African Journal of Rural Development, Vol. 9 (2):pp.149-163 ISSN 2415-2838

Green gram yield prediction using linear regression

TUMUSIIME, R.[1*], MABIRIZI, V.[1], MIREMBE, D.P., [2] ARINANYE TUGUME, R.[3] AND LUBEGA, J.[4]

[1]Department of Open and Distance Learning, Directorate of ICT Services, Kabale University., P.O. Box 317, Kabale, Uganda
[1]Department of Information Technology, Faculty of Computing, Library and Information Science, Kabale University, P.O. Box 317, Kabale, Uganda
[2]Department of Networks, College of Computing and Information Sciences, Makerere University
[3]Directorate of Graduate Training, Kabale University, P.O. Box 317, Kabale, Uganda
[4]Department of Information Technology, School of Computing and Informatics, Nkumba University, P.O. Box 237, Entebbe, Uganda
***Corresponding author:** rtumusiime@kab.ac.ug

ABSTRACT
Predicting crop yields before harvest is key in enabling farmers make critical decisions as far as post-harvest management is concerned. Besides, yield prediction plays a critical role in agriculture enterprise selection hence promoting food and nutrition security in a community. It is worth noting that various factors including ecological zones characteristics and farm management practices can vary significantly from season to season and farm to farmer, hence affecting crop yields. Given the importance of crop yield prediction in agriculture enterprise development and investments, a number of approaches have been adopted by farmers and breeders alike. These approaches range from controlled ideal condition analysis by breeders to the use of advanced plant physiological feature analysis using satellite image processing techniques. While a number of popular crops like rice and maize have a number of models proposed, limited yield prediction studies have been done on neglected crops like green gram. Therefore, this paper discusses the proposed green gram crop yield prediction model based on a stepwise linear regression technique using ecological zone characteristics, farm management practices and historic crop yield as the key variables. The study used a dataset of 107 records (gardens) and 9 features obtained from National Semi-Arid Research Institute (NaSARRI), Serere, Uganda. The predictor variables used were; soil type, soil PH, soil fertility, rainfall distribution, weeding practice, pest and disease management, fertilizer application, plant spacing, and cropping system. The model was evaluated for precision and evaluation result revealed that, with a mean absolute percentage error (MAPE) of 16.8%, the proposed model had a precision of 96.4% was deemed accurate in predicting green gram yield.

Keywords: Crop yields, Green gram, linear regression, Prediction models, Uganda

RÉSUMÉ
Prédire les rendements des cultures avant la récolte est crucial pour permettre aux agriculteurs de prendre des décisions critiques en matière de gestion post-récolte. De plus, la prédiction du rendement joue un rôle clé dans le choix des entreprises agricoles, favorisant ainsi la sécurité

Cite as: Tumusiime, R., Mabirizi, V., Mirembe, D. P., Arinanye Tugume, R. and Lubega, J. 2025. Green gram yield prediction using linear regression. *African Journal of Rural Development* 9 (2):149-163.

alimentaire et nutritionnelle d'une communauté. Il convient de noter que divers facteurs, tels que les caractéristiques des zones écologiques et les pratiques de gestion agricole, peuvent varier considérablement d'une saison à l'autre et d'un agriculteur à un autre, affectant ainsi le rendement des cultures. Compte tenu de l'importance de la prédiction du rendement des cultures dans le développement des entreprises agricoles et des investissements, un certain nombre d'approches ont été adoptées par les agriculteurs et les sélectionneurs. Ces approches vont de l'analyse dans des conditions idéales contrôlées par les sélectionneurs à l'utilisation de techniques avancées d'analyse des caractéristiques physiologiques des plantes à l'aide du traitement d'images satellites. Alors que de nombreux modèles ont été proposés pour des cultures populaires comme le riz et le maïs, peu d'études de prédiction de rendement ont été réalisées sur des cultures négligées comme le haricot mungo. Par conséquent, cet article discute du modèle proposé de prédiction du rendement du haricot mungo basé sur une technique de régression linéaire par étapes utilisant les caractéristiques des zones écologiques, les pratiques de gestion agricole et les rendements historiques des cultures comme variables clés. L'étude a utilisé un ensemble de données de 107 enregistrements (jardins) et 9 caractéristiques obtenues de l'Institut National de Recherche Semi-Aride (NaSARRI), Serere, Ouganda. Les variables prédictives utilisées sont : type de sol, pH du sol, fertilité du sol, distribution des précipitations, pratique du désherbage, gestion des ravageurs et des maladies, application d'engrais, espacement des plantes et système de culture. Le modèle a été évalué pour sa précision et les résultats d'évaluation ont révélé qu'avec une erreur absolue moyenne de pourcentage (MAPE) de 16,8 %, le modèle proposé a une précision de 96,4 %, jugée excellente par les experts pour prédire précisément le rendement.

