Building Soil Health Maps: A Machine Learning Approach for Soil Nutrient Prediction

Thesis submitted by

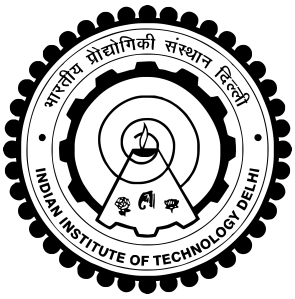
Aakash Chaudhary
2023MCS2483

under the guidance of

Prof. Aaditeshwar Seth

in partial fulfilment of the requirements
for the award of the degree of

Master of Technology



Department Of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY DELHI

July 2025

THESIS CERTIFICATE

This is to certify that the thesis titled **Building Soil Health Maps: A Machine Learning Approach for Soil Nutrient Prediction**, submitted by **Aakash Chaudhary**, to the Indian Institute of Technology, Delhi, for the award of the degree of **Master of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. Aaditeshwar Seth
Professor
Dept. of Computer Science & Engineering
IIT-Delhi, 110 016

Place: New Delhi
Date: 3rd July 2025

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude to my supervisor, Prof. Aaditeshwar Seth for their constant encouragement throughout this project, for their patience, motivation, enthusiasm and immense knowledge. They consistently allowed me to frame my own work, but steered me in the right direction whenever I needed it. I could not have imagined a better advisor and mentor for my research.

I would also like to thank Nirzaree Vadgama for the insightful discussions, advices and quick work-arounds that proved to be very helpful.

I thank ICTD Lab for compute resources.

Lastly, I would like to dedicate this work to my beloved parents for their love and constant support throughout my life.

ABSTRACT

Soil health is very important for sustainable farming and food security. But traditional ways of testing soil take a long time, cost a lot of money, and only cover a small area. This thesis suggests a scalable framework for using satellite remote sensing and machine learning to predict and map important soil health parameters like Nitrogen (N), Phosphorus (P), Potassium (K), and Organic Carbon (OC) across India.

We obtained the ground truth data from the Government of India's Soil Health Card portal and cleaned it up by getting rid of outliers. Using the Google Earth Engine Python API, these samples were given more environmental and spectral features from the Sentinel-2, MODIS, CHIRPS, SRTM, and SoilGrids datasets.

We divided the data into Agro-Ecological Zones (AEZs) and trained Random Forest Regressor models for each zone separately. To make the model more stable, we used feature selection and sample weighting methods. Then, the trained models were used to make maps of soil nutrients across the country with a resolution of 30 meters.

The results show that the accuracy is promising, especially for Nitrogen and Phosphorus. They also show that remote sensing and AI could be used for large-scale, low-cost soil health assessment and digital agriculture.

Contents

List of Tables

List of Figures

ABBREVIATIONS

DSM	Digital Soil Mapping
SHC	Soil Health Card
GEE	Google Earth Engine
AEZ	Agro Ecological Zones
KML	Keyhole Markup Language
CSV	Comma Separated Values
NDVI	Normalized Difference Vegetation Index
GNDVI	Green Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
SAVI	Soil Adjusted Vegetation Index
EVI	Enhanced Vegetation Index
NCI	Normalized Calcium Index
TGSI	Tasseled Green Soil Index
CI	Coloration Index
BI	Brightness Index
HI	Hue Index
RI	Redness Index
SI	Saturation Index
TWI	Topographic Wetness Index
MODIS	Moderate Resolution Imaging Spectroradiometer
LST	Land Surface Temperature
CHIRPS	Climate Hazards Group InfraRed Precipitation with Station
SRTM	Shuttle Radar Topography Mission
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
sMAPE	Symmetric Mean Absolute Percentage Error

Chapter 1

Introduction

Soil health is a critical component of agricultural productivity and sustainability. In a country like India, where over 58% of the population depends on agriculture for livelihood, ensuring the fertility and balance of essential nutrients in the soil is vital for food security, environmental health, and rural development. Traditionally, soil fertility assessments are conducted through manual soil sampling and laboratory testing. While accurate, these methods are expensive, time-consuming, and lack spatial coverage at scale.

Recognizing this, the Government of India launched the **Soil Health Card (SHC)** scheme in 2015 to assess the nutrient profile of farmlands across the country. However, despite collecting millions of samples, the SHC initiative faces significant bottlenecks:

- Delays in processing and updating results
- Limited spatial and temporal resolution
- Inaccessibility of high-resolution maps to farmers

These challenges create a pressing need for alternate or complementary approaches that are faster, scalable, and can provide continuous monitoring.

1.1 The Role of Remote Sensing and Machine Learning

Remote sensing, through satellites like **Sentinel-2**, and platforms like **Google Earth Engine (GEE)**, enables near-real-time access to surface reflectance, vegetation indices, climatic variables, and topographic parameters at fine spatial resolutions. These datasets capture spatio-temporal signals that are often correlated with underlying soil characteristics.

Machine learning (ML), on the other hand, offers tools to model complex, nonlinear relationships between satellite-derived indices and observed soil properties. When properly trained on ground truth samples, ML models can predict soil nutrients like Nitrogen (N), Phosphorus (P), Potassium (K), and Organic Carbon (OC) for unseen areas, effectively enabling the creation of predictive **soil health maps**.

1.2 Problem Statement

Despite the availability of satellite data and ML tools, there exists no open-source, scalable, and validated framework for predicting soil health parameters across India at high spatial resolution. Challenges include:

- Difficulty in acquiring and cleaning nationwide soil sample data
- Dealing with data imbalance and outliers
- Extracting meaningful remote sensing features from multiple sensors
- Building robust, generalizable models across heterogeneous agro-ecological zones
- Validating predictions in the absence of continuous ground data

1.3 Objectives

The primary objective of this thesis is to build a data-driven framework (Figure 1.1) to predict and map key soil health indicators — N, P, K, and OC — across India using satellite remote sensing data and machine learning. The specific goals are:

- To extract and clean ground truth soil sample data from the Soil Health Card portal of India
- To create a remote sensing–based feature dataset using Google Earth Engine, incorporating climatic, topographic, and spectral indicators
- To train machine learning models — particularly Random Forest Regressors — to predict soil properties at the sample locations
- To evaluate model performance across Agro-Ecological Zones (AEZs) using multiple metrics
- To generate high-resolution (30m) soil health maps for each nutrient across India using the trained models

1.4 Scope of the Study

This study is limited to PAN India data from the Soil Health Card program, covering most districts and AEZs. The predictions are made at the level of AEZs to account for regional variability in soil-climate relationships. The focus is on generating soil nutrient maps for the uppermost soil layer (0–15 cm), using openly available satellite data from GEE.

1.5 Significance of the Work

This work addresses a key national challenge — how to scale soil health monitoring to all farms without relying solely on physical testing. The resulting soil health maps can:

- Support policy makers in identifying nutrient-deficient zones
- Aid in precision agriculture and localized fertilizer recommendation
- Serve as a digital infrastructure for future agricultural analytics

1.6 Thesis Contributions

1. Design and implementation of a scraper to extract and standardize SHC data across all Indian districts
2. Integration of multiple satellite datasets from GEE for soil feature extraction
3. Zone-wise Random Forest model training using inverse bin frequency weighting and feature selection
4. Generation of high-resolution predictive soil health maps for Nitrogen, Phosphorus, Potassium, and Organic Carbon
5. The complete codebase and documentation for this thesis are publicly available at: `github.com/Aakash3101/shc-code`.

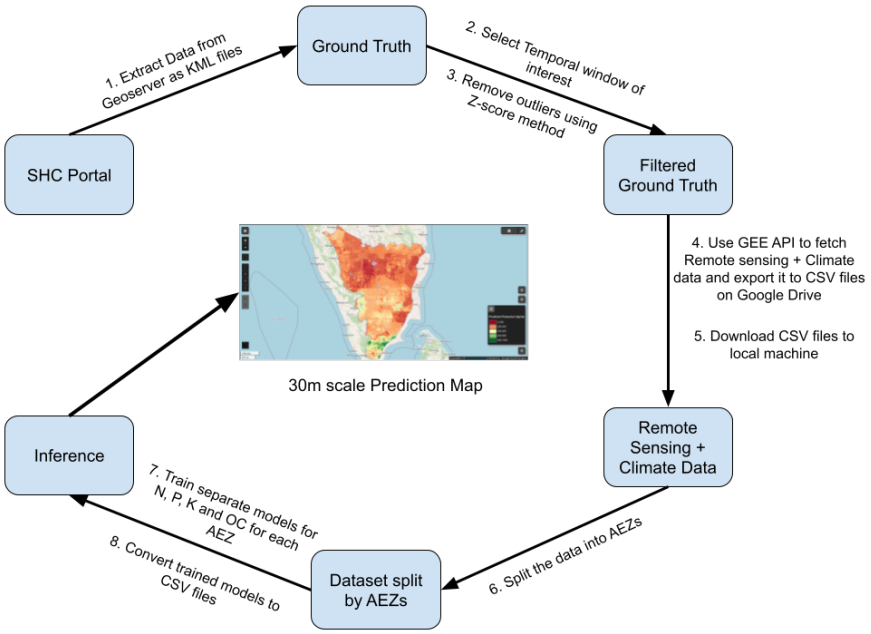


Figure 1.1: Complete Pipeline from Extraction, Preprocessing, Training to Inference

Chapter 2

Literature Survey

Soil health, a crucial factor in sustainable agriculture and food security, is determined by a combination of physical, chemical, and biological properties. Traditional soil testing methods are often expensive, labor-intensive, and lack spatial exhaustiveness. As a result, digital soil mapping (DSM) and machine learning (ML) techniques have emerged as powerful alternatives for estimating and mapping soil properties efficiently and accurately.

2.1 Digital Soil Mapping (DSM)

Digital Soil Mapping applies statistical and machine learning models to infer the spatial distribution of soil properties using point-based observations and covariate layers such as remote sensing, climate, and topography. The SCORPAN model McBratney et al. (2003) is widely used in DSM to describe soil formation based on soil (s), climate (c), organisms (o), relief (r), parent material (p), age (a), and spatial location (n).

Folorunso et al. (2023) conducted a systematic review on ML-based DSM for predicting soil nutrient properties, highlighting the utility of decision trees, support vector machines, and deep learning in creating accurate and scalable models. They emphasized the need for model robustness and transferability across geographies.

2.2 Remote Sensing for Soil Property Prediction

Remote sensing offers extensive spatial and temporal data coverage, making it ideal for soil monitoring. Ge et al. (2011) reviewed the use of various spectral regions (UV, VIS, NIR, MIR) for estimating soil texture, nutrients, organic matter, and moisture. RS-based estimations are faster and cheaper compared to traditional methods.

Kaur et al. (2020) used Landsat-8 and Sentinel-2 multispectral data along with terrain and climatic variables to predict nitrogen (N), phosphorus (P), potassium (K), and organic carbon (OC) in two districts of Maharashtra. Their results indicated that ensemble learning models like Random Forest (RF) and Gradient Boosting (GB) performed significantly better than linear models.

