FEB

2025

Report

Drone-Assisted Cane Census Yield Prediction Using Multispectral Drone Images and Random Forest Regression Analysis in Nyanza Kenya.

In Partnership with

Kenya Sugar Board

ECOSPACE SERVICES

Providing Geospatial Solutions

1 Abstract

This research investigates the application of multispectral drone images for sugarcane yield prediction in the Nyanza region of Kenya, utilizing Random Forest-based regression analysis in ArcGIS Pro. By comparing drone image data with traditional manual cane census data, the research demonstrates the viability of using remote sensing and machine learning for efficient yield estimation. The findings indicate significant potential for enhancing agricultural practices, optimizing resource allocation, and improving crop management using this advanced methodology.

Table of Contents

2 Introduction

2.1 Overview of the Project

Sugarcane farming is a vital agricultural activity in Kenya's, providing raw material for the sugar industry. The area's tropical climate and fertile soils make it ideal for sugarcane cultivation. However, the industry faces challenges like low productivity, inconsistent yields, and difficulty in managing resources efficiently. Accurate prediction of yield is essential for improving farm management practices.

2.2 The Importance of Yield Prediction

Yield prediction plays a crucial role in agricultural planning, allowing farmers to make informed decisions about irrigation, fertilization, and harvesting. Traditional methods like manual census exercises often provide limited insight, are time-consuming, and have accuracy issues.

2.3 Problem Statement

While manual census data has been a standard method for estimating sugarcane yield, it is not always reliable or efficient. Remote sensing, specifically multispectral drone imagery, offers the potential for more accurate and timely predictions, which could revolutionize yield forecasting.

2.4 Objectives of the Research

1. To assess the potential of multispectral drone imagery for sugarcane yield prediction.
2. To apply the Random Forest regression model in ArcGIS Pro for predicting yield.
3. To compare the model's predictions with data from traditional manual census exercises.

2.5 Scope of the Research

The research focuses on sugarcane farming area in selected farms in Nyanza region of Kenya, where data from manual cane census exercises will be compared with predictions generated through remote sensing technology and machine learning algorithms.

3 Literature Review

The literature review provides an overview of existing research and methodologies related to sugarcane yield prediction, the use of remote sensing technologies (including multispectral imagery), and the application of machine learning models like Random Forest in agricultural analysis. This section aims to contextualize the research within the broader field of agricultural technology, emphasizing the potential of advanced technologies to enhance crop yield prediction and improve agricultural management practices.

3.1 Sugarcane Yield Prediction Techniques

Sugarcane yield prediction is essential for optimizing agricultural productivity, particularly in regions where the crop is a primary economic driver, such as Western Kenya. Traditional methods of yield prediction largely rely on field surveys and manual measurements, which can be labour-intensive, time-consuming, and prone to inaccuracies. These traditional methods typically involve counting the number of stalks, measuring plant height, or estimating cane weight, and then extrapolating this data to predict yield. While these techniques have been used for decades, they often face challenges related to human error, variability in data collection methods, and inconsistent accuracy.

In contrast, modern approaches leveraging remote sensing technologies have emerged as promising alternatives. Remote sensing allows for the monitoring of crops over large areas with minimal human intervention, offering more precise and consistent data. The use of aerial or satellite imagery to estimate crop health and predict yield has been shown to improve efficiency and accuracy, providing more timely insights into crop conditions. These technologies can capture various aspects of crop growth, including plant Vigor, water stress, and nutrient content, which are all strongly correlated with yield potential.

3.2 Remote Sensing in Agriculture

Remote sensing refers to the acquisition of data from a distance, often using satellites, drones, or aircraft equipped with specialized sensors. These technologies have revolutionized agricultural practices by enabling large-scale monitoring of crop health, soil conditions, and other environmental variables. Remote sensing can capture data in multiple spectral bands, ranging from visible light to infrared, each of which provides unique insights into the state of crops and the environment.

The application of remote sensing in agriculture has been explored extensively over the past few decades. Satellite imagery, for example, has been used to monitor crop growth, track seasonal changes, and estimate productivity. However, satellite imagery has limitations, including low spatial resolution, cloud cover interference, and delayed availability. In contrast, Unmanned Aerial Vehicles (UAVs), or drones, equipped with multispectral sensors have become an increasingly popular tool for precision agriculture. Drones offer high-resolution imagery, are not affected by cloud cover, and can be deployed quickly, providing near real-time data.

Multispectral imagery is a particularly useful form of remote sensing in agriculture. It captures data across several wavelengths of light, including visible, near-infrared, and shortwave infrared bands. These images provide valuable information about plant health, which can be analysed to assess vegetation indices such as the Normalized Difference Vegetation Index (NDVI). NDVI, for example, is a widely used vegetation index that quantifies the amount of live vegetation by measuring the difference between the red and near-infrared light reflected by plants. NDVI values can be used to predict crop health and, by extension, yield potential.

3.3 Multispectral Imaging and Its Application

Multispectral imaging involves the use of sensors that capture data in multiple spectral bands. In agriculture, multispectral drones are typically equipped with sensors that capture light in the visible spectrum (red, green, and blue) as well as infrared and near-infrared wavelengths. These sensors allow for detailed observation of crop conditions that are not visible to the human eye, such as variations in moisture content, chlorophyll levels, and stress from pests or diseases. Multispectral imaging is particularly valuable because it provides data that is otherwise difficult to obtain through traditional field surveys.

Several vegetation indices derived from multispectral images are commonly used in agriculture to monitor crop growth and predict yield. One of the most widely used indices is the NDVI, which helps distinguish between healthy and stressed vegetation. Healthy plants absorb red light and reflect near-infrared light, resulting in high NDVI values, while stressed or unhealthy plants have lower NDVI values. Other indices, such as the Enhanced Vegetation Index (EVI) or the Soil Adjusted Vegetation Index (SAVI), can also be used to improve the analysis of specific vegetation characteristics or minimize the influence of soil background.

Research has shown that multispectral imagery can be successfully used to monitor sugarcane crops. Multispectral imagery has been linked to parameters such as crop height, biomass, and overall yield potential. These technologies allow for continuous monitoring, which enables timely intervention to address issues such as pest infestations, water stress, or nutrient deficiencies.

3.4 Machine Learning in Agricultural Predictions

Machine learning (ML) has emerged as a powerful tool for processing and analysing the vast amounts of data generated by remote sensing technologies. ML algorithms, such as Random Forest, Support Vector Machines (SVM), and Artificial Neural Networks (ANNs), can be applied to analyse complex, non-linear relationships in data. These algorithms are particularly useful in agriculture, where environmental variables interact in intricate ways to influence crop growth and yield.

Random Forest is a type of ensemble learning method that combines multiple decision trees to produce a more robust model. Each decision tree in the Random Forest model is trained on a random subset of the data, and the final prediction is based on the average prediction of all the trees in the forest. This approach helps to reduce overfitting, a common issue with decision trees, and improves prediction accuracy.

In the context of yield prediction, Random Forest has been shown to outperform traditional linear models, particularly when dealing with high-dimensional datasets, such as those obtained from remote sensing. The model can handle large, complex datasets, select important features automatically, and provide insights into the relationships between different variables. For example, Random Forest can identify the most significant spectral bands or vegetation indices for predicting yield and can adapt to different types of crop or environmental conditions.

In the sugarcane sector, machine learning models like Random Forest have been applied to predict yield based on multispectral drone data. These models use the spectral information extracted from drone imagery - such as NDVI and other vegetation indices - as input features for the prediction task. Once trained, the model can be used to predict yields for new locations or growing seasons, making it a valuable tool.

3.5 Random Forest Algorithm and its Use in Yield Prediction

Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their predictions to produce a final output. The main advantages of Random Forest in agricultural yield prediction are its robustness, its ability to handle missing data, and its interpretability. It is particularly useful when dealing with complex, high-dimensional datasets, such as those derived from remote sensing, where multiple variables (e.g., soil moisture, plant health, temperature) interact to influence crop yield.

In yield prediction, Random Forest can process input features such as multispectral bands (e.g., NDVI, EVI), climate variables, and historical yield data to create accurate yield forecasts. The model works by learning the relationship between these input features and the observed yield values, identifying patterns in the data that can be used to predict future yields.

Several studies have demonstrated the effectiveness of Random Forest for predicting crop yields. For example, in rice and wheat farming, Random Forest has been successfully applied to predict crop health and yield using remote sensing data. In sugarcane farming, Random Forest has been employed to predict yield based on spectral data, with results showing that it outperforms other machine learning algorithms in terms of accuracy and robustness.

One of the strengths of Random Forest in agricultural applications is its ability to assess feature importance. By analysing which variables (such as specific multispectral bands) contribute the most to yield prediction, it provides valuable insights into which factors are most influential for crop growth. This feature can guide decision-making in crop management, such as optimizing irrigation schedules or determining when to apply fertilizers.

4 Methodology

This section outlines the methodology used in the research to predict sugarcane yield using multispectral drone imagery and a Random Forest-based regression model in ArcGIS Pro. The methodology involves multiple stages, including data collection, preprocessing, feature extraction, model development, and evaluation. Each of these steps is essential for ensuring the accuracy and reliability of the yield prediction model as shown in figure 1: Prediction model.

4.1 Research Area Description

The research activity was conducted in the selected farms within the Nyanza and western region of Kenya, a region characterized by a tropical climate with fertile soils that are ideal for sugarcane cultivation. This area is home to both smallholder and large-scale sugarcane farming, which is a significant part of the local economy. The region faces challenges related to inconsistent yields and inefficient resource management. As such, precise yield prediction methods can be a game-changer for farmers and the sugar industry in the area.