Mots clés: Rendements des cultures, Haricot mungo, Régression linéaire, Modèles de prédiction, Ouganda

INTRODUCTION

Green gram or Mung bean (Vigna radiate) is a high-value legume crop and accounts for 8% of the total legume and fibber crop production globally (Bali and Singla, 2022). Green gram is a valuable source of nutrients and minerals for reducing the risk of heart disease, preventing cancer, controlling blood pressure, maintaining healthy skin, and regulating sugar levels among others (Nadia et al., 2022). Besides, the crop has the ability to fix atmospheric nitrogen, has a short maturity period (55 – 70 days), has low input and minimum care requirements and high resistance to drought (Mbeyagala *et al.*, 2017) thus making it suitable for incorporation into different cropping systems**.** Uganda is one of the leading producers of the crop in Africa with northern and eastern regions of the country as the main growing areas (Mbeyagala *et al.*, 2017; Nair *et al.*, 2019). Given the crop characteristics, green gram production, trading and consumption contribute greatly to the achievement of Sustainable Development Goals (SDGs) by providing food and nutrition security as well as increasing household incomes and creating sustainable wealth (Pérez-Escamilla, 2017). However, green gram production in Uganda is constrained by many challenges including; poor management of pests and diseases, poor agronomical practices, climate change, limited access to quality seeds, limited access to quality extension services and declining soil fertility (Ajio *et al.*, 2016; Mbeyagala *et al.*, 2017). Thus, reliable and timely prediction of green gram yield is crucial for both the famers and governments in addressing issues of market access, post-harvest handling and food and nutrition security policies (Talwana *et al.*, 2010). Crop yield prediction helps in identifying the attributes or factors that may significantly affect the crop yield so that early intervention can be enforced (Raju *et al.*, 2019).

Generally, a number of approaches for crop yield prediction exist including plot-by-plot analysis which using experts or farmer tacit knowledge based on plant observation to predict yield

(Fermont and Benson, 2011). Using this method a farmer or expert relay on their past experiences combined with plant physiological presentation to predict the yields (Anjitha *et al*., 2021). Additionally, Fuzzy logic (FL) which works on the principle of assigning a particular output depending on the probability of the state of the input (Chopra *et al.*, 2021), Adaptive Neuro Fuzzy Inference System (ANFIS) and Multiple Linear Regression (MLR) which attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data (Bazrafshan *et al.*, 2022; Joensuu *et al*., 2020) have been proposed. However, FL, ANFIS, MLR and Artificial Neural Network approaches where developed based on data sets from other countries and do not effectively represent the unique ecological zones (Setzer and Higham, 2021) and farm management practices of Uganda (Agnolucci *et al*., 2022). Thus, focus of this study was to develop a green gram yield prediction model suitable for Ugandan agro-ecological zone, climate variation and farm management practices which can be effectively used by rural farmers.

Related Studies

Two approaches to agricultural yield prediction i.e.; process and statistical based modelling techniques exists. Process based models are the basic models for predicting agricultural output that take into account various environmental factors, such as soil qualities, and specific physiological traits of plants, such as information on photosynthesis per unit leaf area (Maestrini *et al.*, 2022). Because these models replicate the physiological processes of crop growth and development in response to environmental factors and management techniques, they are particularly effective for predicting crop output at the field size (Maestrini *et al.*, 2022). However, it is still problematic to apply process-based crop yield predication models because they are data intensive and require calibration hence the need for alternative solutions to crop yield prediction (Mariano and Balzarini, 2021).

On the other hand, without taking into account the underlying mechanisms in crop physiology and ecology, statistical modelling calculates direct connections between predictor variables (such as climate, agronomic practice, and soil characteristics) and crop production in a given data set (Cedric *et al*., 2022). When the training data used is sufficient and reliable, statistical models provide more accurate predictions compared to processed models (Elavarasan and Durairaj Vincent, 2020). The most popular statistical models for crop yield prediction are those based on Machine Learning (ML) algorithms and Linear Regression.

A study by Sitienei *et al.* (2017) proposed a multiple linear regression model to predict tea yield using climatic variables. In their study, multiple regression analysis was used. The study used a contingency table for model verification and revealed that 70% of mode forecasts were correct. They revealed that the regressions were weak, suggesting that tea crop yields don't respond strongly to changes in climate variables. Balakrishnan and Muthukumarasamy (2016) developed a Support Vector Machine (SVM) and Naive Bayes based models called AdaSVM and AdaNaive as the ensemble models for crop yield prediction. The models were built on Year/-Month, Average Temperature, Cloud Cover, Evapotranspiration, Vapour Pressure, Wet day frequency and precipitation variables. The study obtained Black gram yield prediction accuracy of 86.70% (SVM), 89.42% (AdaSVM), 82.4% (Naive Bayes) and 92.6% (AdaNaive).

In an effort to improve the performance of yield predication models, Elavarasan and Durairaj, (2020), a reinforcement Machine Learning algorithm to predict crop yields. They, used deep reinforcement techniques to predict crop yields and provide recommendations to farmers about suitable crops to grow in order to improve the quality of produce. Srinivasa *et al*, (2023) applied both

Artificial Neural Network (ANN) and Convolutional Neural Network (CNN) deep learning techniques in their method for crop yield prediction. Jorvekar *et al.* (2024) carried out a comparative study to investigate crop yield prediction accuracy with MLR, SVR, ANN, KNN, Random forest using various performance metrics on a hybrid model. From their findings, the developed hybrid model exhibited the 92.7% accuracy on yams and 92.43% accuracy on cassava as the highest compared to other transfer leering models.

