Modeling county level maize yields using artificial neural network,

Trans Nzioa and Nakuru County

Joshua Irungu Mwaura

A thesis submitted in partial fulfillment for the degree of Master of Science in Geospatial Information Systems and Remote Sensing in the Jomo Kenyatta University of Agriculture and Technology

2021

DECLARATION

This report is my original work and has not been presented for a degree in any other university.

Sign: _____ Date: _____

Joshua I. Mwaura
ENC324-2981/2016

This report has been submitted for examination with my approval as the university supervisors.

Sign: _____ Date: _____

Dr. Benson K. Kenduiywo
Jomo Kenyatta University of Agriculture and Technology, Kenya

DEDICATION

This work is dedicated to my family.

ACKNOWLEDGEMENTS

I wish to sincerely thank Dr. Benson K. Kenduiywo for his guidance and invaluable input to this research.

I also wish to thank Mr. Bundotich, Technical Manager, Agricultural Development Corporation and the Ministry of Agriculture, Kenya, for their assistance with reference data.

Lastly, thank the department of Geomatic Engineering and Geospatial Information Systems, Jomo Kenyatta University of Agriculture and Technology for support and guidance that has made this research a success.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Listings

LIST OF ACRONYMS

ANN Artificial Neural Network

RBFN Radial Basis Function Networks

NDVI Normalized Difference Vegetation Index

MLR Multiple Linear Regression

GNA Gauss Newton algorithm

RMSE Root Mean Squared Error

EVI Enhanced Vegetation Index

MODIS Moderate Resolution Imaging Spectroradiometer

SAR Synthetic-aperture radar

AHP Analytic Hierarchy Process

RF Random Forest

BA Boruta Algorithm

MCDA Multi-Criteria Decision Analysis

ELM Extreme Learning Machine

SVM Support Vector Machine

SVR Support Vector Regression

LST Land Surface Temperature

GBM Gradient Boosting Model

MT Metric Tonnes

Ha Hectares

LM Linear Regression

ML Machine Learning

CNN Convolutional Neural Network

LMA Levenberg Marquardt algorithm

GDP Gross Domestic Product

Abstract

The rate population growth in the world continues on a linear trajectory. Africa will experience the highest rate of these growth. Thus, this call for collaborative efforts to plan the future needs of the expected population today. One way of objectively planning the future is working with realistic prediction of population demand. Food production is one of the major focus of the these needs. Reliable food prediction will provide decision makers with the necessary knowledge for proper planning. In local levels, maize yield estimates are useful for county food security preparedness, as these administrative units are mandated to ensure its population is food secure.

The study reviewed various feature selection algorithm. This is because the the efficiency of the prediction model algorithms directly depends on sorting variable importance, and removing the unimportant variable whose contribution to prediction results is insignificant. Based on studies in similar fields, the study adopted boruta algorithm for feature selection against analytical hierarchy process and random forest. The choice of this algorithm was primary because boruta algorithm is an improvement of random forest algorithm and analytical hierarchy process is prone to subjectivity of the respondents.

There are various modelling techniques in the field, from statistical to simulations. Statistical algorithms such as regression models have been widely used in agriculture for modelling various aspects including crop yields. Most of statistical techniques used even national for maize yields prediction lacks the spatial components and assumes yields and influencing factors have a linear relationship. This assumption has been contested by researchers over the world and concluded that, crop yields and their associated factors have a complex non-linear relationships.

Phenomenon that exhibit non-linear relationship are best studied using using advanced level of statistical models categorized as machine learning techniques. This study reviewed three of these models; decision trees, sup-

port vector regression and neural networks. Techniques such as decision trees and regression have been used by various studies to model and predict maize yield. The study adopted an artificial neural network (ANN) for prediction. Neural networks have a number of structures and algorithms within the network for learning and later prediction. This study used a feed-forward, back propagation artificial neural network with levenberg-marquardt algorithm (LMA) for training. Levenberg-marquardt algorithm interpolates between the Gauss Newton algorithm (GNA) and the method of gradient descent. LMA was used for learning.

Artificial neural networks framework was chosen because its a data driven method that is relatively less widely used in county level yield prediction. Moreover, neural networks has key merits, such as require less formal statistical training, ability to detect nonlinear relationships by identifying likely interactions between variables and the availability of multiple training algorithms. We modelled historical maize yield between 2005–2016 as function of satellite derived precipitation, temperature, reference crop evapotranspiration, soil moisture and normalized difference vegetation index (NDVI) to predict maize yields at pixel level. The data was obtained with a spatial resolution of $\approx$ 4 km and subsequently, the predictions was done at $\approx$ 4 km pixel size. The historical reference maize yield data was divided into two sets for model training and validation. The model predicted maize yield with R^2 and root mean square error of 0.76 and $0.038MT/ha$ in Trans-Nzoia county and 0.86 and $0.016MT/ha$ respectively in Nakuru county. These findings shows a promising future for applications targeting to rapidly assess county level food preparedness in Kenya because maize is a major staple food.

Chapter 1: Introduction

1.1 Background

The world population has increased from 1 billion in 1800 to 7.7 billion today. Globally the average population density is $25 people per km^2$, but there are very large differences across countries Global-Change-Data-Lab (2020). In recent years, an increasing number of African people are being added every year. These population increases are unprecedented in history. The problem of population is not a problem of numbers. It is a problem of human welfare and of development rapid population growth can have serious consequences for the well-being of humanity worldwide e.g. food supply OAU and ECA (1994). Since land is limited, the methods and science of food production needs to cope with increased demand. If food production remains constant then its hunger for the world population. Global food demand is increasing driven by population, economic growth and urbanization, particularly in developing countries. At the same time dietary patterns are changing towards more livestock products, including fish, vegetable oils and sugar; a trend that is accentuated by the increasing homogeneity of life habits between urban and rural population facilitated by communications technology. The trendy patterns includes; an increasing demand, changing consumption patterns, food losses and waste, imbalances and changes in food systems and consumers' demands FAO (2010). Kenya is an African country with increasing population.

There is a linear growth in world population according to data and projections published by United Nations. This data also gives 1.18% as the current world's population growth per year, which approximates to annual population increment of 83 million people. More than half of global population growth between now and 2050 is expected to occur in Africa. Africa has the highest rate of population growth among major areas, growing at a pace of 2.55% annually in 2010-2015. Consequently, of the additional 2.4 billion people projected to be added to the global population between 2015 and 2050,

1.3 billion will be from Africa United-Nations (2015). This growth in population will directly impact food supply systems.

Africa relies heavily on weather dependant agriculture. It also experiences short-term changes in climate Stige et al. (2006). These two factors increases stress on food production and food systems. According to Ahmed et al. (2009) a higher percentage of African population is expected to be below poverty line. The consequential effects of this stress on food production is, hunger and poverty, which is prevalent in sub-saharan Africa. Therefore, there is need to prioritize strategies and policies to resolve stress and avert poverty in Africa. As well as measure the impact of policies on set objectives, and protect food production from destructive impacts of future climate changes Lobell and B (2010); Schmidhuber and N. (2007). Predicting food production is of significance in solving food problems, but is not easy. This is because food production or yield is a product of climatic and management factors. Weather is a constituent of climate according to Islam et al. (2020) definition of climate as average weather of an area analyzed for a period of 25-30 years. while weather as the atmospheric conditions of an area at a given day. According to Budyko and V (1996) almost half the total losses in all economic sectors is attributed to unfavorable weather conditions. The management factors are also vital in food production, but these data is not readily available in developing countries. Therefore, it is essential to develop a reliable model of food production using weather parameters.

The inspiration of crop yield prediction is based on needs in food security. The need to achieve competing policy objectives while also protecting public investment in agriculture. Crop yield models help in realizing an equilibrium between various needs such as: increased food production, environment protection, decreased resources, higher farming incomes and climate change mitigation Lobell et al. (2017). Crop yield prediction has two main categories namely: statistical and simulation models. The significance of predicting crop yields has been observed. As many studies have modelled yields using statistical methods with various parameters as a means to food security. According to Zhang et al. (2010) statistical models such as linear regression which is based on ordinary least square (OLS) and autoregressive model can be used for yield prediction. In this study, autoregressive model provided a better performance that was

attributed to this model ability to adjust for spatial autocorrelation inherent in the data. The only weakness to this model is it's linear combination of variables to a process understood to be quite complex and dynamic in nature and thus not easily modelled into a regression framework Zhang et al. (2010). A study by Sellam and E. (2016) established that variables such as annual rainfall, area under cultivation and food price index explains 70% variability in crop yields.

The study by Zhang et al. (2010) also demonstrated that Normalized Difference Vegetation Index (NDVI) and precipitation are the major predictors in modelling corn yield. Studies also demonstrated the use of satellite images in agriculture to improve food production and food security. Satellite images provide both extensive spatial coverage and high temporal resolution. These images brought new possibilities such as to map land cover, detect irrigation, estimate biomass, and survey crop health Chen and H. (2006). Moreover, multiple satellite missions have the capability to regularly monitor phenomena on the Earth surface. These satellites provides a rich source of data that can be ingested into crop yield prediction models Fieuzal et al. (2017).

In recent times, statistical models offer better predictions, but still are not effective with complex data set. These limitations has driven crop yields modelling to adopt data driven models Dahikar and V (2014) such as machine learning algorithms. In line with this, W. and J. (2008) found a robust nonlinear relationship between weather and yields that is consistent across space, time, and crops. This introduced non-linear models in crop yields modelling.

1.2 Problem statement

Maize yield estimation provide a beneficial tool to both farmers and businesses to make decisions and amend or introduce policies before harvest. Maize farming in Kenya is done mostly in highland areas. Highland zones have favourable weather factors for maize farming. Trans nzoia and Nakuru are found in highland zone of Kenya. Yield is function of combination multiple factors. These factors are soil, weather, management and random factors. Random and management data at county level is not readily available, soil data is also limited. The relationship between these factors and yields are

complex, and thus requires non-linear approaches such as ANN for estimation.

Remote sensing data have been found to be useful to monitor crop growth and performance, to estimate cropped area and predict crop yields. Developed world has practiced the use of remote sensing data for agricultural management. However, in the smallholder agricultural areas of sub-Saharan Africa and southern Asia, the potential benefits of remote sensing data in agricultural management have not been sufficiently exploited ITC (2020). According to World-Bank (2016) climate change is causing shifts in weather patterns. These shifts are causing seasonal and weather event-based shocks. Productivity of farms suffer from these shocks, decreasing food security. This uncertainty coupled with the lack of ability to fully predict the impacts, poses significant challenges in county food preparedness.

The use of satellite data and data-driven models can help address challenges of food production uncertainty. Especially by utilizing the computational capacity of machine learning algorithms such as artificial neural networks (ANNs) to model the relationships between predictors (inputs) and objective variables (outputs) Deo and M (2015). The advantages of ANNs in yield prediction are: (1) faster and flexible modeling approach, (2) proper and easy to work non-linear relationship, and (3) the model structure incorporates expertise and user experiences Barzegar and A. (2016). McNairn et al. (2014) concluded that ANNs can be used to predict yields using satellite images as long as models are created for unique crop types. Hota (2014) established that the neural network-based estimation has technical efficiency that may lead to improved results. In this study, radial basis function networks (RBFN) outperformed other estimation techniques in consideration. The study also established ANNs as a beneficial model for crop yield prediction based on sensing various soil and atmospheric parameters Dahikar and V (2014). Africa lacks sufficient in-situ data, but satellite data provides a relatively low cost solution. To ensure timely interventions, yield prediction can provide an early warning on imminent food crisis that may face countries in Africa. Data and information models are necessary to sustain all the dimensions of food security; availability, accessibility, utilization and food systems stability. Reports have shown that without the prior information on expected yields with the relevant stakeholders, country suffers from food scarcity shocks annually. The motivation behind this study is to use satellite

data and ANNs model to predict maize yields prior to harvesting period for sustainable food security. We adopt ANNs of multilayer perceptron, feed forward back propagation to predict maize yield at pixel level as function of weather data derived from satellite in Trans Nzoia and Nakuru counties in Kenya.

1.3 Research objectives

The main objective of this study is to estimate county level maize yields using artificial neural network in Trans nzoia and Nakuru counties, Kenya. The specific objectives of this study will be:

1. To study and assess the most significant factors for maize yield estimation using artificial neural network.

2. To develop and evaluate a maize yield estimation model based on artificial neural network.

3. To estimate maize yield for Trans Nzoia and Nakuru county, Kenya.

Chapter 2: Literature Review

2.1 Introduction

The purpose of this chapter is to provide a detailed review of feature selection algorithms and crop yield estimation methods. Feature selection is an important part of using machine learning for estimation. Some of the key advantages of feature selection to machine learning are to avoid overfitting - curse of dimensionality, to have simple and explainable model, and to avoid garbage in garbage out phenomenon by using non-informative features. This study reviewed feature selection algorithms such as analytical hierarchical model, random forest and boruta algorithm. Likewise the reviewed both statistical or classical and machine learning methods of yield estimation.

2.2 The need for yield estimation

Predicting yields accurately offer an opportunity to decision makers to combat food insecurity. Estimation of yield for main crops such as wheat, corn, rice is of importance to counties. Yield estimation assist in developing plans for food production, distribution and consumption in preparation for food shortages and supply shocks. Food shortages results from various combination of factors, either human induced or natural occurrences Khairulzaman et al. (2014). There are various methods and models that have been used for yield estimation. These methods and models includes; regression, simulation, expert systems, and artificial neural networks O'Neal et al. (2002). The models maybe either linear or non-linear systems. The linear systems assumes linear relationships among the input parameters,while non-linear assumes non-linearity. Therefore, most of the linear models are not able perform well because of complexity and non-linear nature of the data Khairulzaman et al. (2014).

Linear regression approaches have been widely used Medar and Rajpurohit (2014).

Mainly, because of ease of use and standard accepted tests of reliability. Which tends to favour regression, despite problems with predictive accuracy caused by dependence on the specific conditions of the input data used to develop the regression O'Neal et al. (2002). Multiple linear regression (MLR) modelling is also very powerful technique and is widely used to estimate linear relationship. Its assumption of linearity in variable relationship is also its limitation. Which in real situation is rarely satisfied. Also, if there are several predictors, it is well nigh impossible to have an idea of the underlying non-linear functional relationship between response and predictor variables Singh Rama Krishna (2008).

Simulation has the advantage of being based on physical relationships, specifying relevant factors affecting yield, and allowing researchers in different areas to use the same sophisticated model that is widely accepted. However, simulation requires many biophysical inputs that often must be estimated rather than measured, and calibration can be time consuming for areas without established sets of parameters. Expert systems rely on human expertise and characterize yield by sets of logical rules, but the initial formation of rules requires extensive communication with the expert, is not readily automated, can be highly subjective, and relies on the limits of the input data.

Artificial neural networks (neural nets), on the other hand, are easily automated, display remarkable accuracy for new situations not represented in the input data, consist of objective mathematical functions instead of subjective rules, do not require preestablished physical relationships, and can be built with readily available input data Singh Rama Krishna (2008). Soil factors, weather factors and management factors directly or indirectly influence crop yield. This influence is either either of a linear or nonlinear relationship to yield.Mahabadi (2018) designed a feed-forward back-propagating ANN model to predict yield for rice. The results of this study showed high performance of a trained neural network to predict the yield of rice.