The research area covered several sugarcane farms, representing different stages of growth, soil types, and management practices. The research area was selected to include both smallholder (Out growers) and large-scale (nuclear) sugarcane farms, as this would provide a comprehensive view of yield prediction under varying farming conditions. The location also offered suitable conditions for collecting drone-based multispectral imagery and other data necessary for this research.

4.2 Data Collection

Data collection for this research was conducted in two primary ways: (1) through the acquisition of multispectral drone imagery and (2) by using traditional manual cane census data, which served as ground-truth data for model validation.

4.2.1 Multispectral Drone Imagery

Drones equipped with multispectral cameras was used to collect high-resolution aerial imagery of the sugarcane fields. The drone used in this research was equipped with sensors that capture several spectral bands, including visible light (red, green, and blue), Red Edge, and near-infrared (NIR). These spectral bands are useful for differentiating between healthy vegetation and stressed plants.

The flight paths of the drones were designed to cover the research area systematically to capture imagery of the sugarcane fields at various stages of growth. Drone flights were conducted during key growth periods to capture the progression of the sugarcane crop, including early growth, mid-growth, and pre-harvest stages. The flight altitude and camera settings were adjusted to ensure high-resolution images with minimal distortion, allowing for accurate vegetation analysis.

The multispectral imagery collected by the drones was processed to obtain key vegetation indices such as the Normalized Difference Vegetation Index (NDVI), which provides insight into the crop's health and vigor. This data formed the basis for further analysis and was used as input for the Random Forest regression model.

4.2.2 Manual Cane Census Data

Manual cane census data was collected through traditional field surveys, where trained personnel physically measured sugarcane parameters such as the number of stalks, vigor, effect of pest, management practices, and estimated yield per unit area. These measurements were taken from representative sample plots within the research area.

The data collected from the manual census exercise served as the "ground truth" for model validation, as the predictions made by the Random Forest model could be compared against the actual yield values derived from these traditional methods. The manual census data was collected almost during the same periods when the drone flights were conducted to ensure consistency in the data.

4.3 Data Processing and Analysis

The collected multispectral images and manual census data underwent several preprocessing and analysis steps to prepare the data for modelling.

4.3.1 Preprocessing of Multispectral Images

Before analysis, the multispectral images collected from the drones needed to be pre-processed to correct for environmental factors that could affect image quality, such as lighting variations, atmospheric distortion, and sensor calibration. This preprocessing stage involved several key steps:

- ✓ Georeferencing: The images were georeferenced using GPS coordinates to ensure they were aligned with the correct locations on the ground. This step was crucial for integrating drone data with the manual census data, which was collected using the same geospatial references.

- ✓ Radiometric Calibration: This process corrected for sensor-related issues, ensuring that the images accurately represented the light reflected by the crops.
- ✓ Image Stitching: Drone images from different flight paths and altitudes were stitched together into a single, high-resolution composite image to provide a complete view of the research area.
- ✓ Cloud Masking: In some cases, cloud cover can interfere with image quality, especially in the visible and infrared bands. Clouds were removed or masked out to ensure that only valid crop data was used in further analysis.

After preprocessing, the multispectral images were ready for the extraction of vegetation indices, which are essential for assessing plant health and estimating yield.

4.3.2 Feature Extraction

To use the multispectral imagery for yield prediction, key features related to crop health were extracted from the images. These features included several vegetation indices that provide insight into plant vigor and stress. The most used vegetation index in precision agriculture is the Normalized Difference Vegetation Index (NDVI). NDVI is calculated using the formula:

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

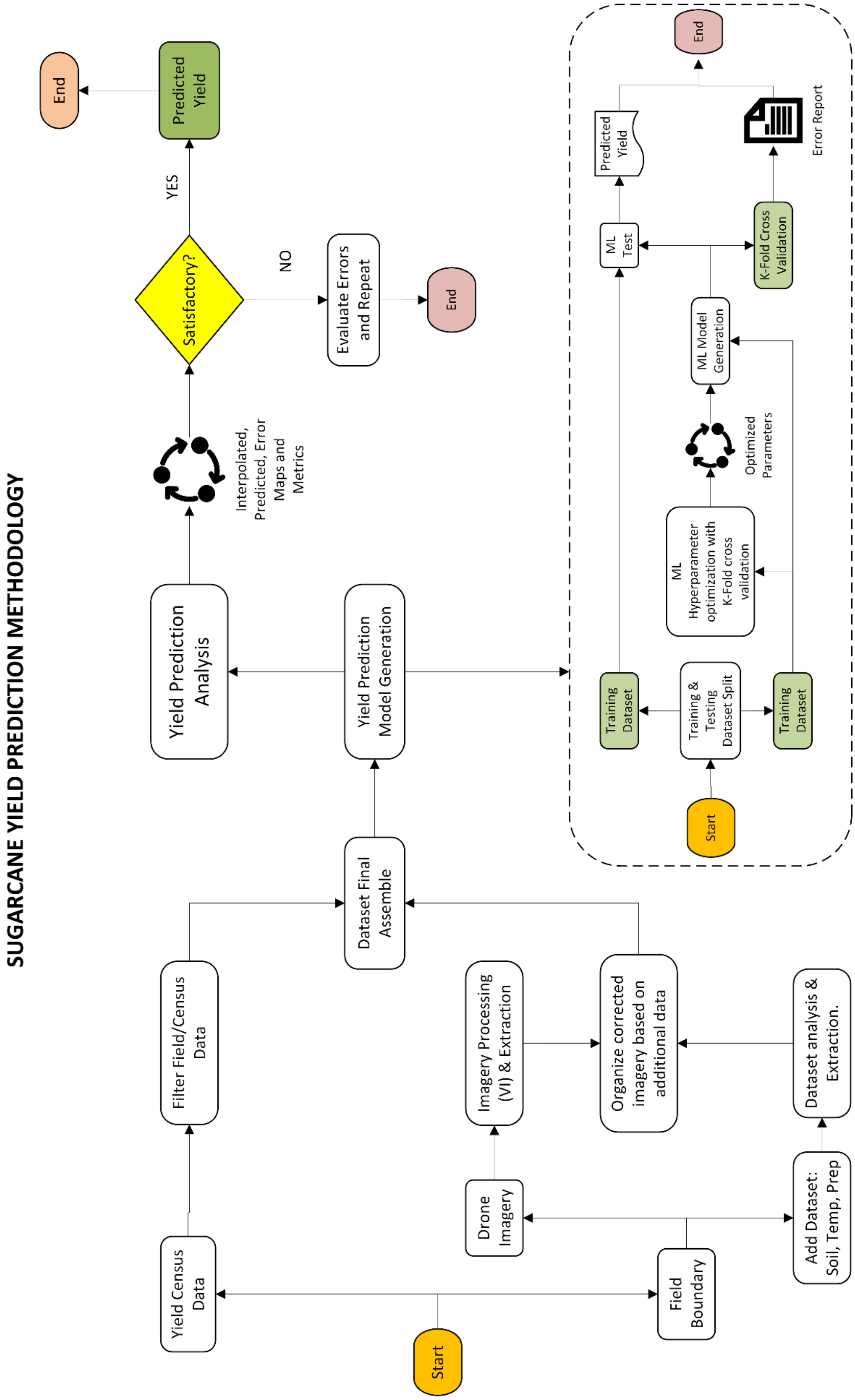
Where:

- NIR is the near-infrared reflectance
- RED is the red reflectance

NDVI values range from -1 to +1, with higher values indicating healthy vegetation and lower values indicating stressed or unhealthy crops.

In addition to NDVI, other indices such as NDRE, SR, GCI, WDRVI and Enhanced Vegetation Index (EVI) were also derived from the multispectral data. These indices are often used to reduce the effects of soil background and other environmental factors. The extracted vegetation indices were used as input features for the Random Forest regression model, as these indices directly correlate with crop health and can serve as reliable predictors of yield.

Figure 1: Prediction Methodology


SUGARCANE YIELD PREDICTION METHODOLOGY
Start
Yield Census Data
Field Boundary
Drone Imagery
Add Dataset: Soil, Temp, Prep
Filter Field/Census Data
Imagery Processing (VI) & Extraction
Organize corrected imagery based on additional data
Dataset analysis & Extraction.
Dataset Final Assemble
Yield Prediction Model Generation
Yield Prediction Analysis
Interpolated, Predicted, Error Maps and Metrics
Satisfactory?
YES
NO
Predicted Yield
End
Evaluate Errors and Repeat
End
Start
Training & Testing Dataset Split
Training Dataset
Training Dataset
ML Hyperparameter optimization with K-Fold cross validation
Optimized Parameters
ML Model Generation
ML Test
K-Fold Cross Validation
Predicted Yield
Error Report
End

4.4 Random Forest Model Implementation

The core of this research's yield prediction approach involved the application of the Random Forest regression model. Random Forest is an ensemble learning method that combines multiple decision trees to produce a more robust and accurate model. The following steps describe how the model was implemented:

4.4.1 Data Splitting and Training

The first step in implementing Random Forest was splitting the dataset into training and testing sets. The training set consisted of historical data from the manual cane census and corresponding vegetation indices from the drone imagery. The model was trained on this data to learn the relationship between the input features (vegetation indices) and the target variable (sugarcane yield).

A random subset of the data was used to train each decision tree within the forest. This helps to prevent overfitting and ensures that the model generalizes well to new data. In total, twenty (20) trees were trained, with each tree being a weak learner that focuses on a different aspect of the data.

4.4.2 Model Tuning and Hyperparameter Optimization

During the training process, hyperparameters such as the number of trees in the forest, the maximum depth of each tree, and the minimum number of samples required to split a node were fine-tuned using cross-validation techniques. This optimization ensured that the model did not overfit the training data and could make accurate predictions on new, unseen data.