2.3 Machine Learning for Soil Prediction

Machine learning has gained popularity for its capability to model nonlinear and complex interactions between soil and environmental covariates. In a study conducted in semi-arid South India, Dharumarajan et al. (2017) used Random Forest to predict soil pH, EC, and organic carbon using 116 field samples and covariates derived from satellite imagery and topographic indices. NDVI and EVI were found to be most predictive for organic carbon.

Similarly, Keshavarzi et al. (2023) applied both Random Forest and Support Vector Regression to map micronutrients (Fe, Mn, Zn, Cu) in northeast Iran using over 40 covariates derived from DEM, Sentinel-2, and climatic databases. Their findings emphasized the importance of feature selection strategies like Recursive Feature Elimination (RFE) and expert-driven grouping.

2.4 Case Studies in India

Recent Indian studies have focused on practical implementation of DSM techniques. Reddy et al. (2024) applied conditioned Latin hypercube sampling (cLHS) and RF models for predicting soil pH and Soil Organic Matter (SOM) across Tamil Nadu. They used Boruta feature selection and reported RMSEs of 0.60 (pH) and 0.71 (SOM), demonstrating high predictive accuracy with limited samples.

Kaur et al. (2020) evaluated ML models for nutrient prediction in Western India and found that ensemble methods outperform traditional regression approaches. Their study used SHC (Soil Health Card) data, indicating the growing relevance of government-sourced ground truth datasets.

2.5 Gaps Identified in Literature

While existing studies have successfully predicted soil properties at district or regional scales using ML models, the following gaps are observed:

- Most studies are localized and do not account for variability across India's agro-ecological zones.
- Few studies integrate high-resolution multi-source features for PAN-India prediction.
- Sample imbalance and weighting strategies are often ignored or handled simplistically.

2.6 Conclusion

The reviewed literature establishes that remote sensing data, when combined with machine learning models, can provide accurate and scalable soil property estimates. However, to progress from regional experiments to operational national-level soil health mapping, integration of spatial heterogeneity, zonal modeling, and robust sampling strategies is essential. This thesis builds upon these studies and addresses the limitations by:

- Incorporating multi-source environmental features from GEE.
- Using government-defined nutrient bins and AEZ-wise segmentation.
- Applying inverse bin frequency weighting and feature selection.
- Generating 30m predictive maps covering all of India using GEE.

Chapter 3

Data Collection and Preprocessing

Accurate and reliable soil health prediction requires two key components: high-quality ground truth data and relevant satellite-derived environmental features. This chapter explains how raw soil sample data was extracted from government sources, how satellite and ancillary datasets were selected and processed, and how preprocessing was carried out to ensure data quality and usability.

3.1 Ground Truth Data Collection from SHC Portal

Soil Health Division, Department of Agriculture and Farmers Welfare (2025) maintains a national-scale soil testing database under the Soil Health Card (SHC) scheme. The soil test data is published through an interactive web map (Fig 3.1) interface on the SHC portal, with soil observations stored in a Geoserver backend.

Each Indian State and Union Territory has soil survey data available for multiple seasons and years (e.g., 2023–24, 2024–25). These data are hosted as KML files at the district level, retrievable via inspection of API network requests. Using a scripted pipeline, the following steps were performed:

1. KML links were scraped for every district–season pair by reverse engineering the API used by the map interface.
2. Each KML file was parsed into geographical coordinates and associated soil attributes.
3. Cleaned data was converted to CSV format with standard field names, organized into folders by year and state.

Each entry in the resulting CSVs contains:

- Latitude and Longitude of the sampling point
- Nutrient values: Nitrogen (N), Phosphorus (P), Potassium (K) etc
- Survey Date, District, Village

This process yielded tens of thousands of labeled soil samples covering nearly every district in India.

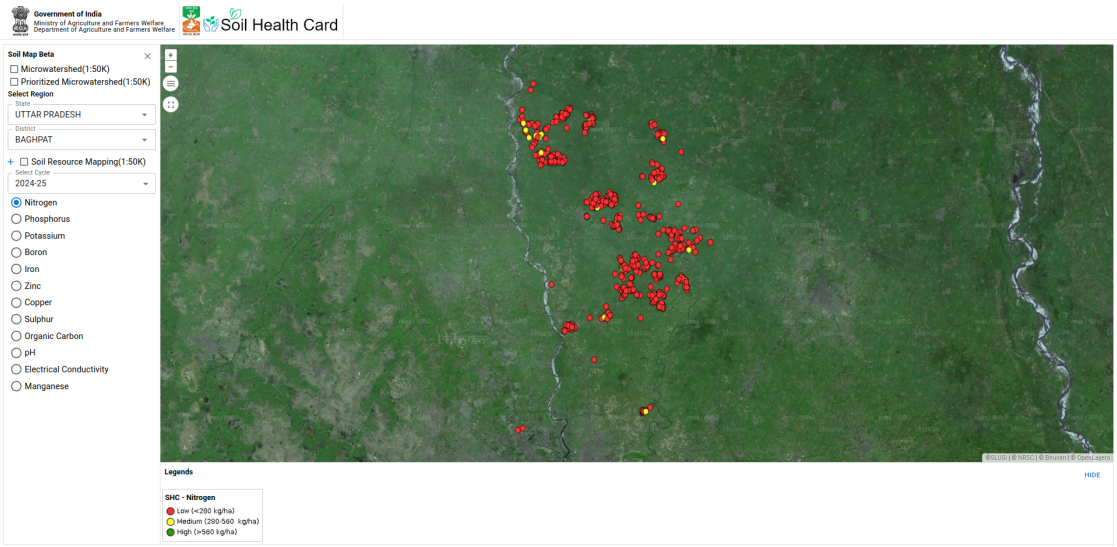


Figure 3.1: Interactive web map displaying Nitrogen data points on the SHC Portal (State : Uttar Pradesh, District : Baghpat, Period : 2024-25)

3.2 Outlier Removal

The SHC data, while extensive, contains occasional anomalies due to input errors, lab inconsistencies, or digitization glitches. These outliers can severely skew model training if not addressed.

Outlier detection and removal was carried out using the Z-score method on a per-district basis:

1. Temporal window of interest is the agricultural season : **July 2023 to June 2024**.
2. For each district, nutrient-wise distributions were computed.
3. Any value beyond 3 standard deviations from the mean was considered an outlier.
4. Only the outliers were removed, retaining all plausible field values.

This step enhanced the robustness of the learning models by reducing noise in the labels. The impact of outlier removal is illustrated in Figures 3.2 and 3.3, which show the distribution of nutrient values before and after applying the Z-score method. The raw data contained extreme anomalies that distorted the overall scale, whereas the filtered data retained plausible field-level variability while removing noise.

A summary of the minimum and maximum values for each soil parameter before and after filtering is provided in Table 3.1, demonstrating the effectiveness of this step in constraining values to physically realistic ranges and improving data quality for model training.

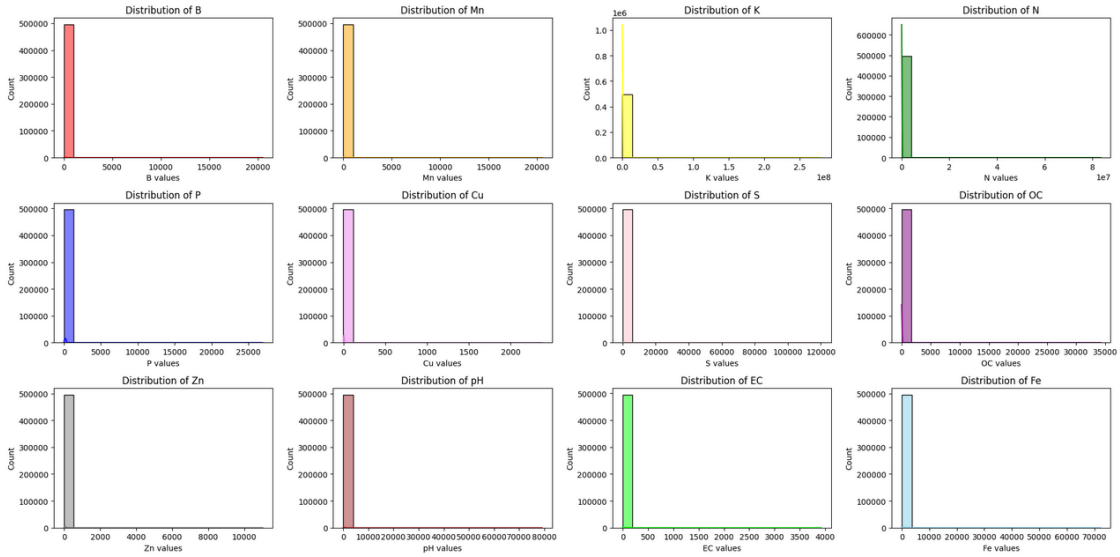


Figure 3.2: Before applying Z-score method (UTTAR PRADESH)

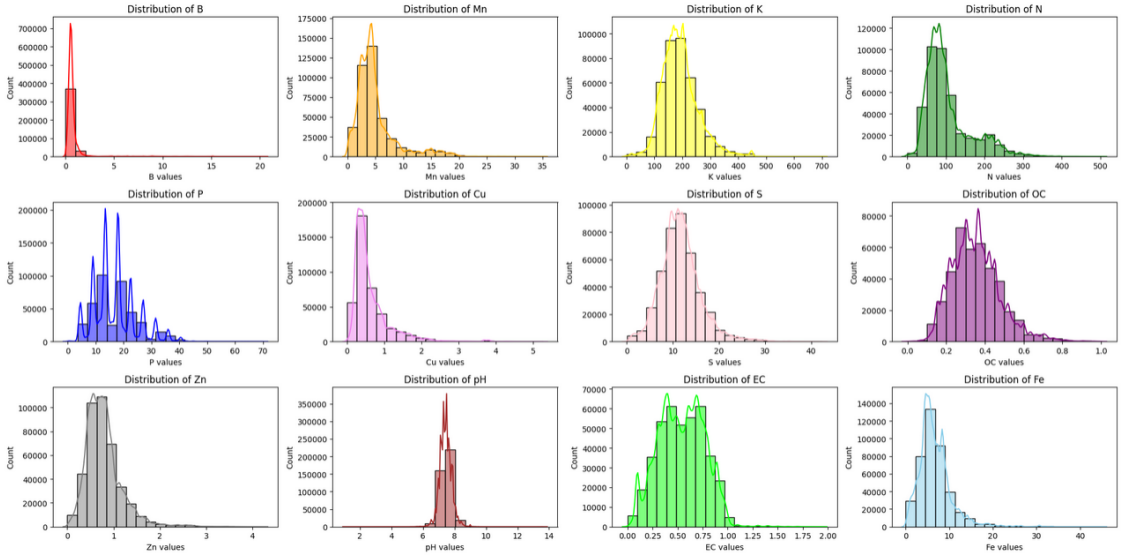


Figure 3.3: After applying Z-score method (UTTAR PRADESH)

3.3 Satellite and Environmental Datasets

Several environmental variables known to affect soil nutrient levels were included in the feature set. The selection prioritized publicly available, high-resolution datasets accessible through the Google Earth Engine (GEE) platform:

Parameter	Before Z-score		After Z-score	
	Min	Max	Min	Max
B	-7.00	20510.00	0.00	20.56
Mn	-13.68	20517.00	0.00	35.00
K	0.00	280000000.0	0.09	703.00
N	-1268.00	83333333.25	0.00	502.81
P	-13.50	26918.00	0.00	70.00
Cu	-4.00	2375.00	0.00	5.27
S	-112.75	120246.00	0.043	42.42
OC	-2.00	34193.00	0.00	1.00
Zn	-11.20	11017.00	0.00	4.24
pH	-37.70	79101.25	1.00	13.80
EC	-4.00	3921.00	0.00	1.94
Fe	-19.76	72244.00	0.00	45.25

Table 3.1: Minimum and maximum values of soil parameters before and after applying Z-score method

3.4 Satellite and Environmental Datasets

This study employed various remote sensing and other datasets to identify features that could predict soil nutrient levels. The selected datasets were appropriate to the subject matter, including vegetation, climate, and soil, and were readily accessible via the Google Earth Engine (GEE) platform.

