S.E = 0.721

Methodology

This section is divided into four subsections. In the first subsection, we have provided a brief description about the sources from where the data is compiled for this study and some background information regarding the same. In the second subsection, we have mentioned the process of data preparation from the source data. Later preprocessing and the model selection process, refer (Appendix-1).

Data Sources

The region of interest is located in the south of Bagru and north of Jaipur-Tonk border, named as Phagi. The geochemical data of this region was provided by the Geological Survey of India, Jaipur, Rajasthan. In total 182 composite samples were taken by GSI to test their soil chemistry. XRF stands for X-Ray Fluorescence technique used to estimate the concentration of various nutrients (in %). See (Figure-2) the region is divided into small squares with each square covering an area of 1 square km. The numbers ranging from 001 to 182 on the toposheet represent composite samples that are prepared by taking and combining samples from 4 adjacent squares. Therefore, each composite sample represents the soil in a 2 square km area.

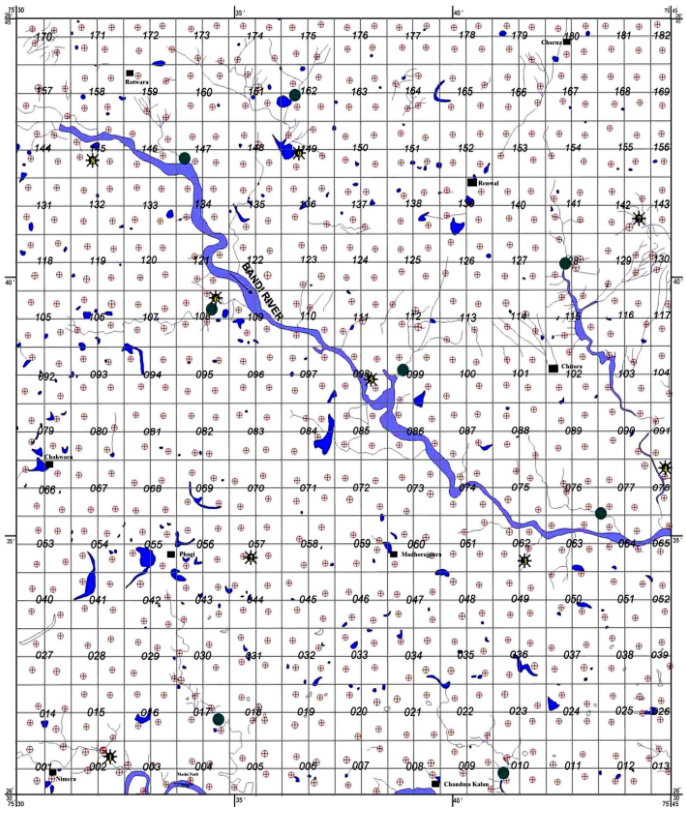


Figure-2: Toposheet of the survey region phagi showing 182 composite sample numbers on grid with 728 sub samples with red marker with geographical coordinates.

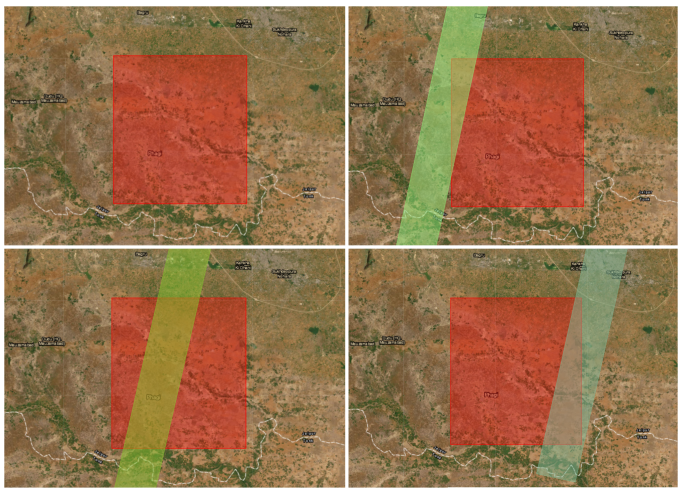


Figure-3: The three hyperspectral strips (in green) cover some area of the survey region (in red).