In another study, Kross et al. (2018) noted the complexity interactions that exists between crop growth and the interrelated variables. To model yields, the study noted the usefulness of artificial neural networks (ANNs) for such complex interactions in system. ANNs models are capable of capturing the non-linear relationships of data with little understanding of the underlying processes. ANN models superiority was observed with

consistent and more accurate yield predictions than regression models by Kaul et al. (2015) for accurate corn and soybean yields prediction under typical Maryland climatic conditions. In a study done in India, Veenadhari et al. (2014) developed a prediction tool based on ANN that used climatic factors only to predicted yield in various regions. The model achieved an accuracy above 75% in all the crops and districts selected. Seshadri et.al integrated an optimization technique with artificial neural network for prediction of paddy yield at 3 different districts in different climatic zones based on 10 years of historical data sets of yields of paddy. The error range was 8% which according to Baral et al. (2011) is within the accuracy requirements.

In variable sensitivity and wheat prediction panah (2008) used an output matrix of wheat yield during the 1999-2005 period. The analysis shown that for the ANN model the most important climatic factor determining wheat yield is the amount of rainfall. Excluding this factor from the input matrix, had an increase in the models RMSE. The use of ANNs based models in crop yield production and prediction is realizing substances gains. This study also shows ANN's potential in predicting maize yield in two counties in Kenya, where maize farming is a key agricultural activity.

2.3 Crop calendar

The time of planting is the most critical factor in farming. To realize high yields from maize, planting should be done at the onset of rainfall. This allows the germinating seed to benefit from the nitrogen flux effect which occurs within the first rains Atlas (2013). In Kenya, different regions have different planting times. Trans Nzoia is a highland - receives high rainfall. The planting time is between March to Mid-April and the harvesting is between September-December. For the highlands, the growing duration takes about 180-270 days Atlas (2013). The Figure 2.1 shows the calendar seasons experienced in Kenya.

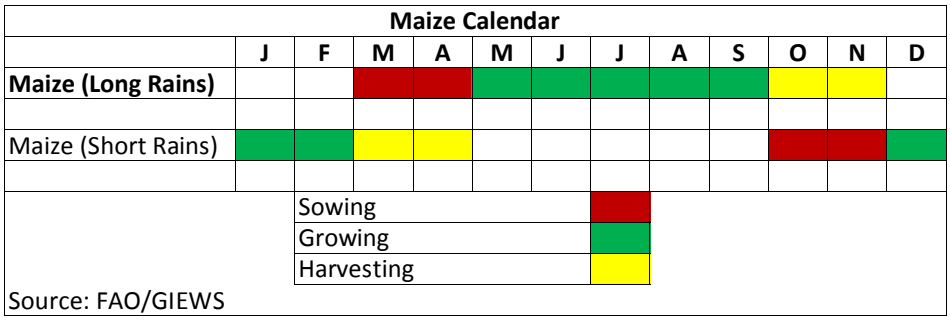


Figure 2.1: Maize Calendar Seasons in Kenya. *Source: FAO/GIEWS, FEWSNET*

The two study areas shared this maize farming calendar in the long rains.

2.4 Yield estimation

Crop production is a complex phenomenon. Agriculture input parameters varies across fields and farmers. This complexity and input parameters variation is influenced by agro-climatic input parameters Veenadhari et al. (2014). The climatic information in Kenya isn't readily available, and thus rely on satellite derived data and yield data collected by *Ministry of Agriculture*.

The aim of agricultural production is to achieve high crop yields. The recognition and management of factors that influence crop yields assist farmers in decision making. There are a number of crop yield prediction models which use either statistical or crop simulation models. The last decade has seen that artificial intelligence techniques provide a more effective approach to predicting crop yield under varying cropping scenarios Niketa et al. (2016). This is mainly because these techniques are able to model complexity from inputs.

2.4.1 Field surveys

The National Agricultural Statistics Service of the US Department of Agriculture uses both phone interviews and field surveys to forecast the yield of several commodities, in-

cluding maize, soybean, and wheat. The enumerators conduct monthly phone interviews with pre-determined growers during the growing season. The growers' assessment of the yield prospects reflects farmers' opinions about the impact of weather events and growing-season conditions on the final yield. In the field surveys, trained enumerators conduct field surveys to collect plant status at close to harvest stage. The stratified sampling technique is where land is categorized by its intensity of cultivation, and where fields from each category are sampled with a greater frequency for intensely cultivated land than marginally cultivated land. Maize, soybean, and wheat yield forecast at field level is done through various mathematical equations estimating the number of fruit per plant and the weight of each fruit based on the plant characteristics collected in the field surveys Branch et al. (2012).

2.4.2 Statistical methods

Simple linear regression

Regression method is common method for forecasting. The method use agrometeorological data as inputs to a statistical regression as a form of seasonal yield forecast. Simple statistical models can be built as a matrix with historical yield data and any number of agrometeorological parameters (e.g., precipitation and temperature). From the matrix, regression equations are derived as a function of the inputs to generate a seasonal yield forecast. The simplicity of a statistical regression model is also the driving force for its applicability, but at the same time its limitation in extrapolating results to other areas outside the boundaries of the observed data. More recently, given the increase in climate variability, and more frequent extreme events, these models are poorly suited to forecast or estimate future yields Basso and Liu (2019). A regression describes the underlying relationship between y_i and x_i involving this error term e_i by

$$y_i = a + bx_i + e_i \tag{2.1}$$

Multiple linear regression

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data.

Combinations of various climatic variables, including minimum and maximum temperatures, relative humidity in the morning and evening, and rainfall, have been used as predictors to forecast maize, wheat, and rice yield Basso and Liu (2019). Every value of the independent variable x is associated with a value of the dependent variable y.

$$y_i = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + ... + b_n x_n \tag{2.2}$$

2.4.3 Remote sensing

There have been an uptake in using satellite images for crop yield estimation. Some of the derivations used yield estimation from images includes, Leaf area index (LAI), vegetation indices (VIs), and satellite derived weather data. These datasets includes; precipitation, temperature, evapotranspiration, soil moisture, wind speed, among others. Remote sensing data has inspired many greenness index for yield estimation such as Enhanced vegetation index (EVI), Two-band enhanced vegetation index (EVI2), Greenness normalized difference vegetation index (GNDVI), Soil adjusted vegetation index (SAVI), Vegetation health index (VHI). Chlorophyll index such as Chlorophyll index red-edge and Fraction of absorbed photosynthetically active radiance (FAPAR) Basso and Liu (2019).

2.4.4 Machine learning

Machine learning (ML) is a branch in the field of artificial intelligence that assists computers in modeling based on historical data and accurately predicting future outcomes. ML approaches are mostly classified into two main categories: supervised learning and unsupervised learning. Classification and clustering are examples of problems in supervised learning and unsupervised learning respectively. The widely used techniques for classification include neural networks, support vector machines, and decision trees, and the most widely used clustering technique is k-means Dogan and Birant (2020). Machine learning is generally more efficient than traditional mathematical and statistical models in various fields since they remain capable of understanding complex relations among features of data samples and predicting unknown feature values for a new sample. This distinguishing characteristic has made ML techniques applicable in a wide

range of scientific disciplines as well as in agriculture. Feature selection is an important part of machine learning algorithm to perform optimally. This study reviewed various feature selection algorithms and adopted boruta algorithm as discussed below.

Feature selection

The *curse of dimensionality* refers to various phenomena that arise when analyzing data in high-dimensional space such as increasing computational burden from existence of voluminous data in the big data era. As the data become sparser in high-dimensional space, and a large number of samples are needed to train models, which greatly decrease the efficiency of data processing. Feature selection provides one of the most effective tools to reduce the dimensionality and increase data processing efficiency Wei et al. (2020b). Feature selection play a vital role in preventing over-fitting, facilitating data visualization, reducing storage requirements and computational costs, and improving the accuracy of pattern recognition algorithms Wei et al. (2020b). Feature selection helps discover which of these inputs is more impactful in modeling for best performance. The model accuracy, stability, and effectiveness directly depends on the relevance of input dataset. Therefore, it is necessary to select the best set of features by identifying the important variables from the overall input data set Kim et al. (2020). Feature selection methods can be split into three major classifications such as filter, wrapper and embedded methods. Filter methods requires definition of some metrics such as correlation/chi-square. Wrapper methods consider the selection of a set of features as a search problem such as recursive feature elimination. Embedded methods use algorithms that have built-in feature selection methods such as, Lasso and RF. The study considered 3 feature selection approaches which are Analytic Hierarchy Process (AHP), Random Forest (RF) and Boruta Algorithm (BA). Based on literature review, the study adopted feature selection method with ground scientific basis with no little human interference to identify the relationship among features and thus ascertain both relevant and redundant features.

Analytic Hierarchy Process For the complex multi-criteria analysis task, a number of Multi-Criteria Decision Analysis (MCDA) models exist. The difference occurs in type of decision criteria, type and number of alternatives, approach to compensation

among decision criteria, and preference ordering. The choice of the method depends on the practical fact, the criteria used for assessing the matrix calculation, possible alternatives), and on how the decision is made. The AHP methodology consists of weighting and ranking procedures considering involved factors defined as criteria layers. AHP is a realistic method organizing and analyzing multi-criteria decisions quantifying the index weight by comparing relative factors with each other Wei et al. (2020a). The weakness of AHP is that the decisions are based on knowledge of respondents, whose knowledge on the subject may not be objective.

Random Forest Multivariate linear regression (MLR) and artificial neural network (ANN) are common black-box modelling methods. Model accuracy is the primary concern on selecting the modelling methods. Accuracy of the model depends on feature selection by the identification of variables importance, to ease of application and time of computation. Depending on the nature of operating variables, various algorithms embedded in the modelling methods may be considered to reach optimal results of the operating variables. Regarding ANN models, the relationship between the input and output variables are hidden and cannot be identified explicitly. While the MLR models can illustrate the correlation between the input and output variables by regression coefficients, the accuracy of MLR models depends on which input variables and their multivariate terms. The degree of input variables affect the accuracy of a model Yua et al. (2017). The random forest (RF) models make use of classifying input variables by decision tree methods to identify variables importance and predict output variables with high accuracy. The RF algorithm can handle a huge set of input variables subject to certain outliers and noise in data Yua et al. (2017). An improved Random forest is Boruta Algorithm.

Boruta Algorithm The study adopted boruta algorithm for feature or variable selection because it's based two important concepts namely: shadow features and binomial distribution. Boruta is a feature selection algorithm that works with various data and is capable of working with any prediction method to determine the importance of variable Kursa and Rudnicki (2010). Boruta implements the first concept by randomly creating shadow features to compete with original features. The shadow feature with highest recorded importance becomes the threshold. The importance of each original features

is compared with this threshold. The original features with performance above this threshold, are selected. Secondly, the variable importance is obtain through an iteration process that follows binomial distribution. The maximum level of uncertainty about the feature is expressed by a 50% probability for selection or elimination S (2020). Boruta works best with transformed data. Data transformations takes various forms such as *maxmin* and *log* O'Neal et al. (2002).

Maxmin: This transformation method maps maximum and minimum value to the desired limits. An example of this type of scale, transforms data between 0 and 1, using the equation below

$$x_{output} = \{\frac{x_{input} - x_{min}}{x_{max} - x_{min}}\}$$

Log: This transformation method scales data using the equation below;

$$g(\varphi_i) = \frac{ln\varphi_i - ln(\varphi_i - 0.001)}{ln(\varphi_{min} - 0.001)}$$

The range of normalized values using log is valid (0-1) only for data between $(\varphi_{min} - 0.001)$ and $(\varphi_{min} - 0.001)^2$. Negative values are possible when the minimum is lower than or equal to 1. If the maximum values does not exceed the square of the minimum value, and all values are greater than 1, it may be suitable. Its suitable to data with large minimum values, to decrease the number of nodes needed for inputs.O'Neal et al. (2002)

Decision trees

Decision trees is one class of classical classification methods or is a hierarchical classifiers which determines a class by multi-level discrimination. Decision trees can be categorized into univariate decision trees, multivariate decision trees, and omnivariate decision trees Wang et al. (2020).

Univariate decision trees refer to those where only a single attribute participates in node splitting, multivariate decision trees are those in which multiple attributes participate in node splitting and Omnivariate decision trees are decision trees in which the splitting at each node is univariate, linear multivariate, or nonlinear multivariate Wang et al. (2020). Although the classical classification approaches are good, literature review has shown that performance of data driven approach is better.

Support Vector Machine Regression

Support vector machine is a pattern classification technique or an algorithm that implements non linear boundaries between classes by transforming the input data into a high dimensional space. SVM employs an approach that attempts to minimize the upper bound of the error by maximizing the separation of the boundary between the hyperplane and the training data. A key feature of SVM is that training SVM is equivalent to solving linear constrained quadratic programming problems. Therefore, SVM solutions are always unique and globally optimal. In SVM, the process of solving the problem depends only on a series of the training data, which is called a support vector. Use only the support vector to get the same solution as using all training data points Shao et al. (2020). After classification, SVM has an extension for prediction called support vector regression.

Support vector regression (SVR) is popularly known for prediction and attempts to minimize the generalization error bound so as to achieve generalized performance. The idea of SVR is based on the computation of a linear regression function in a high dimensional feature space where the input data are mapped via a non linear function. Similar to ANN, support vector regression (SVR) uses a regression analysis to regress dependent variables on explanatory variables based on a weight vector and a bias term. However, the SVR only uses a subset of the data called support vectors to establish relationship between the desired variable and the explanatory variables. The SVR regression equation is determined by optimizing a quadratic weight objective function - an optimization that make use of Lagrange multipliers and quadratic programming-based numerical optimization. Unlike the ANN, SVR do not get stuck in local minimum during optimization, and it always give global minimum when the optimization is complete. SVR is less prone to overfitting the regression function because it uses insensitive loss function and structural risk optimization Achieng (2019). Related studies shows that support vector machine (SVM) and artificial neural network (ANN) produce the most successful results.

Neural networks

Artificial neural networks is one of machine learning techniques which consists of an interconnected assembly of simple processing elements (nodes/neurons) Singh Rama Krishna (2008). Those neurons are connected to each other using weights Jiang et al. (2004). The processing ability of the neural network is stored in the interconnected weights obtained through learning from a set of training patterns Singh Rama Krishna (2008). Neural networks have an advantage with complex relationships because of the interconnections of weights within its network. Thus, it is a reasonable model to attempt to predict maize yield O'Neal et al. (2002). In addition, neural networks do not require a specific distribution. The networks are tuned to reach a certain threshold of error by training iterations O'Neal et al. (2002). The nature of the activation functions used in back-propagation networks dictate on the scaling of the inputs O'Neal et al. (2002). The values of updated neurons need to be between 0 and 1. The study adopted the sigmoid activation function, i.e.

$$f(x) = \frac{1}{(1 + e^{-x})} \tag{2.3}$$

where x is the input for the respective input layer of the neural. In the neural networks, back propagation computes the gradient. The neural learns (during model training) by adjusting the weight (w) and bias (b) for each layer using these gradients.

