OPTIMAL SCALING INTEGRATED WITH PRINCIPAL COMPONENT REGRESSION: MODELING CASSAVA YIELDS,A CASE OF WESTERN KENYA

ALULU VINCENT HARRY

REG NO: SC384-C012-0005/2015

A Research Project Submitted In Fulfillment Of The Requirement For The Award Of The Degree Of Master Of Science In Applied Statistics Of Jomo Kenyatta University Of Science And Technology

DECLARATION

This is my own work and has not been presented for a degree in any other University.

Signature· · · · · · · · · · · · · · · · · · ·.. Date· · · · · · · · · · · · · · · · · · ·..

Alulu Vincent Harry

REG NO: SC384-C012-0005/2015

This project has been submitted for examination with our approval as University supervisors

Signature· · · · · · · · · · · · · · · · · · ·.. Date· · · · · · · · · · · · · · · · · · ·..

Prof. George Orwa

JKUAT, Kenya

Signature· · · · · · · · · · · · · · · · · · ·.. Date· · · · · · · · · · · · · · · · · · ·..

Mr. Henry Athiany

JKUAT, Kenya

Signature· · · · · · · · · · · · · · · · · · ·.. Date· · · · · · · · · · · · · · · · · · ·..

DEDICATION

To my beloved mother Fridah Busolo.

ACKNOWLEDGMENT

This rigorous process has been made possible and easier due to the help and support from a number of people. I acknowledge my two supervisors, Mr. Henry Athiany and Prof. George Orwa for their insight in my work, advise and general guidance. Without them, things could have been more difficult. I also thank Jomo Kenyatta Univeristy of Science and Technology-Kisumu CBD campus fraternity for making it possible in coordination of the entire process. Many thanks to Dr. Woyengo Vincent Were, Senior Research Officer at Kenya Agriculture and Livestock Research Organization-Kakamega, for giving me access to breeding data and support throughout. Last but not least, I thank God for giving me the strength to walk this academic path.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DEFINITION OF KEY TERMS

Optimal Scaling

This is the procedure applied to categorical variables in a dataset as a first transformation tool to generate continuous variables with correspondence between the two sets of variables observed.

Principal Component Analysis

Principal Component Analysis refers to the technique applied to reduce the number of variables in a dataset so as to remain with a subset that contains much information and in the process eliminate redundancy.

Principal Component Regression

This is the procedure of fitting a regression model on a dataset by first identifying the key components and then applying the fundamental regression procedure.

ABSTRACT

Cassava is a major food crop grown in the tropical and subtropical parts of the world. Cassava yields have been estimated, among other methods, based on weather factors,Fisher (1925) and based on plant parameters Lohse et al. (1985) .Most of the existing models do not incorporate all factors of production, a few that attempt this only puts into account the plant and weather factors, leaving out pests and diseases.

In this research work, a model for predicting cassava yield based on all factors of production using the principal component regression integrated with optimal scaling is developed. All factors of production are considered in this model. Moreover, the relationship between the different factors of production is established and the yield estimated based on the key components adduced to the factors of production in trial data in Western region of Kenya. Principal component regression and optimal scaling are used. Pearson correlation prior to principal component analysis indicated significance correlation among the factors of production. A prior to principal component regression, analysis using the variance inflation factor also indicated correlation in key factors of yield forecasting, variance inflation factor of 1666.667 (R^2=0.999). The coefficients derived from this model were unstable and therefore not reliable for yield prediction .Using the amount of explained variance criterion (70%-80%),the first eight principal components which accounts for almost 70% of total model variance are selected. Eight (8) key components are obtained as key determinants of yield; the most vital component having an eigen value of 2.149 and the least important having an eigen value of 1.005. The post principal component regression model was fitted. The PCR model indicates non-correlation among the eight principal components with the VIF attributed to the overall PCR model being2.564

,(R^2=0.610 (Adj R^2=0.590) .

The model developed incorporates all factors of production, regardless of whether the variables are continuous or categorical. It can factor in pests and diseases which are key factors of crop yield that have been neglected in existing models. The model developed will serve as a vital statistical tool in the crop production industry and impact on policy making.

Chapter 1

INTRODUCTION

1.1 Background Of The Study

Cassava (Manihot esculenta Crantz) is a root tuber plant which is grown in tropical and subtropical parts of the world. The starchy tuberous roots of cassava are a major source of carbohydrates and are consumed by 800 million people in Sub-Saharan Africa, Latin America and Asia Benesi (2005) Cassava is grown virtually in most parts of Kenya Karuri et al. (2001) and is a major source of income to farmers in agro-climatically disadvantaged regions and high potential areas of Coast, Central and Western Kenya Githunguri et al. (2007). The Western, Coastal and semi-arid Eastern regions of Kenya have the highest cassava production in that order Karuri et al. (2001). In Kenya, cassava is an important food security and income generating crop for farmers. It supports livelihood of approximately 8.6 million people in the lake basin region.

Most of the cassava is produced by small scale farmers using traditional farming systems Githunguri et al. (2007). About 38% of the cassava produced in the coastal lowlands of Kenya is consumed at household level and 51% of the farmers make chips for domestic use, sale to starch and feed factories or as an intermediate for production of flour Kiura et al. (2005).Cassava is considered as a crop for poor farmers due to its ability to be productive in low nutrient soils, where cereals and other crops perform poorly. Other advantages of cassava include drought tolerance and flexibility in planting and harvesting time. Cassava is also a low input crop and can be incorporated in various cropping systems. These attributes make cassava a mainstay of smallholder farmers in the tropics with limited access to agricultural inputs Aryee et al. (2006); Benesi (2005). As a result of recurrent droughts and subsequent food shortages in Africa, New Partnership for Africa's Development (NEPAD) has identified cassava as one of its key mandate commodities in order to reduce dependence on maize Fermont et al. (2009). In Kenya, the crop is grown on 77,502 ha with an output of 841,196 tons FAO (2007). A crucial impediment to cassava production in most nations in Africa is the Cassava mosaic disease (CMD) caused by single stranded DNA viruses in the family Geminiviridae and genus begomovirus Fauquet et al. (2005).

Cassava yield is measured as the number of tubers in tonnes per hectare (ton/ha) CFSAM (2006).The main factors affecting yield of cassava are inputs and weather. There are several existing models for crop yield forecasting that are also extended to cassava yield forecasting. For instance:

1.1.1 Crop Yield Forecasting Based On Weather Parameters

Weather affects crops differently during different stages of crop growth. Thus the extend of weather influence on crop yield depends not only on the magnitude of weather variables but also on the distribution pattern of weather over the crop season which, as such, calls for the necessity of dividing the whole crop season into fine intervals. This will increase number of variables in the model and in turn a large number of parameters will have to be evaluated from the data. This will require a long series of data for precise estimation of the parameters which may not be available in practice. Thus, a technique based on a relatively smaller number manageable parameters and at the same time taking care of entire weather distribution may solve the problem. Fisher (1925) has suggested a technique which requires small number of parameters to be estimated while taking care of distribution pattern of weather over the crop season. He assumed that the effect of change in weather variable on crop in successive weeks would not be an abrupt or erratic change but an orderly one that follows some mathematical law. He assumed that these effects are composed of the terms of a polynomial function of time. Further, the value of weather variable in w$^{\text{th}}$ week,X_w was also expressed in terms of orthogonal functions of time.

Substituting these in the multiple linear regression equation

$$Y = A_0 + A_1X_1 + A_2X_2 + ... + A_nX_n \tag{1.1.1}$$

Where Y denoted yield and X_w rainfall in w^{th} week , $w = 1, 2, \cdots n$ and utilizing the properties of orthogonal and normalized functions, he obtained

$$Y = A_0 + a_0\rho_0 + a_1\rho_1 + ... + a_k\rho_k \tag{1.1.2}$$

where

$A_0, a_0, a_1, a_2, ...a_k$ are constants to be determined and ρ_i,$i = 1, \cdots k$ are distribution constants of X_w. Fisher has suggested to use $k = 5$ for most of the practical situations. In fitting this equation for $k = 5$, the number of constants to be evaluated will remain 7, no matter how finely growing season is divided. This model was used by Fisher for studying the influence of rainfall on the yield of wheat. However, Hendrick and Scholl (1943) have modified Fisher's technique. They divided the crop season into n weekly intervals and have assumed that a second degree polynomial in week number would be sufficiently flexible to express the relationship. Under this assumption, the model was obtained as:

$$Y = A_0 + a_0 \sum_w X_w + a_1 \sum_w W X_w + a_1 \sum_w W^2 X_w \tag{1.1.3}$$

In this model number of constants to be determined reduces to 4, irrespective of n. This model was extended for two weather variables to study joint effects. Since the data for such studies extended over a long period of years, an additional variate T representing the year was included to make allowance for time trend. Another important contribution in this field is by Baier (1977). He has classified the crop-weather models in three basic types.