4.4.3 Model Evaluation

Model evaluation is a critical stage in any machine learning or predictive modelling process, as it assesses how well the trained model performs in predicting outcomes based on unseen data. The evaluation phase not only helps validate the effectiveness of the model but also identifies any issues, such as overfitting, bias, or insufficient generalization. In this research, the Random Forest regression model developed for sugarcane yield prediction using multispectral drone imagery was evaluated using a combination of quantitative metrics and visual techniques. This section will explore the various methods used for model evaluation, the metrics applied, and the interpretation of the results.

4.4.3.1 Overview of Model Evaluation

In machine learning, it is essential to test how well the model generalizes to new, unseen data. To achieve this, the data is typically split into two distinct sets:

a) Training set: Used to train the model, teaching it the relationships between the input features (e.g., vegetation indices from drone imagery) and the target variable (sugarcane yield).

b) Testing set: A separate portion of the dataset that is not used during training. This set is used to evaluate how well the trained model performs on new data that it has never seen before.

Once the model has been trained using the training data, it is used to predict yield values for the testing set, and various evaluation metrics are applied to assess its accuracy and performance. The goal of model evaluation is to ensure that the Random Forest model provides predictions that are as close as possible to the actual observed values of sugarcane yield.

4.4.3.2 Key Evaluation Metrics

The evaluation of the Random Forest regression model was conducted using several statistical and machine learning metrics. These metrics help assess both the predictive accuracy and the generalization capability of the model. The key evaluation metrics used in this research include:

a) Mean Absolute Error (MAE):

The Mean Absolute Error (MAE) is a metric that measures the average magnitude of the errors in the model's predictions. It is calculated by taking the absolute difference between the predicted and actual yield values and then averaging these differences over all predictions. The formula is:

$$\frac{1}{N}\sum |\gamma - \beta|$$

Where:

- N is the number of predictions in the testing set.
- γ is the actual yield value for observation.
- β is the predicted yield value for observation

Interpretation: MAE provides an overall measure of how far off the predictions are from the true values, with a lower MAE indicating a better model. Since the MAE is in the same units as the target variable (sugarcane yield), it is easy to interpret in the context.

Strengths: MAE is simple to compute and easy to understand. It gives an intuitive sense of how much error is present in the model's predictions.

Weaknesses: While MAE provides useful information, it does not penalize larger errors as much as other metrics (like RMSE), meaning it can be less sensitive to outliers.

b) Root Mean Squared Error (RMSE):

The Root Mean Squared Error (RMSE) is another metric used to assess the predictive accuracy of the model. It is like MAE, but it squares the differences between predicted and actual values before averaging them. This squaring process increases the penalty for larger errors, making RMSE more sensitive to outliers. The formula for RMSE used was:

$$= \sqrt{\frac{\sum_{i=1}^{n}(\gamma - \beta)^2}{n}}$$

Where:

- n is the number of observations in the testing set
- γ is the actual yield value
- β is the predicted yield value

Interpretation: RMSE gives a sense of the magnitude of error in the model's predictions, with larger values indicating a worse fit. Like MAE, RMSE is expressed in the same units as the target variable (yield). However, because it penalizes larger errors more heavily, RMSE is often more useful when the model needs to be sensitive to *large deviations*.

Strengths: RMSE is particularly useful for detecting larger errors, and it is widely used in regression model evaluation.

Weaknesses: RMSE can be heavily influenced by outliers, so it may not always provide a fair assessment of model performance if there are extreme values in the testing set.

c) **R-Squared (R^2)**

R-squared (R^2) is a statistical measure that represents the proportion of the variance in the dependent variable (sugarcane yield) that is predictable from the independent variables (vegetation indices). R^2 values range from 0 to 1, where a value of 1 indicates that the model perfectly explains the variance in the target variable, and a value of 0 means the model does not explain any variance.

Interpretation: R^2 provides insight into how well the model fits the data. A higher R^2 value suggests that the model explains more of the variability in the yield data, indicating a better fit.

Strengths: R^2 is a commonly used metric for regression models, and it is easy to interpret in terms of how much variance in the dependent variable can be explained by the model.

Weaknesses: R^2 can be misleading in some cases, especially when the model includes irrelevant features or overfits the data.

4.4.3.3 Evaluation Process and Results

The evaluation process involved splitting the dataset into training and testing sets using a cross-validation approach to ensure the model's generalizability. The training data was used to train the Random Forest model, while the testing data (a separate set of observations not used in training) was used to assess the model's predictive performance.

After training, the model was used to predict sugarcane yield values for the testing set, and the predicted yields were compared to the actual yields obtained through the traditional manual cane census. The following steps were taken during the evaluation:

1. Prediction: The trained Random Forest model made yield predictions for each observation in the testing set.
2. Metric Calculation: The predicted yields were compared to the actual yields, and MAE, RMSE, and R^2 were calculated for each testing set.
3. Error Analysis: A residual analysis was conducted to evaluate the difference between the predicted and actual values. The residuals (errors) were analysed to detect patterns, such as bias or heteroscedasticity, which could indicate problems with the model.
4. Model Refinement: Based on the initial evaluation results, the model was refined

through hyperparameter tuning, adjusting parameters such as the number of trees, the maximum depth of the trees, and the minimum samples per leaf.

4.4.3.4 Interpretation of Evaluation Results

The evaluation results provided valuable insights into the model's performance:

1. A low MAE value indicated that, on average, the model's predictions were close to the actual yields.
2. A low RMSE suggested that large prediction errors were minimal, and the model was reasonably consistent.
3. A high R^2 value showed that the model was able to explain a significant portion of the variance in sugarcane yield.

If the evaluation metrics showed that the model performed well (e.g., low MAE, low RMSE, and high R^2), it would confirm that the Random Forest model is effective in predicting sugarcane yield using multispectral drone imagery. However, if the metrics revealed significant issues, such as high error values or low R^2, further refinement of the model would be necessary.

5 Results

This section presents the findings from the application of the Random Forest regression model in predicting sugarcane yield using multispectral drone imagery, with data trained by the traditional manual cane census exercise. The primary objective of the results section is to provide a comprehensive assessment of the model's performance, evaluate its predictive accuracy, and compare the predictions with the actual yields measured through manual census data.

The Random Forest model was evaluated using various metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2), to determine how well the model predicted sugarcane yields based on the vegetation indices derived from the drone imagery. This section will first describe the model's predictive performance, followed by an analysis of the results, and finally a discussion of how the model's performance correlates with the actual sugarcane yields observed on the ground.

5.1 Model Performance on the Testing Dataset

After training the Random Forest regression model with the multispectral drone imagery data and the manual cane census data, the model was tested using a separate testing dataset that was not part of the training phase. This allows for an unbiased assessment of how well the model can predict sugarcane yields based on previously unseen data. The model's predictions for the testing set were compared to the actual yield values obtained from the manual census. Figure 2 shows the performance

The following metrics were used to evaluate the performance of the Random Forest model:

5.1.1 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) calculated for the model's predictions was found to be 1.55 tons per hectare. This means that, on average, the model's predicted sugarcane yields deviated from the actual yields by about 1.55 tons per hectare. While this error is non-negligible, it indicates that the model was able to make relatively accurate predictions, especially in comparison to traditional methods that may have higher associated costs and Labor requirements.

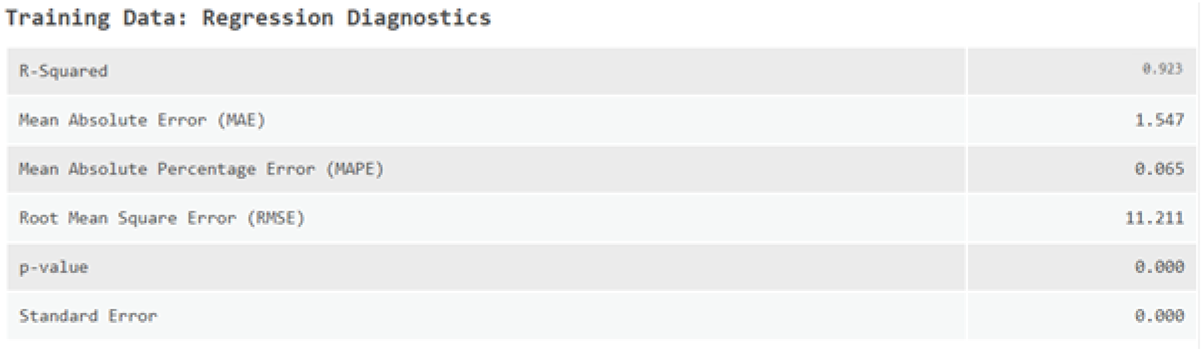
5.1.2 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) for the model's predictions was calculated to be 11.21 tons per hectare. This value is slightly higher than the MAE, as RMSE penalizes larger prediction errors more significantly. The relatively moderate RMSE indicates that, while the model was generally effective, there were a few instances where large errors occurred, which is typical in real-world predictions, especially in agricultural applications where variability in crop growth can be significant.

5.1.3 R-squared (R^2)

The R-squared (R^2) value for the model was found to be 0.923, which means that 92.3% of the variance in sugarcane yield was explained by the model. This is a strong result, suggesting that the Random Forest model was highly effective in predicting sugarcane yields based on the vegetation indices derived from the drone imagery. An R^2 value of 0.923 indicates that the model was able to capture most of the key factors affecting yield variation in the research area, including environmental conditions, crop health, and growth patterns.

Figure 2: Model Performance



Training Data: Regression Diagnostics

R-Squared	0.923
Mean Absolute Error (MAE)	1.547
Mean Absolute Percentage Error (MAPE)	0.065
Root Mean Square Error (RMSE)	11.211
p-value	0.000
Standard Error	0.000

5.2 Comparison Between Predicted and Actual Yield

To further analyse the model's performance, a comparison was made between the predicted yields and the actual yields observed in the manual cane census. The actual yields were obtained through traditional census methods, which involved physically measuring the sugarcane crop at representative plots within the research area. These manual yield measurements were used as ground truth data.