The ANNs model was trained for prediction using the Levenberg-Marquardt algorithm. This is a hybrid technique that uses both Gauss-Newton and gradient descent approaches to achieve optimal solution Wilson and A. (2013). The hybrid approach uses the best characteristics of these two techniques. Gauss-Newton technique is normally faster when the initial guess is relatively close to the optimum, otherwise the algorithm uses the gradient descent technique to find an optimum,

$$x_{k+1} = x_k - \left[J^T J + \mu I\right]^{-1} J^T e \tag{2.4}$$

where J is the jacobian matrix of performance, $J^T J$ is an approximation of the matrix, μ is the adaptive value, x is the variable, e is the error and $J^T e$ is the gradient descend computation. The small values of the parameter μ result in a Gauss-Newton update and large values of μ result in a gradient descent update. This algorithm adaptively varies the

parameter updates between the gradient descent update and the Gauss-Newton update making it an efficient method for weights adaptations Wilson and A. (2013).

2.5 The research gaps

There exist a robust non-linear relationship between weather factors and maize yields. The statistical algorithm mostly used for yield estimation only relies on linear relationship. The non-linear relationships still needs more research using the ANN models as they do not require knowledge of the underlying process to understand relationships of inputs and output parameters. This research intends to use non-linear machine learning algorithm to model yields based on weather factors.

Chapter 3: Materials and Methods

3.1 Introduction

This chapter provides the description of the research process. Provides information about the study area to bring an understanding of the area for research. The chapter also provides a description of the data used in this research, methodology and validation of the results.

3.2 Study area

The study adopt two counties (Trans-Nzoia and Nakuru as shown in Figure 3.1) in Kenya for maize prediction due data availability.

Trans Nzoia county covers an area of about 2,495 km^2, with a population of approximately 1 million KNBS (2019). The climate in Trans Nzoia is mild temperatures, with rainfall of around 1097 mm per year. The main activity is largely agriculture and livestock rearing. Large-scale agriculture is mainly on wheat, maize and dairy farming, while small-scale agriculture is on maize, beans and potatoes.

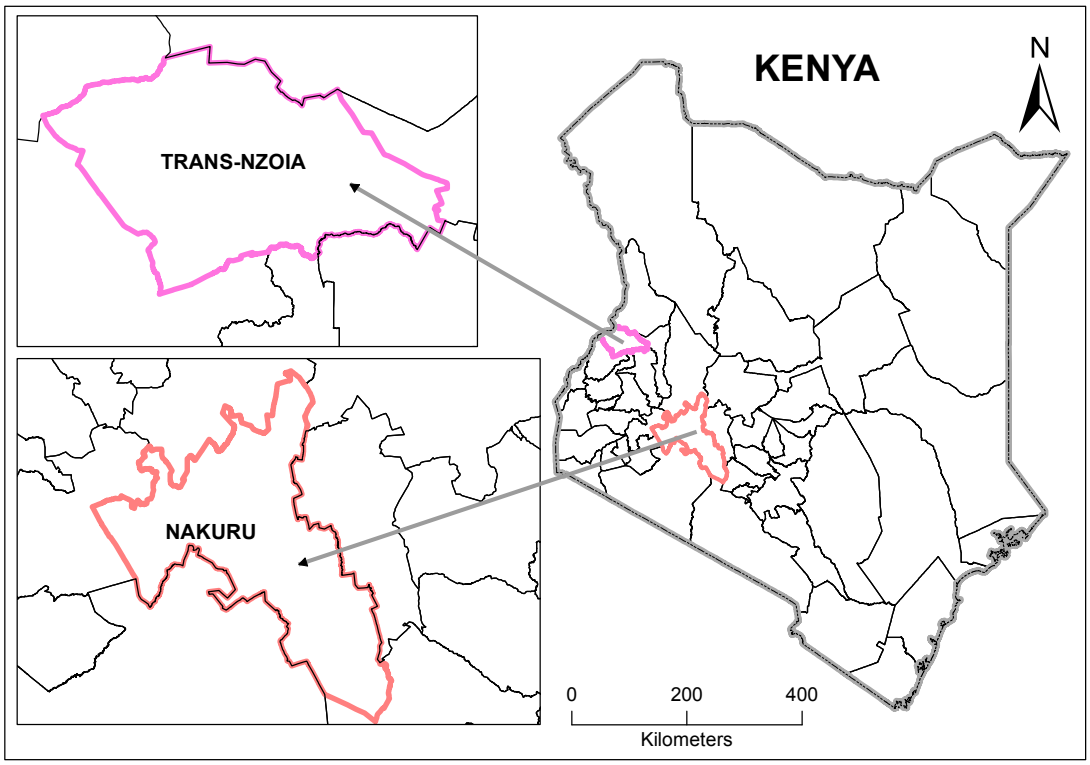


Figure 3.1: Trans Nzoia and Nakuru County.

On the other hand, Nakuru county lies south east of Trans Nzoia and covers an area of about 7,505 sq km, with a population of approximately 2 million people KNBS (2019). The county has also mild temperatures with rainfall of around 895 mm per year. The main activity is agriculture and livestock rearing. Large-scale agriculture is mainly on barley, maize and dairy farming, while small-scale agriculture is on maize, peas and potatoes. Maize is rain-fed in the two counties with the sowing period in March and harvesting in November to December.

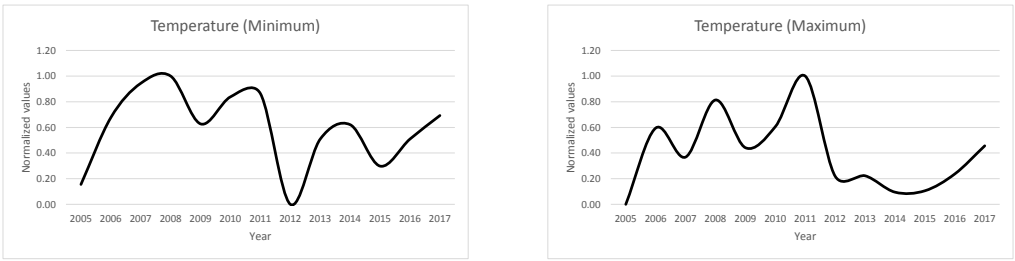
3.3 Data

This study used precipitation, minimum temperature, average temperature, maximum temperature, reference crop evapotranspiration, and NDVI derived from Landsat 7 USGS (1990). All the primary weather factors Islam et al. (2020) such as precipitation, minimum temperature, maximum temperature, and the derived factors such as average temperature, evapotranspiration, and soil moisture were obtained from climatology lab as

multi-band raster images. The data has been validated with a number of station-based observations from a variety of networks including the global historical climate Network, SNOTEL, and RAWS Abatzoglou et al. (2018). All data have monthly temporal resolution and a spatial resolution of ≈4 km. The data cover the period from 1958–2019. The historical maize yield data was obtained from the Ministry of Agriculture in Kenya.

3.3.1 Minimum, Maximum Temperature

Temperature is an essential agro-climatic parameter influencing the rate of plant development. Climate change will influences environmental change and consequently variation in temperature events will affect crop production. Thus the use of minimum, and maximum temperature for each year from 2005 to 2016. The minimum, and maximum temperature from *March to August* was derived from monthly temperatures of Trans Nzoia.



(a) Minimum (b) Maximum

Figure 3.2: Temperature.

3.3.2 Precipitation (mm)

Precipitation replenishes the water cycle and fresh water on the planet. The aggregate precipitation from *March to August* for each year was computed from monthly mean precipitation from the year 2005 to 2016 of Trans Nzoia County.

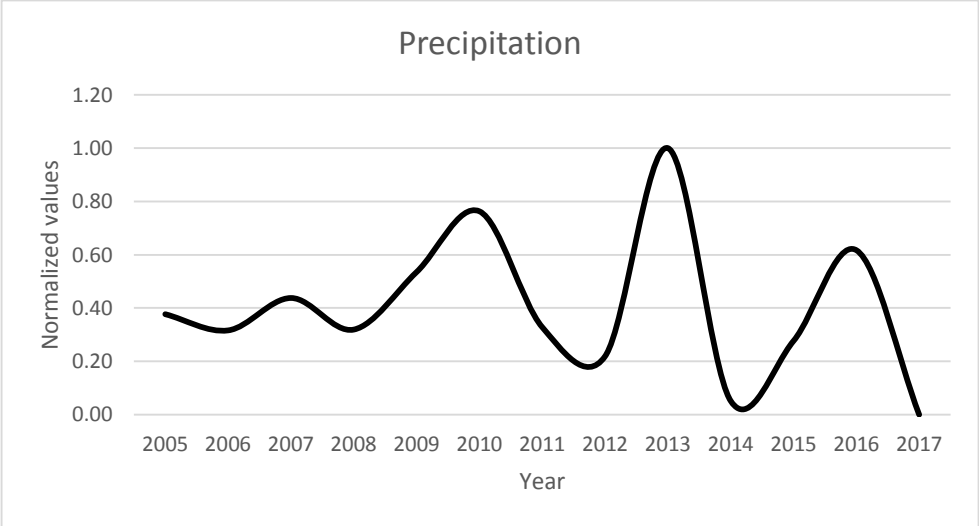


Figure 3.3: Precipitation

3.3.3 Evapo-transpiration (mm)

Evapo-transpiration is the sum of evaporation and plant transpiration from the Earth and sea surface to the air. Evapo-transpiration was calculated on the monthly basis from year 2005 to 2016 of the study area.

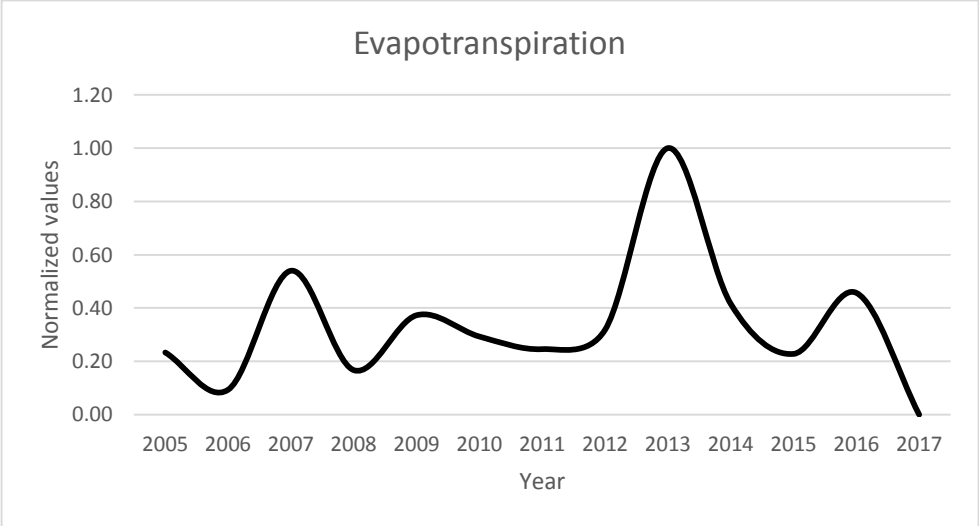


Figure 3.4: Evapotranspiration

3.3.4 Soil Moisture

Soil moisture generally means the water that is held in the spaces between soil particles. Soil moisture within the root zone is the water available to plants, which is considered to be about 200 cm of topsoil. Soil moisture was calculated on the monthly basis from year 2005 to 2016 for the study area.

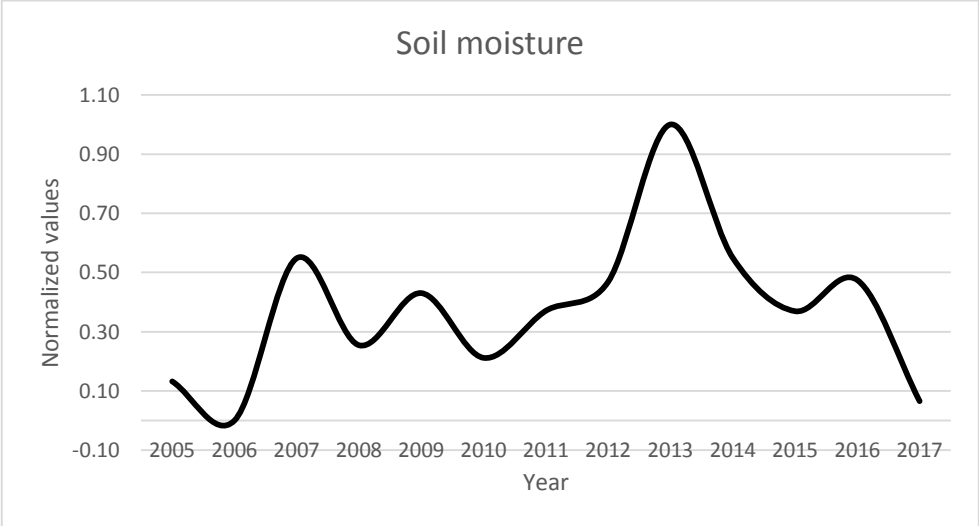


Figure 3.5: Soil moisture

3.3.5 Vegetation indices (e.g. NDVI)

Normalized difference vegetation index is a measure of greenness of crops. It indicate the health of vegetation. The research used the Landsat vegetation NDVI images on monthly basis from the year 2005 to 2016.

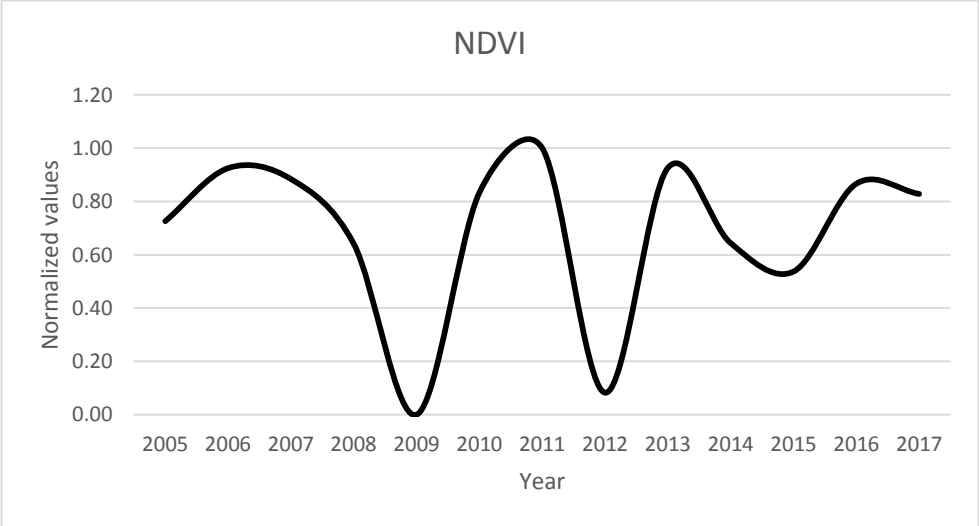


Figure 3.6: Normalized difference vegetation index

3.3.6 Area (Hectares)

The maize cultivated area of Trans Nzoia County in the long rains season (March to November) from the year 2005 to 2016 was considered in the study.

3.3.7 Production (Tonnes)

The maize production for Trans Nzoia County for March to November was considered in this research.

3.3.8 Maize yield (tonnes/hectare)

An amount of maize produced and the area cultivated for maize in Trans Nzoia in long rains season (March to November). The yields considered from the year 2005 to 2016 in tonnes per hectare.