The Hyperspectral images of the same region were taken from the USGS earth explorers website. The three parallel strips on this region denote the hyperspectral data of this region (Figure-3). The strips were already processed for atmospheric correction and radiometric correction along with geographical referencing with actual coordinates on earth. These hyperspectral images are taken from EO-1 Hyperion satellite (Figure-4) with spectral resolution of 10 nm and spatial resolution of 30 meters captures in 242 spectral bands in the range of 0.4 µm to 2.5 µm i.e., from visible to short-wave infrared region.

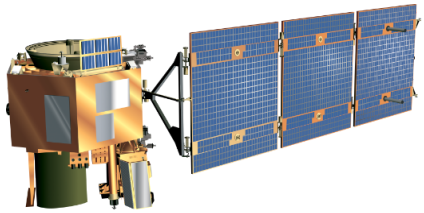


Figure-4: EO-1 satellite (Earth observation) with hyperion imager attached on board that captures the hyperspectral images.

Dataset Preparation

The software used to process the hyperspectral data is ENVI 5.3. Since our images were already corrected for atmospheric interference and radiometric calibration along with geo-referencing, the next step in the processing is to convert the radiance value denoted by the pixel of the hyperspectral image to reflectance using the radiance to reflectance conversion factor given in the meta data of the hyperspectral image. As we already mentioned, each composite sample is made of 4 subsamples taken and combined from adjacent areas. The coordinates of these four sub samples were taken from the soil survey report of

GSI and then using ENVI the pixels that were falling on the same coordinates were selected. The wavelength vs reflectance curve (Figure-5) of those four pixels were generated using the statistical techniques provided by the ENVI. Finally the wavelength vs reflectance curve for the composite sample was generated by taking the mean reflectance value of these four pixels. This mean reflectance curve represents the spectral information of the composite sample. The same process is repeated for all composite samples and the spectral data for all the samples coming under hyperspectral images were generated. The geochemical data from the report for our selected nutrients were taken and the concentration data for all composite samples is also generated. Therefore, our final dataset is the combination of both of these features that contains the sample name, all bands and concentration of the target nutrients for that sample. The data points of features of 242 bands represent the reflectance value at each band for all the samples coming under the hyperspectral and toposheet overlapping region and data points of features of target nutrients represent concentration (in %) for the same sample.

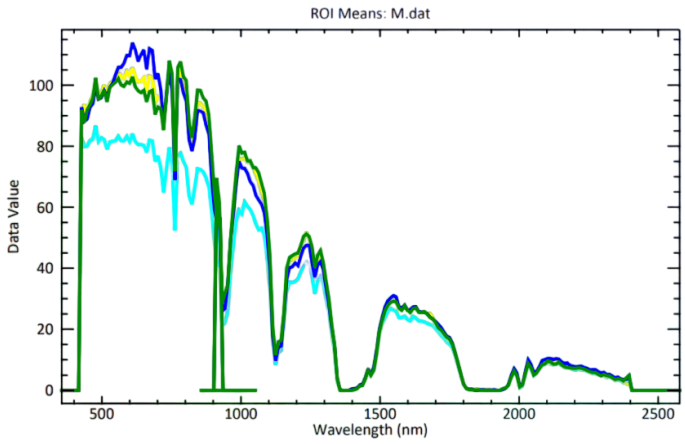


Figure-5: Wavelength vs Reflectance graph of one of the initial samples, four different curves represent reflectance of 4 subsamples using which the composite sample is prepared.