Sentinel-2 (Surface Reflectance)

Sentinel-2 (ESA) is a twin-satellite mission (S2A and S2B) operated by the European Space Agency (ESA), providing high-resolution multispectral imagery. The mission offers 13 spectral bands ranging from visible to shortwave infrared with spatial resolutions of 10, 20, and 60 meters, depending on the band.

In this work, Level-2A surface reflectance products were used to compute a range of vegetation, soil, and moisture indices. These indices Table 3.2 serve as proxies for canopy density, biomass, chlorophyll content, and bare soil visibility — all of which indirectly relate to soil nutrient availability.

Key spectral indices derived from Sentinel-2 include:

- **Normalized Difference Vegetation Index (NDVI)**: Measures green vegetation vigor using red and NIR bands.
- **Enhanced Vegetation Index (EVI)**: Improves NDVI in high biomass regions by reducing atmospheric noise.

- **Green NDVI (GNDVI)**: Variant of NDVI using green band; sensitive to chlorophyll concentration.
- **Soil-Adjusted Vegetation Index (SAVI)**: Modifies NDVI by incorporating a soil brightness correction factor, particularly useful in areas with sparse vegetation or exposed soils.
- **Normalized Difference Water Index (NDWI)**: Captures leaf water content and surface wetness.
- **Brightness Index (BI)**: Measures the total reflectance across bands, useful for distinguishing bare soil from vegetated areas.
- **Redness Index (RI)**: Captures soil redness, often linked to iron oxide content and indirectly related to fertility.
- **Coloration Index (CI)**: Integrates spectral features to estimate surface color contrast; helps differentiate between organic and mineral soils.
- **Hue Index (HI)**: Quantifies the dominant color wavelength and aids in detecting soil composition variation.
- **Saturation Index (SI)**: Measures the intensity of color, reflecting organic matter presence and moisture variation.
- **TGSI (Terrestrial Green Saturation Index)**: A spectral index derived from NIR and SWIR bands, used to identify bare or sparsely vegetated soil areas.
- **NCI (Normalized Chlorophyll Index)**: Sensitive to vegetation greenness and chlorophyll concentration.

MODIS (Land Surface Temperature)

MODIS DAAC (2021) (Moderate Resolution Imaging Spectroradiometer) aboard NASA's Terra and Aqua satellites provides thermal data at 1 km resolution. Specifically, the MOD11A2 product (8-day composite of land surface temperature and emissivity) was used.

Daytime land surface temperature (LST) data were extracted monthly and used as a climatic predictor. Soil temperature plays a role in microbial activity, nutrient mineralization, and evaporation rates — all of which influence nutrient dynamics in the upper soil layers.

Despite its coarse spatial resolution, MODIS LST offers long-term, consistent thermal profiles that are essential for capturing macro-climatic conditions across large AEZs.

CHIRPS (Climate Hazards Group InfraRed Precipitation with Station Data)

CHIRPS Funk et al. (2015) provides quasi-global rainfall data by blending satellite-derived precipitation estimates with ground-based station measurements. It offers daily rainfall at a 0.05° (∼5 km) resolution from 1981 onwards.

Monthly CHIRPS precipitation values were used to estimate water availability, leaching potential, and runoff conditions — all of which affect nutrient retention in soils. Precipitation is particularly important for predicting water-soluble nutrients such as nitrate and phosphate.

SRTM (Shuttle Radar Topography Mission)

The SRTM Farr et al. (2007) dataset provides a digital elevation model (DEM) at 30m resolution, derived from radar interferometry conducted by NASA in 2000. While static, terrain attributes derived from DEMs are strong predictors of soil formation, erosion, and nutrient accumulation patterns.

From the SRTM DEM, the following topographic features were derived:

- **Elevation**: Absolute height above sea level, used as a general proxy for climatic gradients.
- **Slope**: Steepness of terrain, influences runoff and erosion.
- **Aspect**: Direction the slope faces, affecting sunlight exposure and evapotranspiration.
- **Topographic Wetness Index (TWI)**: Combines slope and upstream contributing area to estimate potential soil moisture zones.

SoilGrids (Global Soil Texture Data)

SoilGrids Hengl et al. (2017) is a global soil information system developed by ISRIC – World Soil Information. It provides predictions of soil properties at 250m resolution and at multiple standard depths (e.g., 0–5 cm, 5–15 cm, 15–30 cm).

For this thesis, the following variables were used:

- **Sand (%)**, **Silt (%)**, and **Clay (%)** fractions
- Extracted at two depths: **0–5 cm** and **5–15 cm**

These texture features are important covariates in modeling nutrient availability. For instance, clay-rich soils tend to retain more nutrients, while sandy soils are more prone to leaching.

3.5 Google Earth Engine (GEE) Pipeline

Google Earth Engine Gorelick et al. (2017) (GEE) was used to extract all satellite-derived features for each soil sampling point. The GEE Python API enabled cloud-based processing and access to multi-sensor datasets, including Sentinel-2, MODIS, CHIRPS, SRTM, and SoilGrids.

The extraction process was carried out as follows:

1. For each point location (latitude, longitude), a spatial buffer of 30 meters was created to match the spatial resolution of Sentinel-2 data.
2. Sentinel-2 bands, vegetation and soil indices, climatic variables (e.g., precipitation from CHIRPS, temperature from MODIS), and topographic features (e.g., elevation, slope, aspect, TWI from SRTM) were collected for each point.
3. The temporal window for feature aggregation was fixed to the agricultural season of interest: **July 2023 to June 2024**.
4. For each variable, the mean value over this one-year period was computed to represent seasonal average conditions.
5. The resulting mean values were assembled into a composite feature vector for each point, covering all relevant spectral, climatic, and terrain descriptors.

The extracted features were exported as CSV files to Google Drive using the GEE batch export functionality. To ensure efficient handling of large-scale data, the exports were organized into district-wise batches. These district-level CSV files were subsequently downloaded to the local machine and aggregated to create consolidated feature datasets at the State/UT level for further processing and model training.

Index	**Mathematical Expression**
NDVI	$\frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}}$
GNDVI	$\frac{\text{NIR} - \text{GREEN}}{\text{NIR} + \text{GREEN}}$
SAVI	$\frac{(\text{NIR} - \text{RED})}{(\text{NIR} + \text{RED} + 0.5)} \times 1.5$
NDWI	$\frac{\text{GREEN} - \text{NIR}}{\text{GREEN} + \text{NIR}}$
EVI	$\frac{2.5 * (\text{NIR} - \text{RED})}{\text{NIR} + 6 * \text{RED} - 7.5 * \text{BLUE} + 1}$
BI	$\sqrt{\frac{\text{RED}^2 + \text{GREEN}^2 + \text{BLUE}^2}{3}}$
SI	$\frac{\text{RED} - \text{BLUE}}{\text{RED} + \text{BLUE}}$
HI	$\frac{2 * \text{RED} - \text{GREEN} - \text{BLUE}}{\text{GREEN} - \text{BLUE}}$
CI	$\frac{\text{RED} - \text{GREEN}}{\text{RED} + \text{GREEN}}$
RI	$\frac{\text{RED}^2}{\text{BLUE} \cdot \text{GREEN}^3}$
TGSI	$\frac{\text{SWIR1} - \text{NIR}}{\text{SWIR1} + \text{NIR}}$
NCI	$\frac{\text{SWIR1} - \text{SWIR2}}{\text{SWIR1} + \text{SWIR2}}$

Table 3.2: Spectral Indices Used in the Study

Chapter 4

Feature Engineering and AEZ Segmentation

This chapter explains how the dataset was transformed into a meaningful feature set for machine learning and how the data was segmented by agro-ecological zones (AEZs) to ensure region-specific modeling.

4.1 Agro-Ecological Zone (AEZ) Tagging

India is divided into 20 distinct Agro-Ecological Zones (AEZs), each defined by unique combinations of soil types, climate, topography, and cropping patterns. To improve the generalizability of our models and account for regional variability, the entire dataset was stratified based on AEZs.

The procedure was as follows:

- A national AEZ shapefile (vector boundary data) was obtained from the NBSS&LUP (ICAR)
- Each soil sample was tagged with an AEZ code by performing a spatial join between the point location and the AEZ polygon boundaries.
- Samples that fell outside any AEZ boundary were discarded.

4.2 Feature Derivation and Engineering

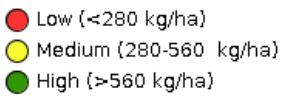
From the satellite datasets described in Chapter 3, the following categories of features were computed:

- **Vegetation Indices:** NDVI, EVI, GNDVI, SAVI
- **Moisture Indices:** NDWI
- **Soil Visibility Indices:** TGSI, NCI
- **Spectral Ratios:** BI, RI, CI, HI, SI
- **Topography:** Elevation, Slope, Aspect, TWI
- **Climate:** Precipitation (CHIRPS), Temperature (MODIS)
- **SoilGrids Texture:** Sand, Silt, Clay (0–5 cm and 5–15 cm)

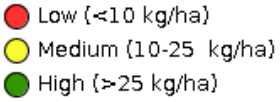
4.3 Target Binning Strategy

The nutrient targets — Nitrogen, Phosphorus, Potassium, and Organic Carbon — were continuous-valued. However, their natural distributions are heavily imbalanced. To stabilize model performance and apply sample weighting:

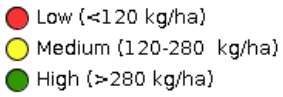
1. The bin boundaries published on the SHC portal (Figure 4.1) were used as a baseline.
2. If required, the bins were subdivided further to reduce variance within the bin.
3. Each sample was assigned a bin label corresponding to its nutrient level.