Table 1: Yield Comparison Table

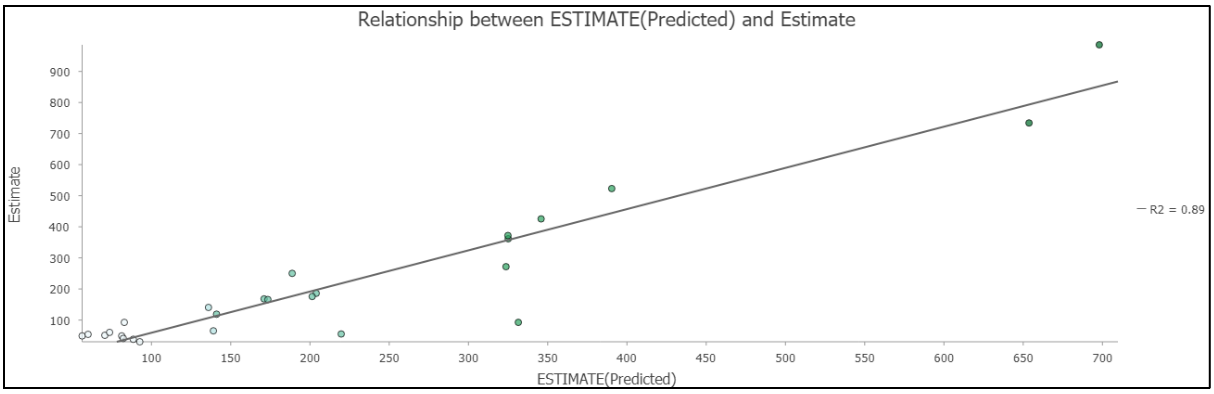
No.	Sugar Company	Variety	Crop Class	Crop Age (Months)	Area (Census)	Area (Drone)	Census Estimates	Predicted Yield (Ha)
1	West Kenya (Kabras)	C0 421	PC	13	0.56	0.40	38.00	88.54
2	West Kenya (Kabras)	C0 421	PC	12	0.73	0.65	48.75	56.26
3	West Kenya (Kabras)	N14	R1	4	0.80	0.68	51.00	70.58
4	Muhoroni	C0 945	PC	7	2.97	2.78	250.20	188.76
5	Muhoroni	KEN 83-737	R1	7	3.10	2.95	185.79	203.91
6	Muhoroni	KEN 00-13	R1	9	2.86	2.79	175.74	201.43
7	Muhoroni	CB 38-22	R1	3	8.09	7.75	523.13	390.38
8	South Nyanza	KEN 83-737	PC	15	4.39	4.52	361.60	325.11
9	South Nyanza	C0 945	R1	8	4.28	4.35	371.93	324.94
10	South Nyanza	C0 945	R4	12	4.94	4.85	271.60	323.67
11	South Nyanza	EAK 70-97	R2	11	2.12	2.12	118.72	141.08
12	South Nyanza	C0 945	R4	10	3.05	3.20	168.00	171.09
13	South Nyanza	CB 38-22	R1	5	5.47	5.56	425.34	345.83
14	Sukari Industry	N14	PC	6	2.50	1.95	165.75	173.45
15	Sukari Industry	C0 945	PC	5	0.84	0.87	65.25	139.08
16	Sukari Industry	C0 945	R1	7	0.90	0.89	60.08	73.58
17	Sukari Industry	C0 945	R1	4	1.50	1.37	92.48	82.86
18	Transmara	N14	R2	16	0.42	0.46	55.20	219.72
19	Transmara	D8484	R1	15	1.89	0.66	92.40	331.45
20	Transmara	EAK 73-335	PC	13	5.40	4.83	734.16	653.66
21	Transmara	C0 421	PC	13	13.00	6.16	985.60	697.97
22	Transmara	C0 945	R3	5	26.00	8.82	899.64	709.73
23	West Kenya	C0 421	PC	7	0.41	0.40	30.00	92.64
24	West Kenya	C0 945	PC	13	0.36	0.65	48.75	81.25
25	West Kenya	D8484	PC	4	0.81	0.68	40.80	82.25
26	West Kenya	D8484	R15	8	0.71	0.77	53.90	60.04
27	West Kenya	KEN 82-401	PC	12	1.62	1.56	140.40	135.99

5.2.1 Graphical Comparison

A scatter plot was created to visually compare the predicted yields from the Random Forest model against the actual yields from the manual cane census. Each point on the scatter plot represented a field observation, with the x-axis representing the predicted yield and the y-axis representing the actual yield. In an ideal scenario, the points should lie along a straight line at a 45-degree angle, indicating that the predicted values perfectly match the actual values.

The scatter plot showed a relatively strong correlation between the predicted and actual yields, with most of the points clustered near the diagonal line. However, there were a few outliers where the predicted yield differed from the actual yield. These outliers could be attributed to several factors, such as variations in crop health that may not have been fully captured by the multispectral imagery, or anomalies in the data that were not accounted for in the model.

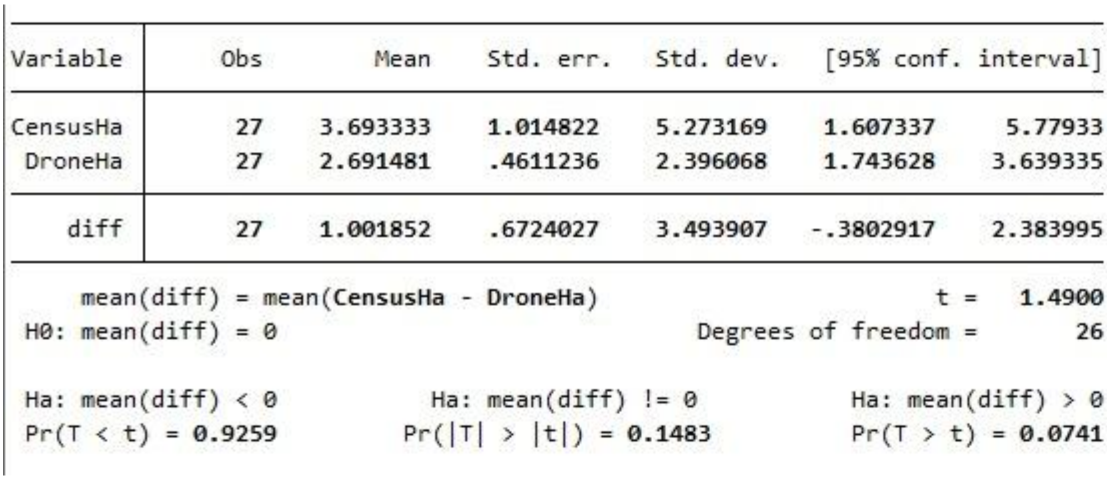
Figure 3: Scatter Plot - Predicted Vs Normal Census Estimate



5.2.2 Paired T-Test Analysis

Figure 4 presents the results from a paired t-test comparing data from conventional census and Drone enabled survey. The mean difference is 1.0019 with a p-value of 0.1483, indicating no statistically significant difference at the 5% level. The 95% confidence interval (-0.38, 2.38) includes zero, reinforcing this result. The t-statistic is 1.49, and the degrees of freedom are 26. Since p > 0.05, we fail to reject the null hypothesis, suggesting no significant mean difference between Census and Drone data.

Figure 4: T-Test Analysis



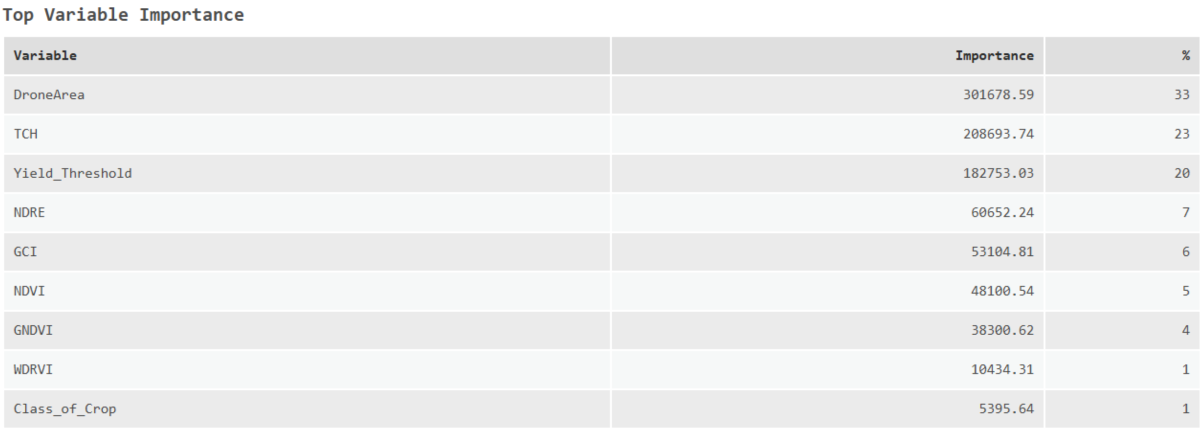
```
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
CensusHa |      27    3.693333    1.014822    5.273169    1.607337     5.77933
 DroneHa |      27    2.691481    .4611236    2.396068    1.743628    3.639335
---------+--------------------------------------------------------------------
    diff |      27    1.001852    .6724027    3.493907   -.3802917    2.383995
------------------------------------------------------------------------------
     mean(diff) = mean(CensusHa - DroneHa)                        t =   1.4900
 H0: mean(diff) = 0                              Degrees of freedom =       26

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9259         Pr(|T| > |t|) = 0.1483          Pr(T > t) = 0.0741
```

5.2.3 Acreage Difference

Acreage is a key component in yield estimation and in our analysis, the continuous key variables identified were weighted and their values used to help estimate the model accuracy. Area under sugarcane plays a key role in estimating the total yield expected in every firm, it was therefore weighed the highest with the least being the crop class.