1. Crop growth simulation models
2. Crop-weather Analysis models
3. Empirical statistical models

The most commonly used models in crop forecasting are Empirical Statistical models. In this approach, one or several variables (representing weather or climate, soil characteristics or a time trend) are related to crop responses such as yield. The weighting coefficients in these equations are by necessity obtained in an empirical manner using standard statistical procedures, such as multi-variable regression analysis. Several Empirical Statistical models were developed all over the world. The independent variables included weather variables, agrometeorological variables, soil characteristics or some suitably derived indices of these variables. Water Requirement Satisfaction Index (WRSI), Thermal Interception Rate Index (TIR), Growing Degree Days (GDD) are some agroclimatic indices used in models.

Southern Oscillation Index(SOI) has also been used with other weather variables to forecast crop yield Ramakrishna et al. (2003). To account for the technological changes year variable or some suitable function of time trend was used in the models. Some workers have also used two time trends. Moving averages of yield were also used to depict the technological changes. In contrast to empirical regression models, the Joint Agricultural Weather Information Centre employs the crop weather analysis models that simulate accumulated crop responses to selected agro-meteorological variables as a function of crop phenology. Observed weather data and derived agro-meteorological variables are used as input data. USDA and FAO are the two organizations that systematically forecast world agricultural production and global crop information based on weather. Daily monitoring of satellite weather images and meteorological data provides the framework for agricultural weather analysis. Daily, weekly and seasonal summaries are processed and merged with historical weather and crop data for evaluation of the crop-yield potential. FAO has also carried out number of studies using agro-meteorological models. The methodology consists of developing an index depending on water deficit / water surplus in successive periods of crop growth. These models have good potential for early crop yield assessment for rainfed crops Frere and Popov (1979).

In India, major organizations involved in developing methodology for forecasting crop yield based on weather parameters are IMD and IASRI. The methodology adopted by IMD involves identification of significant correlations between yield and weather factors during successive overlapping periods of 7 to 60 days of the crop growing season. By analyzing the correlation coefficients for statistical and phenological significance, the critical periods when the weather parameters have significant effect on yield are identified. The weather parameters in critical periods along with trend variables are used through multiple regression analysis to obtain forecast equations. Using this methodology models were developed for principal crops on meteorological subdivisions basis. Data from various locations are averaged to get the figures for meteorological sub-divisions and these are utilised to develop the forecast model. Monthly forecasts are issued from these models by taking the actual data upto time of forecast and normal for the remaining period.

In some models yield Moisture Index, Generalised Monsoon Index, Moisture Stress, aridity anomaly Index are also used Sarkar (2003); Sarwade (1988). At IASRI, the model suggested by Hendricks and Scholl has been modified by expressing effects of changes in weather variables on yield in the w^{th} week as function of respective m correlation coefficients between yield and weather variables. This will explain the relationship in a better way as it gives appropriate weightage to different periods. Under this assumption, the models were developed for studying the effects of weather variables on yield using complete crop season data whereas forecast model utilised partial crop season data. These models were found to be better than the one suggested by Hendricks and Scholl.

The forecast model finally recommended was of the form:

$$Y = A_0 + \sum_{i=1}^{p} \sum_{j=0}^{1} a_{ij} Z_{ij} + \sum_{i \neq i'=1}^{p} \sum_{j=0}^{1} a_{ii`} Z_{ii'j} + cT + e \tag{1.1.4}$$

where

$$Z_{ij} = \sum_{w=1}^{m} r_{iw}^{j} X_{iw} \qquad and \qquad Z_{ii'j} = \sum_{w=1}^{m} r_{ii'w}^{j} X_{iw} X_{i'w} \tag{1.1.5}$$

Here Y is yield,$r_{iw}/r_{ii'w}$is correlation coefficient of yield (adjusted for trend effect) with i_{th} weather variable (X_{iw}) /product of i^{th} and i'^{th} weather $(X_{iw}X_{i'w})$ variables in w^{th} week, m is week of forecast, p is number of weather variables used and e is error term.

Models were successfully used for forecasting yields of various crops at district level as well as agroclimatic zone level Agrawal et al. (1980, 1983, 1986, 2001); Mehta et al. (2000). These models were used to forecast yield of paddy and wheat in different situations; (i) rainfed area having deficient rainfall (paddy), (ii) rainfed area having adequate rainfall (paddy) and (iii) irrigated area (wheat). The results revealed that reliable forecasts can be obtained using this approach when the crops are 10-12 weeks old. This approach was also used to develop forecast model for sugarcane at district level Mehta et al. (2000). However, these studies were carried out at district level and required a long series data of 25-30 years which are not available for most of the locations. Therefore, the study has been undertaken to develop the model on agro-climatic zone basis by combining the data of various districts within the zone so that a long series could be obtained in a relatively shorter period. Previous years yield, moving averages of yield and agricultural inputs were taken as the variables taking care of variation between districts within the zone. Year variable was included to take care of technological changes. Different strategies for pooling district level data for the zone were adopted. Results revealed that reliable forecasts can be obtained using this methodology at 12 weeks after sowing i.e. about 2 months before harvest. The data requirement reduced to 10-15 years as against 25-30 years for district level models. The approach has been successfully used for forecasting yields of rice, wheat and sugarcane for Uttar Pradesh Agrawal et al. (2001). At district level, model based on time series data on weather parameters has also been developed using technique of discriminant function analysis. The long series of 25-30 years has been classified into three groups - congenial, normal and adverse with respect to crop yields. Using weather data of these groups, linear / quadratic discriminant functions were fitted. These functions were used to find weather scores for each year at different phases of crop growth and were used as regressors in forecast model Rai (2000).

In another approach based on water balanced technique, models for rainfed crops using weighted stress indices have been developed. In this approach, water deficit / surplus has been worked out at different phases of crop growth and using suitable weights, accumulated weighted stress index has been developed for each year which was used as regressor in the forecast model Saksena et al. (2001)

1.1.2 Crop Yield Forecasting Based On Plant Parameters

Effects of weather and inputs are manifested through crop stand, number of tillers, leaf area, and number of earheads just to mention but a few which ultimately determine crop yield. As such, plant characters can be taken as the integrated effects of various weather parameters and crop inputs. Thus the other approach to forecast crop yield is to use plant characters. In USDA, the net yield per acre for each sample plot is computed as Lohse et al. (1985).

$$y_i = (F_i \times C_i \times Wi) - L_i \tag{1.1.6}$$

where ;

Fi = Number of fruits harvested or forecast to be harvested in the ith sample plot

Ci = Conversion factor using the row space measurement to inflate the plot counts to a per acre basis

Wi = Average weight of fruit harvested or forecast to be harvested

Li = Harvest loss as measured from post-harvest gleanings (the historic average is used during the forecast season)

$$\bar{y} = \left(\frac{\sum y_i}{n}\right) \tag{1.1.7}$$

for the n sample plots.

Separate models are used to forecast the number of fruits (F_i) to be harvested and the final head weight (W_i). This method cannot be followed in India/tropical countries as time period from head emergence to maturity is hardly one to two months for most of the crops whereas in USA this takes two to three months. Forecast of head weight at maturity therefore cannot be obtained much in advance in India, as such this will not be useful for obtaining early forecast in such countries. In India, yield is directly regressed on plant counts and yield contributing characters for obtaining forecast model. Considerable work has been done at IASRI using this approach. The data are collected at different periodic intervals through suitable sampling design for 3 to 4 years from farmers fields. Two types of approaches have been attempted namely: Between year model and Within year model.

Between Year Models

These models are developed taking previous year(s) data. Objective yield forecasts are obtained by substituting the current year plant data into a model developed from the previous year(s). An assumption is made that the present year is a part of the composite population of these (previous) years. Most commonly used models are based on regression approach.

Different between year models

In this category, the following models are used.