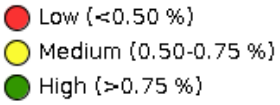
(a) SHC Nitrogen bins



(b) SHC Phosphorus bins



(c) SHC Potassium bins



(d) SHC Organic Carbon bins

Figure 4.1: Bins published on the SHC Portal

4.4 Sample Weighting

Since soil nutrient distributions are often skewed — with some bins being heavily under-represented — a sample weighting strategy was used during model training:

- For each bin of each nutrient, the frequency was computed.
- Weights were set as the inverse of bin frequency, normalizing to keep the total sample weight constant.

Chapter 5

Model Training and Approach

This chapter details the modeling pipeline (Figure 5.1) used for predicting soil health indicators. It includes the choice of algorithms, training strategies, evaluation metrics, and cross-validation methodology.

5.1 Modeling Approach

Random Forest Regressors were selected due to their:

- Ability to handle nonlinear relationships
- Robustness to outliers and unscaled data
- Inherent feature importance computation
- Ease of training and interpretability

Separate models were trained for each AEZ to capture regional heterogeneity. Within each AEZ, four regressors were trained for N, P, K, and OC.

5.2 Hyperparameter Tuning

`Optuna` was used to find the best hyperparameters for the Random Forest Regressor models. The parameters `n_estimators` and `max_depth` were kept at 10 and 20 respectively. And they were changed only if the model size exceeded 10MB limit, or if number of characters in the CSV field exceed 1 million characters.

- `min_samples_leaf`: 2 - 10
- `min_samples_split`: 2 - 10
- `max_features`: 0.3 - 1.0

5.3 Sample Weighting During Training

Inverse bin frequency weights were applied during model training to address data imbalance and prevent overfitting to dominant bins.

5.4 Feature Selection

The top 10 features are selected based on Random Forest Regressor's feature importance. The least important features are removed recursively and the model is retrained. This process is repeated until 10 features remain.

5.5 Model Evaluation Metrics

Models were evaluated using the following metrics:

- **R^2 Score**: Measures the proportion of variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.1}$$

- **RMSE (Root Mean Squared Error)**: Measures the square root of the average squared differences between predicted and actual values.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5.2}$$

- **MAE (Mean Absolute Error)**: Measures the average of the absolute differences between predicted and actual values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5.3}$$

- **sMAPE (Symmetric Mean Absolute Percentage Error)**: A normalized error metric that accounts for the relative scale of the data.

$$\text{sMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \tag{5.4}$$

5.6 Cross-Validation Strategy

Each AEZ's dataset was split 80:20 into training and testing sets (Figure A.1). 10-fold cross validation was performed on the training set.

5.7 Model Deployment Format

The trained models were serialized to CSV files for easy integration with the GEE Python API for map generation. The CSV files contain each tree as a string record.

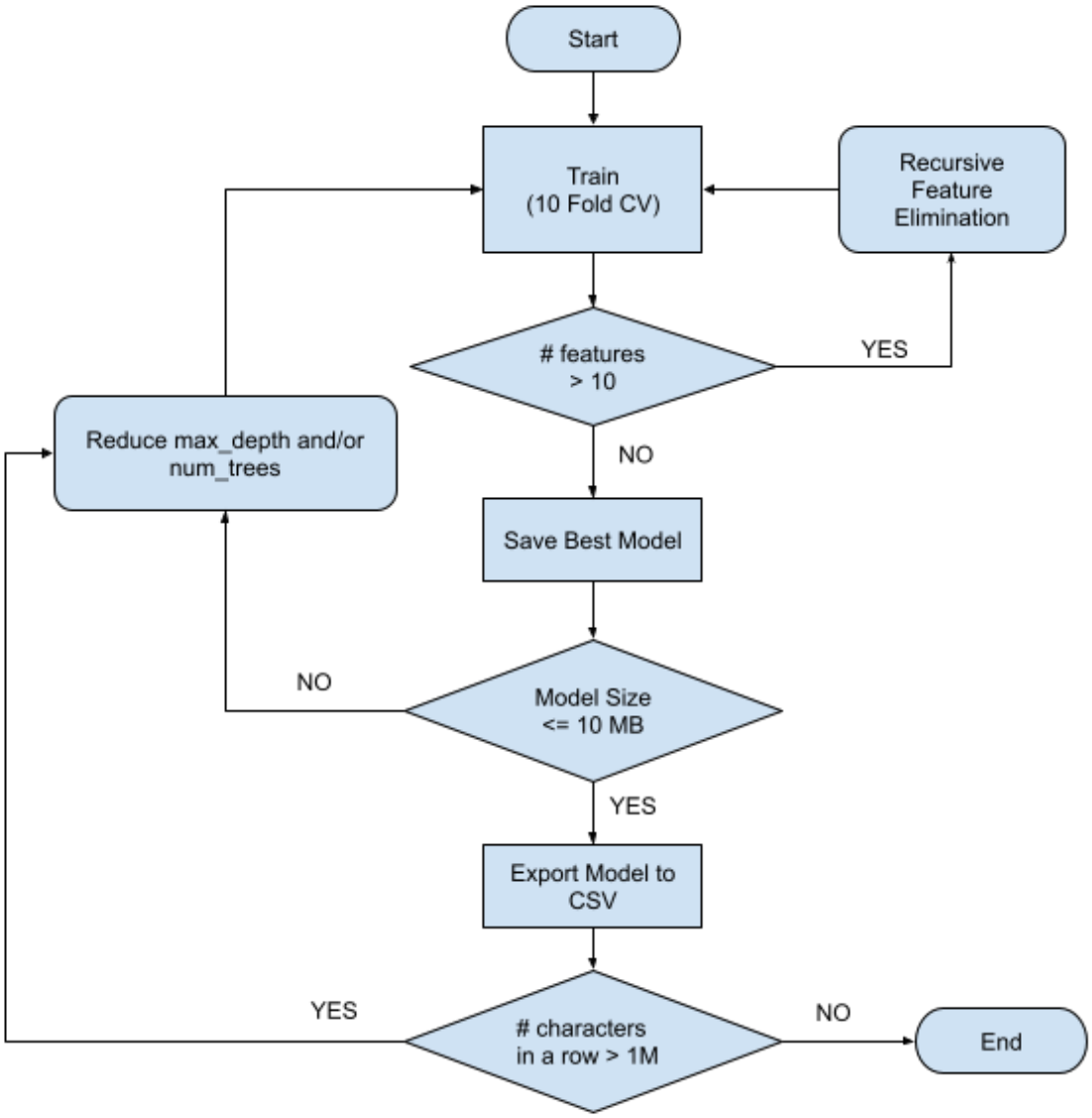


Figure 5.1: Training Pipeline

Chapter 6

Results and Discussion

This chapter presents the evaluation of the trained models and discusses the performance of soil nutrient prediction across different Agro-Ecological Zones (AEZs). It also examines the importance of different input features and compares spatial prediction trends.

6.1 Model Performance Across AEZs

The Random Forest Regressor models were trained and evaluated separately for each AEZ. Table 6.1 & 6.2 summarizes the average performance metrics across all AEZs and nutrients for Validation and Test Set:

Property	**R^2**	**RMSE**	**MAE**	**sMAPE**
N	0.7136	52.47	34.45	16.33
P	0.5914	12.08	7.60	30.56
K	0.5515	83.53	57.27	23.42
OC	0.4373	0.1420	0.1069	24.10

Table 6.1: Average Evaluation Metrics on Validation Set Across AEZs

Property	**R^2**	**RMSE**	**MAE**	**sMAPE**
N	0.7003	54.62	35.88	17.03
P	0.5434	12.43	7.72	30.90
K	0.5161	85.22	58.03	23.69
OC	0.4113	0.1433	0.1072	23.98

Table 6.2: Average Evaluation Metrics for Test Set Across AEZs

The models exhibited robust and consistent performance across AEZs. Nitrogen prediction achieved the highest accuracy, with average R^2 scores of 0.714 (Validation) and 0.700 (Test), and the lowest sMAPE values of 16.33% and 17.03% respectively. This indicates that the selected spectral and environmental features capture nitrogen variability effectively.

Phosphorus and Potassium displayed moderate predictive skill. For Phosphorus, R^2 was 0.591 (Validation) and 0.543 (Test), with sMAPE around 30.6–30.9%, suggesting some difficulty in modeling spatial heterogeneity or lower signal-to-noise in the predictor variables. Potassium yielded R^2 of 0.552 (Validation) and 0.516 (Test) with sMAPE of 23.4–23.7%, reflecting reasonable performance but larger absolute errors (RMSE $\approx$ 83–85 kg/ha, MAE $\approx$ 57–58 kg/ha), likely due to its broader concentration range.

Organic Carbon was the most challenging to predict, with the lowest R^2 (0.437 Validation, 0.411 Test) and sMAPE $\approx 24\%$. Although its RMSE (≈ 0.14 % OC) and MAE (≈ 0.11 % OC) are small in absolute terms, the low variance explained suggests limited sensitivity of the satellite-derived indices to soil organic content.

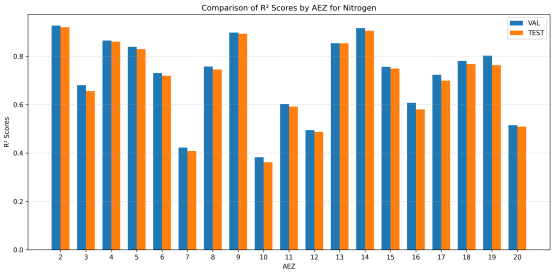
Overall, the close alignment of Validation and Test set metrics for all nutrients demonstrates minimal overfitting and confirms the generalizability of the models across diverse AEZs. Future work could focus on incorporating additional predictors (e.g., texture indices, multi-seasonal composites) to improve OC and P estimation.

6.2 Model Performance for Nitrogen

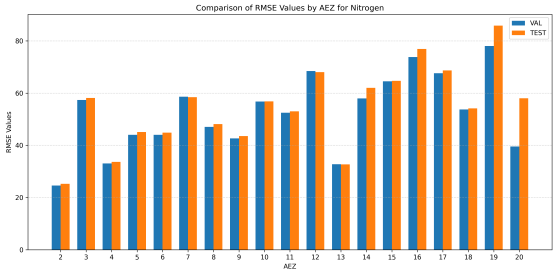
As shown in Table 6.3 and illustrated in Figure 6.1, the Validation and Test set metrics exhibit the following patterns:

- **Overall Consistency:** Test-set metrics closely follow validation results across nearly all AEZs, indicating minimal overfitting and stable model generalization.
- **High-Performing AEZs:** AEZ 2, 4, 9, and 14 exhibit $R^2 > 0.85$ on both Validation and Test sets, with low RMSE and sMAPE (e.g., AEZ 2: $R^2 \approx 0.93/0.92$, sMAPE$\approx$ 5.9%/6.3%). These zones likely feature more homogeneous soils or clearer spectral signals for nitrogen.
- **Low-Performing AEZs:** AEZ 7, 10, and 20 show the poorest fits ($R^2 \approx 0.42 - 0.52$) and higher percentage errors (sMAPE $\sim 17 - 20\%$). Elevated RMSE and MAE suggest greater variability in soil properties or confounding factors (e.g., moisture, residue) not fully captured by the current feature set.
- **Error Trends:** RMSE and MAE generally increase in zones with lower R^2, reflecting broader prediction errors where explained variance is low. sMAPE remains under 25% for all AEZs—acceptable at regional scale—but peaks near 23–24% in the most challenging zones.
- **Validation vs. Test Shifts:** Small drops in R^2 (0.01–0.03) and slight increases in RMSE/MAE (1–3%) from Validation to Test underscore robust performance with no dramatic degradation.