The importance of soil moisture information to predictive yield models was noted by

Filippi et al. (2019), and soil moisture products are now becoming more readily available at finer spatial and temporal resolutions, particularly through the use of the Sentinel satellites Torres et al. (2012). Both received rainfall were highly important variables in the models, and the inclusion of other climatic data variables, such as temperature could also improve the model predictions McMaster and W (1997). A study used historic data of wheat yield and associated plantation area, rainfall, and temperature, but also noted that incorporating statistics and artificial neural networks can produce highly satisfactory forecasting of wheat yield Guo and Xue (2014).

3.4 Data Processing

The Figure 3.7 shows the data processing workflow adopted in this study.

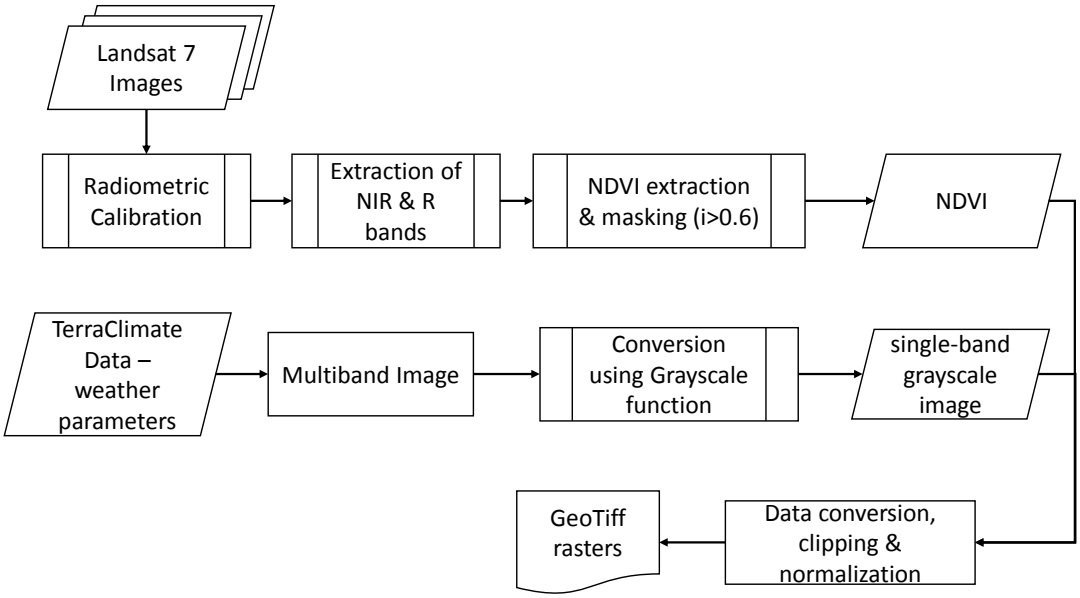


Figure 3.7: Data Processing.

The Landsat images were calibrated so as to convert digital numbers to spectral radiance (Figure 3.7). We then used the Near Infra-Red (NIR) and red (R) bands to compute NDVI as

$$\text{NDVI} = \frac{\text{NIR} - \text{R}}{\text{NIR} + \text{R}}. \tag{3.1}$$

NDVI values between -1 to 0.6 were masked out yielding a raster with values from 0.6 up to 1.0 which represents vegetation. The multiband rasters from for weather parame-

ters were converted to single-band raster to unit weights using grayscale function

$$\text{Output} = (B1 \times W2) + (B2 \times W2) \tag{3.2}$$

where $B1$ is the first raster and $B2$ is the second raster in the multi-band raster, $W1$ and $W2$ were set to 1. The resultant raster, in netCDF file format, was converted to geotiff format. The data was normalized using the min-max transformation, i.e.

$$X = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.3}$$

where X is a variable representing one of the data sets used.

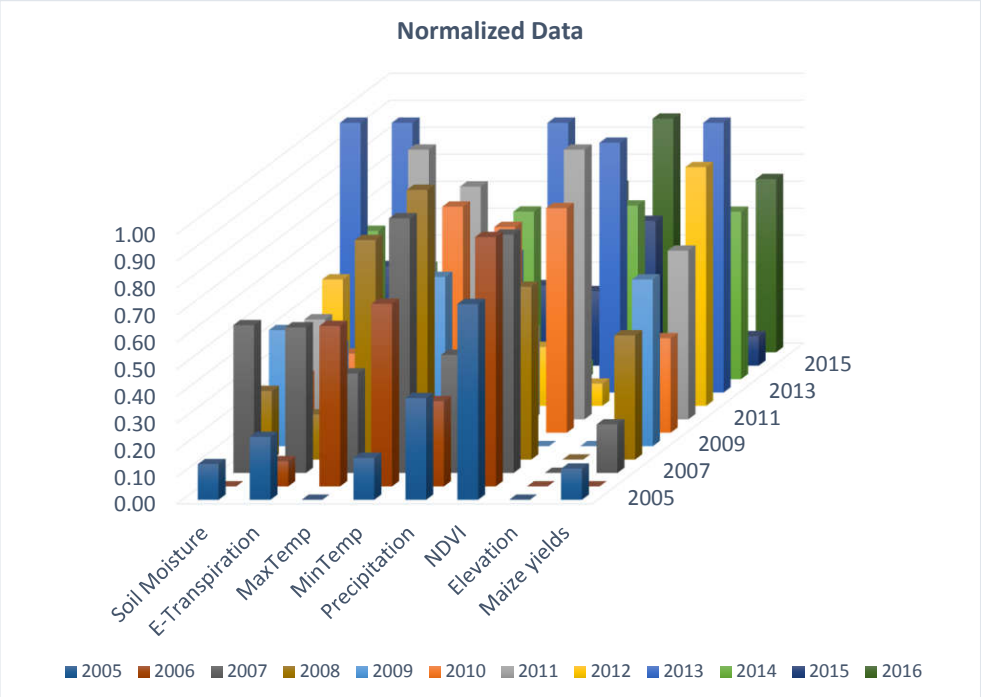
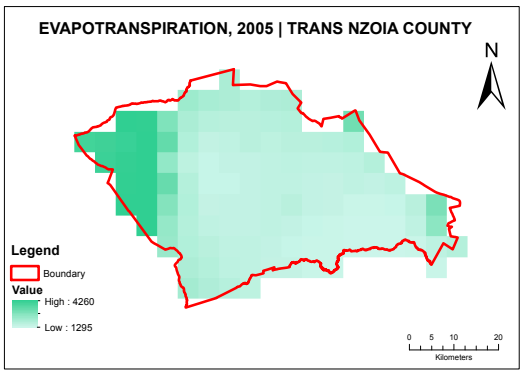
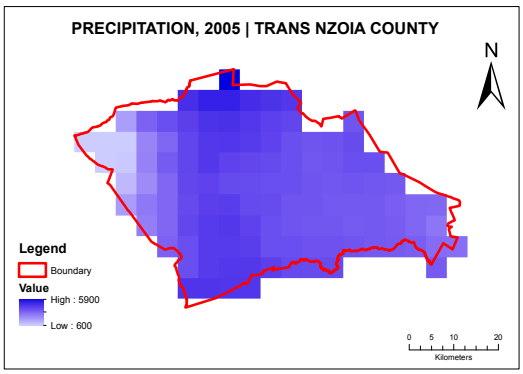


Figure 3.8: Normalized variables in Trans Nzoia county: *precipitation, temperature (min, max), evapotranspiration, ndvi, soil moisture, elevation, and yields*.

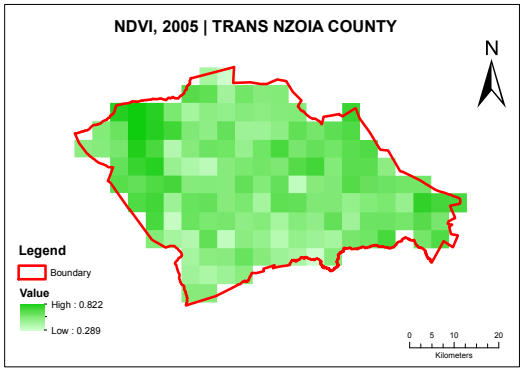
The figures below shows some of the pre-processed inputs are shown for Trans Nzoia and Nakuru county respectively.



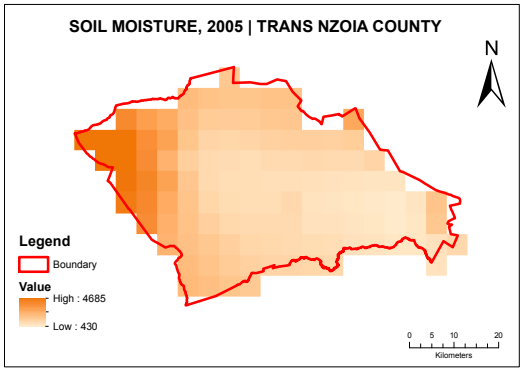
(a) Evapotranspiration



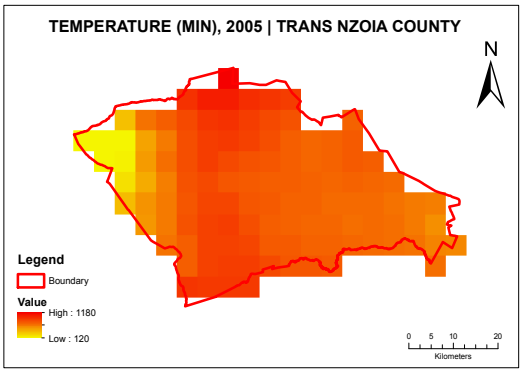
(b) Precipitation



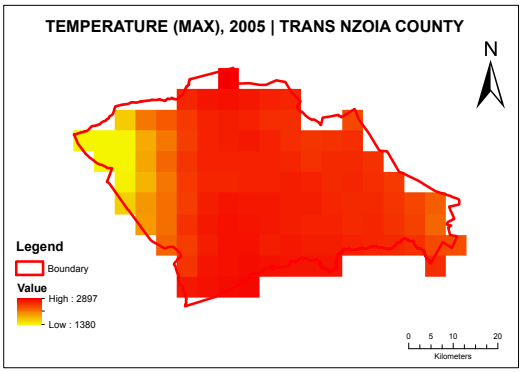
(c) NDVI



(d) Soil moisture



(e) Minimum temperature

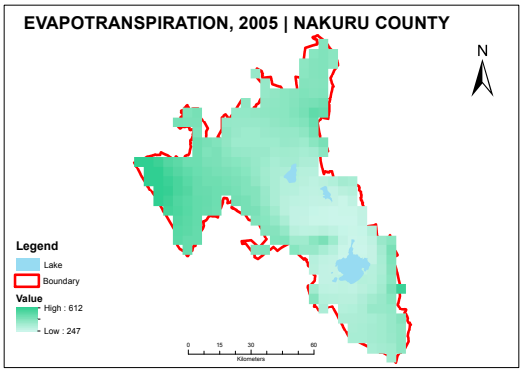


(f) Maximum temperature

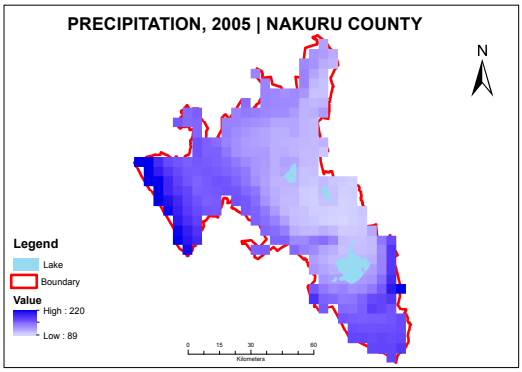
Figure 3.9: Trans Nzoia County

The climate in Trans Nzoia is mild temperatures, with rainfall of around 1097 mm per year. A closer look and the spatial distribution of the factors shows that, the temperature can go to a high of 2897° Celsius in the lowlands and low of 120° Celsius in the mountainous zone. Soil moisture and evapotranspiration are high in the mountainous zone and decreasing south-east to the lowlands. Precipitation with Trans-Nzoia is

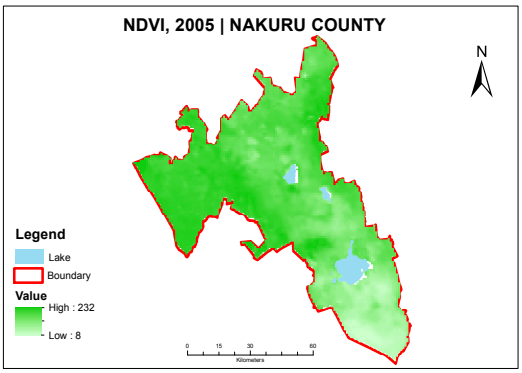
relatively across the county.



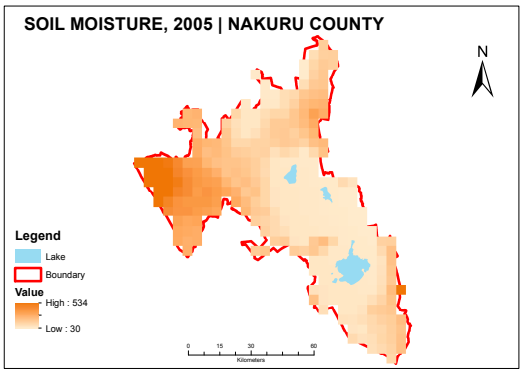
(a) Evapotranspiration



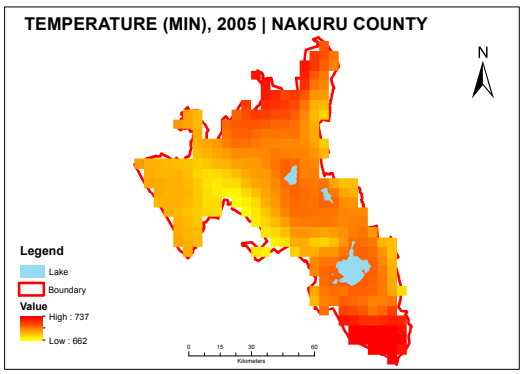
(b) Precipitation



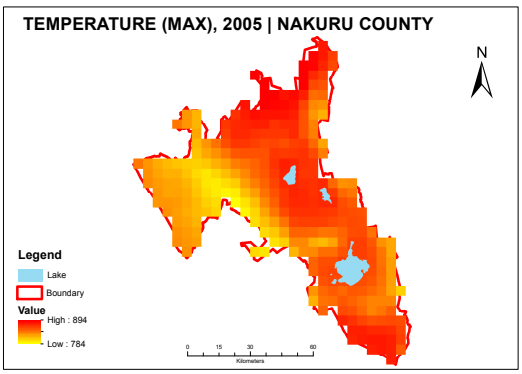
(c) NDVI



(d) Soil moisture



(e) Minimum temperature



(f) Maximum temperature

Figure 3.10: Nakuru County

Nakuru county has also mild temperatures with rainfall of around 895 mm per year. Using the Figure 4.2b above, there is spatial variability of these factors. Nakuru has high temperatures along a direction from north-west to south-east. Soil moisture, precipitation and evapotranspiration are high in the north-west regions.