Linear Regression Models

This model uses the regression equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + e \tag{1.1.8}$$

where Y and X_i are yield and plant characters respectively. These may be used in original scale or some suitably transformed variables of these can be used. β_0 and β_i are constants to be estimated and e is random error. These models utilise data at one point of time only during the crop growth Jha et al. (1981); Sardana et al. (1972); Singh and Bapat (1988); Singh et al. (1976). These models were improved taking regressors as principal components of plant characters Jain et al. (1984) or growth indices based on plant characters observed on two or more points of time during the crop growth Jain et al. (1985). The growth indices are obtained as weighted accumulations of observations on plant characters in different periods, weights being respective correlation coefficients between yield and plant characters. The model can be written as:

$$Y = \beta_0 + \sum \beta_i G_i + e \tag{1.1.9}$$

Where

$$G_i = \sum_{w=n_1}^{n_2} r_{iw} X_{iw}$$

G_i is the index of the i^{th} character, w is period identification, n_1 and n_2 are the initial and final periods considered in developing the index of the character, r_{iw} is simple/partial correlation coefficient between yield and i^{th} character in w^{th} period (X_{iw}).

Probability Model

Multiple regression technique has been extensively used in developing models for crop yield forecasting. Least squares technique is used for estimating the parameters of the regression model. The optimality properties of these estimates are described in an ideal setting which is not often realised in practice. It has been observed that regression based on different subsets of data produce very different results, raising questions of model stability. To overcome some of the drawbacks of regression model probability model for forecasting crop yield using Markov Chain theory has been developed. This method, being completely model free, does not require any assumption about independent and dependent variables. Markov Chain method has the advantage of providing non-parametric interval estimates and is robust against outliers/extreme values. In this method, growth process of the crop is divided in s phenological stages. A markov chain model is constructed by defining a set of states, which describe the condition of an individual plant (or average condition of a group of plants) at specified time within the phonological stages. Individual states are defined on the basis of available qualitative and quantitative information to describe plant condition. Let n_i, for $i = 1, 2, \cdots, s$ denote the number of states at the commencement of stage i. Let $A_{i,i+1}$, and for $i = 1, 2, \cdots, (s-1)$ denote the $(n_i \times n_{i+1})$ transition matrix which gives the transition probabilities of a plant (or group of plants) moving from any possible state of stage i to any possible state of stage i+1. As a property of transition matrices, each row of an $A_{i,i+1}$ matrix has a summation of unity. Let F denote the matrix of transition probabilities from each of the (n-ns) states of (s-1) intermediate stages to each of the ns states, the last (harvest) stage. The ns states are defined as quantitative intervals of yield.

F matrix can be obtained as:

$$\mathbf{F} = \begin{pmatrix} \prod_{i=1}^{s-1} A_{i,i+1} \\ \prod_{i=2}^{s-1} A_{i,i+1} \\ A_{s-2,s-1}\, A_{s-1,s} \\ A_{s-1,s} \end{pmatrix} \tag{1.1.10}$$

F matrix can be used to forecast crop yields. Each row of F represents a crop condition (state) at a certain crop stage. The ns states of the final stage are defined as quantitative intervals of yield. Each column of F represents a different yield interval. The values of each row of F are the estimated probabilities of the crop producing a final yield within each of the ns intervals. Thus, each row of F is a predicted yield distribution for a given stage and state. Each of the (n-ns) forecast yield distributions in the F matrix may be analysed to get mean and standard error of the forecast. In particular, transition probability matrix As-1, s will give mean and standard error of forecast at stage(s-1). This method was applied to forecast yield of corn and cotton by USDA Matis et al. (1985) and sugarcane Agrawal and Jain (1996); Jain and Agrawal (1992). Models using higher order markov chain and using principal components and growth indices of plant characters in markov chain approach were also developed Jain and Ramasubramanian (1998); Ramasubramanian and Jain (1999); Ramasubramanian et al. (2004).

Within year models

The 'between year models' while performing satisfactorily in typical years may falter in atypical years. A model which uses data from the current growing season only may be beneficial in improving forecasts during a year with atypical growing conditions. These models are developed to provide forecasts of pertinent components of crop yield relying entirely on growth data collected from plant observations during the current growing season. A logistic model having some yield components as dependent variable and an independent 'time' variable generally fits well to the growth process of crop yield components like dry matter accumulation etc. The model uses repeated observations from the current year to estimate the parameters needed to forecast the dependent variable at maturity. The model is:

$$Y_i = \frac{\alpha}{1 + \beta\rho^{ti}} + e_i \tag{1.1.11}$$

and $\quad i = 1, 2, \cdots n; \quad \alpha > 0, \quad \beta > 0, \quad 0 < \rho < 1$

where

Y_i = dependent growth variable

t_i = independent time variable

e_i = error term

Partial crop season data are utilised to fit the curve and the value at harvest is predicted through this curve which in turn is used to forecast yield House (1977); Jain and Agrawal (1992); Larsen (1978); Nealon (1976). The parameter α is the most important parameter to be estimated as it gives average amount of yield component (eg. dry matter) at maturity. It is likely to be overestimated when partial crop season data based on small data points that too falling on the lower side of the curve where the growth has steep rise are used to fit the model to forecast the yield component at maturity. This may need suitable modification in the model so as to capture (dry matter at maturity) from partial crop season data. The modified logistic model Jain and Agrawal (1992) is as follows:

$$Y_i = \frac{\alpha\sqrt{\dfrac{t_m}{t_f}}}{1 + \beta\rho^{ti}} + e \tag{1.1.12}$$

where t_m is time of maturity and t_f is the time of forecast.

1.1.3 Models Using Spectral Data

Since the approach using plant characters requires collection of data from farmers' fields, the data can be used on characters which can be measured easily without involving much expertise, cost and sophisticated instruments. Some characters contributing significantly towards yield may not find place in the model due to these limitations. This calls for the necessity of including some other variables in the model along with biometrical characters which could take care of such variables indirectly.

During the last three decades, considerable work has been carried out in India in the spectral response and yield relationships of different crops at Space Applications Centre, Ahmedabad, under the remote sensing applications mission called Crop Acreage and Production Estimation (CAPE). Spectral indices such as ratio of infra-red (IR)/Red(R) and Normalised difference (ND) = (IR-R) / (IR+R) are calculated from remotely sensed data and are used as regressors in the model Singh et al. (2012); Space Application Centre (1990).

The scheme needed further improvement. Project has been formulated to integrate Agrometeorology and Land-based observations along with remote sensing data.

The experience in this context is that remote sensing can supplement the existing data collection system but never completely replace it. The two data collection systems must be integrated through rigorous statistical methodology. At Space Application Centre, methodology has been developed which provides multiple forecasts for rice and wheat using remotely sensed data for acreage forecast whereas forecasts for productivity are obtained using meteorological and agro- meteorological indices Patel et al. (2004) .

1.1.4 Models Using Farmers' Appraisal

Farmer is the best judge of the likely production in the field. Farmers' appraisal, therefore, could serve as a good input for forecasting the yield and replace some of the characters requiring expertise or use of sophisticated instruments for their measurements and thus reducing the cost on data collection. A study has been carried out to study the feasibility of using farmers' appraisal in the forecast model for sugarcane Agrawal and Jain (1996). The results revealed that a reliable forecast could be obtained using plant population and farmers appraisal.

Another methodology based on farmers appraisal data has been developed using Bayesian approach. The study has been carried out for wheat in Muzaffarnagar district. Expert opinion data were collected in a number of rounds in a year by interviewing the selected farmers regarding their assessment about the likely crop production and chance of occurrences in yield classes. From these responses average prior probabilities were computed. Actual harvest yield and farmers appraisal data on yield for previous year(s) were taken into account to obtain posterior probabilities which were then used for obtaining Bayesian forecast of crop yield for current year Chandrahas and Rai (2001).

1.1.5 Principal Component Regression

Forewarning models can be developed using the principal component techniques as normally relevant weather variables are large in number and are expected to be highly correlated among themselves. Using the first few principal components of weather variables as independent variables forecast models can be developed.

1.1.6 Discriminant Function Analysis

The methodology is similar to the one used for yield forecasting, replacing yield by the character depicting pests and diseases Johnson et al. (1996).

1.1.7 Artificial Neural Network (ANN)

ANN provides an attractive alternative tool for forecasting purposes. ANNs are data driven self-adaptive methods in that there are few apriori assumptions about the models for problems under study. They learn from examples and capture subtle functional relationships among the data even if the underlying relationships are unknown or hard to describe. After learning the data presented to them, ANNs can often correctly infer the unseen part of a population even if data contains noisy information. As forecasting is performed via prediction of future behaviour (unseen part) from examples of past behaviour, it is an ideal application area for ANNs, at least in principle Agrawal et al. (2004); De Wolf and Francl (2000); De Wolf and Franel (1997). However, the technique requires a large data base.