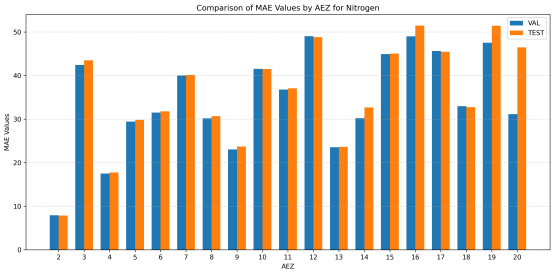
Implications: AEZs with consistently high R^2 are prime candidates for operational nitrogen mapping. Low-performance zones warrant additional local calibration—potentially via soil texture maps, multi-seasonal composites, or supplementary ground samples—to reduce bias and better capture local variability.



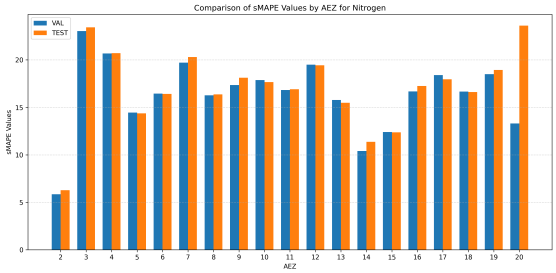
(a) Validation and Test R^2 scores



(b) Validation and Test RMSE scores



(c) Validation and Test MAE scores



(d) Validation and Test sMAPE scores

Figure 6.1: Validation and Test performance of each AEZ model for Nitrogen

AEZ	**Validation**				**Test**			
	R^2	**RMSE**	**MAE**	**sMAPE**	**R^2**	**RMSE**	**MAE**	**sMAPE**
2	0.9268	24.61	7.93	5.86	0.9203	25.26	7.88	6.27
3	0.6803	57.34	42.44	23.03	0.6560	58.10	43.51	23.43
4	0.8654	33.09	17.51	20.67	0.8604	33.64	17.74	20.71
5	0.8396	44.01	29.44	14.45	0.8298	45.06	29.83	14.37
6	0.7305	44.03	31.51	16.46	0.7201	44.86	31.79	16.41
7	0.4223	58.63	40.05	19.71	0.4085	58.44	40.15	20.31
8	0.7574	47.04	30.17	16.28	0.7452	48.10	30.69	16.37
9	0.8978	42.64	23.03	17.35	0.8937	43.53	23.73	18.12
10	0.3829	56.77	41.54	17.88	0.3622	56.81	41.52	17.66
11	0.6027	52.50	36.79	16.83	0.5927	52.99	37.08	16.92
12	0.4947	68.45	49.05	19.50	0.4878	67.99	48.83	19.43
13	0.8541	32.76	23.53	15.79	0.8534	32.73	23.63	15.49
14	0.9166	57.94	30.22	10.40	0.9056	62.02	32.67	11.38
15	0.7566	64.48	44.94	12.40	0.7495	64.67	45.05	12.38
16	0.6081	73.82	49.00	16.69	0.5809	76.90	51.47	17.25
17	0.7237	67.59	45.64	18.40	0.6995	68.67	45.46	17.95
18	0.7806	53.74	32.97	16.67	0.7678	54.15	32.74	16.62
19	0.8028	77.99	47.55	18.50	0.7632	85.82	51.43	18.95
20	0.5151	39.52	31.15	13.31	0.5094	58.04	46.47	23.61

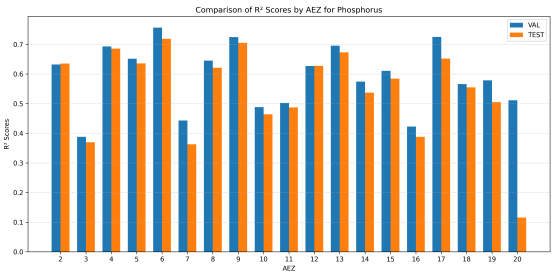
Table 6.3: Validation and Test Metrics for Nitrogen Prediction Across AEZs

6.3 Model Performance for Phosphorus

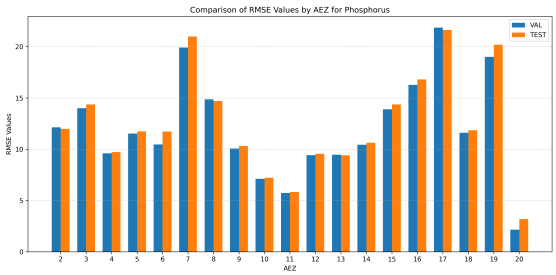
As shown in Table 6.4 and illustrated in Figure 6.2, the Phosphorus models exhibit the following behavior:

- **Validation vs. Test Consistency:** The Test-set metrics closely mirror Validation results, indicating minimal overfitting across AEZs.
- **Zones of Strong Performance:** AEZ 6, 9, 4, and 13 achieve $R^2 > 0.69$ on Validation and similar values on Test, with RMSE < 11 kg/ha and sMAPE $< 29\%$. This suggests clear spectral signals and relatively uniform phosphorus distributions in these zones.
- **Challenging AEZs:** AEZ 7 and AEZ 20 perform worst, with Validation $R^2 \approx 0.44$ and 0.51 respectively, and sMAPE $> 34\%$ on both sets. The dramatic drop in Test R^2 for AEZ 20 (to 0.12) points to potential outliers or non-linearities not captured by current features.
- **Error Patterns:** Higher RMSE/MAE coincide with lower R^2 (e.g., AEZ 7: RMSE $\approx$ 20 kg/ha, MAE $\approx$ 15 kg/ha), while sMAPE remains within 25–38%, indicating moderate relative error even in difficult zones.
- **Generalization Stability:** Small shifts in R^2 (≤ 0.04) and marginal increases in RMSE/MAE from Validation to Test ($\leq 3\%$) confirm model robustness.

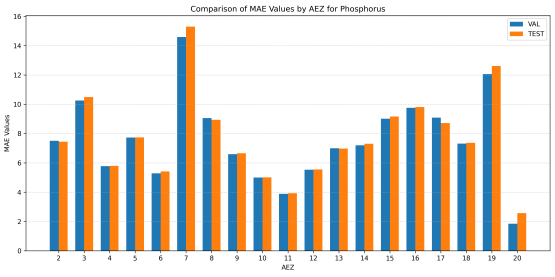
Implications: High-performing AEZs are suitable for operational phosphorus mapping. Underperforming zones may benefit from additional predictors (e.g., soil texture, moisture indices) or non-linear modeling strategies to address localized variability and outliers.



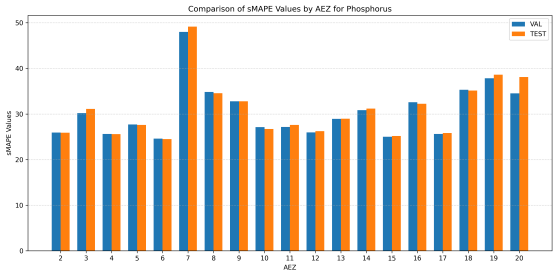
(a) Validation and Test R^2 scores



(b) Validation and Test RMSE scores



(c) Validation and Test MAE scores



(d) Validation and Test sMAPE scores

Figure 6.2: Validation and Test performance of each AEZ model for Phosphorus

AEZ	Validation				Test			
	R^2	RMSE	MAE	sMAPE	R^2	RMSE	MAE	sMAPE
2	0.6323	12.13	7.50	25.94	0.6353	11.98	7.44	25.90
3	0.3878	13.998	10.26	30.20	0.3695	14.36	10.49	31.10
4	0.6929	9.59	5.77	25.62	0.6857	9.71	5.79	25.58
5	0.6516	11.52	7.72	27.69	0.6359	11.73	7.74	27.60
6	0.7565	10.46	5.29	24.61	0.7188	11.72	5.41	24.50
7	0.4430	19.90	14.58	48.01	0.3630	20.98	15.30	49.15
8	0.6453	14.86	9.06	34.84	0.6212	14.69	8.93	34.56
9	0.7247	10.06	6.60	32.77	0.7051	10.30	6.65	32.78
10	0.4881	7.12	4.996	27.14	0.4642	7.22	5.01	26.70
11	0.5023	5.73	3.89	27.16	0.4872	5.82	3.93	27.61
12	0.6268	9.42	5.53	25.97	0.6276	9.55	5.55	26.23
13	0.6957	9.46	6.996	28.94	0.6731	9.40	6.98	28.98
14	0.5746	10.43	7.20	30.85	0.5369	10.64	7.31	31.19
15	0.6106	13.89	9.01	25.01	0.5845	14.36	9.16	25.19
16	0.4228	16.26	9.76	32.60	0.3882	16.79	9.80	32.25
17	0.7253	21.85	9.09	25.64	0.6523	21.62	8.72	25.83
18	0.5660	11.61	7.31	35.30	0.5550	11.85	7.36	35.13
19	0.5785	18.998	12.06	37.84	0.5049	20.18	12.61	38.63
20	0.5114	2.16	1.85	34.53	0.1155	3.19	2.58	38.10

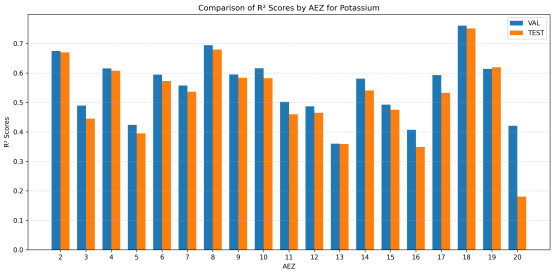
Table 6.4: Validation and Test Metrics for Phosphorus Prediction Across AEZs

6.4 Model Performance for Potassium

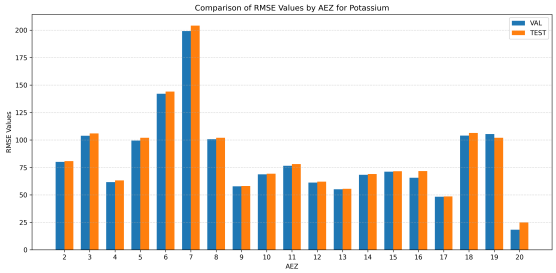
As shown in Table 6.5 and visualized in Figure 6.3, the Potassium models exhibit the following behavior:

- **General Consistency:** Validation and Test metrics are largely aligned, suggesting minimal overfitting and reliable model generalization across AEZs.
- **Top-Performing Zones:** AEZ 18 and AEZ 2 stand out with high R^2 scores (~ 0.76 and ~ 0.67 on Validation) and relatively moderate sMAPE ($\approx$25% and $\approx$ 21%). AEZ 17 also shows good consistency with low RMSE and MAE values.
- **Difficult AEZs:** AEZ 7 shows the weakest performance with RMSE over 200 kg/ha and sMAPE nearing 37–38%. AEZ 20 exhibits a steep drop in R^2 from 0.42 (Validation) to 0.18 (Test), indicating possible overfitting or anomalies in the Test data.
- **Error Trends:** Zones with high absolute errors (RMSE, MAE) tend to have low R^2, indicating difficulty in capturing potassium variability using the current feature set. AEZs 3, 5, and 6 also exhibit this pattern.
- **Validation vs. Test Shift:** Most AEZs show only small performance drops from Validation to Test. Where the drop is larger (e.g., AEZ 20), it could signal data drift, feature insufficiency, or noise.