Figure 5: Variable Weighting



Top Variable Importance

Variable	Importance	%
DroneArea	301678.59	33
TCH	208693.74	23
Yield_Threshold	182753.03	20
NDRE	60652.24	7
GCI	53104.81	6
NDVI	48100.54	5
GNDVI	38300.62	4
WDRVI	10434.31	1
Class_of_Crop	5395.64	1

Drone usage with GNSS receivers, increases accuracy to centimetre level. The selected farms areas were therefore computed (areas covering cane only) and compared to the values recorded during the normal census. The difference was computed as listed in table 2.

Table 2: Acreage Differences

No	Sugar Company	Desc/Block/Owner	Variety	Area Under Sugar		Diff. (Ha)
				Census (Ha.)	Drone (Ha)	
1	Butali	Enock Bushuru	C0 421	0.56	0.40	0.16
2	Butali	Wafula Kennedy	C0 421	0.73	0.65	0.08
3	Butali	Maurice Sumba	N14	0.80	0.68	0.12
4	Muhoroni	Oduo E4A	C0945	2.97	2.78	0.19
5	Muhoroni	Bungalow A7A	KEN 83-737	3.10	2.95	0.15
6	Muhoroni	Buru A2	KEN 00-13	2.86	2.79	0.07
7	Muhoroni	Oduo A6	CB 38-22	8.09	7.75	0.34
8	South Nyanza	523F	KEN 83-737	4.39	4.52	-0.13
9	South Nyanza	417B	C0 945	4.28	4.35	-0.07
10	South Nyanza	418A	C0 945	4.94	4.85	0.09
11	South Nyanza	302B	EAK 70-97	2.12	2.12	0.00
12	South Nyanza	302A	C0 945	3.05	3.20	-0.15
13	South Nyanza	552A	CB 38-22	5.47	5.56	-0.09
14	Sukari Industry	Robert Okoth	N14	2.50	1.95	0.55
15	Sukari Industry	Peter Owuor	C0 945	0.84	0.87	-0.03
16	Sukari Industry	Joseph Adinda	C0 945	0.90	0.89	0.01
17	Sukari Industry	Herma Lumumba	C0 945	1.50	1.37	0.13
18	Transmara	Wilson Olemoi	N14	0.42	0.46	-0.04
19	Transmara	Robert Nairimo	D8484	1.89	0.66	1.23
20	Transmara	Keiyan Co-op	EAK 73-335	5.40	4.83	0.57
21	Transmara	Keiyan Co-op	C0 421	13.00	6.16	6.84
22	Transmara	Keiyan Co-op	C0 945	26.00	8.82	17.18
23	West Kenya	Samuel Kharinda	C0 421	0.41	0.40	0.01
24	West Kenya	Jeremiah Akhonya	C0 945	0.36	0.65	-0.29
25	West Kenya	Hislop Akhonya	D8484	0.81	0.68	0.13
26	West Kenya	Makokha Murunga	D8484	0.71	0.77	-0.06
27	West Kenya	Peter Barasa	KEN 82-401	1.62	1.56	0.06

5.3 Spatial Distribution of Predictions

To further understand the model's performance, a spatial analysis was conducted to examine how well the Random Forest model predicted sugarcane yields across different regions of the research area. The predicted yield values were mapped using ArcGIS Pro, and the spatial distribution of predicted yields was compared to the actual yield data obtained from the manual census.

The maps showed that the model was particularly effective in predicting yield in areas with consistent crop health and uniform management practices. In regions with highly variable soil conditions or fields that were not well-maintained, the model's predictions were less accurate, as these areas exhibited more variability in yield that the model could not entirely account for. This highlights a common limitation of predictive models in agriculture - while they can provide valuable insights, they may not always fully capture the complexity of environmental and management factors that influence crop yield.