Spectral Information Preprocessing

Principal Component Analysis, this is a dimensionality reduction technique often used to reduce the feature space of a dataset while retaining most of the information as of original feature space. The optimal number of components are selected and then projected on the original feature space to get the new reduced feature space dataset. This reduces the computation time and enhances the performance of our model. Since our dataset contains data values for 242 bands, there is a presence of bands that have no information. Therefore, optimal bands are selected using PCA. Before applying PCA, the bands with

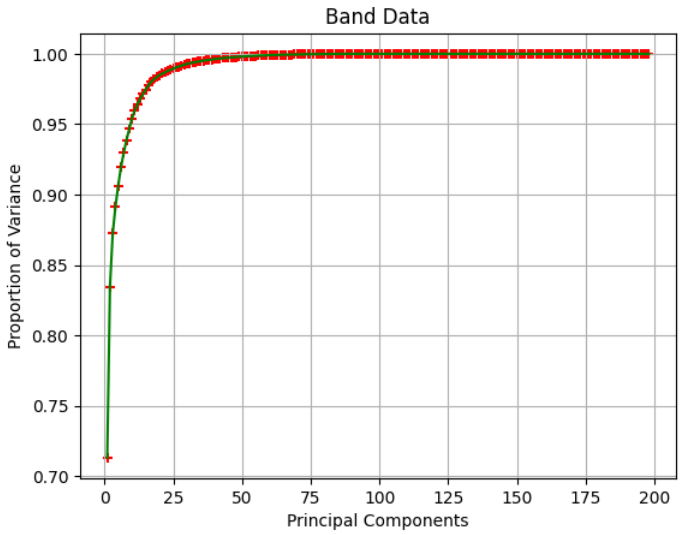


Figure-6: Proportion of Variance vs Principal Components graph showing principal components and respective variance in the data they cover.

0 reflectance are removed and then the reflectance in the rest of the bands are normalized. The first step in PCA is to find the optimal number of principal components that can summarize around 95% variance in the dataset. (Figure-6) It can be seen that around 10 Principal Components cover 95% variance of our dataset. Therefore, the required number of components are selected. Second step is to project those components on the actual data so that the number of features can be reduced to the number of optimal components. Now, this new dataset works as a replacement for the original dataset to train the model.

Model Selection and Performance Analysis

The particular model that is used for regression is SVR stands for Support Vector Regressor. It is a type of SVM that performs regression tasks to predict continuous values unlike SVM that is used for classification purposes. The support vector regression comes with various kernel functions. Kernel functions are used to convert non linear patterns in data into linear patterns so that regression can be done. Each kernel function has its own hyperparameter that can be tuned during training to select the best model for that kernel function. The set of optimal hyperparameters are chosen using Grid Search Cross Validation. In this process, it trains and evaluates the model on all possible combinations of hyperparameters using cross-validation. It generates a score for each combination based on the evaluation metric. After evaluating all combinations, GridSearchCV identifies the combination of hyperparameters that achieved the highest score on the evaluation metric. After getting the best set of hyperparameters, the specific kernel function and its best set of hyperparameters are passed along with the data in a

model. Since our sample space was quite low due to less availability of hyperspectral data in the survey region, the training and testing was done using a technique called LOOCV stands for Leave one out cross validation. It is another type of cross validation used when a dataset does not contain a huge amount of instances. The technique takes a data point as testing and the rest (N-1) data points are assigned for training. This is done for each and every data point in the dataset iteratively. After splitting the dataset in this way, the maximum possible training instance can be created at a time. After training the model, that remaining one test point is used for testing and its prediction is stored outside. After doing this for all points, a complete set of predicted values is obtained for each sample. Later, Root Mean Squared Error (R.M.S.E), which is our evaluation metric for analysis and comparison of performance of various kernel functions, is computed using the predicted and actual concentration data. The graph between predicted and actual concentration data is also plotted to visualize the effectiveness of the model in predicting concentration accurately. The diagonal line over the graph representing the reference line for precise prediction, and the points lying closer to this line represent that the model is performing better. A diagonal trend of points on the graph in the direction of line and closeness that depicts models efficiency and overall prediction accuracy.