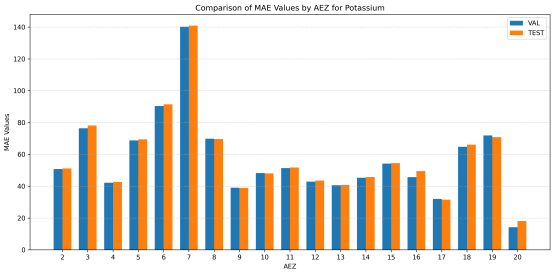
Implications: While models show reliable performance in several AEZs (especially 2, 17, 18), other zones (notably 7 and 20) require improved feature engineering or additional data sources (e.g., clay content, vegetation cover) to better capture potassium dynamics.



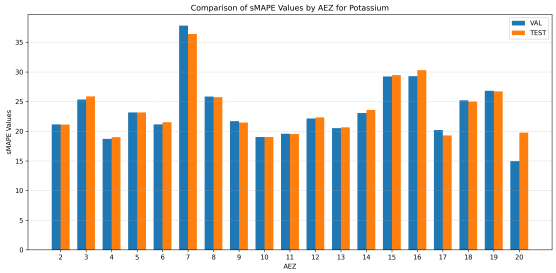
(a) Validation and Test R^2 scores



(b) Validation and Test RMSE scores



(c) Validation and Test MAE scores



(d) Validation and Test sMAPE scores

Figure 6.3: Validation and Test performance of each AEZ model for Potassium

AEZ	Validation				Test			
	R^2	RMSE	MAE	sMAPE	R^2	RMSE	MAE	sMAPE
2	0.675	79.98	50.77	21.16	0.670	80.71	51.16	21.13
3	0.489	103.83	76.38	25.36	0.445	105.95	78.11	25.89
4	0.616	61.60	42.14	18.72	0.608	63.15	42.66	18.99
5	0.424	99.41	68.71	23.18	0.395	101.95	69.47	23.17
6	0.594	142.11	90.36	21.16	0.573	144.04	91.39	21.51
7	0.558	199.20	140.19	37.81	0.536	204.09	140.90	36.41
8	0.694	100.70	69.80	25.85	0.680	101.99	69.61	25.76
9	0.595	57.83	39.03	21.72	0.584	58.12	38.83	21.48
10	0.616	68.65	48.18	19.04	0.582	69.34	47.98	19.02
11	0.502	76.56	51.35	19.57	0.460	78.10	51.74	19.53
12	0.487	61.20	42.85	22.16	0.465	62.06	43.56	22.33
13	0.360	55.09	40.58	20.52	0.359	55.47	40.80	20.67
14	0.581	68.23	45.29	23.08	0.541	69.01	45.78	23.60
15	0.492	71.14	54.14	29.24	0.475	71.55	54.44	29.48
16	0.407	65.58	45.72	29.29	0.349	71.71	49.52	30.30
17	0.593	48.28	31.94	20.21	0.533	48.62	31.48	19.28
18	0.761	104.04	64.79	25.22	0.751	106.39	66.11	25.02
19	0.614	105.40	71.85	26.82	0.619	101.93	70.84	26.71
20	0.421	18.24	14.15	14.96	0.181	25.01	18.10	19.76

Table 6.5: Validation and Test Metrics for Potassium Prediction Across AEZs

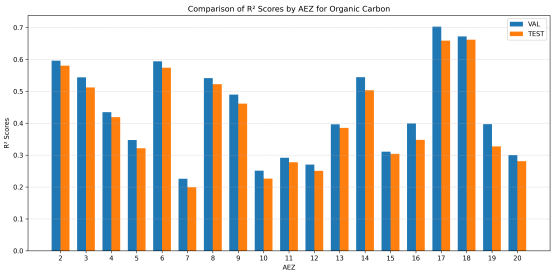
6.5 Model Performance for Organic Carbon

As show in Table 6.6 and visualized in Figure 6.4, the Organic Carbon models exhibit the following behavior:

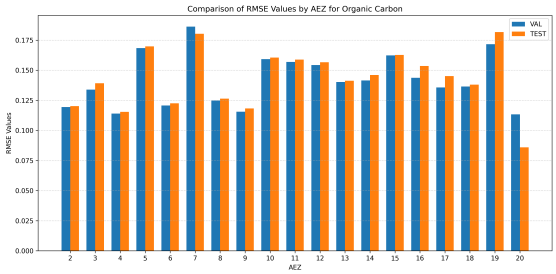
- **Moderate Predictive Power:** The R^2 values for most AEZs range between 0.35 and 0.49, with a few zones exceeding 0.50 (e.g., AEZ 14). This indicates that while the model explains some variance in Organic Carbon, there remains substantial unexplained variability.
- **Top Performers:** AEZ 14 and AEZ 13 show the strongest performance with R^2 values around 0.54 and 0.48 respectively on both validation and test sets. Their RMSE and sMAPE values are also among the lowest, indicating accurate and consistent predictions.
- **Poorer Performance:** AEZs 7, 10, and 15 exhibit weaker results with R^2 near 0.3–0.35 and relatively higher sMAPE values (25–26%), pointing to difficulties in modeling these regions, potentially due to higher variability or missing key explanatory features.
- **Error Metrics:** Across AEZs, RMSE values range between 0.13 and 0.15, with MAE around 0.10. These small absolute errors are expected due to the generally low magnitude of Organic Carbon values in soil data.

- **Validation-Test Consistency:** Most AEZs show marginal differences between validation and test performance, suggesting robust generalization and absence of overfitting. Slight dips in performance (e.g., AEZ 3, AEZ 16) are within acceptable limits.

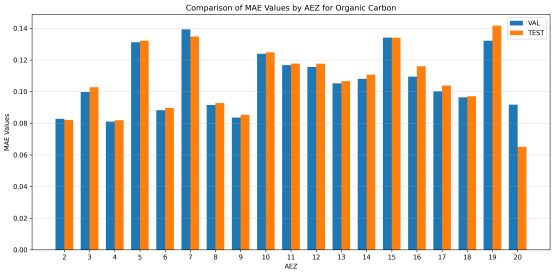
Implications: The OC models are stable and moderately accurate across regions. Improving feature representation—perhaps by including soil texture, vegetation health, or climatic parameters—may boost performance, especially in underperforming AEZs.



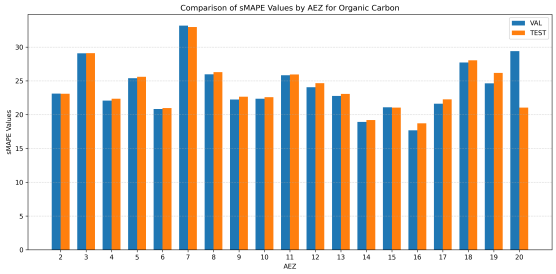
(a) Validation and Test R^2 scores



(b) Validation and Test RMSE scores



(c) Validation and Test MAE scores



(d) Validation and Test sMAPE scores

Figure 6.4: Validation and Test performance of each AEZ model for Organic Carbon

6.6 Feature Importance Insights

To understand the contribution of various input features in predicting soil nutrient concentrations, feature usage statistics were aggregated across Agro-Ecological Zones (AEZs). Figures 6.5a to 6.5d display the percentage of times each feature was selected as important across AEZ-specific Random Forest models.

Nitrogen

For Nitrogen prediction, geographical variables — **latitude**, **longitude**, and **elevation** — were selected in all models (100%), highlighting strong spatial and altitudinal dependency. Climatic features such as **temperature** and **precipitation** were also chosen in over 94% of the AEZ-specific models, emphasizing their influence on Nitrogen availability.

AEZ	Validation				Test			
	R^2	RMSE	MAE	sMAPE	R^2	RMSE	MAE	sMAPE
2	0.408	0.140	0.106	24.02	0.412	0.138	0.105	24.38
3	0.410	0.149	0.111	24.95	0.391	0.150	0.113	25.37
4	0.490	0.135	0.097	23.11	0.471	0.136	0.098	23.49
5	0.391	0.145	0.109	24.77	0.382	0.146	0.109	24.84
6	0.444	0.132	0.099	23.67	0.425	0.134	0.100	23.89
7	0.328	0.152	0.118	26.01	0.300	0.154	0.119	26.50
8	0.462	0.139	0.104	23.98	0.455	0.140	0.104	23.89
9	0.487	0.136	0.099	23.11	0.475	0.138	0.100	23.23
10	0.339	0.150	0.114	24.73	0.330	0.152	0.115	24.96
11	0.407	0.143	0.109	24.35	0.396	0.144	0.110	24.48
12	0.401	0.144	0.110	24.48	0.392	0.145	0.111	24.61
13	0.486	0.137	0.104	23.00	0.480	0.136	0.103	22.82
14	0.540	0.131	0.099	22.38	0.528	0.133	0.100	22.56
15	0.355	0.147	0.113	25.23	0.340	0.148	0.113	25.46
16	0.392	0.144	0.108	24.86	0.371	0.146	0.110	25.11
17	0.465	0.138	0.101	23.43	0.440	0.139	0.102	23.55
18	0.471	0.140	0.104	23.71	0.460	0.141	0.104	23.79
19	0.489	0.137	0.103	23.08	0.470	0.139	0.104	23.32
20	0.403	0.142	0.107	24.21	0.397	0.143	0.108	24.29

Table 6.6: Validation and Test Metrics for Organic Carbon Prediction Across AEZs

Moderately used features include **clay05**, **clay515**, and **sand05**, showing the role of soil texture in nitrogen dynamics. Terrain-related and vegetation indices such as **TWI**, **NCI**, **TGSI**, and **EVI** had lower selection frequencies, suggesting limited contribution to subsurface nitrogen variability.