Study Approach and Methodology

The study was conducted using a mixed methods research approach which applied both qualitative and quantitative methods of data collection and analysis as illustrated in the Figure 1 below.

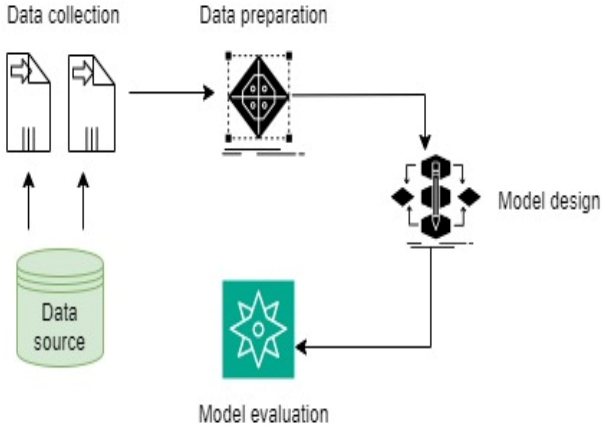

Figure 1. Model design process

Data collection. A total of 117 respondents participated in the study, comprising 110 green gram farmers and 7 legume breeders or agronomy experts. Prior to data collection, all respondents were provided with detailed information about the purpose, objectives, and scope of the study. Participation was entirely voluntary, and written informed consent was obtained from all participants before they engaged in the study. To ensure the confidentiality of respondents, all data collected was anonymised and securely stored. Unique identifiers were assigned to participants instead of using names or other personal details. Access to the data was restricted to the research team, and all electronic records were password-protected. The study complied with applicable data protection regulations. Different tools were administered to the two groups of respondents to ensure relevance and accuracy. Farmers were probed using a structured questionnaire tailored to their farming practices, challenges, and perspectives. On the other hand, the legume breeders and agronomy experts participated in semi-structured interviews or filled out an expert-focused questionnaire to capture their specialized insights on legume cultivation, breeding, and agronomy practices.

The inclusion of legume breeders and agronomy experts was critical to provide technical insights and validate the findings from farmers. This group contributed knowledge on best agronomic practices, breeding techniques, and challenges from a scientific and industry perspective, enriching the overall analysis and ensuring a balanced understanding of green gram production and its challenges.

The respondents to the study were purposively selected based on their knowledge and experience in green gram production. Given the nature of the crop which main grown by women to supplement household nutrition majority of the farmers who participated in the study were women (70), men where 47. The respondents were from Serere, Ngora and Kumi districts. The data was collection using online data collection tools to minimise errors of data entry and cleaning. The interview protocol focused on the key variables i.e. crop management practices (including; weeding, plant spacing, crop spraying, fertilizer application, cropping system), rainfall distribution experienced (uniform or non-uniform) and the characteristics of farmland (soil type, soil PH, soil fertility), collected by direct observation, and laboratory analysis to ensure a comprehensive assessment of the farmland

characteristics, enhancing the reliability and validity of the findings.

Data processing

The aim of this step was to clean the data by addressing missing values and outliers. Missing values (often called NaN) cannot be handled directly by regression models, and outliers (values that lie at an abnormal distance from other values) are likely to mislead the model. The collected data was queried and loaded as a data frame using Pandas, a Python library. For the numerical data, missing values were handled using Python's inbuilt fillNa () function, which imputed the mean of each column for missing entries. Outliers were detected and removed using a robust regression method based on the median robust estimator to ensure the model was not unduly influenced by extreme values. For the qualitative (categorical) data**,** pre-processing was performed to prepare it for regression analysis. The following steps were undertaken:

i. **Encoding Categorical Data:** Categorical variables were converted into numerical format using one-hot encoding or label encoding, depending on the nature of the variable. For instance, nominal variables (e.g., farm type) were one-hot encoded, while ordinal variables (e.g., soil quality ratings) were label-encoded to preserve the inherent order. All traits were numerically encoded, as they were either categorical (e.g., soil type: clay=0, sand=1, loamy=2) or binary (e.g., soil fertility: fertile=1, infertile=0), making them suitable for numerical correlation analysis without further transformation.

ii. **Handling Missing Qualitative Data:** Missing values in categorical variables were imputed using the mode (most frequent category) to maintain consistency.

iii. **Feature Transformation:** Once encoded, the qualitative data was integrated with numerical data to form a comprehensive dataset ready for regression analysis. This transformation ensured that all variables were in a numerical format suitable for statistical modelling.

The Green Gram Yield Prediction Model design

A regression analysis was conducted to identify variables impact on green gram yield (variable weighted factor). Thus, a regression model was used to calculate by how much the yield of green gram (dependent variable) changed when the contributing (independent) variables changed. The process determined which factors signified most (variable coefficient), which factors could be ignored, and how these factors influenced each other. In the first step of the model design, we defined dependent variable (yield of green gram) that was hypothesized to be influenced by several independent variables. Conventionally, the model is represented as;

$$y_i = \beta_0 + \beta_i x_i + \epsilon \quad \text{.....(1)}$$

Where y_i represents the i^{th} observation on the yield (t ha^{-1}) of the green gram. Independent variables hypothesized to influence y_i are represented by x_i, while β_i represents the variables coefficients. Where applicable, β_0 represents the intercept and ϵ represents the error term. Conceptually, the proposed regression model is represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \epsilon \quad \text{.....(2)}$$

Where y = yield (t ha^{-1}), x_1 = soil type, x_2 = soil Ph, x_3 = soil fertility, x_4= rainfall distribution, x_5 = weeding, x_6= crop sprayed, x_7= fertilizer application, x_8 = cropping system, x_9 = plant spacing.