3.5 Yield Prediction using ANN

The study developed an ANN yield prediction model based on selected variables/features discussed above, mainly, satellite data. Neural network adopts the parallel architecture of our brain and the operation of biological neural networks Puig-Arnavat and C. (2015). The algorithm is designed to recognize patterns in complex data optimally. Neural networks have neurons with connections. A neuron contain a value and activation function whereas connection holds a weight and bias. The neurons are divided into input, hidden and output layers. Neural networks have three parts; feed forward, activation functions, and back propagation Kadir et al. (2014).

The term feed forward in neural network refers to the process of updating the neuron in the next layer, by multiply the activations by weights. Activation functions are the logistic functions. They scale the values of updated neurons to be between 0 and 1. The study adopted the sigmoid activation function.

The ANNs model was trained for prediction using the Levenberg-Marquardt algorithm. This is a hybrid technique that uses both Gauss-Newton and gradient descent approaches to achieve optimal solution Wilson and A. (2013). The hybrid approach uses the best characteristics of these two techniques. Gauss-Newton technique is normally faster when the initial guess is relatively close to the optimum, otherwise the algorithm uses the gradient descent technique to find an optimum. This algorithm adaptively varies the parameter updates between the gradient descent update and the Gauss-Newton update making it an efficient method for weights adaptations Wilson and A. (2013).

In our study, the learning algorithm was based on feed forward back propagation multi-layer neural networks. The ANN model used in the study has following types of activation functions: tangent sigmoid function, sigmoid function and linear function. In back-propagation, sigmoid function and linear function are used as the activation functions. In the predictive model, the study used the Levenberg-Marquardt algorithm with linear and tangent sigmoid functions as activation functions. The initial step included assigning of model weights and thresholds, followed by neuron activation using the activation functions. The weights were updated based on the 6 neurons for input and hidden layer, and 1 neuron for output layer.

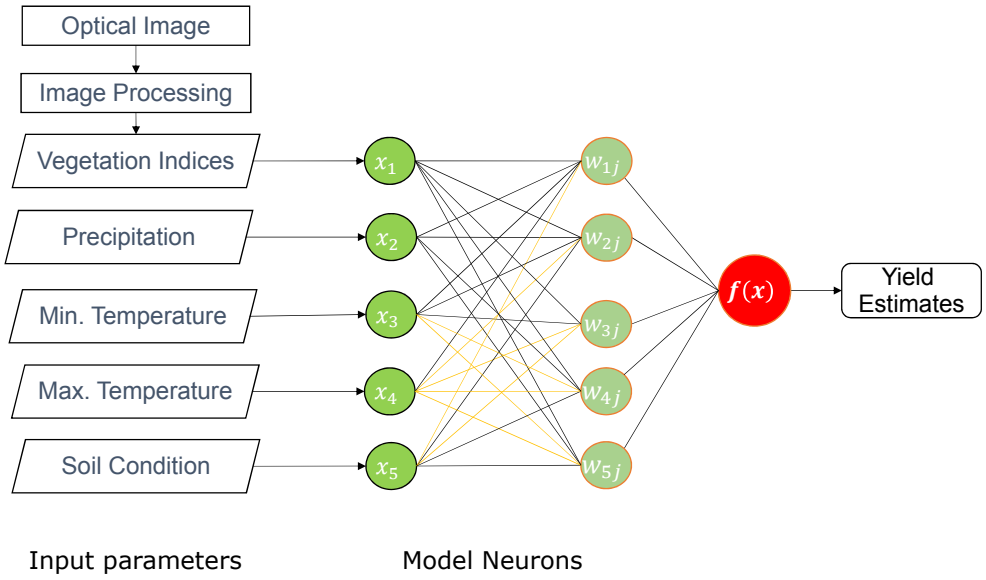
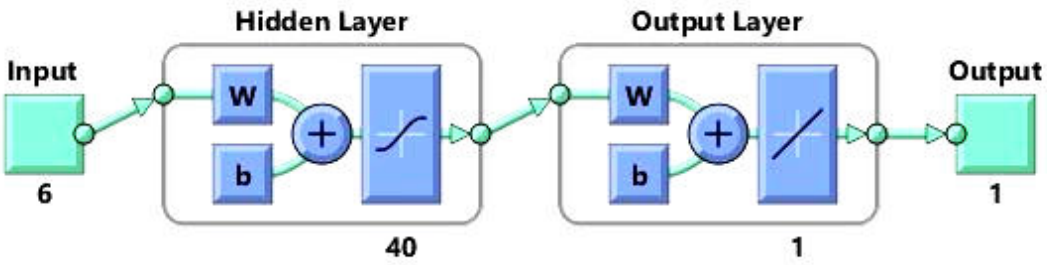


Figure 3.11: Artificial neural network (ANN) architecture.

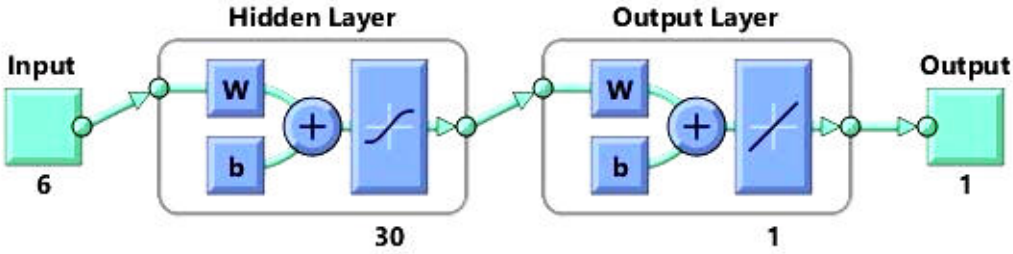
The prediction model (Figure 3.11) has six input variables which results in total of 6960 ($6 \times 116 \times 10$) data points. The data was normalized between 0 to 1, to neutralize the effect of influence by large data. The input variables were selected based on their influence to yields using boruta algorithm. The data was then divided randomly; 70% for training, 20% for validation and the remaining 10% for testing the model to determine optimum performance in modeling maize yields.

In retrospect, we designed two ANNs model configuration for the two counties. Figure 3.12a shows the model structure adopted for maize yield prediction in Trans Nzoia county. The best model fit was achieved in 40 iterations and attained a max performance. On the other, hand Figure 3.12b illustrates the model structure designed for Nakuru county.

The prediction process starts from the smallest network architecture and is gradually the number of hidden neurons increased. Based on literature review the ANN architecture of the activation function (nonlinear) was determined, which in this study is a sigmoid function. Once the activation functions and the number of hidden layers is set and adjusted respectively, ANN learning process was carried out. This process of adjusting hidden layers and re-learning the ANN is repeated till best the statistics are achieved.



(a)



(b)

Figure 3.12: Designed ANNs architecture for yield prediction in (a) Trans Nzoia and (b) Nakuru counties.

The best ANN architecture for estimation of maize yields was obtained using a trial-and-error approach by varying the number of neurons in the hidden layer. As is shown in Figure 3.12a, the developed ANN prediction model consists of 40 neurons in the hidden layer and the transfer function pair tansig–purelin, generating a 6-40-1 ANN architecture for Trans nzoia county. Similarly, in Figure 3.12b, the developed ANN prediction model consists of 30 neurons in the hidden layer and the transfer function pair tansig–purelin, generating a 6-30-1 ANN architecture for Nakuru county.

3.6 Validation

This study used R^2 and $RMSE$ for validation of prediction results. R^2 is also known as the coefficient of determination, and shows the proportion of variance in dependent variables that is predictable from the independent variables. The equations below shows

the mathematical notation of these statistics.

$$R^2 = 1 - \frac{(n-1)}{(n-p)} \times \frac{\text{SSE}}{\text{SST}} \tag{3.4}$$

where *SSE* is the sum of squared error, *SST* is the sum of squared total, n is the number of observations, and p is the number of regression coefficients. The Root Mean Square Error ($RMSE$) is the difference between the predicted and actual values or the deviation of the residuals (prediction errors) i.e.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{2}(y_t - y)^2}{n}} \tag{3.5}$$

where y_t is the predicted value, y is the actual value and n is the number of samples Shastry et al. (2017). $RMSE$ depicts the concentration of data around the line of best fit.

Chapter 4: Results and Discussion

4.1 Introduction

This chapter presents the results of the study as well as the discussion of the results. The chapters also compares these results with other research in the same area.

4.2 Results

Maize yield prediction is a crucial function in planning for food security of the population of a county level or even of the whole country. Agriculture sector continues to play a vital role in the rural economy. Agriculture in Kenya and Africa countries as well was developing countries, is the backbone of the economies. Agriculture provides a substantial portion of their Gross Domestic Product (GDP). Thus, the possibility to obtain yield estimates with reasonable accuracy prior to harvest is important, since timely interventions can take place in case low yields are predicted ITC (2020). Agriculture in Kenya is a fully devolved function of service provision to the county governments underscoring the importance of County Governments' role in ensuring food security. Estimating crop yields at county level is thus very important, as well as offering a solution to the right decision making level. Agriculture is key to Kenya's economy, it directly contribute 26% of the Gross Domestic Product (GDP) and indirectly 27% of GDP through linkages with other sectors. The sector employs more than 40% of the total population and more 70% of Kenya's rural people. Agriculture in Kenya is large and complex, with a multitude of public, parastatal, non-governmental and private sectors FAO (2020).

The study used historical maize yield from Trans Nzoia and Nakuru counties, in Kenya. Trans Nzoia was used for training and Nakuru to test the model performance as both counties have similar maize growing seasons. Figure 3.8 shows the normalized variables after min-max transformation. Generally, the maize yields were high in the

year 2012 which is also picked by the variables. In other years the variables more or less show the same trend as the historical yields. Normally the data set were acquired with different value ranges. Data set with high value range from similar studies have shown that results to biased influence in prediction results. The Figure 3.8 represents a sample of normalization variables using maxmin transformation.

The normalized variables in Figure 3.8 were the inputs to boruta algorithm for feature selection. The variable selection results in Figure 4.1 shows that weather variables and soil moisture have more influence to maize yield. Elevation was not considered as a significant factor. Feature selection is an important part of prediction process as it reduces memory storage, training time, computational cost and increases the performance of the predictive mode. Al-Qerem (2020) noted that relevant features have the useful information to prediction model. Feature selection allows removal of redundant, noisy or irrelevant features. Feature selection also prevents model over-fitting, facilitating data visualization, reducing storage requirements and computational costs, and improving the accuracy of modelling algorithms. Feature selection provides one of the most effective tools to reduce the dimensionality by selecting important featues from the data and increase model processing efficiency Wei et al. (2020b).

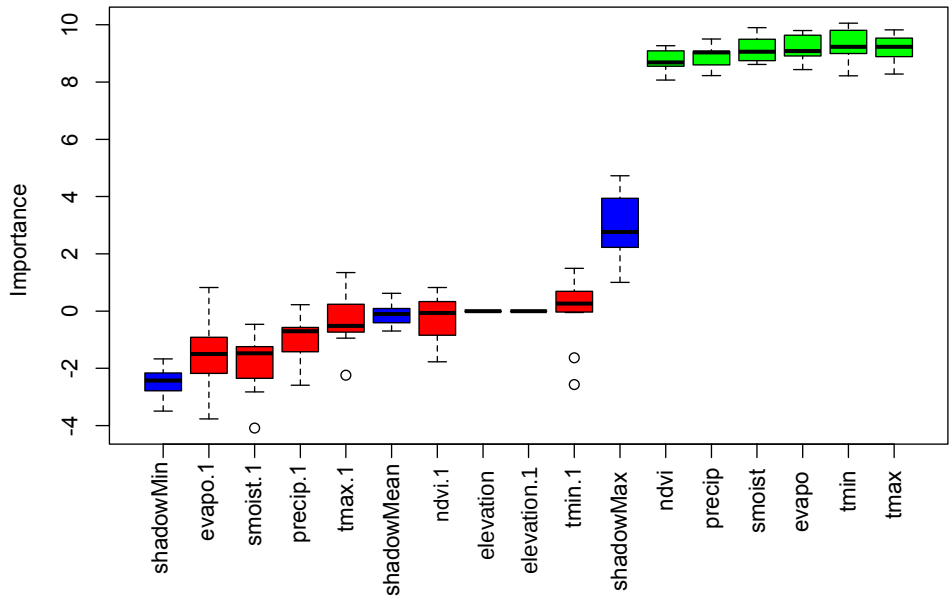


Figure 4.1: Feature Selection using boruta algorithm: *where precip (precipitation), smoist (soil moisture), evapo (evapotranspiration), tmin (minimum temperature), tmax(maximum temperature), ndvi and their respective shadow features.*

The blue bars represents the randomized shadow features for the minimum, mean and max thresholds. The red bars shows the shadow features of the respective features. The highest blue box plots - shadow maximum, defined the threshold of the features for this study. The green bars are the important features, as they are above the threshold. Consequently, features selected for yield prediction were: NDVI, precipitation, soil moisture, evapotranspiration, minimum and maximum temperature.

Analysis of remote sensing data in combination with other ancillary data such as soil moisture, allows the determination of crop yield prior to harvest period. NDVI offers the ability of remote sensing to provide information on crop status and health which is a key contributor to the estimation of potential crop yield ITC (2020). Weather factors such as temperature and precipitation also plays a similar key role to maize yield prediction. Cognizant of the variability of precipitation and temperature within the maize growth season, the study used the average values for this two factors. Other factors such as crop cultivated, fertilizer application, and other biophysical and management factors were not considered in the study due to unavailability. Boruta algorithm was effective in the identification of this factors.

The selected features in Table 4.1 below formed the input to the ANN algorithm.

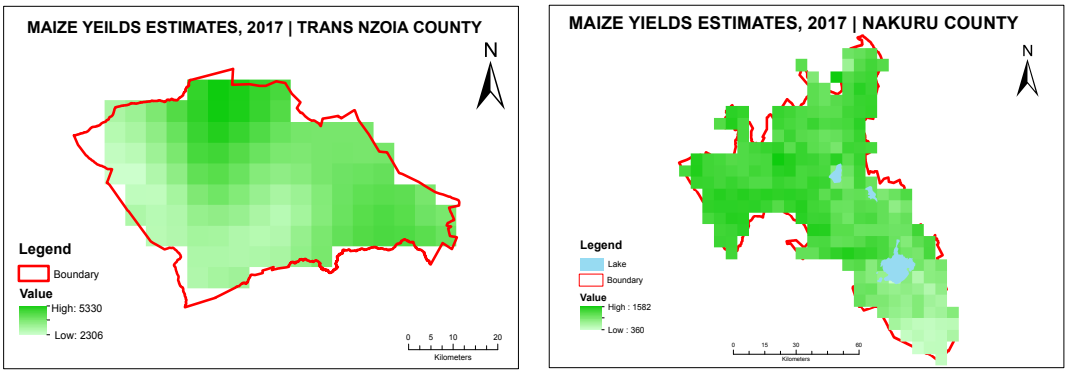
Table 4.1: Significant factors

Factors	Minimum	Maximum	Significance (cm)
Max Temp	18	22	9
Min Temp	15	18	9
Precipitation	315	1960	8.5
Soil Moisture	525	4045	9
Vegetation Index	0.28	0.71	8.5
Evapotranspiration	390	5150	9

In a study by Kursa et al. (2010) using biology data to select significant features. The use boruta algorithm showed that the selection of the important attributes can reveal important information. In comparison with random forest classifier, boruta allowed much more efficient selection of the important attributes than than former with low z-score. The absolute value of Z-score is proved to be not very informative of the attribute importance. Boruta provided a criterion for a variable selection which is based on simple statistical test. These six factors for each study area formed the inputs for yield

prediction using the developed artificial neural network algorithm. The neural network processing depends on its interconnected weights obtained through learning from a set of training patterns. Neural networks is able to deal with complex relationships because of the interconnections of weights within its network. This feature provides guarantees that neural networks achieves better performance.