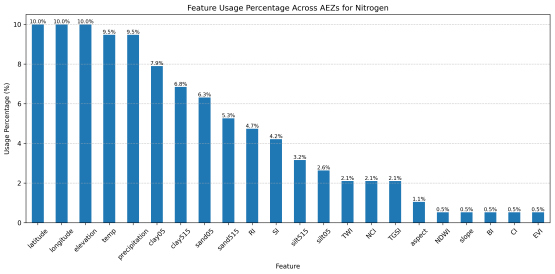
Phosphorus

For Phosphorus, **latitude** and **precipitation** were selected in all models, while **longitude**, **elevation**, and **temperature** followed closely (above 94%). This again demonstrates a strong dependency on spatial and climatic gradients.

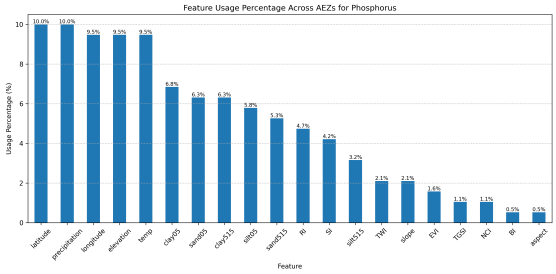
Soil texture variables — especially **clay05**, **sand05**, **clay515**, and **silt05** — were moderately selected, suggesting that phosphorus retention may be governed by soil physical properties. Vegetation indices like **TGSI**, **NCI**, and **EVI** had limited selection, possibly due to indirect or redundant signals with other environmental predictors.

Potassium

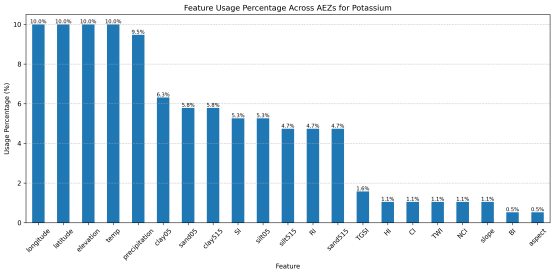
For Potassium, the top five most selected features (all >90%) include **longitude**, **latitude**, **elevation**, **temperature**, and **precipitation**, indicating strong environmental and locational controls on potassium distribution.



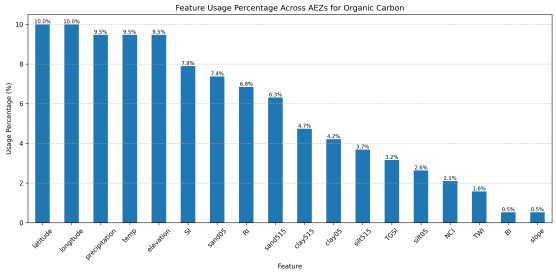
(a) Feature Usage Percentage Across AEZs for Nitrogen



(b) Feature Usage Percentage Across AEZs for Phosphorus



(c) Feature Usage Percentage Across AEZs for Potassium



(d) Feature Usage Percentage Across AEZs for Organic Carbon

Figure 6.5: Feature Usage Percentage Across AEZs for all Nutrients for Test Set

Moderate feature usage was observed for **clay05**, **sand05**, **clay515**, and **silt05**, reflecting potassium interactions with soil texture. Topographic and vegetation indices such as **TGSI**, **CI**, **NCI**, and **TWI** were selected in fewer models, highlighting their relatively lower influence.

Organic Carbon

For Organic Carbon, **latitude** and **longitude** again emerged as top predictors (100%), followed by **precipitation**, **temperature**, and **elevation** (each 95%), reaffirming the spatial and climatic nature of soil carbon content.

Textural and soil information features such as **SI**, **sand05**, **clay515**, and **silt515** showed moderate relevance. Vegetation and terrain indices like **TGSI**, **NCI**, and **TWI** had lower selection percentages, suggesting a relatively minor role in predicting organic carbon at regional scales.

6.7 Comparison with Ground Truth

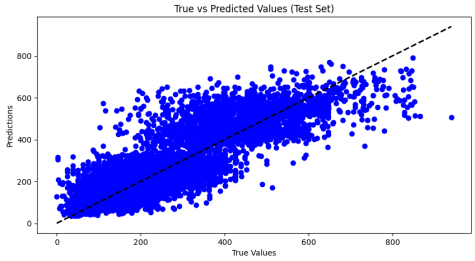
Figure 6.6 presents the predicted versus true plots for all four nutrients for AEZ 8.

The alignment of points along the diagonal line indicates the degree of accuracy. The **Nitrogen** model performed the best in AEZ 8, achieving an R^2 score of 0.7452 and a sMAPE of 16.37. Predictions closely track the ground truth values, especially in the mid-range.

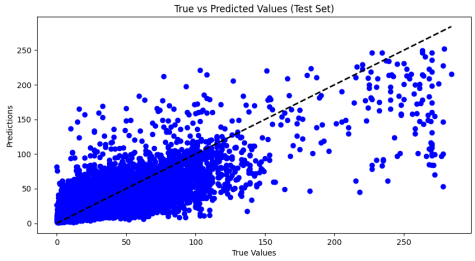
Potassium and **Phosphorus** showed moderate performance, with R^2 scores of 0.6797 and 0.6211, respectively. While both capture the overall trends in the data, dispersion increases at higher true values, especially for Phosphorus, which also had the highest sMAPE of 34.56, suggesting underprediction in high-value regions.

The model for **Organic Carbon** had the lowest R^2 of 0.5221 and an sMAPE of 26.28, reflecting higher residual variance. Still, it broadly captured the distribution of the test data, with underestimation at the lower end and saturation at higher values.

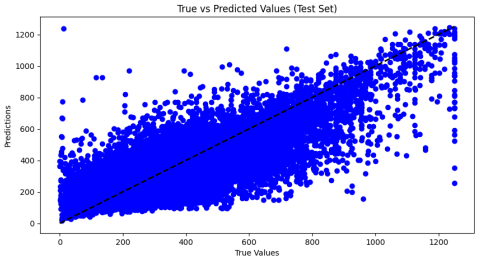
These comparisons underscore that spatial variability and nutrient-specific characteristics affect prediction accuracy. Nevertheless, the models generalize reasonably well across test points, especially for Nitrogen and Potassium.



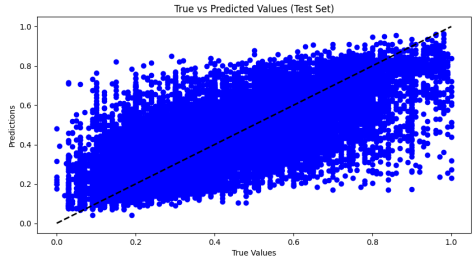
(a) Nitrogen, R^2 : 0.7452



(b) Phosphorus, R^2 : 0.6211



(c) Potassium, R^2 : 0.6797



(d) Organic Carbon, R^2 : 0.5221

Figure 6.6: True vs Predicted values on the test set for Nitrogen, Phosphorus, Potassium, and Organic Carbon in AEZ 8. Diagonal dashed lines indicate ideal predictions.

6.8 Predicted Soil Nutrient Maps

The spatial distribution of predicted soil nutrients - Nitrogen, Phosphorus, Potassium, and Organic Carbon was visualized using the trained Random Forest models for the southern region of India. The predictions were made at a 30-meter spatial resolution using Sentinel-2

and other remotely sensed indices.

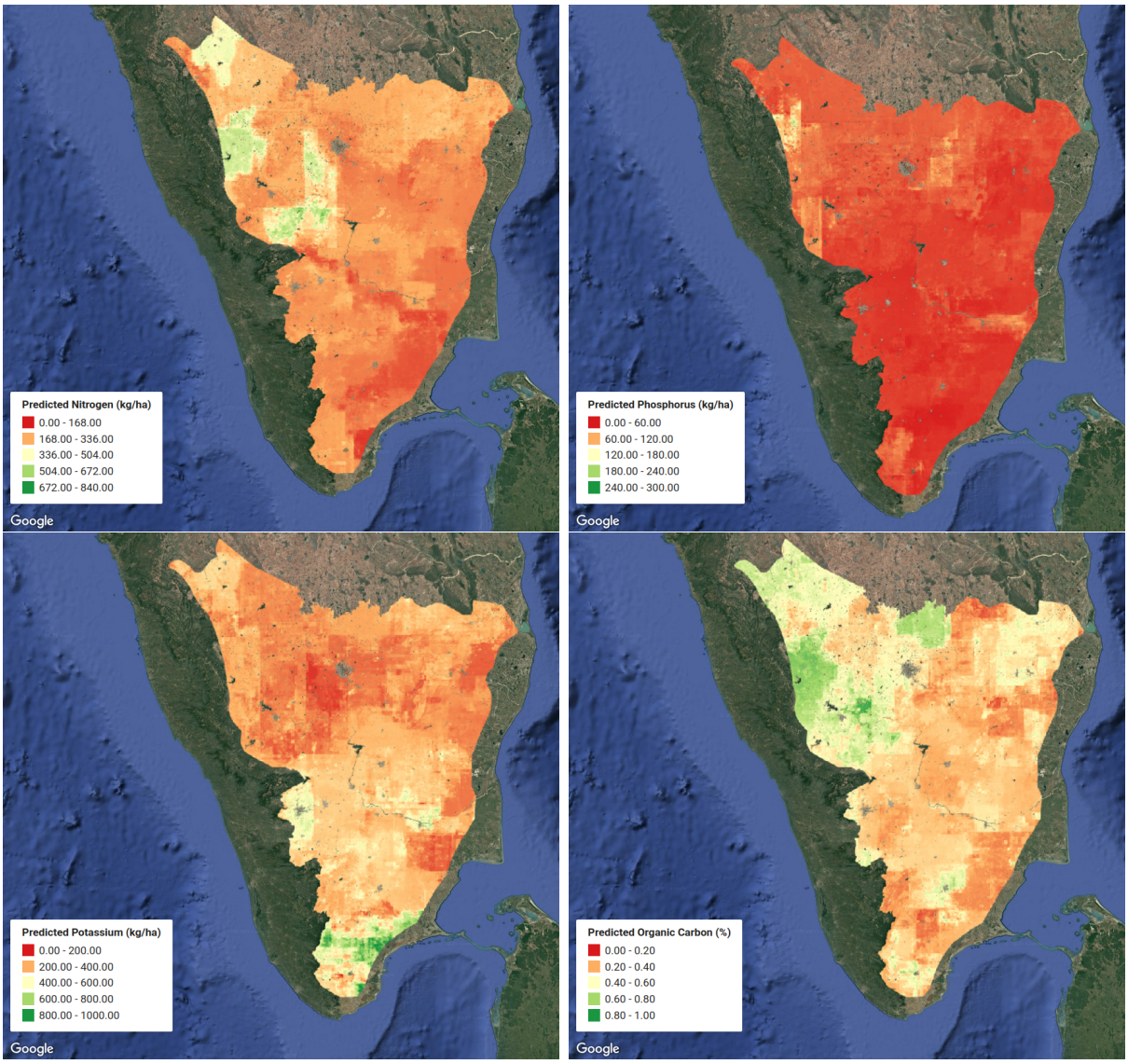


Figure 6.7: Predicted soil nutrient maps for AEZ 8: (Top-Left) Nitrogen (kg/ha), (Top-Right) Phosphorus (kg/ha), (Bottom-Left) Potassium (kg/ha), and (Bottom-Right) Organic Carbon (%).