1.1.8 Within Year Model

Sometimes, past data on pests and diseases are not available but the pests and diseases status at different points of time during the crop season are available. In such situations, within years growth model can be used for forewarning maximum disease severity / pest population, provided there are 10-12 data points between time of first appearance of pest / disease and maximum or most damaging stage. The methodology is similar to yield forecast model as eplained in the aforementioned within year models Agrawal et al. (2004).

1.1.9 Developments In Cassava Industry

Key developments have been observed in the cassava industry in recent times as pertains yield forecasting. Downscaled climate scenarios can be used to produce seasonal crop yield forecasts, which are needed in order to effectively and efficiently plan and allocate agricultural resources to reduce risk and uncertainties due to seasonal climate and weather variability. In Thailand and Southeast Asia, rice and cassava are two major food and energy crops. The majority of their production areas are under rainfed conditions and are very sensitive to weather and climate variability. Under DSS-SCY4cast (Decision Support System for Seasonal Crop Yield Forecast) framework, soil, climatic, crop genetic coefficients and crop management data sets were incorporated and linked to the CSM-CERES-Rice (Crop System Model-Crop-Environment Resource Synthesis-Rice) and CSM-CropSim-Cassava (Crop System Model Crop Simulation-Cassava) process-oriented models. The framework was designed and deployed as an online tool to produce seasonal crop yield forecasts at monthly intervals, i.e., MayDecember (8 month forecast), JuneDecember (7 month forecast), JulyDecember (6 month forecast), AugustDecember (5 month forecast), SeptemberDecember (4 month forecast), OctoberDecember (3 month forecast), NovemberDecember (2 month forecast), and monthly forecast for December (1 month forecast). Local communities are the integral component of the approach as the provider of field-based observed data and as the beneficiary of the seasonal crop yield forecasts. The described approach can be tailored to support local and policy communities in Thailand and Southeast Asian to produce seasonal forecasts of various crop production systems and learn to adapt and to sustain the environment and society.

There are a number of issues affecting the cassava industry in Kenya. For instance, the Agricultural Sector Development Strategy (ASDS) 2010-2020 still considers cassava as a food crop rather than an industrial crop hence does not address the challenges and constraints facing industrial production and processing of cassava Aberi (2012). The industrial potential of cassava has not been tapped to a great extent and broadening the pre-harvest yield modeling techniques utilization of cassava remains unexploited. Cassava is associated with labour intensive farming and low-cost produce unlike other root and tuber crop Benesi (2005). Therefore, it is vital to develop low cost, energy effective and adequate statistical model for yield prediction and subsequent industrial utilization of cassava. This model should be able to easily incorporate both the disease and pest aspects as vital components that affect yield.

The conventional Principal component regression is a well known technique to reduce number of explanatory variables in the model. The technique involves conversion of explanatory variables into a set of uncorrelated variables with variances in descending order (known as principal components). The whole variation of the system explained by explanatory variables is explained by first few principal components which are used as regressors in the model in place of original variables. Besides solving the problem of number of explanatory variables more than number of observations, the technique also solves the problem of multicollinearity. The approach has been attempted for forecasting yields of rice, wheat and sugarcane in Uttar Pradesh in India but the approach was not found to be successful Singh (2010). The success of PC relies on linearity and good transformation of variables which is not always assured partly due to the presence of regressors that are on other scales mostly ordinal and nominal. It is worth noting that the transformations and standardization in PCR are always underpinned on the assumption that all the regressors are linear. This is untrue in cases where we merely have data in form of ranks.

In this research work a specic form of PCR in which the conventional PCR was integrated with optimal scaling technique that converted the non-metric (ordinal and nominal) variables to a continuous scale. PCR is a linear technique, in the sense that observed variables are approximated by linear combinations of principal components. It can also be a bilinear technique, in the sense that elements of the data matrix are approximated by inner products, which are bilinear functions of component scores and component loadings. The nonlinearities in the forms of PCR in this case were nonlinear transformations of the variables when applying the optimal scaling on non-metric variables, and we still preserved the basic (bi) linearity of PCR. In their research work,Mair et al. (2009) applied optimal scaling on all the variables in their dataset, before fitting a homogenity analysis model using the homals package in R. However, in this research project, only the nominal and ordinal factors in the dataset were subjected to the Gifi system of descriptive multivariate analysis-optimal scaling technique (a non-linear technique) Michailidis and de Leeuw (1998) for conversion to a continuous scale before the conventional PCR was applied to the final dataset of continuous variables. Therefore, the optimal scaling was integrated into the conventional PCR to develop an prediction model that incorporated soil and disease data. The major benefit of this model is the fact that the PCR was finally applied to variables that were on same measurement scale (continuous) as opposed to application of the conventional PCR on data on different scales.

1.2 Statement of the problem

There has been increased demand for export of cassava to China, US and other developed nations of the world whose production of currency papers, starch related products, recently fast increasing bio-fuel energy and others, largely depends on Africa's production of cassava. The demand for Cassava has globally increased and it has overshot supply; the occurrence of drop in yield has put a lot of pressure on production of Cassava and the present increase in cultivation is not enough to curb demand, according to Food and Agriculture Organization of the United Nations database FAOstat (2009). The available statistical models for yield forecasting are not efficient either, since most do no consider all factors of production in the model at the same time. A few for instance the integrated model Mehta et al. (2000) take more factors into account. No satisfactory forecasting model which has universal validity exists till date Singh (2010). Moreover the existing multiple linear models are unstable in terms of prediction when many covariates are included Stevens (2002). Therefore, research scientists incur losses in time when they attempt to forecasts yield using these models that apply to inputs, weather, pests and diseases as separate entities. There is an inadequacy in models that take into account all factors at the same time. This is partly due to the fact that the data collection scales are different and also how different the factors of production affect yield. The existing few models that attempt to combine these factors are so time-consuming since they require a lot of calculations and data transformations. This can only be made easier via crop simulation models (CSMs) which are not being used in Kenya due to the expertise and technology required. Therefore, most research scientists rely on the traditional models for yield forecasting.

Achieving the production efficiency of cassava was the problem that this research project intended to address through a statistical tools approach. The project was focusing on coming up with a model for cassava yield forecasting for different varieties as affected by different factors of production that is capable of delivering the cassava production efficiency.

1.3 Objectives

1.3.1 General Objective

To use optimal scaling integrated with principal component regression in modeling of cassava yields in Western Kenya.

1.3.2 Specific Objectives

1. To establish the relationship amongst different factors of cassava production.
2. To determine key principal components in the factors of cassava production using the optimal scaling integrated with principal component regression
3. To predict cassava yields using the PCR model integrated with optimal scaling

1.4 Justification Of The Study

An incorporated model, that utilizes two techniques that address all the aspects of the underlying data set is key to precise prediction of cassava yield. This model incorporates PCR and optimal scaling which is a transformation technique for categorical data. This therefore makes it easier to incorporate pests and diseases in the prediction model.

The key to better policies by the government in the agriculture sector lies within the ability to understand the major components of crop yield. Therefore, an efficient model for cassava yield prediction will be very important in policy-making and coming up with interventions that can improve production. The model will also be a knowledge base and a tool that will help the farmer population to understand the key factors of cassava production.

1.5 Limitations

This research work is limited to identifying existence of relationship between factors of production, however it does not go beyond the key factors as reported in the dataset. The key principal components derived herein therefore pertains to the factors of this very data set. Inclusion of more diverse factors will therefore result in different key principals. The prediction thus made is as a result of this PCR and a different PCR model based on different factors would not guarantee similar results.

1.6 Organization of the project

The rest of this project is organised as follows. In the chapter that follows, there is literature review in which literature relating to the specific objectives have been reviewed and gaps therein exposed through a critique. Thereafter, the chapter that follows has the methodology in which the statistical method and model have been discussed. The subsequent chapter after methododlogy has the empirical sudy in which the actual data analysis has been carried out. The last chapter has the findings are summarised, conclusions and recommendations made.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

It is generally accepted that the earliest descriptions of the singular value decomposition technique now known as PCA were given by Pearson Pearson (1901) and Hotelling (1933) in the context of a two-way analysis of an agricultural trial.

Pearson (1901) developed the principal component analysis technique. He observed that in many physical, statistical and biological investigations it is desirable to represent a system of points in a plane, three or higher (p) dimensional space by the best fitting straight line or plane. He also noted that the geometric optimization problems he considered lead to principal components (PCs).

Hotelling (1933) motivation was that there may be a smaller fundamental set of independent variables which determine the values of the original p variables. He notes that such variables have been called factors in the psychological literature, but introduces the alternative term components to avoid confusion with other uses of the word factor in mathematics.

Hotelling chooses his components so as to maximize their successive contributions to the total of the variances of the original variables, and calls the components that are derived in this way the principal components.