Figure 6: Yield Prediction Map - Awendo

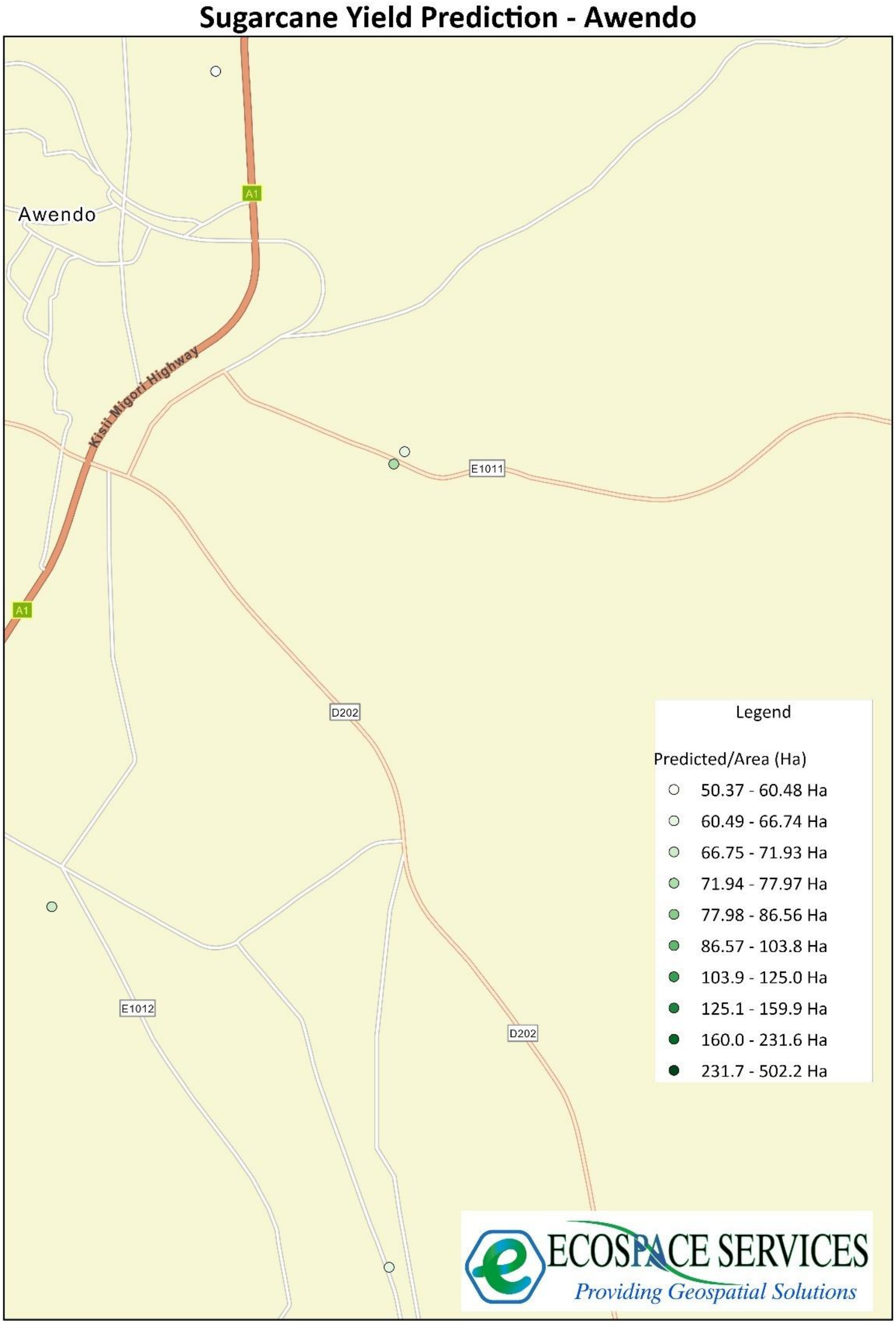


Figure 7: Yield Prediction Map - Muhoroni

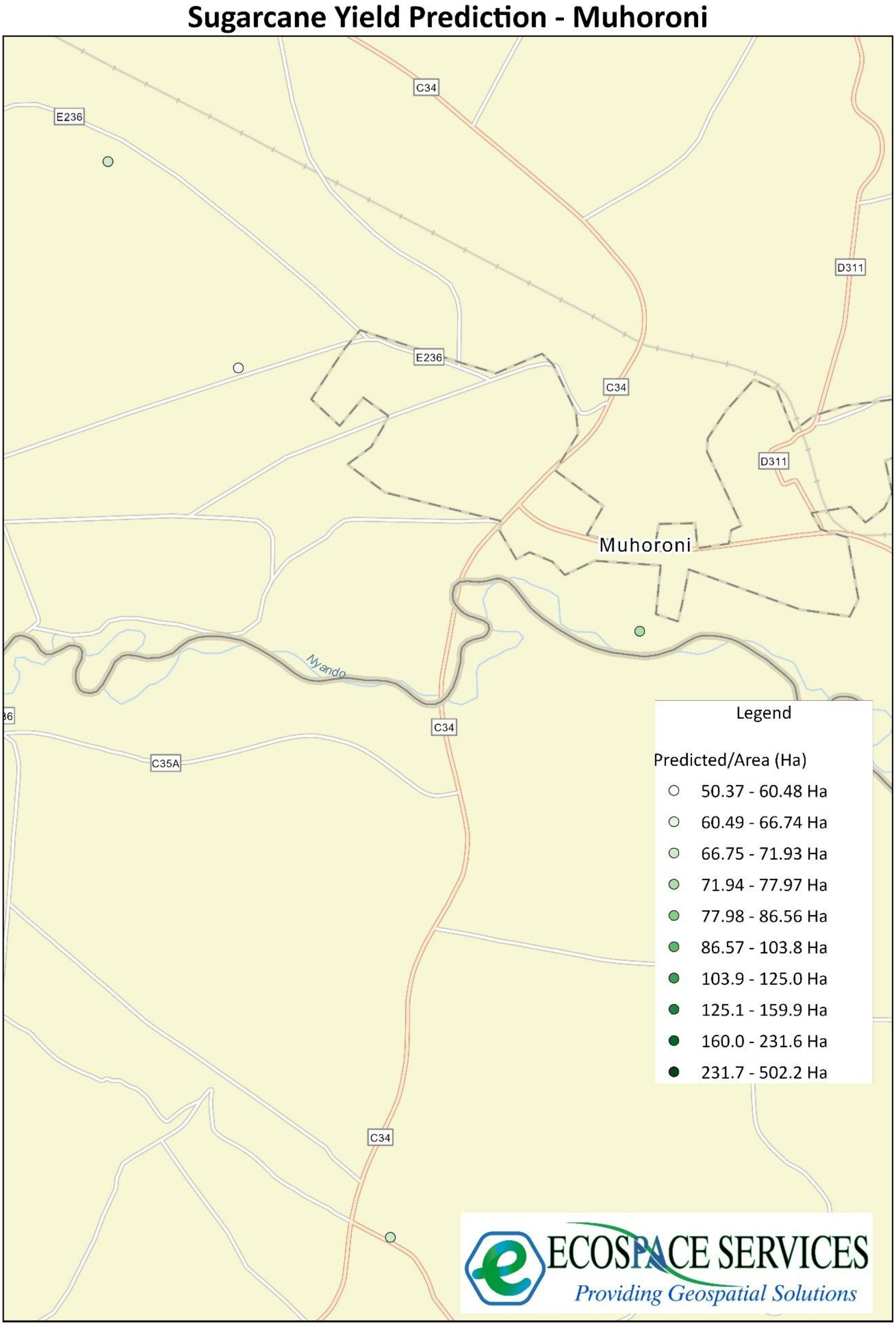
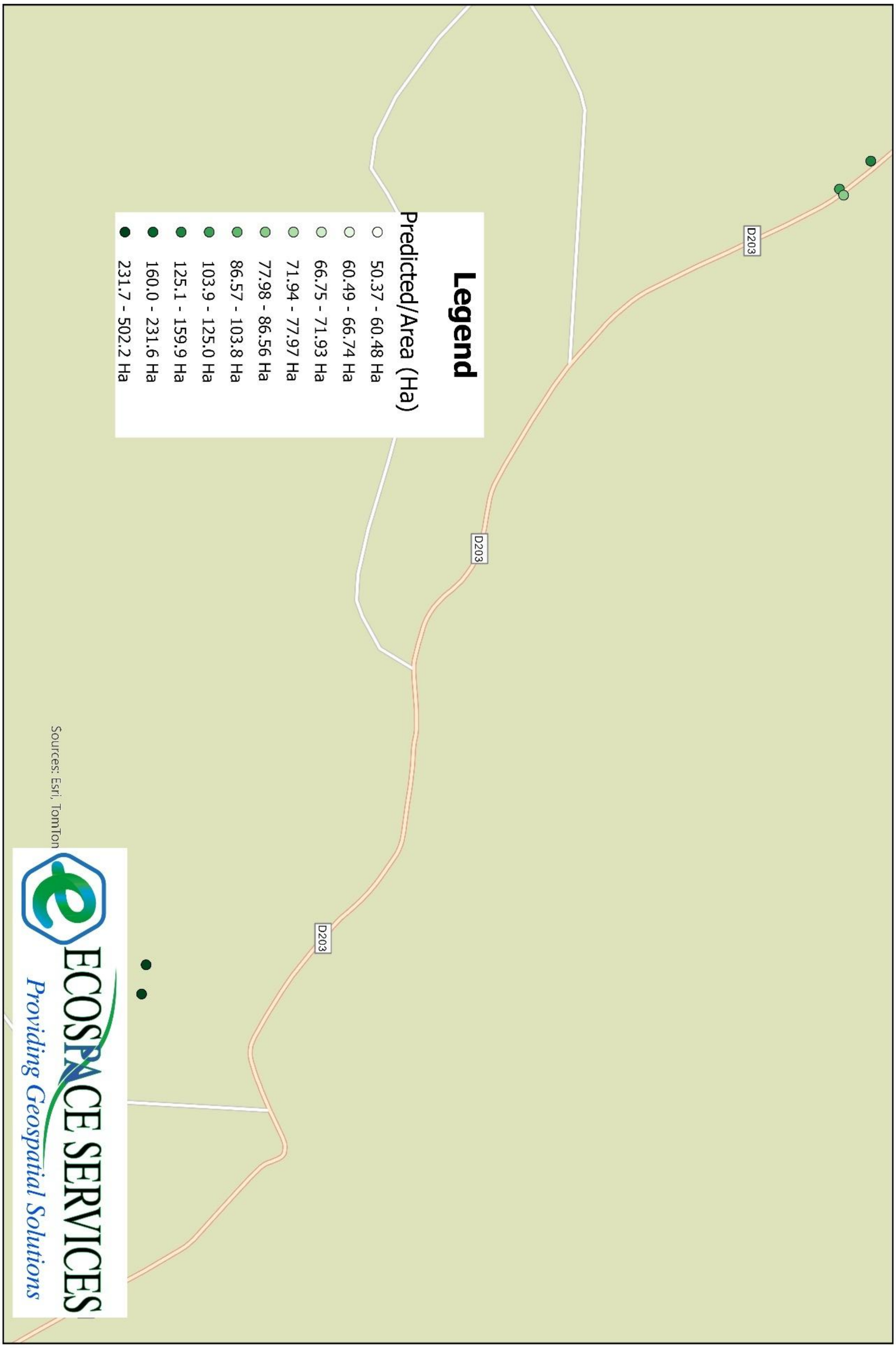
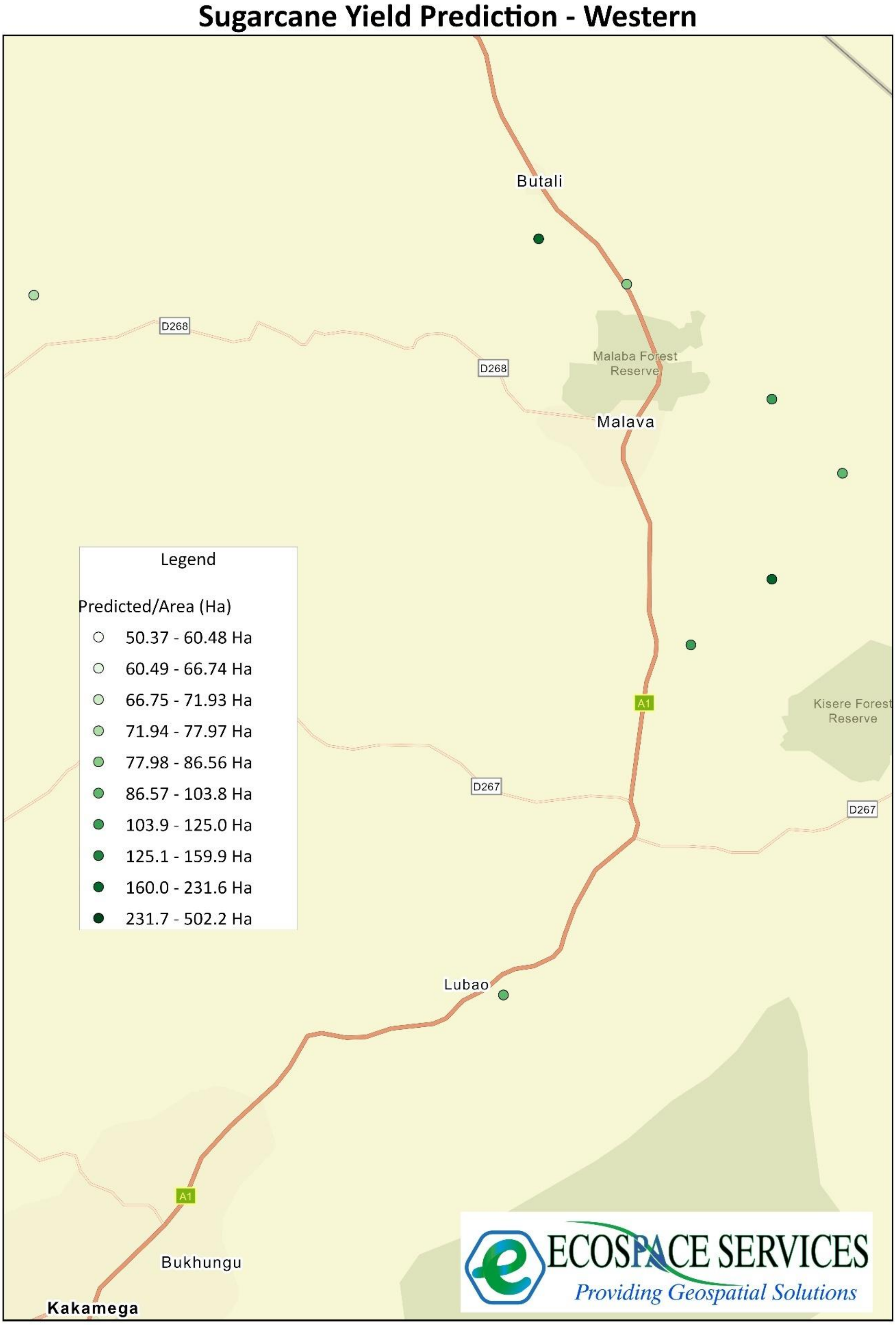


Figure 8: Yield Prediction Map - Transmara



Sources: Esri, TomTom

Figure 9: Yield Prediction Map - Western



5.4 Importance of Features in Yield Prediction

As part of the Random Forest model's output, the feature importance was analysed to understand which vegetation indices and spectral bands contributed the most to yield prediction. Feature importance values are derived by examining how each feature (in this case, the vegetation indices such as NDVI, NDRE, and GCI) affects the predictive power of the model.