In the next step of model design, different models were developed by backward elimination of variable(s) with high p values (>0.05). The p value is a representation of the statistical significance of

the variable. These variables have little contribution to the target variable (yield) and may cause overfitting, increase training time and reduce prediction accuracy of the model. Each time a variable was eliminated, the developed model was fitted with the data to check if its coefficient of determination (R^2) value was greater than or equal to the previous R^2 value. The R^2 value is the representation of the amount of the variance in the yield which is explained by the model specified (or measure of goodness of fit of the model). The model with R^2 value less than the previous value was ignored. The algorithm for the model development and selection is illustrated in the following steps;

Step 1: The hypothetical model; $y = \beta_0 + \beta_i x_i + \epsilon,\ i = 1,2,\dots,12$ was fitted with all the predictors, check R^2 value if high enough then it is model one eliminate variable with the highest p value and obtain model two;
$y = \beta_0 + \beta_i x_i + \epsilon,\ i = 1,2,\dots\dots\dots\dots\dots\dots\dots\dots,11$
else stop, the model is not good fitting

Step 2: Fit model two;
$y = \beta_0 + \beta_i x_i + \epsilon,\quad i = 1,2,\dots,11$ with data check its R^2 value if high enough or is equal to the previous value then it is model two eliminate variable with the highest p value and obtain model three; $y = \beta_0 + \beta_i x_i + \epsilon,\quad i = 1,2,\dots\dots\dots,10$ else the stop, the model is not good fitting.

Step3: Repeat step 2 for model three, model four …. and stop when no good fitting model is reached.

Model Evaluation Parameters

The model performance was assessed using a number of parameters including; Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) to establish the models' accuracy.

i. Mean Absolute Error (MAE)

In order to establish the absolute difference between the real data and the model's prediction, MAE was the most ideal metric to use. MAE does not indicate underperformance or over performance of the model (whether or not the model under or overshoots actual data). A modest MAE (5-10) indicates that the model performs well in predictions, whereas a big MAE (above 10) indicates that the model might struggle in some situations. Although this is nearly seldom the case, a zero MAE value indicates that the model is an excellent predictor of the outputs. The formula for MAE is;

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{……..……..…(3)}$$

Where;
y_i= true yield
$\hat{y}_i$ = predicted yield
n = number of observations

ii. Root Mean Squared Error (RMSE)

This metric measures the average magnitude of the error by taking the square root of the average of squared differences between prediction and actual observation. Because model errors are likely to have a normal distribution rather than a uniform distribution, the RMSE is a better metric to present for such a type of data.
The formula of RMSE is;

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{…….…….. … (4)}$$

iii. Mean absolute percentage error (MAPE)

MAPE measures the prediction accuracy of the model in percentage calculated as the average absolute percent error for every time period minus the forecast values divided by the actual values. Mathematically MAPE is calculated using equation (5).

$$MAPE = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{|Actual - Forecast|}{|Actual|}\right) \times 100 \quad \text{……..……(5)}$$

MAPE is preferred over other metrics since it offers clear data interpretation for easy conceptualization

because an absolute value is used, MAPE and MAE are both resistant to the effects of outliers.

Model Evaluation

The overall model evaluation steps are described in the following steps;

Step 1: Partition the original training data set into *k* equal subsets. Each subset is called a **fold**. Let the folds be named as f_1, f_2…… f_k

Step 2: For i = 1=1, 2, 3……..... k

(a) Keep the fold fi as validation set and keep the remaining k-1 folds in the Cross-validation training set.

(b) Train the model using the cross-validation training set and calculate the model accuracy by validating the predicted results against the validated set.

(c) Estimate the accuracy of the learning model by the accuracies derived in all the k cases of cross validation.

To explore the relationships among the traits, a correlation analysis was conducted. The variables analysed included soil type, soil pH, soil fertility, rainfall distribution, weeding, crop spraying, fertilizer application, cropping system, plant spacing, and yield. This analysis aimed to provide insights into the interdependencies among traits before modelling. The results were visualized in a heat map.

RESULTS

Trait correlations

The correlation analysis revealed positive correlations between soil fertility (r = 0.68) and plant spacing (r = 0.54). Improper weeding (r = −0.45) and non-uniform rainfall distribution (r = −0.37) exhibited negative correlations with yield. Variables such as soil pH and cropping system showed weak correlations with yield (r < 0.2) (Figure 1).

Model performance

The model with lowest MAE, RMSE and MAPE is selected as the best regression model for predicting green gram yields. Results presented in this section are based on the three regression models that were created including; Model one, Model two and Model three discussed below:

Model one:

The first model utilizes all the nine features that were identified during data collection. The regression model was fitted with the data and Ordinary Least Square (OLS) summary results is shown in Table 1.