The analysis that finds such components is then christened the method of principal components.

Tipping and Bishop (1999) introduces prior distributions for the parameters of a model for PCA. His main motivation was to provide a means of deciding the dimensionality of the model for prediction rather than the measurement scale of the components being considered. This is a key element that this research project will exploit.

Lanterman (2000) and Wang and Staib (2000) each use principal components in quantifying prior information in (different) image processing contexts.

Raychaudhuri et al. (2000) used the PCA to analyze DNA microarray data sets. They stated that clustering genes based on expression information, it can be important to determine if the experiments have independent information or are highly correlated. In their analysis genes were considered as variables or the experiments as variables or both. They observed that PCA can find a reduced set of variables that are useful for understanding the experiments. Moreover, PCA Reduction of dimensionality in the sporulation data aids in data visualization. However, application of PCA to time series is somewhat controversial because of the problems with uneven time intervals and the dependencies between data points. Jamak et al. (2012) used PCA in authorship detection. In their research work, function words are counted from selected text (short words used by the author) and then this data is transferred into principal components form. Principal components method compresses data without losing its descriptive power. In conclusion, they proposed a method of authorship detection using principal component analysis for texts written in Bosnian language, which has proven as an efficient tool. However, they observed that future research should focus on the short words, and describe the extend of the shortness of words significant for the authorship detection. Furthermore, the research should test the behaviour of the method among larger amount of text samples.

2.2 Factors Guiding Cassava Production

Cassava yields are affected by socioeconomic factors, market conditions and abiotic constraints. Nonetheless pests and diseases are well known to substantially reduce yields, resulting in multi-billion-dollar crop losses Anderson (2005); Coulibaly et al. (2004); Fondong et al. (2000); Hillocks and Jennings (2003); Hillocks et al. (2002); Legg et al. (2004); Maruthi et al. (2004); Renkow and Byerlee (2010); Waddington et al. (2010). In plant breeding experiments, the yield attained at a certain time is dependent on environmental factors, genetic factors, diseases and pests. The effects of weather and inputs are manifested through crop stand, number of tillers, leaf area, and number of earheads just to mention but a few which ultimately determine crop yield. As such, plant characters can be taken as the integrated effects of various weather parameters and crop inputs Therefore, all these factors are key while coming up with a model for yield prediction. Models built based on data on plant characters along with agricultural inputs have been found to be better than models based on plant characters alone in jowar and apple Jain et al. (1985). However all these production factors are key and a model built on either one only has obvious shortfalls.

Incorporating all factors is ideal, but often it is not possible to include all the variables in a single model. In such situations composite forecast can be obtained as a suitable combination of forecasts obtained from different models. Various strategies for combining forecasts have been suggested under different situations so as to take into account the many factors of production Mehta et al. (2000). All these strategies attempt to model yield based on different factors of production. Walker et al. (2007) used pecific climatic, edaphic, crop and fallow growth data from five sites in Western Kenya to calibrate and validate simulations of maize and improved fallow growth using the Water, Nutrient and Light Capture in Agroforestry Systems (WaNuLCAS) model.Although their model was effective in modelling maize yield after cross validation, the model left out diseases and pests which are major factors determining yield levels.

2.3 Dimension Reduction

Amadei et al. (1993) introduced PCA in molecular dynamics (MD). They proposed that except for the degrees of freedom that belong to the essential subspace of proteins, all the other modes are largely irrelevant Gaussian fluctuations, therefore necessitating the choice of the key components in an MD model. Nonetheless an inadequacy they observed was that a single MD trajectory may not entail all possible modes that are essential towards dynamical conformational changes. A direct consequence of such considerations is that even for a single trajectory, the principal modes obtained during one observation window may differ from another window. While this remains true, a very long (few hundreds of ns) MD simulation may not necessarily yield highly convergent eigenvectors from PCA as compared to simulations with a time span on the order of tens of ns. Even though efforts have been put in devising methods of enhanced sampling of essential dynamics Amadei et al. (1999); Hess (2002), convergence of eigenmodes remains a critical issue.

Zhu et al. (2012) integrated the credibility model (semi-linear credibility, and regression credibility models) in the PCA to develop an improved crop forecasting model through the incorporation of weather data. Empirical results showed that their model was able to provide better in-sample and out-of-sample yield forecasting results after cross-validation. However, their model left out other key factors like pests and diseases which are important determinants of yield. In this research project, optimal scaling will be incorporated with PCR technique so that the pest and disease aspects are included in the model. Hansen et al. (2009) incorporated the PCA in General Circulation Models (GCMs) in developing simulation models for maize yield prediction in Semi-Arid Kenya. In their analyses they combined downscaled rainfall forecasts, crop yield simulation, stochastic enterprise budgeting and identification of profit-maximizing fertilizer N rates and stand densities .Pest and disease scores which are vital factors in crop production were not incorporated in their model.

Chapter 3

METHODOLOGY

In fitting the cassava yield prediction model, optimal scaling integrated with principal component regression approach was used. Yield (Y), the regressed variable was predicted based upon cassava genotype,soil, pest and disease factors.

3.1 Statistical Method and Model Review

3.1.1 Optimal Scaling

Optimal scoring assigns numeric values to the observations in a way that simultaneously fulfills two conditions: (I) The assigned scores strictly maintain the specified measurement characteristics for the data, and (2) they fit the statistical model as well as possible Jacoby (1999). This optimal scaling strategy provides the best set of numerical assignments for the data, where "best" is defined in terms of goodness of fit between an analytic model and a set of empirical observations Young (1981).

A data matrix of the non-metric independent variables $x_1, x_2, \cdots, x_n$ is transformed through optimal scaling to convert them from a discrete scale to a continuous scale. This is done in correspondence of a dependent quantity y say yield. The elements of y have a one-to-one correspondence with the elements of x; that is, x_1 corresponds to y_1, x_2 corresponds to y_2, and so on. Letting the optimally scaled vector be x*, then optimal scaling is simply the procedure of obtaining x* from x, y, and the measurement assumptions that the analyst makes about x. The optimal scaling procedure simply takes the conditional means of the y_is within the observational categories of x, and assigns those means to the entries in x* corresponding to their respective categories. In this research project, reaction to diseases and pests was recorded as scores on a research scientists predetermined scale of 0-5. Therefore, such variables were subjected to transformation using optimal scaling before they were incorporated into the PCA procedure.

3.1.2 Principal Component Regression

Principal Component Analysis

PCA is a standard statistical technique that can be used to reduce the dimensionality of a data set. The method is mainly concerned with identifying variances and correlations in the data set. We meet the goal of reducing the dimensionality by maximizing the variance of a linear combination of the variables Rencher (2002). We presented the mathematics behind the method of PCA by considering a general case. More details on technical aspects can be found in Cadima and Jolliffe (1995); Cooley and Lohnes (1971); Hyvarinen et al. (2001); Jolliffe (2002). Consider a data set consisting of p variables observed on n subjects. Variables are denoted by $(x_1, x_2, \cdots, x_p)$. In general, data are in a table with the rows representing the subjects (individuals) and the columns the variables. The data set can also be viewed as an $n \times p$ rectangular matrix X. Note that variables are such that their means make sense. The variables are also standardized. In this case, the PCA is called normalized principal component analysis, and was based on the correlation matrix (and not on variance-covariance matrix). The variables were to lie on the unit sphere; their projection on the subspace spanned by the principal components is the "correlation circle". Standardization allowed the use of variables which are not measured in the same units (e.g. temperature, weight, distance, size, etc.). The PCA gave us a subspace of reasonable dimension so that the projection onto this subspace retained "as much as possible" of the information present in the data set, i.e., so that the projected clouds of points would be as ”dispersed” as possible. In other words, the goal of PCA was to compute another basis that best re-expressed the dataset. The hope was that this new basis would filter out the noise and reveal hidden structure.

$$\mathbf{x}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ . \\ . \\ . \\ x_{pi} \end{pmatrix} reducesdimensionality \to z_i = \begin{pmatrix} z_{1i} \\ z_{2i} \\ . \\ . \\ . \\ z_{qi} \end{pmatrix} ; q < p \tag{3.1.1}$$

Dimensionality reduction implies information loss. How do we represent the data in a lower dimensional form without losing too much information? Preserve as much information as possible is the objective of the mathematics behind the PCA procedure.