Spatial distribution of final yield estimates are shown in Figure 4.2a for Trans nzoia and Figure 4.2b for Nakuru. In Trans-Nzoia county, the Northern and Eastern regions have high estimates of maize yield than Southern and Western regions. On the other hand, Nakuru has high yields in the North western and eastern parts. These areas were noted to receive relatively high rainfall, while areas with low estimates experience high temperature.



(a) Trans Nzoia (b) Nakuru

Figure 4.2: Counties yield estimations.

4.3 Discussion

The performance of designed ANNs model was evaluated using coefficient of determination (R^2) and Root Mean Squared Error (RMSE). R^2 is a statistical measure of the goodness of fit of a model with values between 0 and 1. The higher the R^2 the better the model fits the data. For instance, $R^2 = 1$ means the model fits the data perfectly.

RMSE is a good measure of how well the model predicts the response, and it is the most important criterion for fit. The lower the RMSE values the better the fit.

Table 4.2 shows the coefficient of determination R^2 and RMSE obtained from dif-

ferent model structure configurations of ANNs model (Figure 3.12a) in Trans Nzoia county based on the selected variables. The highest R^2 value which corresponds to the least RMSE was obtained at 40 iterations. There was however no clear trend on R^2 and RMSE with increase in the number of iterations.

Table 4.2: ANN structures with corresponding R^2 and RMSE (MT/ha) in Trans Nzoia County.

Case	Inputs	No. Neurons	Structure	R^2	RMSE
1	6	20	06:20:1	0.67	0.698
2	6	30	06:30:1	0.57	0.344
3	6	40	06:40:1	0.76	0.038

As a test, similar variables for Nakuru county were subjected to the model architecture in Figure 3.12b which gave the results in Table 4.5. In this case, the model gave the highest R^2 but coincidentally the RMSE was not the lowest in this case. The least RMSE was obtained at 30 iterations where the R^2 value decreased by 4%.

Table 4.3: ANN structures with corresponding R^2 and RMSE (MT/ha) in Nakuru County.

Case	Inputs	No. Neurons	Structure	R^2	RMSE
1	6	10	06:10:1	0.85	0.886
2	6	20	06:20:1	0.90	0.127
3	6	30	06:30:1	0.86	0.016

Overall, the best model quality in maize yield prediction is achieved at R^2 of 0.76 and 0.86 with corresponding RMSE values of 0.038 MT/ha and 0.016 MT/ha in Trans Nzoia and Nakuru county respectively (Tables 4.2–4.3). This means the model explained a minimum of 76% of maize yield variability based on the NDVI and weather data. This is quite significant given that the highest deviations observed from the ANNs models is $\pm$ 0.038 MT/ha of maize yield at county level on average. The lack of standardized and comprehensive reporting of the yields at county levels may have influence model performance. Nonetheless, ANNs computed reasonable yield estimates in the two counties as shown in Table 4.4.

Table 4.4: Estimated yield and R^2 values, 2017.

County	Estimated yield (MT/HA)	R^2
Trans Nzoia	4.13	0.76
Nakuru	2.26	0.86

Yield prediction in this study provides promising estimates. Using deep neural networks Khaki and Wang (2019) found the best balance between prediction accuracy and limited overfitting in the training process. The study adopted a neural network with 21 hidden layers and 50 neurons in each layer. The adopted model was found to have a superior prediction accuracy, with a root-mean-square-error (RMSE) being 12% of the average yield and 50% of the standard deviation for the validation dataset using predicted weather data. The RMSE reduced to 11% of the average yield and 46% of the standard deviation using perfect weather data.

In a research undertaken by Khan et al. (2019). They used satellite derived spectral indices as proxies to factors influencing mentha crop biomass. The study achieved an R-squre value of 76.2% with a root mean squared error of 2.74 t/ha, an indication that there is good correlation between the feld-measured biomass and estimated biomass using multi-layer perceptron (MLP) neural network. Comparatively combining ground measurement and finer resolution satellite data greatly improves the machine learning algorithms prediction ability.

In another similar research, Jefries et al. (2020) mapped sub-field maize yields in Nebraska, USA by combining remote sensing imagery, crop simulation models, and machine learning. The Weather parameters needed to capture the inter-annual effects of weather on maize yields were derived from daily weather station observations.The research study also noted aggregating yield model prediction over time improved performance relative to single year metrics. In similar fashion this study adopted an averaging approach to factors. Jefries et al. (2020) study findings on the other hand used linear regression models in yield predictions. A proportional bias to yield predictions were noted from derived statistics. We further compared ANNs model in Trans Nzoia with ordinary regression and established that ANNs results are better by an R^2 of 0.12 (Table 4.5). It is probably because the regression model adopts a linear interaction between the factors, e.g., temperature, humidity, rainfall which affects the crop yield. So ANNs

still remains a favourable yield estimation tool.

Table 4.5: R^2 and $RMSE(MT/ha)$ for prediction models.

Model	R^2	$RMSE$
Ordinary regression	0.64	0.089
Artificial neural network	0.76	0.038

In a study done in western Australia, Filippi et al. (2019) predicted yield for wheat, maize and canola and achieved high r-squared of between 0.89 and 0.91. The study attributed this to use of high resolution data set. The remote sensing imagery used were of 100 m spatial resolution as well as remotely-sensed EVI images sourced from MODIS and are at a 250 m spatial resolution. Our study achieved a lower r-squared of about 0.76 using remote sensing imagery of $\approx$ 4 km. Filippi et al. (2019) also emphasised the need to include freely-available data at finer spatial scales, such as Landsat at 30 m or Sentinel 2 at 15 m resolution, as this would give more detailed information.

A vineyard yield estimation by using remote sensing, computer vision and artificial neural network techniques by Ballesteros et al. (2020). The study similarly noted that machine learning techniques used resulted in much more accurate results than linear models. Also, more precise yield predictions were obtained using images taken near to the harvest date, while promising results were obtained at earlier stages. Our study also noted some similar resulted as indicated in (Table 4.5) where artificial neural network performed better the ordinary regression by a difference of $\approx$ 0.12 in r-squared value.

Miao et al. (2006) also did a study to identify important factors that influence corn yield and grain quality variability using artificial neural networks. The study found that relative elevation, landscape and soil factors to be among the most important for corn yield and quality. Miao et al. (2006) noted that selecting suitable inputs, estimating their responses to environmental factors and managing them as required for optimum yield and grain quality are key to precision crop management. This study result found relative elevation as an important factor while as our study eliminated elevation as an important factor using boruta algorithm. The interpretation to this difference can be attributed to coarse spatial resolution of digital elevation model.

There is a consistent better performance of machine learning algorithm compared to statistical approach. This was also observed in a study by Charoen-Ung and Mittrapiya-

nuruk (2018) where they used machine learning algorithms such as Random Forest with Forward Feature Selection and Hyper-parameter Tuning for sugarcane yield grade prediction. The accuracy of their study was 71% compared to non-machine learning approach accuracy of 51%. This observation is similar to our finding where Artificial neural networks performed better that ordinary regression.

The conventional ANN model randomly assign weighs and bias of input and hidden layer. A study by Gopal and Bhargavi (2019) used the machine learning algorithm Artificial Neural Network (ANN) to predict yields. The model achieved a root mean square error of 0.098. This study went ahead and developed a hybrid Multiple Linear Regression (MLR)-Artificial Neural Network (ANN) model. The hybrid model ANN's input layer weights and bias were initialized by using MLR's coefficients and bias. Using this approach, the study noted an improved accuracy in yield prediction. While this study achieved promising results as show in (Table 4.5), to prediction county level yield estimates, A hybrid approach can still be adopted to find out whether similarly observations to Gopal and Bhargavi (2019) can be achieved.

Another perspective to this study was demonstrated by Ranjan and Parida (2019) when they conducted an acreage mapping of paddy and also yield prediction using sentinel-based optical and SAR data. The study used random forest classifier. Random forest is a machine learning classifier, which efficiently considers the large database processing. It is able to handle thousands of input variables without assumption of variables. Random forest classifier provide optimal accuracy of classification by assembling bulk of decision trees during the training. From this study, it was observed that machine learning algorithm yields improved prediction results using SAR data as opposed to optical data from remote sensing. The overall accuracy for this study using sentinel 1A and sentinel-2B were 89% and 87% respectively.

Rao and Manasa (2019) focused on predicting the crop yield using the Artificial Neural Networks. The set-forth to predict soil quality and suggest inputs such as fertilizer, as well as estimate the expected yields. Rao and Manasa (2019) were able to implement a model using Artificial Neural Networks (ANN) which predicts the soil quality taking input as several important parameters related to soil.

There exist a strong climate-yield relationship. The influence of most meteorolog-

ical factors show high significance to yields. However, the importance varies as Xu et al. (2019) observed differences in the relationship between the two zones under study. The study method was able to evaluate the levels of wheat yields affected by different weather conditions.

Wang et al. (2018) conducted a research using deep learning. The transfer learning model outperformed all other statistical models. The results in Argentina and Brazil demonstrate that the study approach successfully learn effective features from raw data and achieve improved performance compared to traditional methods. Wang et al. (2018) also noted that a successful crop yield prediction with deep learning in regions with little training data relies on the ability to fine-tune pre-trained models. The study was able to showcase that remotely sensed data, such as satellite imagery, potentially provide a cheap, equally effective alternative.

In a study Monga (2018) used convolutional neural networks (CNN) to develop models that can estimate the weight of grapes on a vine using an image. Convolutional Neural Networks (CNN) is an advancement in many machine learning applications such as computer vision, speech recognition and natural language processing. With limited data set of 60 images of grape vines, the approach achieved a 87% accuracy for predicting harvest grape yield. This shows that convolutional neural network has a great potential in agricultural yield estimation. Monga (2018) concludes that with large data sets, it would be possible to achieve even better accuracies as well as yield prediction. Similarly, our maize yield prediction models also gives similar indications. This observation in machine learning approaches is concludes that, with large data sets their products improves.

A region-specific crop yield analysis was studied by Shah et al. (2018) using machine learning approaches such as multivariate polynomial regression, support vector machine regression and random forest models to predict the crop yield per acre. This approach presents an intelligent way to predict crop yield and suggest the optimal climatic factors to maximize crop yield. The outlined methods uses yield and weather data collected from United States Department of Agriculture. The weather parameters included in the data set are humidity, yield, temperature and rainfall. The objective of this study was to help the farmers choose the most suitable temperature and moisture

content at which the crop yield will be optimal. This was demonstrated with support vector machine regression achieving a maximum R-squared value of 96% being the best performance. Although the study didn't discuss how the variables were selected, the use of weather factors to predict yields is gives a good prediction. Shah et al. (2018) also noted that there was need to include more variables for a robust yield prediction.

In support of regional prediction. Ahmad et al. (2018) study notes that for regional yield forecasting, remote sensing has a greater advantages of less input data set. In this study for crop modeling, the CERES-Maize model was calibrated and evaluated with the field experiment data and after calibration and evaluation, this model was used to forecast maize yield. In remote sensing, Landsat 8 images for the peak season were classified using machine learning algorithm. After classification, time series normalized difference vegetation index (NDVI) and land surface temperature (LST) of the surveyed 64 farms were calculated. All machine learning algorithms showed the accuracy greater the 90%, however support vector machine (SVM-radial basis) showed the higher accuracy of 97%, for classification of maize area. In conclusion, Ahmad et al. (2018) stated that, the overall strength of relationship between estimated and actual grain yields were good with R-squared of 94% in both techniques. It is important to note that the combination of in-situ data and remote sensing data was the key differentiator in achieving good prediction in this study. Due to unavailability of these key data, our study concentrated on soil and weather factors, but their contribution cannot be underestimated especially for region or small acre yield prediction.

Data driven crop production is the future of precision agriculture. A study by Ghazvinei et al. (2018) tried to establish an integrate model using extreme learning machine (ELM) to predict the concluding growth amount of sugarcane. The predicted yields of extreme learning machine (ELM) were evaluated and compared with artificial neural network (ANN) and genetic programming models. The study used six input parameters selected for analysis including; maximum temperature (degrees celsius), evaporation (mm), wind speed (m/s), sunshine (hour), rainfall (mm), humidity (%), irrigation (mm), and soil electrical conductivity (EC) (ds/m). These parameters were considered potentially influential for the sugarcane growth as the modelling output parameter. The ELM model achieved the best r-squared of 92% and a root mean squared error of 0.21.

Based on this results, ELM model showed interesting and notable abilities against the gradient-based procedures of neural networks. Moreover, the study also revealed that ELM is considerable quicker in learning rapidity in comparison with the conventional ANNs algorithm. Furthermore, it was also noted that ELM model performed with the minimum norm of weights and the least training error while such an applicability did not appear in the traditional learning algorithms. Although, out study using ANNs shows satisfactory results for county level yield prediction, there is need to also try ELM and compare the results as well.

In a site-specific study. Khanala et al. (2018) integrated high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. In this study, remotely sensed image-derived variables were integrated with field collected data to develop models. The main objective is this study was to compare the performance of various machine learning algorithms and identify the importance of remotely sensed image-derived variables, in spatial prediction of soil properties and corn yield. The corn yield was monitored using multispectral aerial images and topographic data. The models developed for prediction of soil properties and corn yield using linear regression (LM) and five machine learning algorithms i.e., Random Forest (RF); Neural Network (NN); Support Vector Machine (SVM) with radial and linear kernel functions; Gradient Boosting Model (GBM); and Cubist (CU)). These mmodels were evaluated in terms of coefficient of determination (R^2) and root mean square error (RMSE). The machine learning algorithms were found to outperform the LM algorithm. The use of high spatial resolution mapping of soil properties and crop yield is important for proper management of crop. Another important observation was that soil and vegetation indices based on bare-soil imagery played a more significant role in demonstrating in-field variability of corn yield and soil properties than topographic variables. This observation agrees with our study where elevation as topographic variable was also considered not important for maize yield estimation.

Girish et al. (2018) also demonstrated the capability of machine learning in pecision agriculture. Predictive analysis can help the farmers to choose whether a particular crop is suitable for specific rainfall and crop price values.

Chapter 5: Conclusions and Recommendations

5.1 Introduction

This chapter concludes this report. The conclusions are deduced from the findings and discussion in previous chapter of this report. The conclusion notes the significance of the research in the field of agriculture and also recommends further research at the end of the chapter.

5.2 Conclusions

Yield prediction is beneficial to both farmers and businesses as it provides an opportunity to make decisions and amend or introduce policies before harvest. The aim of this study was to use satellite data and ANNs model to predict maize yields at county level prior to harvesting period for sustainable food security.