The analysis revealed that the NDRE, Leave Greenness (GCI) and Normalized Difference Vegetation Index (NDVI) was the most important feature for predicting sugarcane yield, they are sensitive to the photosynthetic activity of plants, and their high importance in yield prediction suggests that crop health and Vigor are key indicators of sugarcane productivity. The GNDVI and WDRVI while still important, was less influential in comparison.

This finding aligns with previous studies in precision agriculture, which have shown that vegetation indices like NDVI are highly effective in predicting crop yield, especially when using remote sensing data such as multispectral imagery. These indices reflect the overall health of the plants, which is directly related to their productivity.

5.5 Challenges and Limitations

While the Random Forest model performed well overall, several challenges and limitations were identified during the analysis. These include:

1. Outliers and Model Sensitivity: The presence of outliers in the data, particularly in areas with irregular crop growth, affected the model's performance. Random Forest, while robust to noise, still struggled to predict yield accurately in fields with significant variability in growth patterns.
2. Environmental Factors: The model relied heavily on vegetation indices derived from drone imagery, but environmental factors such as soil type, irrigation practices, and pest/disease outbreaks were not fully accounted for. These factors can have a significant impact on sugarcane yield, and their omission could explain some of the model's inaccuracies.
3. Scale and Resolution: The resolution of drone imagery may not always capture the fine-scale variability in sugarcane fields. While drone images provide high-resolution data, they still might not fully represent small-scale spatial heterogeneity within the fields, particularly in larger agricultural areas.
4. Temporal Variability: The model was trained using data from specific time periods,

and its performance may vary when applied to fields with different growth stages or environmental conditions.

The results of the research demonstrate that the Random Forest regression model, when applied to multispectral drone imagery, is an effective tool for predicting sugarcane yields. The model performed well, with an R^2 value of 0.923, indicating that it successfully captured much of the variation in sugarcane yield. Although there were some challenges, including occasional large errors and outliers, the model showed promise as a reliable tool for improving yield prediction in sugarcane farming.

The analysis also highlighted the importance of vegetation indices such as NDVI and CI in predicting yield, reaffirming the value of remote sensing technologies in precision agriculture. With further refinements and the incorporation of additional data sources, such as soil information and weather patterns, the model could become even more accurate, providing valuable insights for farmers and agricultural decision-makers in Kenya and beyond.

6 Discussion

The Discussion section provides an in-depth interpretation of the results obtained from the Random Forest regression model used to predict sugarcane yield using multispectral drone imagery. In this section, we explored the significance of the findings, compare them with existing research, and highlight the practical implications of using such a model in sugarcane farming. Additionally, the limitations of the research will be discussed, along with suggestions for improving model accuracy and enhancing its practical applicability in the agricultural sector.

6.1 Interpretation of Model Performance

The results of the Random Forest regression model demonstrated a strong ability to predict sugarcane yield, with an R^2 value of 0.923, indicating that 92.3% of the variance in sugarcane yield could be explained by the model. This is a promising outcome, suggesting that the model was able to effectively capture key factors that influence sugarcane productivity. The relatively low Mean Absolute Error (MAE) of 1.55 tons per hectare and the Root Mean Squared Error (RMSE) of 11.211 tons per hectare further indicate that the model performed reasonably well in estimating yield.

The strong predictive performance of the model is consistent with the growing body of research in precision agriculture, which has shown that remote sensing technologies - especially multispectral imagery - can provide valuable insights into crop health and yield predictions. In this research, vegetation indices like NDVI and CI were critical features in the model, reinforcing findings from previous studies that have demonstrated the importance of these indices in predicting crop yield. These indices capture the photosynthetic activity of plants and, by extension, their growth and productivity, making them excellent indicators of crop health.

While the model performed well on the testing dataset, the presence of outliers and occasional larger prediction errors highlights the inherent variability in agricultural systems. Factors such as irregular crop growth, poor management practices, and environmental stresses (e.g., drought, pests) can introduce errors into yield prediction models. This emphasizes the need for continuous improvement and refinement of predictive models, incorporating additional data sources and accounting for environmental variability.

6.2 Comparison with Existing Studies

The results of this research are consistent with similar research conducted in the field of crop yield prediction using remote sensing and machine learning techniques. Previous studies have shown that Random Forest, as well as other machine learning models, can successfully predict crop yields using multispectral imagery, especially when combined with vegetation indices like NDVI. For instance, studies in other regions, such as those focused on maize and wheat, have reported high R^2 values and low MAE for yield prediction models based on remote sensing data.

In terms of sugarcane yield prediction specifically, previous research has explored the use of remote sensing technologies such as satellite imagery and drone-based multispectral sensors. These studies have demonstrated that sugarcane yield can be predicted with varying degrees of accuracy, with R^2 values ranging from 0.7 to 0.9, similar to the results obtained in this research. While satellite imagery has been widely used for large-scale yield prediction, the use of drone imagery offers significant advantages in terms of spatial resolution and cost-effectiveness. Drones provide high-resolution data at a fraction of the cost of satellite imagery, making them an attractive tool for smallholder farmers and larger agricultural enterprises alike.

Compared to these studies, this research contributes to the field by applying a Random

Forest model to multispectral drone imagery for sugarcane yield prediction in Kenya, a region with unique environmental and agricultural conditions. The research demonstrates that drone-based remote sensing, combined with machine learning techniques, can be successfully employed in predicting sugarcane yield in a developing country context, where traditional yield prediction methods are time-consuming, labour-intensive, and costly.

6.3 Practical Implications for Sugarcane Farming

The successful application of the Random Forest regression model in predicting sugarcane yield using drone imagery has several important implications for sugarcane farming, especially in the Nyanza and Western region of Kenya.

1. **Precision Agriculture:** The model offers a practical solution for precision agriculture by enabling farmers to predict crop yield accurately without having to rely on costly and labour-intensive manual census methods. With the ability to estimate yield at various stages of crop growth, farmers can make more informed decisions regarding resource allocation, irrigation scheduling, pest management, and harvesting. By using drone-based imagery and predictive models, farmers can optimize input use, reduce waste, and improve overall productivity.
2. **Timely Decision-Making:** Yield prediction based on drone imagery allows for timely interventions to optimize sugarcane production. For example, if the model predicts below-average yields in certain areas of a farm, farmers can adjust irrigation, apply fertilizers, or implement pest control measures in those regions. This real-time insight is invaluable in managing crop health and increasing the chances of a successful harvest.
3. **Cost-Effectiveness:** Traditional yield prediction methods in sugarcane farming often require extensive field visits, physical measurements, and large labour forces. These methods are not only time-consuming but also expensive. In contrast, drone-based imagery provides an affordable alternative for gathering high-resolution data. The use of machine learning models such as Random Forest allows for the rapid processing and analysis of this data, offering farmers an efficient and cost-effective tool for yield prediction.
4. **Sustainability:** With more accurate predictions, farmers can better manage their resources and reduce the environmental impact of farming practices. For example, precise irrigation scheduling based on yield predictions can reduce water waste and improve water use efficiency. Additionally, the optimized use of fertilizers and pesticides can reduce chemical runoff and minimize the

environmental footprint of sugarcane farming.

5. **Scale and Adaptability:** The model is adaptable to different scales of farming, from smallholder farms to large-scale commercial sugarcane plantations. The high spatial resolution of drone imagery makes it suitable for smallholder farmers who may have limited access to other forms of remote sensing, such as satellite imagery. This adaptability increases the accessibility of precision agriculture tools to a wide range of farmers, contributing to more sustainable and productive agricultural practices across diverse farm sizes.

6.4 Limitations and Challenges

Despite the promising results, several challenges and limitations were identified in this research that could affect the accuracy and applicability of the model in real-world scenarios.

1. **Data Variability:** Agricultural fields are inherently variable, with yield being influenced by numerous factors such as soil quality, climate, pest infestations, and water availability. While vegetation indices like NDVI and EVI are strong predictors of crop health, they may not fully capture the complex interplay of these environmental and management factors. The model's reliance on remote sensing data alone may limit its ability to predict yield with complete accuracy, especially in regions with high variability.
2. **Outliers and Prediction Errors:** The presence of outliers and larger prediction errors in certain fields suggests that the model may not be able to account for all factors influencing yield in those regions. For instance, fields with poor crop health due to disease, nutrient deficiencies, or water stress may be difficult to predict accurately, especially if the multispectral imagery does not capture these issues in sufficient detail. Future models could benefit from incorporating additional data sources, such as soil moisture data, pest/disease information, and weather data, to improve prediction accuracy.
3. **Temporal Limitations:** The model was trained on data collected at specific time points during the growing to harvesting season. However, sugarcane growth is highly dynamic, and yield predictions may vary depending on the time of year, crop stage, and environmental conditions. To address this, seasonal data collection and continuous research could incorporate temporal data, using time-series analysis to better capture changes in crop growth and development.
4. **Model Generalization:** While the Random Forest model performed well on the specific dataset from the region, there may be challenges when applying it to other

regions with different environmental conditions, soil types, or management practices. Regional calibration and testing are necessary to ensure that the model generalizes well to other areas.