We first of all assume that we want to project the data points on a 1-dimensional space. The principal component corresponding to this axis is a linear combination of the original variables and can be expressed as follows:

$$z_1 = A_{11}x_1 + A_{12}x_2 + \cdots + A_{1p}x_p = Xu_1 \tag{3.1.2}$$

where $u_1 = (A_{11}, A_{12}, \cdots, A_{1p})^T$ is a column vector of weights. The principal component z_1 is determined such that the overall variance of the resulting points is as large as possible. One could make the variance of z_1 as large as possible by choosing large values for the weights $A_{11}, A_{12}, \cdots, A_{1p}$. To prevent this, weights are calculated with the constraint that their sum of squares is one, that is u_1 is a unit vector subject to the constraint:

$$A_{11} + A_{12} + \cdots + A_{1p} = \parallel u \mid^2 \tag{3.1.3}$$

Equation(3.1.2) is also the projections of the n subjects on the first component. PCA finds u_1 so that:

$$Var(z_1) = \frac{1}{n}\sum_{i=1}^{n} z_{1i}^2 = \frac{1}{n} \parallel z_i^2 \parallel = \frac{1}{n} u_1^{'} X^{'} X u_1 \tag{3.1.4}$$

is maximal. The matrix $C = \dfrac{1}{n} X^{'} X$is the correlation matrix of the variables.The optimization problem is therefore:

Max $u_1^{'} Cu1$

s.t $\parallel u_1^2 \parallel = 1$

This implies that we search for a unit vector u_1 so as to maximize the variance of the projection on the first component. The technique for solving such optimization problems (linearly constrained) involves a construction of a Lagrangian function and solving of partial derivatives.

$$\delta_1 = u_1^{'} C u_1 - \Lambda_1 \left(u_1^{'} u_1 - 1 \right) \tag{3.1.5}$$

Taking the partial derivative $\dfrac{\partial \delta_1}{\partial u_1} = C u_1 - \Lambda_1 u_1$ and solving the equation $\dfrac{\partial \delta_1}{\partial u_1} = 0$ yields:

$$C u_1 = \Lambda_1 u_1 \tag{3.1.6}$$

By pre-multiplying each side of this condition by $u_1^{'}$ and using the condition $u_1^{'} u_1 = 1$ we obtain:

$$u_1^{'} C u_1 = \Lambda_1 u_1^{'} u_1 = \Lambda_1 \tag{3.1.7}$$

It is known from matrix algebra that the parameters u_1 and Λ_1 that satisfy conditions (3.1.7) and (3.1.8) are the maximum eigenvalue and the corresponding eigenvector of the correlation matrix C.

Thus the optimum coefficients of the original variables generating the first principal component z_1 are the elements of the eigenvector corresponding to the largest eigenvalue of the correlation matrix. These elements are also known as loadings. The second principal component is calculated in the same way, with the condition that it is uncorrelated (orthogonal) with the first principal component and that it accounts for the largest part of the remaining variance. Using induction, it can be proven that PCA is a procedure of eigenvalue decomposition of the correlation matrix. The coefficients generating the linear combinations that transform the original variables into uncorrelated variables are the eigenvectors of the correlation matrix. We also note that; rather than maximizing variance, it might sound more plausible to look for the projection with the smallest average (mean-squared) distance between the original points and their projections on the principal components. This turns out to be equivalent to maximizing the variance (Pythagorean Theorem). Principal components are all uncorrelated (orthogonal) to one another. This is because matrix C is a real symmetric matrix and as studied under linear algebra, it is diagonalizable and the eigenvectors are orthogonal to one another. Again because C is a covariance matrix, it is a positive matrix in the sense that $u^{'}C_u \geq 0$ for any vector u . Therefore the eigenvalues of C are all non-negative Rencher (2002).

$$\mathbf{V}ar(Z) = \begin{pmatrix} \Lambda_1 & 0 & \cdot & 0 \\ 0 & \Lambda_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \Lambda_p \end{pmatrix} \tag{3.1.8}$$

The eigenvectors are the “preferential directions”of the data set. The principal components are derived in decreasing order of importance; and have a variance equal to their corresponding eigenvalue. The first principal component is the direction along which the data have the most variance. The second principal component is the direction orthogonal to the first component with the most variance. It is clear that all components explain together 100% of the variability in the data. Thus the PCA works like a change of basis and allows us to obtain a linear projection of our data, originally in R_p, onto R_q, where $q<p$. The variance of the projections on to the first q principal components is the sum of the eigenvalues corresponding to these components. If the data fall near a $q-$dimensional subspace, then $p-q$ of the eigenvalues will be nearly zero.

Criteria for determining the number of meaningful components to retain

When determining the number of meaningful components, the subspace of components retained must account for a reasonable amount of variance in the data. Usually the eigenvalues are expressed as a percentage of the total. The fraction of an eigenvalue out of the sum of all eigenvalues represents the amount of variance accounted by the corresponding principal component. The cumulative percent of variance explained by the first q components is calculated with the formula:

$$r_q = \frac{\sum_{j=1}^{q} \Lambda_j}{\sum_{j=1}^{p} \Lambda_j} \times 100 \tag{3.1.9}$$

How many principal components we should use depends on how big an rq we need. This criterion involves retaining all components up to a total percent variance Jolliffe (2002). It is recommended that the components retained account for at least 60% of the variance. The principal components that offer little increase in the total variance explained are ignored; those components are considered to be noise. When PCA works well, the first two eigenvalues usually account for more than 60% of the total variation in the data.

Regression (PCR)

The conventional regression equation is written in matrix form as:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{e} \tag{3.1.10}$$

In ordinary least squares, the regression coefficients are estimated using the formula:

$$\mathbf{B} = (\mathbf{X}^{'}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{'}\mathbf{Y} \tag{3.1.11}$$

Since the variables are standardized $X^{'}X = R$ where R is the correlation matrix of independent variables. To perform principal components (PC) regression, we transform the independent variables to their principal components as outlined above. Mathematically, we write;

$$\mathbf{X}^{'}\mathbf{X} = \mathbf{PDP}^{'} = \mathbf{Z}^{'}\mathbf{Z} \tag{3.1.12}$$

where **D** is a diagonal matrix of the eigenvalues of $X^{'}X$, **P** is the eigenvector matrix of **XX** and **Z** is a data matrix similar in structure to **X** made up of the principal components. **P** is orthogonal so that$P^{'}P =$ I. We have created new variables **Z** as weighted averages of the original variable **X**. This is similar to using transformations such as the logarithm and the square root on our data values prior to performing the regression calculations. Since these new variables are principal components, their correlations with each other are all zero. If we use variables X_1, X_2, and X_p, we will end up with Z_1, Z_2, and Z_p. The principal components are then selected as outlined above to remain with a few that account for maximum variance in the dataset.

When we regress Y on $Z_1, Z_2, ...Z_{p-1}$, multicollinearity is no longer a problem . In this research project PCA was used to identify the key factors that contribute to cassava yield. This was done by applying the PCA technique to the X matrix formed from all the factors of production of cassava (environmental factors, genetic factors, diseases and pests). After identification of the key factors through PCA, the PCR was used to fit a model for predicting the yield of cassava. Regression coefficients were generated and interpreted in a similar manner to multiple linear regression.

3.1.3 Parameter Estimation

Regression Coefficients

Multiple Linear regression coefficients estimation

If data is mean-centred, the regression model in MLR is:

$$\mathbf{Y} = \mathbf{XB} + e \tag{3.1.13}$$

Using OLS, the regression coefficients are determined by minimizing $\mathbf{e^T e}$. This gives a solution that can be expressed as:

$$\mathbf{B} = \mathbf{X^T X}^{-1}\mathbf{X^T Y} = \mathbf{B_{OLS}} \tag{3.1.14}$$

Principal Component Regression Coefficients Estimation

These are the estimated values of the regression coefficients $b_0, b_1, \cdots, b_p$. The value indicates how much change in Y occurs for a one-unit change in x when the remaining Xs are held constant. These coefficients are also called partial-regression coefficients since the effect of the other $X^{'}$s has been removed.

From the obtained principal components, we will transform our results back to the X scale to obtain estimates of **B**.

These estimates will be biased, but the size of this bias is more than compensated for by the decrease in variance. That is, the mean squared error of these estimates is less than that for least squares. Mathematically, the estimation formula becomes:

$$\mathbf{A} = (\mathbf{Z}^{'}\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}^{'}\mathbf{Y} = \mathbf{D}^{-\mathbf{1}}\mathbf{Z}^{'}\mathbf{Y} \tag{3.1.15}$$

because of the special nature of principal components. This is simply the ordinary least squares regression applied to a different set of independent variables. The two sets of regression coefficients, A and B, are related using the formulas

$\mathbf{A} = \mathbf{P}^{'}\mathbf{B}$

and

$\mathbf{B} = \mathbf{P}\mathbf{A}$

Standard Error

These are the estimated standard errors (precision) of the PC regression coefficients. The standard error of the regression coefficient, s_{bj} , is the standard deviation of the estimate.