The regression equation obtained using the coefficient in Table 2 is shown in equation (6)

$$y = 0.27 + 0.22x_1 - 0.01x_2 - 0.09x_3 + 0.04x_4 + 0.05x_5 + 0.06x_6 + 0.13x_7 + 0.05x_8 - 0.01x_9 \quad \ldots\ldots(6)$$

Where;

$x_1 = Soil\_Type$

$x_2 = Soil\_PH$

$x_3 = Soil\_Fertility$

$x_4 = Rainfall\_Dist$

$x_5 = Weeding$

$x_6 = Crop\_Sprayed$

$x_7 = Fertilizer\_App$

$x_8 = Plant\_Spacing$

$x_9 = Cropping$

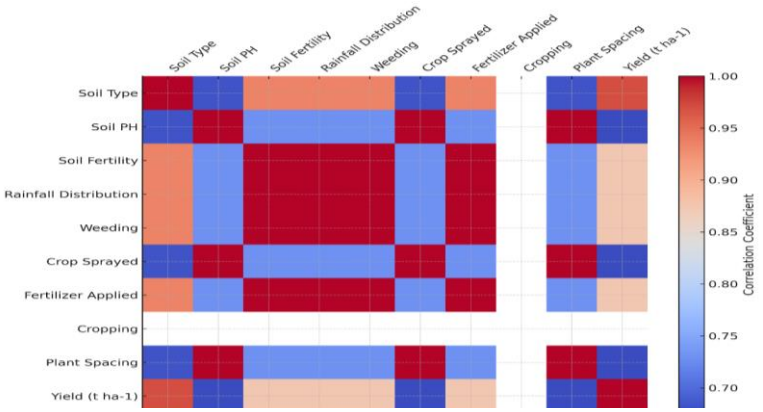

Figure 2. Heat map of correlation of traits evaluated in the study

Table 1. Model one OLS Regression Results

Dep. Variable:	Yield	**R-squared:**	0.729			
Model:	OLS	**Adj. R-squared:**	0.637			
Method:	Least Squares	**F-statistic:**	32.32			
Date	Fri. 05 Jul 2019	**Pro(F-statistic):**				
Time	20:06:02	**Log-Likelihood:**	63.7e-25			
No Observation:	110	**AIC:**	-104.9			
Df Residuals:	100	**BIC:**	-82.74			
Df Model:	9					
Covariance Type:	Non-robust					
	coef	**Std err**	**t**	**p>|t|**	**[0.025**	**0.975]**
Intercept	0.2732	0.041	6.590	0.000	0.191	0.355
Soil_Type	0.2188	0.030	7.283	0.000	0.159	0.278
Soil_PH	-0.0110	0.054	-0.205	0.838	-0.118	0.096
Soil_Fertility	-0.0971	0.092	-1.061	0.291	-0.279	0.084
Rainfall_Dist	0.0397	0.065	0,612	0.542	-0.089	0.168
Weeding	0.0452	0.043	1.058	0.292	-0.040	0.130
Crop_Sprayed	0.0691	0.065	1.058	0.293	-0.061	0.199
Fertilize_App	0.1325	0.060	2.217	0.029	0.014	0.251
Plant_Spacing	0.0522	0.030	1.736	0.088	-0.007	0.112
Cropping	-0.0101	0.060	-0.169	0.868	0.129	0.109
Omnibus:	11.539		**Durbin-Watson:**	1.161		
Prob(Omnibus):	0.003		**Jarque-Bera (JB):**	17.094		
Skew:	0.494		**Prob(JB):**	0.000194		
Kurtosis:	4.660		**Cond. No.**	21.8		

Model two: To obtain model two, the feature with the highest p-value (Soil_PH) was eliminated. The OLS model summary presented in Table 2

Table 2. Model two OLS Regression Results

Dep. Variable:	Yield	**R-squared:**	0.758		
Model:	OLS	**Adj. R-squared**	0.742		
Method:	Least Squares	**F-statistic:**	45.76		
Date:	Fri, 05 Jul 2019	**Prob (F-statistic):**	1.05e-28		
Time:	20:06:08	**Log-Likelihood:**	67.901		
No. Observations:	110	**AIC:**	-119.8		
Df Residuals:	101	**BIC:**	-98.20		
Df Model:	8				
Covariance Type:	nonrobust				
	coef	**std err**	**t**	**P>it]**	**[0.025**
Intercept	0.2700	0.038	7.005	0.000	0.194
Soil_Type	0.2186	0.030	7.315	0.000	0.159
Soil Fertility	-0.0967	0.091	-1.062	0.291	-0.277
Rainfal Dist	0.0393	0.065	0.609	0.544	-0.089
Weeding	0.0433	0.041	1.044	0.290	-0.039
Crop_Sprayed	0.0635	0.059	1.077	0.284	-0.053
Fertilizer_App	0.1325	0.059	2.227	0.028	0.014
Plant Spacing	0.0514	0.030	1.732	0.088	-0.007
Cropping	-0.0098	0.060	-0.161	0.872	-0.128
Omnibus:	11.571	**Durbin-Watson:**	1.153		
Prob (Omnibus):	0.003	**Jarque-Bera (JB):**	17.213		
Skew:	0.493	**Prob (JB):** **0.000183**	0.000183		
Kurtosis:	4.668	**Cond. No.**	20.6		

The regression equation obtained is indicated in equation (7)

$$y = 0.27 + 0.22x_1 - 0.10x_3 + 0.04x_4 + 0.04x_5 + 0.06x_6 + 0.13x_7 + 0.05x_8 - 0.01x_9 \ldots\ldots\ldots\ldots\ldots (7)$$

Model three:

Model three was obtained by further eliminating cropping feature. The OLS results summary is presented in Table 3.