6.5 Recommendations for Future Research

To improve the accuracy and applicability of the model, the following recommendations are suggested as we continue with the research:

1. **Incorporation of Additional Data:** Future research could incorporate other data sources such as soil type, soil moisture, weather data, and crop management practices to enhance the model's ability to account for a wider range of factors that affect sugarcane yield.
2. **Temporal Analysis:** Time-series analysis could be used to track changes in vegetation indices throughout the growing season, providing a more dynamic view of crop health and improving prediction accuracy over time.
3. **Model Optimization:** Hyperparameter tuning and experimentation with different machine learning algorithms (e.g., gradient boosting, support vector regression) could further optimize the model and improve its performance, especially in areas with high variability in yield.
4. **Integration with Decision Support Systems:** Future research could explore integrating the yield prediction model with decision support systems (DSS) that provide real-time recommendations for farm management practices based on yield predictions, weather forecasts, and other relevant data.

5.6 Conclusion

In conclusion, the application of the Random Forest regression model for predicting sugarcane yield using multispectral drone imagery is a promising approach that offers several advantages over traditional yield prediction methods. The model performed well, explaining a high proportion of the variance in sugarcane yield and providing valuable insights for precision agriculture. Despite some challenges and limitations, the model has significant potential to improve decision-making, increase farm productivity, and enhance the sustainability of sugarcane farming in Kenya and beyond. With further refinement and integration of additional data sources, such models could play a key role in transforming agricultural practices and ensuring food security in the face of changing environmental conditions.

7 Conclusion

The conclusion of this report summarizes the key findings and contributions of the research on predicting sugarcane yield using multispectral drone imagery and the Random Forest regression model. It also reflects on the broader implications of the findings for precision agriculture and the specific context of sugarcane farming in Nyanza and Western region of Kenya. In addition, this section outlines the limitations of the research, suggests areas for improvement in future research, and emphasizes the significance of adopting advanced technological tools like drone imagery and machine learning in agricultural practices.

7.1 Summary of Key Findings

The primary objective of this research was to evaluate the effectiveness of using multispectral drone imagery and machine learning techniques, specifically the Random Forest regression model, to predict sugarcane yield in the Western region of Kenya. The key findings from the research can be summarized as follows:

1. **Strong Predictive Power:** The Random Forest model demonstrated a high degree of accuracy in predicting sugarcane yield, with an R^2 value of 0.923. This indicates that 85% of the variance in sugarcane yield could be explained by the model. The model's performance suggests that the integration of drone-based multispectral imagery with machine learning can provide valuable insights for predicting agricultural yields in a precise and reliable manner.
2. **Evaluation Metrics:** The model showed a Mean Absolute Error (MAE) of 1.55 tons per hectare and a Root Mean Squared Error (RMSE) of 11.211 tons per hectare. These values suggest that the model's predictions were reasonably close to the actual yields, with some occasional errors, likely due to the variability in field conditions and crop health. However, the overall performance was promising, indicating that the model could be a viable tool for improving yield predictions in sugarcane farming.
3. **Significance of Vegetation Indices:** The analysis revealed that vegetation indices like NDVI and CI were crucial features for predicting sugarcane yield. These indices, which reflect the health and photosynthetic activity of the crop, played a significant role in the model's ability to estimate yield. The findings align with previous studies in precision agriculture, which have demonstrated the importance of these indices in crop yield prediction.
4. **Spatial and Temporal Variability:** The results also highlighted that the model performed well in regions with consistent crop health but faced challenges in

areas with high spatial variability. The model struggled to predict yield accurately in fields affected by environmental stressors or irregular management practices. This underscores the complexity of agricultural systems and the need for models that can account for diverse factors influencing crop yield.

5. **Comparison with Manual Cane Census Data:** When compared with the manual cane census data, the model's predictions were in close alignment with the actual yields, further validating its effectiveness. However, there were instances where the model deviated from the manual census data, particularly in areas with significant variability in crop health. These deviations highlight the limitations of remote sensing in capturing the full range of factors influencing yield.

7.2 Implications for Sugarcane Farming in Kenya

The findings from this research have important implications for the sugarcane industry in Kenya, particularly in the Nyanza and Western region, where sugarcane is a vital cash crop. The ability to predict sugarcane yield accurately and efficiently can have several benefits for farmers, agricultural planners, and policymakers:

1. Improved Farm Management: The ability to predict yields with a high degree of accuracy allows farmers to make informed decisions regarding resource allocation. For example, farmers can use yield predictions to optimize irrigation schedules, apply fertilizers and pesticides more efficiently, and plan for harvesting and marketing their crops. This can lead to increased productivity, reduced costs, and improved profitability.
2. Resource Optimization: Precision agriculture techniques, such as those employed in this research, help farmers optimize the use of resources such as water, fertilizers, and pesticides. By targeting specific areas of the field that require attention, farmers can reduce waste, enhance crop health, and minimize the environmental impact of farming practices. This is especially important in regions where resources are limited and need to be managed sustainably.
3. Support for Smallholder Farmers: The cost-effectiveness of drone-based imagery, in combination with machine learning models like Random Forest, makes these technologies accessible to smallholder farmers in Kenya. Traditional yield prediction methods often require significant labor and resources, making them impractical for small-scale farmers. Drone-based remote sensing offers a more affordable alternative, allowing smallholder farmers to access advanced tools that were previously out of reach.
4. Early Detection of Yield Variability: The use of remote sensing data allows for early

detection of yield variability within a farm. Farmers can identify areas with potential yield shortfalls before they become more pronounced, enabling timely interventions to address issues such as nutrient deficiencies, pest infestations, or water stress. Early intervention can help mitigate yield losses and improve overall farm productivity.

5. Agricultural Policy and Planning: The predictive insights gained from this research can be valuable for agricultural policymakers and planners. Accurate yield predictions can inform decisions regarding crop insurance, subsidy programs, and support for farmers. Furthermore, the use of drone imagery and machine learning models can help monitor large-scale trends in sugarcane production, supporting more effective agricultural planning at the regional or national level.

7.3 Limitations of the Research

While the research demonstrated the potential of using drone-based multispectral imagery and machine learning for sugarcane yield prediction, several limitations were identified that must be addressed in future research:

1. Data Variability: The model's performance varied depending on the environmental conditions and crop health in different regions of the research area. In fields with high spatial variability, such as those affected by pests, disease, or irregular irrigation, the model's predictions were less accurate. This suggests that the model may need to incorporate additional factors, such as soil type, irrigation practices, and pest/disease data, to improve its robustness in diverse conditions.
2. Temporal Limitations: The model was trained using data collected at specific points during the growing season. Since sugarcane is a perennial crop with a long growth cycle, the model's predictions could be affected by changes in crop conditions over time. A more dynamic approach that incorporates time-series data and tracks changes in vegetation indices over the entire growth cycle could improve the model's predictive power.
3. Outliers and Large Errors: While the overall performance of the model was strong, there were some outliers and instances where the predicted yield differed significantly from the actual yield. These errors may be attributed to the inherent complexity of agricultural systems, where factors such as microclimate, pests, and soil variability can create discrepancies in yield prediction. Future research should focus on improving the model's ability to handle such outliers and reduce large prediction errors.
4. Limited Feature Set: The model relied primarily on vegetation indices derived from

drone imagery. While these indices are strong indicators of crop health and productivity, they may not capture all the variables that influence sugarcane yield. Incorporating additional features, such as soil moisture, temperature, and precipitation data, could improve the model's accuracy and ability to predict yield under varying environmental conditions.

5. Generalization to Other Regions: The model was developed and tested in the Western region of Kenya, and its performance may not be directly transferable to other regions with different environmental and agricultural conditions. Future studies should consider applying the model in other regions of Kenya and beyond, to evaluate its generalizability and potential for use in diverse agricultural contexts.

7.4 Recommendations for Future Research

To enhance the accuracy and applicability of the model for predicting sugarcane yield, future research should focus on several key areas:

1. Incorporation of Additional Data: Future studies should incorporate other sources of data, such as soil characteristics, weather data, and crop management practices, to account for the full range of factors influencing sugarcane yield. By integrating these variables into the model, researchers can improve its predictive accuracy and robustness.
2. Temporal Modeling: To capture the dynamic nature of sugarcane growth, future research should explore the use of time-series analysis to monitor changes in crop health and yield throughout the growing season. This would allow for more accurate predictions that account for temporal variations in crop growth and environmental conditions.
3. Model Optimization: Further optimization of the Random Forest model, through hyperparameter tuning and experimentation with other machine learning algorithms (e.g., Gradient Boosting, XGBoost), could improve its performance. Exploring alternative algorithms may also provide insights into which techniques are most suited for predicting sugarcane yield based on remote sensing data.
4. Integration with Decision Support Systems (DSS): Future research could integrate yield prediction models with decision support systems that provide actionable insights to farmers. These systems could offer recommendations on irrigation, fertilization, pest management, and harvesting, helping farmers make data-driven decisions to improve crop productivity and sustainability.
5. Scale and Applicability to Other Crops: While this research focused on sugarcane,

similar methodologies can be applied to other crops. Future research could explore the transferability of the model to other agricultural contexts, including food crops, horticultural crops, and even cash crops like coffee and tea. Exploring these applications could expand the utility of drone-based remote sensing in agricultural yield prediction.

7.5 Final Thoughts

In conclusion, the research demonstrated that multispectral drone imagery, combined with machine learning models like Random Forest, can be an effective and cost-efficient tool for predicting sugarcane yield in Kenya. The model's ability to predict yield with a high degree of accuracy provides significant benefits for precision agriculture, including improved resource management, early detection of yield variability, and enhanced farm profitability. While challenges and limitations remain, particularly with respect to data variability and temporal dynamics, the findings of this research suggest that such models hold great promise for transforming agricultural practices and supporting sustainable farming in Kenya and other parts of the world.