Standardized Regression Coefficients

These are the estimated values of the standardized regression coefficients. Standardized regression coefficients are the coefficients that would be obtained if you standardized each independent and dependent variable. Here standardizing is defined as subtracting the mean and dividing by the standard deviation of a variable. A regression analysis on these standardized variables would yield these standardized coefficients.

When there are vastly different units involved for the variables, this is a way of making comparisons between variables. The formula for the standardized regression coefficient is:

$$b_{jstd} = b_j = \left(\frac{S_y}{S_{x,j}}\right) \tag{3.1.16}$$

Where S_y and $S_{x,j}$ are the standard deviations for the dependent variable and the corresponding j^{th} independent variable.

Variance Inflation Factor (VIF)

These are the values of the variance inflation factors associated with the variables. When multicollinearity has been eliminated, all these values will be expected to be less than 10.

F-Ratio

This is the F statistic for testing the null hypothesis that all $\beta^{'}s = 0$. This F-statistic has p degrees of freedom for the numerator variance and $n-p-1$ degrees of freedom for the denominator variance. Since PC regression produces biased estimates, this F-Ratio is not a valid test.

Root Mean Square Error

This is the square root of the mean square error. It is an estimate of σ, the standard deviation of the $e^{'}$s.

R-Squared

This is the coefficient of determination. This statistic will explain how much variation in yield is accounted for by the main factors of cassava production. Moreover, this statistic will help in showing the model validity.

Model Accuracy Diagnostics

The model accuracy will be checked by the use of k-fold cross validation. In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k 1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged to produce a single estimation.

This technique will be used in splitting the cassava data set into k-subsets. Each subset will be held out while the model is trained on all the other subsets. This process will be completed until model accuracy is determined for each instance in the data set, and an overall accuracy estimate provided. In this research we chose to perform 10 fold cross-validation and therefore set the validation argument to CV, however there are other validation methods available.Nonetheless it is a robust method for estimating accuracy Xu and Liang (2001).

3.2 Properties Of The Principal Component Regression Model

The fitting process for obtaining the PCR estimator involves regressing the response vector on the derived data matrix W_k which has which has orthogonal columns for any $k \in \{1, \ldots, p\}$ since the principal components are mutually orthogonal to each other.Thus in the regression step, performing a multiple linear regression jointly on the k selected principal components as covariates is equivalent to carrying out k independent simple linear regressions (or, univariate regressions) separately on each of the k selected principal components as a covariate. When all the principal components are selected for regression so that $k = p$, then the PCR estimator is equivalent to the ordinary least squares estimator. Thus, $\widehat{\boldsymbol{\beta}}_p = \widehat{\boldsymbol{\beta}}_{\text{ols}}$. This is easily seen from the fact that $W_p = \mathbf{X}V_p = \mathbf{X}V$ and also observing that V is an orthogonal matrix. There are three vital properties on which the principal component regression model is anchored upon:

Variance Reduction

For any $k \in \{1, \ldots, p\}$, the variance of $\widehat{\boldsymbol{\beta}}_k$ is given by:

$\text{Var}(\widehat{\boldsymbol{\beta}}_k) = \sigma^2 \ V_k(W_k^T W_k)^{-1} V_k^T = \sigma^2 \ V_k \ \text{ diag}\left(\lambda_1^{-1}, \ldots, \lambda_k^{-1}\right) V_k^T = \sigma^2 \sum_{j=1}^{k} \frac{\mathbf{v}_j \mathbf{v}_j^T}{\lambda_j}$. In particular:

$\text{Var}(\widehat{\boldsymbol{\beta}}_p) = \text{Var}(\widehat{\boldsymbol{\beta}}_{\text{ols}}) = \sigma^2 \sum_{j=1}^{p} \frac{\mathbf{v}_j \mathbf{v}_j^T}{\lambda_j}$. Hence for all $k \in \{1, \cdots, p-1\}$ we have:

$\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{ols}}) - \text{Var}(\widehat{\boldsymbol{\beta}}_k) = \sigma^2 \sum_{j=k+1}^{p} \frac{\mathbf{v}_j \mathbf{v}_j^T}{\lambda_j}$.

Thus, for ll $k \in \{1, \cdots, p\}$ we have:

$\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{ols}}) - \text{Var}(\widehat{\boldsymbol{\beta}}_k) \succeq 0$

where $A \succeq 0$ indicates that a square symmetric matrix A is non-negative definite. Consequently, any given linear form of the PCR estimator has a lower variance compared to that of the same linear form of the ordinary least squares estimator.

Addressing Multicollinearity

In multicollinearity, two or more of the covariates are highly correlated, so that one can be linearly predicted from the others with a non-trivial degree of accuracy. As a consequence, the columns of the data matrix $\mathbf{X}$ that correspond to the observations for these covariates tend to become linearly dependent and therefore, $\mathbf{X}$ tends to become rank deficient losing its full column rank structure. More quantitatively, one or more of the smaller eigenvalues of $\mathbf{X}^T\mathbf{X}$ get(s) very close or, become(s) exactly equal to 0 under such situations. The variance expressions above indicate that these small eigenvalues produce the maximal inflation effect on the variance of the least squares estimator, thereby destabilizing (due to a high VIF) the estimator significantly when they are close to 0. This issue can be effectively addressed through using a PCR estimator obtained by excluding the principal components corresponding to these small eigenvalues.

Dimension reduction

PCR may also be used for performing dimension reduction. To see this, let L_k denote any $p \times k$ matrix having orthonormal columns, for any $k \in \{1, \ldots, p\}$. Suppose now that we want to approximate each of the covariate observations $\mathbf{x}_i$ through the linear transformation $L_k\mathbf{z}_i$ for some $\mathbf{z}_i \in \mathbb{R}^k (1 \leq i \leq n)$.

Then, it can be shown that

$\sum_{i=1}^{n} \left\|\mathbf{x}_i - L_k\mathbf{z}_i\right\|^2$

is minimized at $L_k = V_k$, the matrix with the first k principal component directions as columns, and $\mathbf{z}_i = \mathbf{x}_i^k = V_k^T\mathbf{x}_i$, the corresponding k dimensional derived covariates. Thus the k dimensional principal components provide the best linear approximation of rank k to the observed data matrix $\mathbf{X}$.

The corresponding reconstruction error is given by:

$$\sum_{i=1}^{n} \left\| \mathbf{x}_i - V_k \mathbf{x}_i^k \right\|^2 = \begin{cases} \sum_{j=k+1}^{n} \lambda_j & 1 \leqslant k < p \\ 0 & k = p \end{cases}$$

Thus any potential dimension reduction may be achieved by choosing k, the number of principal components to be used, through appropriate thresholding on the cumulative sum of the eigenvalues of $\mathbf{X}^T\mathbf{X}$. Since the smaller eigenvalues do not contribute significantly to the cumulative sum, the corresponding principal components may be continued to be dropped as long as the desired threshold limit is not exceeded. The same criteria may also be used for addressing the multicollinearity issue whereby the principal components corresponding to the smaller eigenvalues may be ignored as long as the threshold limit is maintained.

Chapter 4

EMPIRICAL STUDY

This chapter describes the analysis of data followed by a discussion of the research findings. The findings relate to the research questions that guided the research project. Data were analyzed to identify, describe and explore the relationship between yield and different factors of cassava production in cassava breeding sites, relationships between the factors of production and to reduce dimensionality in the cassava data matrix.

4.1 Source Of Data

This study was focused on modeling the yield of cassava in Western Kenya at any time of the year based on the environmental, genetic, diseases and pest variables. The study was centered upon data collected at various KALRO cassava breeding sites in Kenya. The sites included Busia, Kakamega and Kitale which are in the Western region of Kenya that takes the bulk of cassava production Karuri et al. (2001).