Various field of study have acknowledged feature selection as an indispensable key pre-processing procedure. Although feature selection is a challenging topic, this study aimed to use an algorithm that has little or no human influence in the process of identifying important features. The algorithm considered were Analytic Hierarchy Process (AHP) and Random Forest (RF) and Boruta Algorithm. Based on literature review, the study adopted boruta algorithm feature selection. This study also recognised the key role of feature selection. Feature selection leads to a feature subset, with informative subset by selecting important features or deleting unimportant features from the original feature set. This study identified minimum temperature, maximum temperature, precipitation, soil moisture, vegetation index and evapotranspiration as the important factors

to predict maize yield using boruta algorithm. Elevation as a topographic factor was not considered important, as most studies have also found.

Machine learning (ML) is a branch in the field of artificial intelligence that assists computers in modeling based on historical data and accurately predicting future outcomes. ML approaches are mostly classified into two main categories: supervised learning and unsupervised learning. Classification and clustering are examples of problems in supervised learning and unsupervised learning respectively. The widely used techniques for classification include neural networks, support vector machines, and decision trees, and the most widely used clustering technique is k-means. Based on literature review of similar studies, support vector machine and artificial neural networks produce successful prediction results.

Based on this finding, our study adopted ANNs of multilayer perceptron, feed forward back propagation to predict maize yield at pixel level as function of weather data derived from satellite in Trans Nzoia and Nakuru counties in Kenya. Using matlab software, we developed a maize yield modelling program. The study developed a feed-forward, back propagation artificial neural network with levenberg-marquardt algorithm. We managed to train our model for prediction using the Levenberg Marquardt algorithm.

The model successfully was able to predict maize yield with R^2 and root mean square error of 0.76 and $0.038MT/ha$ in Trans-Nzoia county and 0.86 and $0.016MT/ha$ respectively in Nakuru county for 2017. This study managed to explored the potential of ANN model in addressing the problem of yield prediction, while considering the complex interactions of inputs involved for growth of maize.

This study has demonstrated that maize yield estimation at county level in Kenya can be achieved at a reasonable prediction accuracy using ANNs and satellite data. In developing countries, this combination presents a solution to food insecurity shocks normally experienced.

5.3 Recommendations

The study considered mainly the remotely-sensed satellite weather data, but future research in similar area needs to integrate physical and management factors for maize yield prediction. Similar studies that integrated physical and management factors observed improved yield prediction results.

Feature selection remains of key importance to modelling. This study considered boruta algorithm, but other studies trying to predict results of a future outcome has used random forest with satisfactory success. Another feature selection criterion that can be considered in future studies is use of pearson correlation.

As Radar data continues to gain continues observation. Future research should integrate SAR data in maize yield prediction. From discussion, a study had noted improved results of prediction with SAR data. Also the use of fine spatial resolution data cannot be over-emphasized.

Since the model works county level prediction, we recommend adoption for a farm/site-based approach to be able to include management and random factors.

Use of other machine learning techniques e.g. ELM as well as integrating statistical model with machine learning model needs to be explored further in the equatorial regions. Another branch of machine learning, deep learning has powerful functions and flexibility such as deep neural network (DNN), CNN and SVM. SVM is a very powerful classification model in machine learning. CNN is a type of feedforward neural network that includes convolution calculation and has a deep structure. It is one of the representative algorithms of deep learning.

Deep learning emphasizes the depth of the model structure, usually there are five, six or more hidden layers. Through layer by layer feature space conversion, in-depth learning can get the most excellent expression of features. Examples of deep learning models includes; convolutional neural network, deep trust network model, self coding network model, restricted Boltzmann machine model.

Another option to improve the maize yield prediction results is to integrate prediction algorithms with optimization techniques. For research studies in this area should look for ways to incorporate this aspect.

References

Abatzoglou, Dobrowski S. Z., S. A. Parks, and Hegewisch K. C. Terraclimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Scientific Data*, 5(1):170191, 2018. doi: 10.1038/sdata.2017.191.

Kevin O Achieng. Modelling of soil moisture retention curve using machine learning techniques: Artificial and deep neural networks vs support vector regression models. *Computers and Geosciences*, 133(104320), 2019. doi: 10.1016/j.cageo.2019.104320.

Ishfaq Ahmad, Umer Saeed, Muhammad Fahad, Asmat Ullah, M. Habib ur Rahman, Ashfaq Ahmad, and Jasmeet Judge. Yield forecasting of spring maize using remote sensing and crop modeling in faisalabad-punjab pakistan. *Journal of the Indian Society of Remote Sensing*, 46(10):1701–1711, 2018. doi: 10.1007/s12524-018-0825-8.

Ahmed, Diffenbaugh N. S., and Hertel T. W. Climate volatility deepens poverty vulnerability in developing countries. *Environmental Research Letters*, 4(3):034004, 2009. doi: 10.1088/1748-9326/4/3/034004.

Ahmad Al-Qerem. An efficient machine-learning model based on data augmentation for pain intensity recognition. *Egyptian Informatics Journal*, 2020. doi: 10.1016/j.eij.2020.02.006.

Global Yield Gap Atlas. Information from collected literature for nigeria to be used for crop modelling for nigeria. *Global Yield Gap Atlas*, 2013. URL http://www.yieldgap.org/.

Rocío Ballesteros, Diego S. Intrigliolo, José F. Ortega, Juan M. Ramírez-Cuesta, Ignacio Buesa, and Miguel A. Moreno. Vineyard yield estimation by combining remote sensing, computer vision and artifcial neural network techniques. *Precision Agriculture*, 2020. doi: 10.1007/s11119-020-09717-3.

Seshadri Baral, Asis Kumar Tripathy, and Pritiranjan Bijayasingh. Yield prediction using artificial neural networks. *Computer Networks and Information Technologies*, 142(315-317), 2011. doi: 10.1007/978-3-642-19542-6_57.

Barzegar and Asghari Moghaddam A. Combining the advantages of neural networks using the concept of committee machine in the groundwater salinity prediction. *Modeling Earth Systems and Environment*, 2(1):26, 2016. doi: 10.1007/s40808-015-0072-8.

Bruno Basso and Lin Liu. Chapter four - seasonal crop yield forecast: Methods, applications, and accuracies. *Advances in Agronomy*, 154:201–255, 2019. doi: 10.1016/bs.agron.2018.11.002.

Statistical Methods Branch, Statistics Division, National Agricultural Statistics Service, and Department of Agriculture. The yield forecasting and estimating program of nass. https://www.nass.usda.gov/Education_and_Outreach/Understanding_Statistics/Yield_Forecasting_Program.pdf, 2012.

Budyko and Menzhulin G. V. *Climate Change Impacts on Agriculture and Global Food Production: Options for Adaptive Strategies*. Springer, New York, NY, 1996. doi: 10.1007/978-1-4613-8471-7_16.

Phusanisa Charoen-Ung and Pradit Mittrapiyanuruk. Sugarcane yield grade prediction using random forest with forward feature selection and hyper-parameter tuning. *Advances in Intelligent Systems and Computing*, 769:33–42, 2018. doi: 10.1007/978-3-319-93692-5_4.

Chen and Mcnairn H. A neural network integrated approach for rice crop monitoring. *International Journal of Remote Sensing*, 27(7):1367–1393, 2006. doi: 10.1080/01431160500421507.

Dahikar and Rode S. V. Agricultural crop yield prediction using artificial neural network approach. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, 2(1):683–686, 2014.

Deo and Sahin M. Application of the artificial neural network model for prediction of monthly standardized precipitation and evapotranspiration index using hydrometeorological parameters and climate indices in eastern australia. *Atmospheric Research*, 2015. doi: 10.1016/j.atmosres.2015.03.018.

Alican Dogan and Derya Birant. Machine learning and data mining in manufacturing. *Expert Systems with Applications*, 21(114060), 2020. doi: 10.1016/j.eswa.2020.114060.

FAO. Addressing food crises - towards the elaboration of an agenda for action in food security in countries in protracted crisis, high-level expert forum (hlef). 2010.

FAO. The agriculture sector in kenya. http://www.fao.org/kenya/fao-in-kenya, 2020. (Accessed: 22nd October 2020).

R. Fieuzal, C. Marais Sicre, and F. Baup. Estimation of corn yield using multi-temporal optical and radar satellite data and artificial neural networks. *International Journal of Applied Earth Observation and Geoinformation*, 57:14–23, 2017. doi: 10.1016/j.jag.2016.12.011.

Patrick Filippi, Edward J. Jones, Niranjan S. Wimalathunge, Pallegedara D. S. N. Somarathna, Liana E. Pozza, Sabastine U. Ugbaje, Thomas G. Jephcott, Stacey E. Paterson, Brett M. Whelan, and Thomas F. A. Bishop. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agriculture*, 2019. doi: 10.1007/s11119-018-09628-4.

Pezhman Taherei Ghazvinei, Hossein Hassanpour Darvishi, Amir Mosavi, Khamaruzaman bin Wan Yusof, Meysam Alizamir, Shahaboddin Shamshirband, and Kwok wing Chau. Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network. *Engineering Applications of Computational Fluid Mechanics*, 12(1):738–749, 2018. doi: 10.1080/19942060.2018.1526119.

L Girish, Gangadhar S, Bharath T R, Balaji K S, and Abhishek K T. Crop yield and rainfall prediction in tumakuru district using machine learning. *International Journal*

for Research in Engineering Application and Management, 2018. doi: 10.18231/2454-9150.2018.0805.

Global-Change-Data-Lab. Our world in data. https://ourworldindata.org/world-population-growth, 2020. (Accessed: 2nd November 2020).

Maya Gopal and R. Bhargavi. A novel approach for efficient crop yield prediction. *Computers and Electronics in Agriculture*, 165, 2019. doi: 10.1016/j.compag.2019.104968.

William W. Guo and Heru Xue. Crop yield forecasting using artificial neural networks: A comparison between spatial and temporal models. *Mathematical Problems in Engineering*, 2014. doi: 10.1155/2014/857865.

Hota. Artificial neural network and efficiency estimation in rice yield. *International Journal of Innovative Research in Science, Engineering and Technology*, 3(7):14787–14805, 2014.

Islam, Islam M.M, and Islam M.N. et al. Climate change adaptation strategies: a prospect toward crop modelling and food security management. *Modeling Earth Systems and Environment*, 6:769–777, 2020. doi: 10.1007/s40808-019-00708-6.

ITC. Potential uses of remote sensing in smallholder context. https://www.stars-project.org/en/, 2020. (Accessed: 22nd October 2020).

Graham Jefries, Timothy S. Grifn, David H. Fleisher, Elena N. Naumova, Magaly Koch, and Brian D. Wardlow. Mapping sub-field maize yields in nebraska, usa by combining remote sensing imagery, crop simulation models, and machine learning. *Precision Agriculture*, 21:678–694, 2020. doi: 10.1007/s11119-019-09689-z.

D Jiang, X Yang, N Clinton, and N Wang. An artificial neural network model for estimating crop yields using remotely sensed information. *International Journal of Remote Sensing*, 25:1723–1732, 2004. doi: 10.1080/0143116031000150068.

Kadir, Ayob M. Z, and Miniappan N. Wheat yield prediction: Artificial neural network based approach. *International Conference on Engineering Technology and Technopreneuship*, pages 161–165, 2014. doi: 10.1109/ICE2T.2014.7006239.

Monisha Kaul, Robert L. Hill, and Charles Walthall. Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems*, 85(1):1–18, 2015. doi: 10.1016/j.agsy.2004.07.009.

Muhd Khairulzaman, Abdul Kadir, Mohd Zaki Ayob, and Nadaraj Miniappan. Wheat yield prediction: Artificial neural network based approach. *International Conference on Engineering Technology and Technopreneuship*, 2014. doi: 10.1109/ICE2T.2014.7006239.

S Khaki and L Wang. Crop yield prediction using deep neural networks. *Frontiers in Plant Science,*, 10, 2019. doi: 10.3389/fpls.2019.00621.

Mohammad Saleem Khan, Manoj Semwal, Ashok Sharma, and Rajesh Kumar Verma. An artifcial neural network model for estimating mentha crop biomass yield using landsat 8 oli. *Precision Agriculture*, 2019. doi: 10.1007/s11119-019-09655-9.

Sami Khanala, John Fulton, Andrew Klopfenstein, Nathan Douridas, and Scott Shearer. Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Computers and Electronics in Agriculture*, 153:213–225, 2018. doi: 10.1016/J.COMPAG.2018.07.016.

Woohyun Kim, Yerim Han, Kyoung Jae Kim, and Kwan-Woo Song. Electricity load forecasting using advanced feature selection and optimal deep learning model for the variable refrigerant flow systems. *Energy Reports*, 6:2604–2618, 2020. doi: 10.1016/j.egyr.2020.09.019.

KNBS. County statistical abstract, trans nzoia county. Technical report, Kenya National Bureau of Statistics, 2019.

Kross, Znoj E., Callegari D., Kaur G., Sunohara M., Vliet L., Rudy H.and Lapen D., and Mcnairn H. Evaluation of an artificial neural network approach for prediction of corn and soybean yield. *International Conference on Precision Agriculture*, 06 2018.

Kursa and W. Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software, Articles*, 36(11):1–13, 2010. doi: 10.18637/jss.v036.i11.

Miron Kursa, Aleksander Jankowski, and Witold R. Rudnicki. Boruta – a system for feature selection. *Fundamenta Informaticae*, 101:271–285, 2010. doi: 10.3233/FI-2010-288.

Lobell and Burke M. B. On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11):1443–1452, 2010. doi: 10.1016/j.agrformet.2010.07.008.

Lobell, David B Azzari, and George. Satellite detection of rising maize yield heterogeneity in the u.s. midwest. *Environmental Research Letters*, 0, 2017. doi: 10.1088/1748-9326/aa5371.

Nafiseh Yaghmaeian Mahabadi. Use of the intelligent models to predict the rice potential production. *International Academic Journal of Innovative Research*, 4(2):14–25, 2018.

McMaster and Wilhelm W W. Growing degree-days: One equation, two interpretations. *Agricultural and Forest Meteorology*, 87:291–300, 1997.

McNairn, Kross A, Lapen D, Caves R, and Shang J. Early season monitoring of corn and soybeans with terrasar-x and radarsat-2. *International Journal of Applied Earth Observation and Geoinformation*, 28:252–259, 2014. doi: 10.1016/j.jag.2013.12.015.

Ramesh Medar and Vijay. S. Rajpurohit. A survey on data mining techniques for crop yield prediction. *International Journal of Advance Research in Computer Science and Management Studies*, 2(9):59–64, 2014.

Yuxin Miao, David J. Mulla, and Pierre C. Robert. Identifying important factors influencing corn yield and grain quality variability using artificial neural networks. *Precision Agriculture*, 7(2):117–135, 2006. doi: 10.1007/s11119-006-9004-y.

Tanya Monga. Estimating vineyard grape yield from images. *Lecture Notes in Computer Science*, page 339–343, 2018. doi: 10.1007/978-3-319-89656-4_37.

Gandhi Niketa, Petkar Owaiz, and Armstrong Leisa J. Rice crop yield prediction using artificial neural networks. *IEEE International Conference on Technological Innovations in ICT For Agriculture and Rural Development*, 2016.

OAU and ECA. Population and development in africa, oau and eca, 1994.

O'Neal, Engel B., Ess D., and Frankenberger J. Ae—automation and emerging technologies: Neural network prediction of maize yield using alternative data coding algorithms. *Biosystems Engineering*, 83:31–45, 9 2002. doi: 10.1006/bioe.2002.0098.