Distinct spatial patterns are observed across AEZ 8. Higher nitrogen levels appear concentrated in central Karnataka, while potassium is richer in southern Tamil Nadu. Organic carbon shows a strong gradient with higher values in the forested and hilly western regions, indicating the influence of topography and vegetation cover.

These high-resolution prediction maps can aid targeted interventions in soil fertility management and support sustainable agricultural planning specific to AEZ 8.

Chapter 7

Future Work

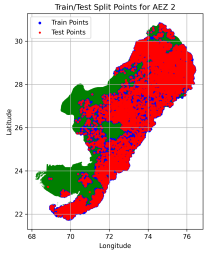
This study has demonstrated the potential of using remote sensing and environmental variables for large-scale soil nutrient prediction. However, several avenues remain open for further exploration and improvement.

- **Incorporation of Time-Series Indices:** Currently, most spectral indices are used as annual means. Future work can explore generating vegetation and moisture indices at a finer temporal scale—such as monthly or seasonal composites. This could help capture crop cycles, rainfall variability, and seasonal vegetation dynamics more accurately, potentially improving prediction performance.
- **Feature Expansion:** Additional features such as land use/land cover (LULC), crop type, irrigation data, historical yield data, or soil microbiome activity can be considered to provide deeper context to the models.
- **Model Enhancement:** Advanced machine learning approaches such as Gradient Boosting (e.g., XGBoost, LightGBM), Quantile Random Forests, Deep Learning, or Hybrid models combining physical and data-driven methods could be explored. These may better capture non-linearities and complex interactions between features.
- **Spatial Generalization:** Cross-regional validation or domain adaptation techniques may help evaluate how well models trained in one AEZ generalize to others, enabling nationwide scale-up with minimal retraining.
- **Integration with Decision-Making Systems:** The generated soil maps can be linked with agronomic advisory platforms to provide actionable insights to farmers. Future work may focus on building such interfaces for practical deployment.

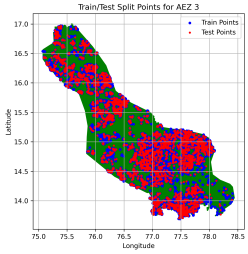
Overall, enhancing the spatiotemporal granularity of inputs, experimenting with new features and models, and integrating uncertainty-aware predictions can further advance the robustness and utility of soil nutrient mapping systems.

Appendix A

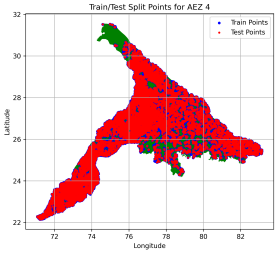
APPENDIX



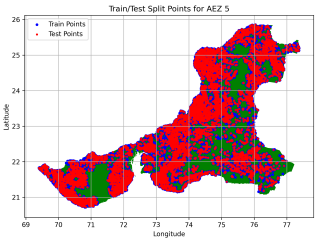
(a) AEZ 2



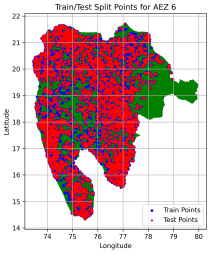
(b) AEZ 3



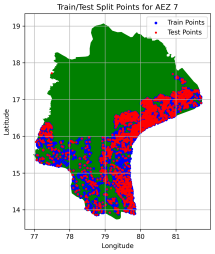
(c) AEZ 4



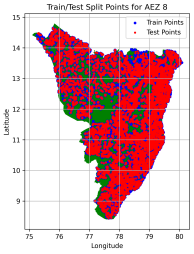
(d) AEZ 5



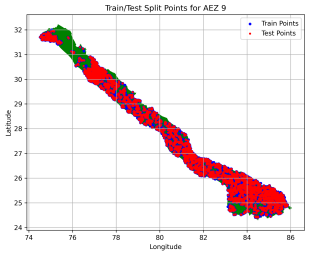
(e) AEZ 6



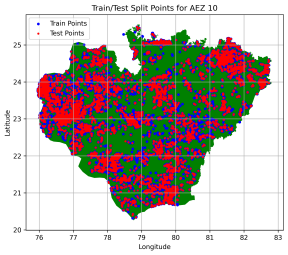
(f) AEZ 7



(g) AEZ 8

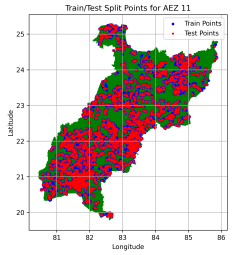


(h) AEZ 9

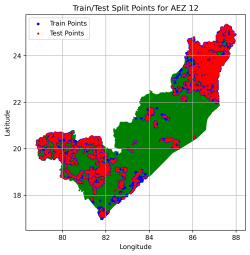


(i) AEZ 10

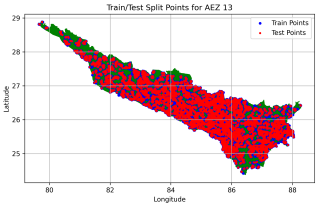
Figure A.1: Train-Test Split Plots for AEZs 2 to 10



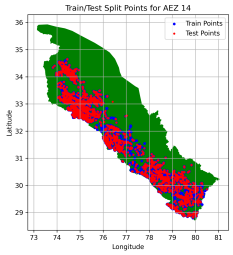
(a) AEZ 11



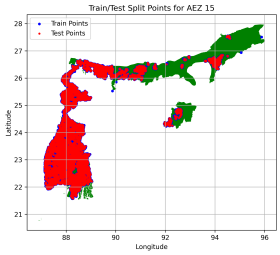
(b) AEZ 12



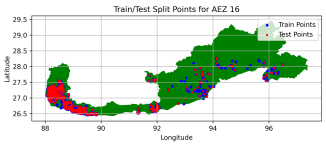
(c) AEZ 13



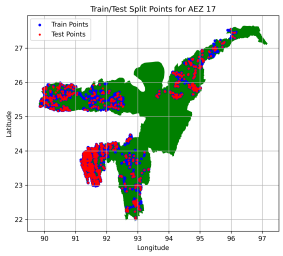
(d) AEZ 14



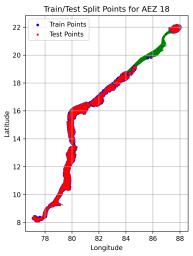
(e) AEZ 15



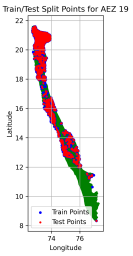
(f) AEZ 16



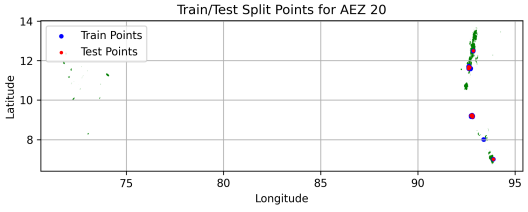
(g) AEZ 17



(h) AEZ 18



(i) AEZ 19



(j) AEZ 20

Figure A.2: Train-Test Split Plots for AEZs 11 to 20

Bibliography

NASA LP DAAC. 2021. Mod11a2.061 terra land surface temperature and emissivity 8-day global 1km. Accessed via Google Earth Engine.

S. Dharumarajan, Rajendra Hegde, and S. K. Singh. 2017. Spatial prediction of major soil properties using random forest techniques—a case study in semi-arid tropics of south india. *Geoderma Regional*, 10:154–162.

European Space Agency (ESA). 2015. Sentinel-2 msi: Multispectral instrument, level-2a. Accessed via Google Earth Engine.

Tom G. Farr, Paul A. Rosen, Edward Caro, Robert Crippen, R. Duren, Scott Hensley, Michael Kobrick, Michael Paller, Ernesto Rodriguez, Lily Roth, David Seal, Stephen Shaffer, Joanne Shimada, Jason Umland, Michael Werner, Michael Oskin, Douglas Burbank, and David Alsdorf. 2007. The shuttle radar topography mission. *Reviews of Geophysics*, 45(2).

Olusegun Folorunso, Oluwafolake Ojo, Mutiu Busari, Muftau Adebayo, Adejumobi Joshua, Daniel Folorunso, Charles Okechukwu Ugwunna, Olufemi Olabanjo, and Olusola Olabanjo. 2023. Exploring machine learning models for soil nutrient properties prediction: A systematic review. *Big Data and Cognitive Computing*, 7(2):113.

Chris Funk, Pete Peterson, Martin Landsfeld, Diego Pedreros, Joel Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, Andrew Hoell, and Joel Michaelsen. 2015. The climate hazards infrared precipitation with stations — a new environmental record for monitoring extremes. *Scientific Data*, 2:150066.

Yufeng Ge, J. Alex Thomasson, and Ruixiu Sui. 2011. Remote sensing of soil properties in precision agriculture: A review. *Frontiers of Earth Science*, 5(3):229–238.

Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27. Big Remotely Sensed Data: tools, applications and experiences.

Tomislav Hengl, Gerard B. M. Heuvelink, Javier M. de Jesus, Ricardo A. Gonzalez, Markus Kilibarda, Bas Kempen, Andrew Shepherd, Bas Lehmann, Edmar R. MacMillan, and Gyorgy Szatmari. 2017. Soilgrids250m: Global gridded soil information based on machine learning. *PLoS ONE*, 12(2):e0169748.

Gunkirat Kaur, Kamal Das, and Jagabondhu Hazra. 2020. Soil nutrients prediction using remote sensing data in western india: An evaluation of machine learning models. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 4677–4680. IEEE.

Ali Keshavarzi, Fuat Kaya, Levent Başayiğit, Yeboah Gyasi-Agyei, Jesús Rodrigo-Comino, and Andrés Caballero-Calvo. 2023. Spatial prediction of soil micronutrients using machine learning algorithms integrated with multiple digital covariates. *Nutrient Cycling in Agroecosystems*, 127:137–153.

Alex B. McBratney, M. L. Santos, and Budiman Minasny. 2003. On digital soil mapping. *Geoderma*, 117(1–2):3–52.

NBSS&LUP (ICAR). 2023. Agro-ecological zones of india: Revised atlas and shapefiles. `https://www.nbsslup.in/`. National Bureau of Soil Survey and Land Use Planning, Indian Council of Agricultural Research (ICAR).

B. Bhanukiran Reddy, Maragatham S., Santhi R., Balachandar D., Vijayalakshmi D., Davamani V., Vasu D., and Gopalakrishnan M. 2024. Predictive soil mapping using random forest models: Applications in ph and soil organic matter assessment. *Plant Science Today*, 11(4):463–474.

Soil Health Division, Department of Agriculture and Farmers Welfare. 2025. Soil health card - slusi visualisation portal.