The regression equation obtained using model three is presented in equation (8)

$$y = 0.27 + 0.22x_1 - 0.01x_3 + 0.04x_4 + 0.04x_5 + 0.06x_6 + 0.13x_7 + 0.05x_8 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Table 3. Model three OLS Regression Results

Dep. Variable:	Yield	**R-squared:**	0.758			
Model:	OLS	**Adj. R-squared**	0.742			
Method:	Least Squares	**F-statistic:**	45.76			
Date:	Fri, 05 Jul 2019	**Prob (F-statistic):**	1.05e-28			
Time:	20:06:13	**Log-Likelihood:**	67.901			
No. Observations:	110	**AIC:**	-119.8			
Df Residuals:	102	**BIC:**	-98.20			
Df Model:	7					
Covariance Type:	nonrobust					
	co-ef	**std err**	**t**	**P>[it]**	**[0.025**	**0.975]**
Intercept	0.038	7.136	0.000	0.194	0.344	0.030
Soil_Type	7.348	0.000	0.160	0.277	0.064	0.167
Soil Fertility	-0.0978	0.090	-1.083	0.281	-0.277	0.081
Rainfal Dist	0.0392	0.604	0.610	0.543	-0.088	0.167
Weeding	0.0416	0.040	1.042	0.300	-0.038	0.121
Crop_Sprayed	0.0581	0.048	1.201	0.233	-0.038	0.154
Fertilizer_App	0.1322	0.059	2.234	0.028	0.015	0.250
Plant Spacing	0.0512	0.030	1.735	0.086	-0.007	0.110
Omnibus:	11.612	Durbin-Watson:	1.155			
Prob(Omnibus):	0.003	**Jarque-Bera (JB):**	17.271			
Skew:	0.495	**Prob(JB): 0.000183**	0.000178			
Kurtosis:	4.669	**Cond. No.**	19.4			

Table 4. Legend for Abbreviations Used

Abbreviation	Full Name
Dep. Variable	Dependent Variable
R-squared	Coefficient of Determination (R^2)
Adj. R-squared	Adjusted Coefficient of Determination
OLS	Ordinary Least Squares
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
Df	Degrees of Freedom
coef	Coefficient
Std err	Standard Error
t	t-statistic
p>|t|	p-value
[0.025, 0.975]	95% Confidence Interval
Omnibus	Omnibus Test Statistic
Durbin-Watson	Durbin-Watson Statistic
Prob(Omnibus)	Probability of Omnibus Test
Jarque-Bera (JB)	Jarque-Bera Test Statistic
Prob(JB)	Probability of Jarque-Bera Test
Skew	Skewness of the Data
Kurtosis	Kurtosis of the Data
Cond. No.	Condition Number (Multicollinearity Measure)

Upon successful formation of the three above mentioned models, each model was fitted with the data and the values for Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) were computed using python functions. The prediction and evaluation results for the three models with the three metrics are shown in the Table 5 and Table 6 respectively.

Table 5. Prediction results

ID	Actual Yield (t ha^{-1})	Model One Predicted Yield (t ha^{-1})	Model Two Predicted Yield (t ha^{-1})	Model Three Predicted Yield (t ha^{-1})
70	0.704	0.726	0.734	0.716
102	0.856	0.877	0.891	0.861
98	0.312	0.401	0.400	0.397
59	0.734	0.754	0.750	0.746
104	0.311	0.288	0.340	0.335
74	1.005	0.891	0.901	0.980
47	0.689	0.710	0.671	0.658
60	0.576	0.692	0.688	0.687
32	0.909	0.800	0.831	0.882
105	0.298	0.311	0.250	0.272

Table 6. Evaluation results

Metrics	Model		
	Model 1	Model 2	Model 3
MAE (t ha^{-1})	0.104	0.102	0.101
RMSE (t ha^{-1})	0.127	0.126	0.122
MAPE (%)	17.88	17.84	16.83

DISCUSSION

Comparing the evaluation results of the three models basing, MAE, RMSE and MAPE metrics, we observe that green gram yield prediction can be achieved using a regression model. The study developed three models, model one has nine features while model two has eight features and model three has seven features. When model one was fitted with the data, the study obtained 0.104, 0.12 and 717.88 as the respective values of MAE, RMSE and MAPE. While for model two, the study obtained 0.102, 0.126 and 17.84 for MAE, RMSE and MAPE respectively. Model three, the respective values of MAE, RMS and MAPE obtained were 0.101, 0.122 and 16.83 respectively. From these results, we conclude that the three models generated acceptable yields estimation. However, Model Three demonstrated superior yield estimation capabilities compared to the other two models, as evidenced by its lower MAE of 0.101, RMSE of 0.122, and MAPE of 16.83. These lower error values indicate that Model Three provided more accurate predictions, making it the most reliable model for forecasting green gram yields in this study.