Hojjatolah Yazdan panah. A neural network based model for rain fed wheat yield forecasting. *Mountain Meteorology*, 2008.

Puig-Arnavat and Bruno J. C. Artificial neural networks for thermochemical conversion of biomass. In *Recent advances in thermo-chemical conversion of biomass*, pages 133–156. Elsevier, 2015.

Avinash Kumar Ranjan and Bikash Ranjan Parida. Paddy acreage mapping and yield prediction using sentinel-based optical and sar data in sahibganj district, jharkhand (india). *Spatial Information Research*, 2019. doi: 10.1007/s41324-019-00246-4.

T. Venkat Narayana Rao and S. Manasa. Artificial neural networks for soil quality and crop yield prediction using machine learning. *International Journal on Future Revolution in Computer Science and Communication Engineering*, 5(1):57 – 60, 2019.

Mazzanti S. Boruta explained exactly how you wished someone explained to you. https://towardsdatascience.com, 2020. (Accessed: 22nd February 2020).

Schmidhuber and Tubiello F. N. Global food security under climate change. *Proceedings of the National Academy of Sciences*, 104(50):19703–19708, 2007. doi: 10.1073/pnas.0701976104.

Sellam and Poovammal E. Prediction of crop yield using regression analysis. *Indian Journal of Science and Technology*, 9(38):1–5, 2016. doi: 10.17485/ijst/2016/v9i38/91714.

Ayush Shah, Akash Dubey, Vishesh Hemnani, Divye Gala, and D. R. Kalbande. Smart farming system: Crop yield prediction using regression techniques. *Proceedings of International Conference on Wireless Communication*, page 49–56, 2018. doi: 10.1007/978-981-10-8339-6_6.

Minglei Shao, Xin Wang, Zhen Bu, Xiaobo Chen, and Yuqing Wang. Prediction of energy consumption in hotel buildings via support vector machines. *Sustainable Cities and Society*, 57(102128), 2020. doi: 10.1016/j.scs.2020.102128.

Shastry, Sanjay H.A, and Bhanusree E. Prediction of crop yield using regression techniques. *International Journal of Soft Computing*, 12(2):96–102, 2017. doi: 10.36478/ijscomp.2017.96.102.

Prajneshu Singh Rama Krishna. Artificial neural network methodology for modelling and forecasting maize crop yield. *Agricultural Economics Research Review*, 21, 2008. doi: 10.22004/ag.econ.47354.

Stige, Stave J., Chan K., Ciannelli L., Pettorelli N., Glantz M., Herren H. R., and Stenseth N. C. The effect of climate variation on agro-pastoral production in africa. *Proceedings of the National Academy of Sciences*, 103(9):3049–3053, 2006. doi: 10.1073/pnas.0600057103.

Torres, Snoeij P, Geudtner D, Bibby D, Davidson M, and Attema. Gmes sentinel-1 mission. *Remote Sensing of Environment*, 120:9–24, 2012.

United-Nations. World population prospects: The 2015 revision, key findings and advance tables. https://www.un.org/en/development/desa/publications/world-population-prospects-2015-revision.html, 2015. (Accessed: 29th July 2020).

USGS. Earthexplorer. https://earthexplorer.usgs.gov/, 1990. (Accessed: 22nd February 2020).

Veenadhari, Misra Bharat, and C. D. Singh. Machine learning approach for forecasting crop yield based on climatic parameters. *International Conference on Computer Communication and Informatics: Ushering in Technologies of Tomorrow, Today*, pages 1–5, 2014. doi: 10.1109/ICCCI.2014.6921718.

Schlenker W. and Roberts M. J. Estimating the Impact of Climate Change on Crop Yields: The Importance of Nonlinear Temperature Effects. Working paper, NA-

TIONAL BUREAU OF ECONOMIC RESEARCH, Cambridge, MA, 2008. URL http://www.nber.org/papers/w13799.

Anna X. Wang, Caelin Tran, and Nikhil Desai. Deep transfer learning for crop yield prediction with remote sensing data. *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*, 2018. doi: 10.1145/3209811.3212707.

Fei Wang, Quan Wang, Feiping Nie, Zhongheng Li, Weizhong Yu, and Fuji Ren. A linear multivariate binary decision tree classifier based on k-means splitting. *Pattern Recognition*, page 107521, 2020. doi: 10.1016/j.patcog.2020.107521.

Cong Wei, Jingyue Wei, Qiaoping Kong, Dan Fan, Guanglei Qiu, Chunhua Feng, Fusheng Li, Sergei Preis, and Chaohai Wei. Selection of optimum biological treatment for coking wastewater using analytic hierarchy process. *Science of The Total Environment*, 20(140400), 2020a. doi: 10.1016/j.scitotenv.2020.140400.

Guangfen Wei, Jie Zhao, Yanli Feng, Aixiang He, and Jun Yu. A novel hybrid feature selection method based on dynamic feature importance. *Applied Soft Computing Journal*, 93(106337), 2020b. doi: 10.1016/j.asoc.2020.106337.

Wilson and Mantooth H. A. *Model-Based Optimization Techniques. Model-Based Engineering for Complex Electronic Systems*. Elsevier, 2013. doi: 10.1016/b978-0-12-385085-0.00010-5.

World-Bank. *World Bank Big Data Innovation Challenge*. world bank, 2016. doi: https://doi.org/.

Xiangying Xu, Ping Gao, Xinkai Zhu, Wenshan Guo, Jinfeng Ding, Chunyan Li, Min Zhu, and Xuanwei Wu. Design of an integrated climatic assessment indicator (icai) for wheat production: A case study in jiangsu province, china. *Ecological Indicators*, 101:943–953, 2019. doi: 10.1016/j.ecolind.2019.01.059.

F.W Yua, W.T. Ho, K.T. Chan, and R.K.Y. Sit. Critique of operating variables importance on chiller energy performance using random forest. *Energy and Buildings*, 139: 653–664, 2017. doi: 10.1016/j.enbuild.2017.01.063.

Zhang, Lei L., and Yan D. Comparison of two regression models for predicting crop yield. *IEEE International Geoscience and Remote Sensing Symposium*, pages 1521–1524, 2010. doi: 10.1109/IGARSS.2010.5652764.

Appendix A: Project scripts

The program below was developed for feature selection in this study. The program was developed using R language and boruta library.

````
---
title: "Variable Selection using Boruta Algorithm"
date: "`r Sys.Date()`"
#output: html_document
---

```{r setup,message=FALSE,warning=FALSE}
library(caret)
library(data.table)
library(Boruta)
library(plyr)
library(dplyr)
library(pROC)

ROOT.DIR <- "."

```
# Data Preparation for Boruta
```{r DataRetrieval}

# retrive data for analysis
binary.df <- read.csv(file.path(ROOT.DIR,"/mdata/varTest.csv"),
    stringsAsFactors = FALSE)
#Column Names
names(binary.df)
```
#Define categorical variables
```{r CategorizeData}
#here
````
````

```
binary.df$yield = as.factor(binary.df$yield)
binary.df$ndvi = as.factor(binary.df$ndvi)
#binary.df$smoist = as.factor(binary.df$smoist)
#binary.df$evapo = as.factor(binary.df$evapo)
#binary.df$tmax = as.factor(binary.df$tmax)
#binary.df$tmin = as.factor(binary.df$tmin)
#binary.df$precip = as.factor(binary.df$precip)
#binary.df$SRTM = as.factor(binary.df$SRTM)
```
#Explore Data
```{r ExploreData}
#Summarize Data
summary(binary.df)

#Check number of missing values
sapply(binary.df, function(y) sum(is.na(y)))
#binary.df <- na.omit(binary.df)
```
#Run Boruta Algorithm
```{r RunBoruta}
#load Boruta package
library(Boruta)

# Run Boruta Algorithm
set.seed(456)
#boruta <- Boruta(yield~., data = binary.df, doTrace = 2)
boruta <- Boruta(yield~ ., data=na.omit(binary.df), doTrace=2)
print(boruta)
#plot(boruta)
plot(boruta, xlab = "", xaxt = "n")
k <-lapply(1:ncol(boruta$ImpHistory),function(i)
  boruta$ImpHistory[is.finite(boruta$ImpHistory[,i]),i])
names(k) <- colnames(boruta$ImpHistory)
Labels <- sort(sapply(k,median))
axis(side = 1,las=2,labels = names(Labels),
       at = 1:ncol(boruta$ImpHistory), cex.axis = 0.7)
```

```
#More random variables added to the original datasets
```{r MoreRandomization}
#Add some random permuted data
set.seed(777)
binary.df.new<-data.frame(binary.df,apply(binary.df[,-1],2,sample))
names(binary.df.new)[5:7]<-paste("Random",1:3,sep="")
binary.df.new$Random1 = as.numeric(as.character(binary.df.new$Random1
    ))
binary.df.new$Random2 = as.numeric(as.character(binary.df.new$Random2
    ))
#Save important variables
head(binary.df.new)
```
#Run Boruta Again
```{r RunBoruta}
set.seed(456)
boruta2 <- Boruta(yield~., data = binary.df.new, doTrace = 1)
print(boruta2)
#plot(boruta2)
plot(boruta2, xlab = "", xaxt = "n")
k <-lapply(1:ncol(boruta2$ImpHistory),function(i)
  boruta2$ImpHistory[is.finite(boruta2$ImpHistory[,i]),i])
names(k) <- colnames(boruta2$ImpHistory)
Labels <- sort(sapply(k,median))
axis(side = 1,las=2,labels = names(Labels),
       at = 1:ncol(boruta2$ImpHistory), cex.axis = 0.7)
```
#Save Important variables
```{r SaveVariables}
attStats(boruta2)
#See list of finalvars
finalvars = getSelectedAttributes(boruta2, withTentative = F)
```
#Compare Boruta with recursive feature elimination (RFE) algorithm in
    Caret
```{r RFE algorithm}
library(caret)
```
```

```
library(randomForest)
library(e1071)
set.seed(456)
control <- rfeControl(functions=rfFuncs, method="cv", number=10)
rfe <- rfe(binary.df.new[,2:7], binary.df.new[,1], rfeControl=control
    )
print(rfe, top=10)
plot(rfe, type=c("g", "o"), cex = 1.0)
predictors(rfe)
head(rfe$resample, 10)
```

Listing A.1: R code for feature selection

The program below was developed for yield estimation. The program was developed using neural networks in matlab.

```
% ================================ %
% Neural Network                   %
% Master's Degree Project, JKUAT   %
% ================================ %
% ================================================================ %
% Inputs.csv & Targets.csv files to trains the networks            %
% The data is randomly splits the supplied data into:              %
% 70% for training, 15% for validation, and 15% for testing        %
% Training is done using the Levenberg-Marquardt algorithm         %
% trainlm updates weight and bias values                           %
% according to Levenberg-Marquardt optimization                    %
% ================================================================ %
%Clear storage and create folder named outputs
clear all;
fclose all;
clc;
help MaizeYieldsEstimation.m
%start
disp ('Computing... will take a few minutes.');

%------------Import data-----------------------------------------%
%inputs=csvread('.\mdata\Maize\MaizeInputs.csv');
```

```
inputs=csvread('.\mdata\Nku\inputsY05Y16.csv');
%targets=csvread('.\mdata\Maize\MaizeYields.csv');
targets=csvread('.\mdata\Nku\outputsY05Y16.csv');
%inputs2016=csvread('.\mdata\Maize\MaizeInputs2016.csv');
inputs2006=csvread('.\mdata\Nku\inputY06.csv');
%targets2016=csvread('.\mdata\Maize\MaizeOutputs2016.csv');
targets2006=csvread('.\mdata\Nku\outputY06.csv');
%inputs2017=csvread('.\mdata\Maize\MaizeInputs2017.csv');
inputs2017=csvread('.\mdata\Nku\inputY17.csv');
%targets2017=csvread('.\mdata\Maize\MaizeOutputs2017.csv');
targets2017=csvread('.\mdata\Nku\outputY17.csv');
%------------Train the networks--------------------------------%
for i=1:30  %vary number of hidden layer neurons from 1 to 30
    %number of hidden layer neurons
    hiddenLayerSize = i;
    %FFBP network trained using Levenberg-Marquardt algorithm
    %mse - approximation(correct-estimated outputs), better if
   smaller
    net = newff(inputs,targets,hiddenLayerSize,{'tansig','purelin'},'
   trainlm','learngdm','mse');
    net.divideParam.trainRatio = 70/100;
    net.divideParam.valRatio = 15/100;
    net.divideParam.testRatio = 15/100;
    net.trainparam.show = 25;
    net.trainparam.time = inf;
    net.trainparam.goal = 0;
    net.trainparam.max_fail = 6;
    net.trainparam.min_grad = 1e-010;
    net.trainparam.mu = 0.001;
    net.trainparam.mu_dec = 0.1;
    net.trainparam.mu_inc = 10;
    net.efficiency.memoryReduction;
    % train the network
    [net,tr] = train(net,inputs,targets);
    %simulate 15% test data
    outputs = net(inputs(:,tr.testInd));

```

```
    %simulate year 2017 for Kitale or 2017 for Nakuru data
    outputs2017 = net(inputs2017);
    outputs2006 = net(inputs2006);

    %predict/simulate year 2017 for kitale & nakuru data
    outputs2017 = net(inputs2017);

    %RMSE for 15% random test data
    rmseTest(i)=sqrt(mean((outputs-targets(tr.testInd)).^2));
    %RMSE for year 2017 & 2006 test data
    rmse2017(i) = sqrt(mean((outputs2017-targets2017).^2));
    rmse2006(i) = sqrt(mean((outputs2006-targets2006).^2));

    %Regression for 15% random test data
    rTest(i)=regression(targets(tr.testInd), outputs);
    %Regression for year 2017 & 2006 test data
    r2017(i)=regression(targets2017, outputs2017);
    r2006(i)=regression(targets2006, outputs2006);

    %save the network in output folder
    %save(['.\mdata\Maize\net' num2str(i)],'net');
    %save(['.\mdata\Nku\net' num2str(i)],'net');
end
%-------------View the Network-----------------------------------%
view(net);
%Plot the RMSEs b*- means blue, Star, solid & ro- means red, circle
    solid
plot(1:30, rmseTest, 'mx-');
hold on;
plot(1:30, rmse2017, 'go-');
legend('Random', 'Year 2006');
xlabel('Number of hidden layer neurons');
ylabel('RMSE (^oC)');

%--------------Save the RMSEs-----------------------------------%
%fid=fopen('.\mdata\Maize\rmse.txt', 'wt');
fid=fopen('.\mdata\Nku\rmse1740.txt', 'wt');
```

```
%fid=fopen('.\mdata\Maize\est2017.txt', 'wt');
est=fopen('.\mdata\Nku\est201740.txt', 'wt');

%--------------print errors---------------------------------------%
fprintf(fid, '%4.0f\t %f\t %f\t %f\n', [1:30; rmseTest; rmse2006;
    rmse2017]);
fprintf(est, '%f\n', [outputs2017]);
%finish
fclose all;
disp ('Done... thank you.');

%-----------------------End-------------------------------------%
```

Listing A.2: Matlab code for yield prediction

Useful links

http://www.climatologylab.org/terraclimate.html

https://www.diagrams.net

https://www.overleaf.com