Results and Discussion

In this section, we have compiled the results of our study and project. The various kernel functions were trained and tested using the suitable methodologies addressed in earlier sections. The evaluation metric that we have used to compare the performance of different kernel functions of SVR is R.M.S.E stands for Root Mean Squared Error.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(predicted_i - actual_i)^2}$$

These R.M.S.E rates are computed for each target nutrient prediction for every kernel function of the SVR model. Additionally, we have also trained and tested two other regression models - Random Forest and Linear Regression. From (Table-2) it can be seen that all models

Table-2: Showing Results compilation - R.M.S.E values of the target nutrients for each model

Model	Fe_2O_3	K_2O	MgO	TiO_2
Linear SVR	0.5118	0.1969	0.3079	0.1969
Polynomial SVR	0.5485	0.1989	0.3412	0.0954
Radial Basis Function SVR	0.5235	0.1978	0.3207	0.0794
Random Forest Regression	0.5066	0.2228	0.3355	0.0947
Linear Regression	0.5210	0.2070	0.3249	0.0957

are performing well on the dataset. But the R.M.S.E values of RBF kernel SVR performs overall best in the SVR model category and it is also observed that Random Forest and Linear Regressor are performing equally well on data. From (Figure-7) and (Figure-8) the scatter plots between actual and predicted values of concentration can be seen for both RBF SVR and Random Forest Models. Linear SVR gives best prediction for K_2O and TiO_2, Polynomial SVR gives best prediction for TiO_2, RBF SVR gives best prediction for TiO_2, Random Forest gives best prediction for TiO_2, Linear Regression gives the best prediction for TiO_2. Overall result is that the best prediction is done for TiO_2 and K_2O.

Conclusion

In conclusion, the use of remote sensing technology, particularly hyperspectral imaging is proved to be efficient in estimating the soil nutrients concentration. Also, these techniques show promising results on basically all the models. Although, there is still a need for refinement and optimization of various aspects of this approach but with a

significant study in this field, this technique can replace all the traditional methods of soil testing. There can be various other machine learning techniques that can be use like ensemble learning to refine the prediction. Therefore, the results obtained conclude that soil testing and analysis can be performed using modern technologies promisingly.