When compared to earlier studies, this work aligns with and builds upon previous findings that support the use of regression models in agricultural yield prediction. However, unlike many earlier models that may rely on larger or more complex datasets, the current study demonstrates that feature optimization can enhance prediction accuracy while reducing computational costs. For instance, earlier research has often highlighted the trade-off between accuracy and feature reduction, but this study successfully strikes a balance by achieving higher accuracy with fewer features. The implications of these findings are substantial. For farmers, these models provide a practical tool for yield forecasting, enabling better planning of planting schedules, fertilizer application, and resource allocation. Breeders can use the insights from these predictions to focus on developing high-yield, resilient green gram varieties by targeting environmental and management factors most correlated with yield. Policymakers and agricultural extension officers stand to benefit as well, using these models to identify trends in crop performance, design intervention programs, and provide precise recommendations to farmers, particularly in resource-constrained settings.

Accurate crop yield prediction plays a vital role in strategic planning, policy formation, and decision-making processes related to import-export, pricing, crop distribution, and procurement. Over the years, various methods have been developed for forecasting agricultural yields, but many face limitation in terms of data availability, accuracy, and adaptability to local contexts. In this study, we investigated green gram yield forecasting in Serere, Uganda, using linear regression models based on weather variable, ecological zones, and farm management practices. The stepwise regression procedure allowed for the identification of significant variables, and models were evaluated using performance metrics such as Mean Absolute Error (MAE), Root Mean Square

Error (RMSE), and Mean Absolute Percentage Error (MAPE). Among the models, Model Three demonstrated superior performance, with the lowest error metrics, making it the most accurate predicting green gram yields.

However, despite the positive results, this approach has limitations. First, linear regression assumes a linear relationship between variables, which may not fully capture complex interactions in agricultural systems. Furthermore, the model's accuracy is influenced by the quality and granularity of the data, which can vary depending on local practices and data collection methods. Additionally, the model does not account for unexpected external factors such as pests, diseases, and socio-economic changes, which could significantly impact yields. To address this limitation, breeders should focus on developing more resilient green gram varieties that can adopt to different ecological zones and changing weather patterns. This will help to minimize the adverse effects of climatic variability on yields. Additionally, breeders should engage in participatory breeding that involve local farmers. By incorporating farmers, these programs can ensure that newly developed varieties address practical challenges faced on the ground.

For farmers, improved farm management practices are critical. Farmers should adopt precision agriculture technique, such as real-time data usage for optimising input application, including soil testing, water management, and pest control measures. Furthermore, farmers can benefit from weather forecasting tools and models like the one developed in this study. These tools can guide planting schedules and other vital farming activities, helping to reduce risks associated with unpredictable weather condition. Policymakers should invest in infrastructure for data collection on weather patterns, soil conditions, and farm management practices, as reliable data will enhance the accuracy for predictive models and enable more targeted intervention. Additionally, creating policies that encourage the adoption of technology-driven farming practices is essential. This may include mobile-based advisory service that provide tailored recommendations for farmers based on yield forecasting models. Public-private partnership should also be promoted to ensure smallholder farmers have access to predictive tools and client-resilient agriculture technologies**.**

CONCLUSION AND RECOMMENDATION

In conclusion, while linear regression models offer a useful method for predicting green gram yields, future research should focus on incorporating more sophisticated techniques, such as machine learning, to capture non-linear relationships and account for external disruption. Additionally, improved data quality and broader stakeholders' engagement will be key to maximizing the potential of predictive models in agriculture decision making.

ACKNOWLEDGEMENT

We thank NASSARI Serere, Uganda, and the ERIGNU Project for their invaluable technical support, which greatly contributed to the success of this work.

DECLARATION OF CONFLICT OF INTEREST

The authors have no conflicts of interest to disclose.

REFERENCES

Agnolucci, P., Rapti, C., Alexander, P., Lipsis, V. and De Robert, A. 2022. *Impacts of rising temperatures and farm management practices on global*. 1–42.

Ajio, F., Talwana, H . and Kagoda, F. 2016. Evaluation of Mungbean plant spacing for optimising yield in smallholder cropping systems. *Makerere University E-Repository* 14 (14): 403–406.

Anjitha, K. S., Sameena, P. P. and Puthur, J. T. 2021. Functional aspects of plant secondary metabolites in metal stress tolerance and their importance in pharmacology. *Plant Stress* 2 100038.

https://doi.org/10.1016/j.stress.2021.100038

Balakrishnan, N. and Muthukumarasamy, G. 2016. Crop Production - Ensemble Machine Learning Model for Prediction. *International Journal of Computer Science and Software Engineering* 5 (7): 148–153.

Bali, N. and Singla, A. 2022. Emerging Trends in Machine Learning to Predict Crop Yield and Study Its Influential Factors: A Survey. *Archives of Computational Methods in Engineering* 29 (1): 95–112. https://doi.org/10.1007/s11831-021-09569-8

Bazrafshan, O., Ehteram, M., Dashti Latif, S., Feng Huang, Y., Yenn Teo, F., Najah Ahmed, A. and El-Shafie, A. 2022. Predicting crop yields using a new robust Bayesian averaging model based on multiple hybrid ANFIS and MLP models: Predicting crop yields using a new robust Bayesian averaging model. *Ain Shams Engineering Journal* 13 (5): 101724. https://doi.org/10.1016/j.asej.2022.101724

Cedric, L. S., Adoni, W. Y. H., Aworka, R., Zoueu, J. T., Mutombo, F. K., Krichen, M. and Kimpolo, C. L. M. 2022. Crops yield prediction based on machine learning models: Case of West African countries. *Smart Agricultural Technology* 2. https://doi.org/10.1016/j.atech.2022.100049

Chopra, S., Dhiman, G., Sharma, A., Shabaz, M., Shukla, P. and Arora, M. 2021. Taxonomy of Adaptive Neuro-Fuzzy Inference System in Modern Engineering Sciences. *Computational Intelligence and Neuroscience 2021*. https://doi.org/10.1155/2021/6455592

Elavarasan, D. and Durairaj Vincent, P. M. 2020. Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications. *IEEE Access*8:86886–86901. https://doi.org/10.1109/ACCESS.2020.2992480

Fermont, A. and Benson, T. 2011. Estimating yield of food crops grown by smallholder farmers. *International Food Policy Research Institute* 1: 68.

Joensuu, T., Edelman, H. and Saari, A. 2020. Circular economy practices in the built environment. *Journal of Cleaner Production*, *276*. https://doi.org/10.1016/j.jclepro.2020.124215

Jorvekar, P. P., Wagh, S. K. and Prasad, J. R. 2024. Predictive modeling of crop yields: a comparative analysis of regression techniques for agricultural yield prediction. *Agricultural Engineering International: CIGR Journal* 26 (2): 125–140.

Maestrini, B., Mimić, G., van Oort, P. A. J., Jindo, K., Brdar, S., van Evert, F. K. and Athanasiados, I. 2022. Mixing process-based and data-driven approaches in yield prediction. *European Journal of Agronomy* 139. https://doi.org/10.1016/j.eja.2022.126569

Mariano, C. and Balzarini, M. 2021. A random forest-based algorithm for data-intensive spatial interpolation in crop yield mapping. *Computers and Electronics in Agriculture* 184.

Mbeyagala, E. ., Amayo, R., Obuo, J. ., Pandey, A. K., War, A. R. and Nair, R. M. 2017. *A manual for mungbean (Greengram) production in Uganda*. *September* 1–24. http://www.nasarri.go.ug/scientific-papers/Mungbean Production Uganda.pdf

Nadia, G., Nicola, L. L., Elke, K. A. and James A, O. 2022. Chickpea protein ingredients: A review of composition, func-tionality, and applications, Comprehensive reviews in food science and food safety. *Comprehensive Reviews in Food Science and Food Safety* 21 (1): 435–452.

Nair, R. M., Pandey, A. K., War, A. R., Hanumantharao, B., Shwe, T., Alam, A. K. M. M., Pratap, A., Malik, S. R., Karimi, R.,

Mbeyagala, E. K., Douglas, C. A., Rane, J. and Schafleitner, R. 2019. Biotic and Abiotic Constraints in Mungbean Production—Progress in Genetic Improvement. *Frontiers in Plant Science* 10:1–24. https://doi.org/10.3389/fpls.2019.01340

Pérez-Escamilla, R. 2017. Food security and the 2015-2030 sustainable development goals: From human to planetary health. *Current Developments in Nutrition* 1 (7): 1–8. https://doi.org/10.3945/cdn.117.000513

Raju, K. ., Hegde, V. R . and Hegde, S. A. 2019. Geospatial Technologies for Agriculture. *Springer Link*.

Srinivasa, B., Priya, R., Seemantini,P., Nadiger Sandeep, R. and Khushal, N. Pathade Kamlesh, S. 2023. A Novel Approach for Crop Yield Prediction based on Hybrid Deep Learning Approach. *IEEE Explorer*. https://doi.org/10.1109/ICCES57224.2023.10192652

Setzer, J. and Higham, C. 2021. Global trends in climate change litigation: 2023 snapshot. *Global Trens in Climate Change Litigation July* 37. https://www.lse.ac.uk/granthaminstitute/wp-content/uploads/2020/07/Global-trends-in-climate-change-litigation_2020-snapshot.pdf

Sitienei, B. J., Juma, S. G. and Opere, E. 2017. On the use of regression models to predict tea crop yield responses to climate change: A case of Nandi East, Sub-County of Nandi County, Kenya. *Climate* 5 (3). https://doi.org/10.3390/cli5030054

Srinivasa, B. Priya, R.,Seemantini,P. Nadiger Sandeep, R., Khushal, N. and Pathade Kamlesh, S. 2023. A Novel Approach for Crop Yield Prediction based on Hybrid Deep Learning Approach. *IEEE Explorer*. https://doi.org/10.1109/ICCES57224.2023.10192652

Talwana, H. L., Elepu, G., Wanyera, N. and Obuo, J. 2010. Improving greengram production for nutritional diversification, income and food security in Uganda. *Proceedings of The Second RUFORUM Biennial Meeting* 20–24.