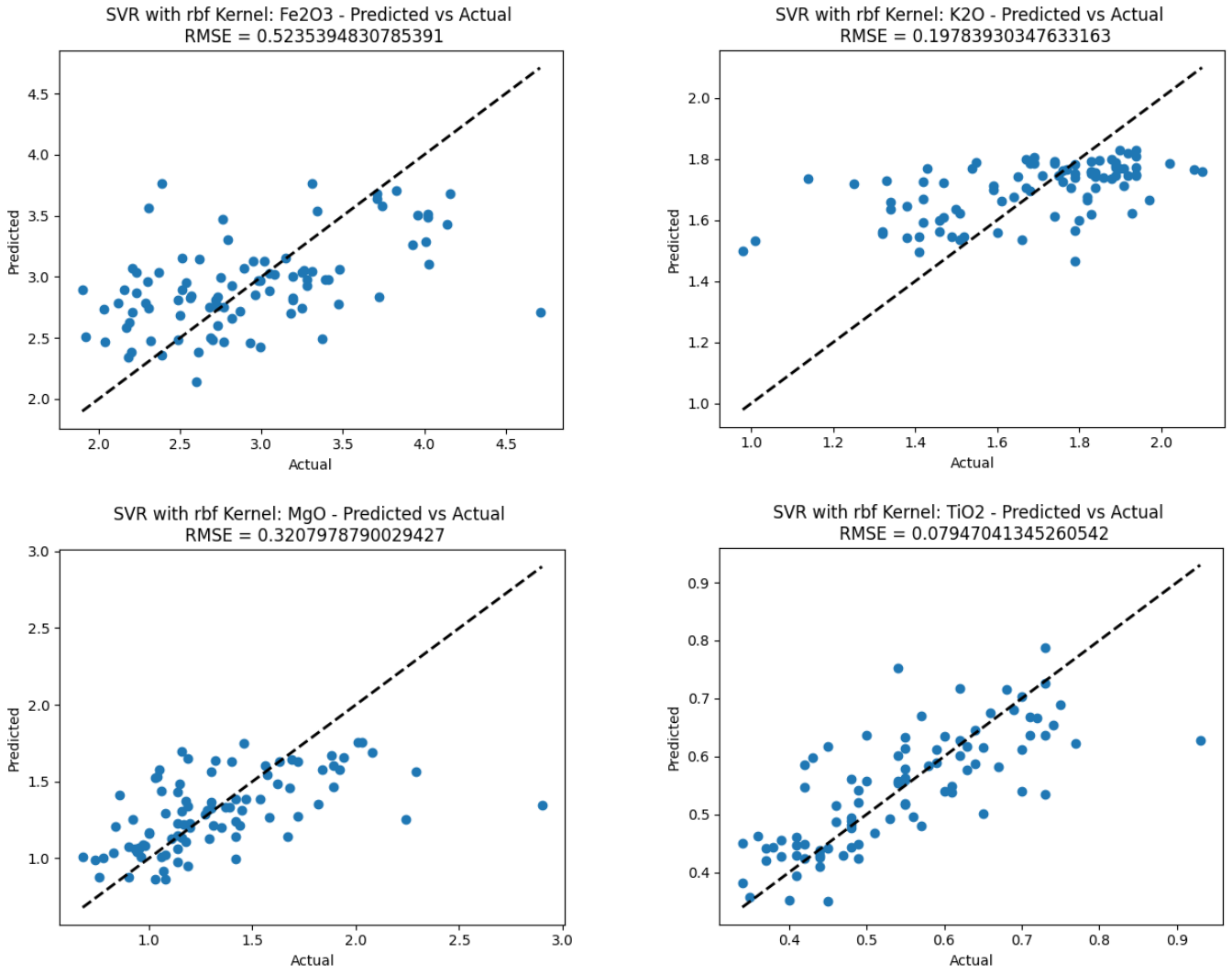
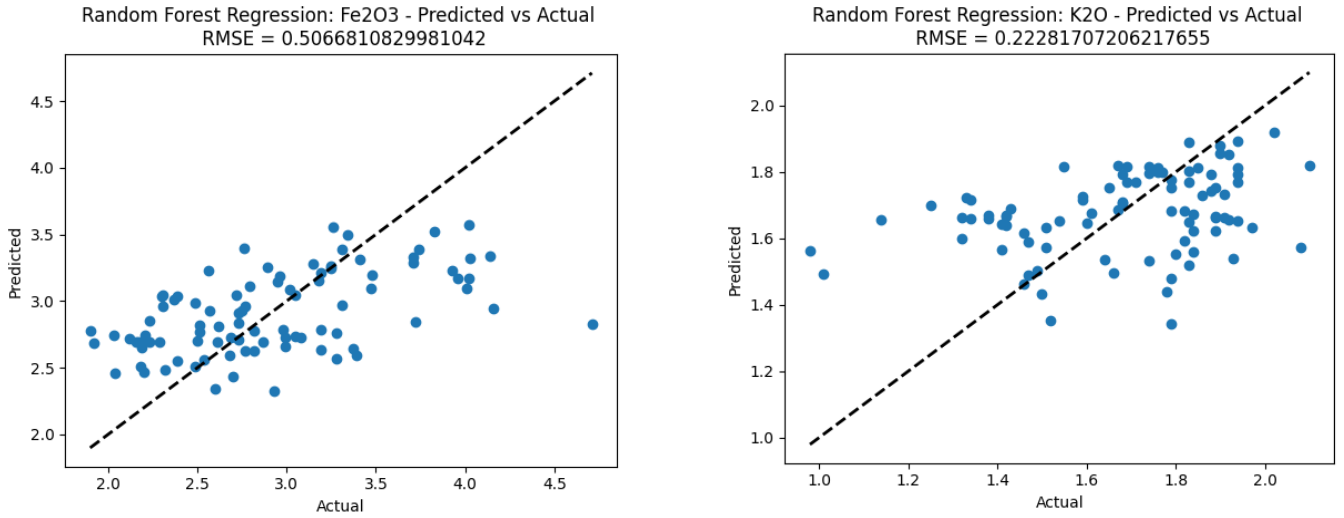


Figure-7: Showing the scatter plot between actual and predicted concentration of target nutrients for RBF SVR Model



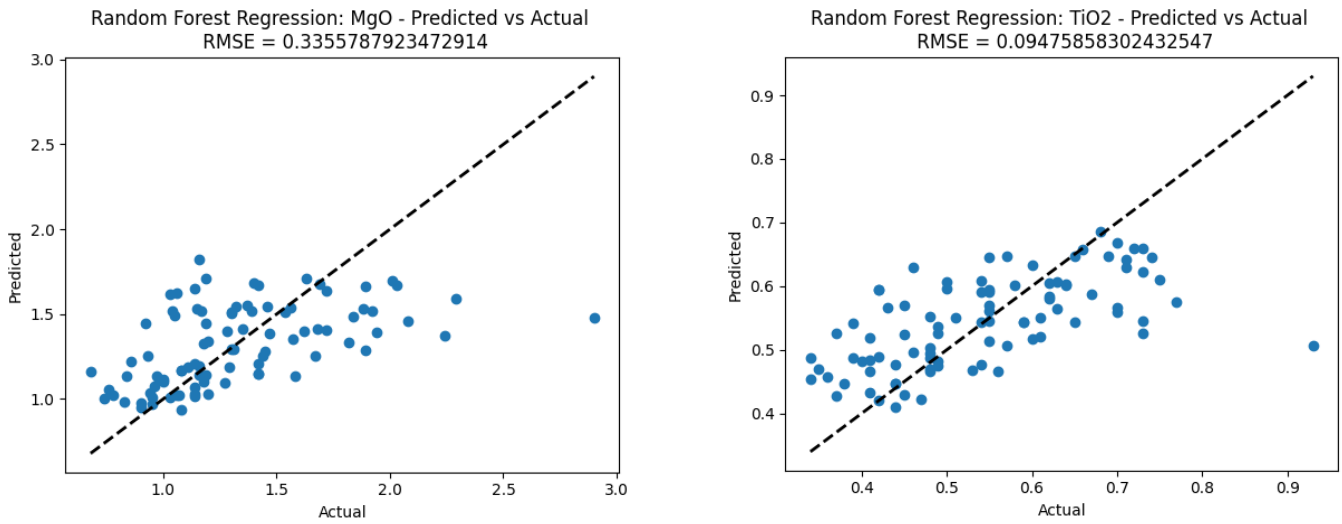


Figure-8: Showing the scatter plot between actual and predicted concentration of target nutrients for RF Model

References

1. Meshram et al. Need of Soil Testing for Improvement of Soil Health and Crop Productivity.

2. Baker et al. Soil Analysis: A Key to Soil Nutrient Management.

3. Nanni et al. Spectral Reflectance Methodology in Comparison to Traditional Soil Analysis.

4. A.S. Reis et al. Detection of soil organic matter using hyperspectral imaging sensor combined with multivariate regression modeling procedures.

5. Lixin lin et al. Hyperspectral Analysis of Soil Total Nitrogen in Subsided Land Using the Local Correlation Maximization-Complementary Superiority (LCMCS) Method .

6. Yiping peng et al. Estimation of Soil Nutrient Content Using Hyperspectral Data.

7. Jingjing Ma et al. Rapid detection of total nitrogen content in soil based on hyperspectral technology.

8. Naveen J.P. Anne et al. Modeling soil parameters using hyperspectral image reflectance in subtropical coastal wetlands

9. Hengliang Guo et al. Mapping Soil Organic Matter Content Based on Feature Band Selection with ZY1-02D Hyperspectral Satellite Data in the Agricultural Region.

Appendix 1: Methodology - Visual Representation