Variable Description Data was collected on data templates from 10 plots in each of the 3 replications in each site leading to 180 cases (n=180) of data as per variables:

Table 4.1: Variable description

Key	Description
SITE	location where the trial was planted
REP	Replications
ENTRY	Variety (genotype) name code
SAH	Plant population in the plot at harvest
BHT	Height to first branch in cm
PHT	Plant height in cm
NTOTAL	Total number of storage roots harvested
WTOTAL	Total weight of storage roots harvested (kg)
YLD	Yield (t/ha)
CYN	Cyanide content of the storage roots on a scores scale of 1-9
RDM	Root dry matter content (%)
CADS	Cassava anthracnose disease severity (score scale of 1-5)
CBBS	Cassava bacterial blight disease severity (score scale of 1-5)
CBSDS	Cassava brown streak disease severity (score scale of 1-5)
CMVS	Cassava mosaic virus disease severity (score scale of 1-5)
CGMS	Cassava green mites severity (score scale of 1-5)
CMBS	Cassava mealy bugs severity (score scale of 1-5)

The data collection was accomplished by research assistants who worked so closely with breeding technologists and the senior cassava breeder at the breeding sites. However, not all plots had data on all variables under study and therefore complete responses were from 176 plots, that is a response rate of 98%.

4.2 Analysis Methods

Multidimensional scaling methods and quantitative data transformation techniques were used to transform the data collected on all variables under the study. Roystons multivariate normality (MVN) test was performed to test normality before the principal component analysis procedure and a chi-square Q-Q plot produced. Inferential statistical analysis was used to identify relationships among the dependent and independent variables. The statistical significance of multiple linear regression model for selected variables and principal components was determined using the Fishers F-test while significance for correlation among variables was based upon the Pearson correlation coefficient r. Multicolinearity testing for multiple linear regression model (MLR) and the principal component regression model (PCR) was based on the variance inflation factor. Significance for individual variables in regression models was based upon Student t-test. The level of significance was set at 0.05. The analyses were done in statistical R package version 3.3.1.

4.3 Empirical Study

4.3.1 Correlation Analysis Among Independent Variables

Table 4.2: Correlation analysis among independent variables

FACTOR	SITE	REP	ENTRY	SAH	BHT	PHT	NTOTAL	WTOTAL	RDM	CYN	CADS	CBBS	CBSDS	CMVS	CGMS	CMBS
SITE	1.000															
REP	0.002	1.000														
	0.980															
ENTRY	0.040	0.004	1.000													
	0.602	0.959														
SAH	-0.107	0.057	0.108	1.000												
	0.158	0.450	0.152													
BHT	-0.009	-0.027	0.180*	0.024	1.000											
	0.911		0.017	0.749												
PHT	0.047	0.024	0.058	-0.063	0.222**	1.000										
	0.532	0.753	0.445	0.409	0.003											
NTOTAL	0.136	-0.028	0.107	-0.002	0.031	0.329**	1.000									
	0.072	0.713	0.158	0.976	0.680	0.000										
WTOTAL	0.105	0.030	0.015	-0.091	-0.076	0.297**	0.470**	1.000								
	0.164	0.690	0.847	0.229	0.314	0.000	0.000									
RDM	0.155*	-0.159*	0.065	-0.022	0.032	0.061	0.028	0.048	1.000							
	0.040	0.035	0.390	0.773	0.677	0.418	0.716	0.523								
CYN	-0.077	-0.037	0.047	-0.106	0.056	-0.140	-0.169*	-0.117	0.006	1.000						
	0.311	0.624	0.540	0.160	0.463	0.063	0.025	0.122	0.936							
CADS	-0.100	0.041	-0.022	0.090	0.150*	0.009	-0.162*	-0.146	0.016	-0.065	1.000					
	0.189	0.589	0.768	0.235	0.048	0.904	0.032	0.054	0.837	0.391						
CBBS	-0.285**	-0.060	-0.014	-0.143	0.073	0.296**	0.110	0.013	0.055	0.124	-0.074	1.000				
	0.000	0.428	0.851	0.059	0.334	0.000	0.147	0.864	0.472	0.101	0.329					
CBSDS	0.063	-0.054	-0.098	-0.081	0.012	0.001	-0.010	-0.040	0.016	0.091	0.122	-0.019	1.000			
	0.405	0.479	0.194	0.284	0.875	0.989	0.895	0.595	0.835	0.230	0.107	0.803				
CMVS	-0.124	-0.115	0.240**	0.083	0.015	-0.199**	-0.066	-0.122	0.011	0.075	0.017	0.048	-0.012	1.000		
	0.101	0.129	0.001	0.271	0.842	0.008	0.388	0.105	0.887	0.319	0.821	0.530	0.874			
CGMS	0.034	-0.111	0.134	0.108	-0.022	0.051	0.239**	0.153*	-0.001	-0.031	-0.089	-0.138	-0.014	-0.088	1.000	
	0.652	0.144	0.076	0.153	0.768	0.506	0.001	0.043	0.990	0.681	0.239	0.067	0.855	0.246		
CMBS	0.065	0.092	0.016	-0.039	0.002	-0.126	-0.124	-0.116	0.004	0.046	-0.021	-0.033	-0.003	-0.021	-0.024	1.000
	0.391	0.224	0.838	0.605	0.978	0.096	0.102	0.126	0.959	0.548	0.780	0.665	0.965	0.783	0.751	

From the table 4.2 above there is significant positive correlation between SITE and RDM, significant negative correlation between SITE and CBBS , significant negative correlation between REP and RDM , significant positive correlation between ENTRY and BHT, significant positive correlation between ENTRY and CMVS, significant positive correlation between BHT and PHT, significant positive correlation between BHT and CADS, significant positive correlation between PHT and N_TOTAL, significant positive correlation between PHT and W_TOTAL, significant positive correlation between PHT and CBBS , significant negative correlation between PHT and CBBS, significant positive correlation between N_TOTAL and W_TOTAL, significant negative correlation between N_TOTAL and CYN, significant negative correlation between N_TOTAL and CADS, significant positive correlation between N_TOTAL and CGMS, significant positive correlation between W_TOTAL and CGMS. Thus it is evident that despite YLD being dependent on the different factors of production, there is correlation among the factors of production themselves. This therefore justifies the need for dimensionality reduction and obtaining a smaller set of independent variables that can best model yield.

Table 4.3: Establishing relationship among the independent variables using multiple linear regression (MLR) statistics and variance inflation factor.

IndepVar	Coeff	Std Error	P-value	VIF
SITE	1.009	0.002	0.0458*	6.541
REP	1.012	0.004	0.166	7.659
ENTRY	0.999	0.001	0.586	5.421
SAH	1.04	0.001	0.000**	214.549
BHT	1	0	0.93	19.931
PHT	1	0	0.118	37.693
NTOTAL	1	0	0.798	11.815
WTOTAL	1.018	0	0.000**	22.305
RDM	1.005	0.001	0.001	46.038
CYN	1.03	0.004	0.0039**	17.156
CADS	0.983	0.009	0.43	13.763
CBBS	1.029	0.009	0.169	14.755
CBSDS	1.198	0.022	0.0004**	58.949
CMVS	1.032	0.008	0.103	11.514
CGMS	1.013	0.01	0.551	14.81
CMBS	1.479	0.034	0.000**	145.737

[*] N/B: Tolerance= (1/VIF) while * and ** indicate significance at 0.05 and 0.01 respectively. F-value=16200 with 160 degrees of freedom

From table 4.3, most of the factors regressed against the rest returned a VIF of above 10 implying multicollinearity. For instance, regressing SAH on the rest of the factors returned a VIF (1/1-R^2) of 214.549, regressing BHT on the rest of the factors had a VIF of 19.931, regressing PHT on the rest of the factors returned a VIF of 37.693, regressing W_TOTAL on the remaining factors of production gave a VIF of 22.305, RDM regressed on the other factors returned a VIF of 46.038, CYN regressed on the rest of the factors of production returned a VIF of 17.156, CBSDS regressed on the other factors of production returned a VIF of 58.949 and CGMS and CMBS regressed on the rest of the factors gave a VIF of 14.810 and 145.737 respectively. Moreover, most of the factors had higher values of standard error and this adds to the evidence of existence of multicollinearity. The overall model returned, F=16200 (DF=160), R^2=0.9994 and VIF of 1666.6667. This high value of VIF indicates presence of multicollinearity in the overall model for predicting cassava yield when all the factors of production are included in the model. Therefore coefficients derived from this model will be unstable and therefore results for yield prediction would be unreliable and invalid. This justifies dimension reduction through principal component analysis.

4.3.2 Principal Component Analysis For Dimension Reduction And Selection Of Key Components

4.3.2.1 Testing For Normality

Before the PCA procedure, the data was tested for multivariate normality. The table below has the statistics from this procedure.

Table 4.4: Testing for normality

H-value	P-value
957.043	p<0.0001

The p-value 0.000<0.05 indicates the results are significant, implying absence of multivariate normality. It is worth noting that the MVN test is very sensitive when the sample is so large (n>30), therefore we will still go ahead with PCA despite the absence of normality and assume normality due to sample size.

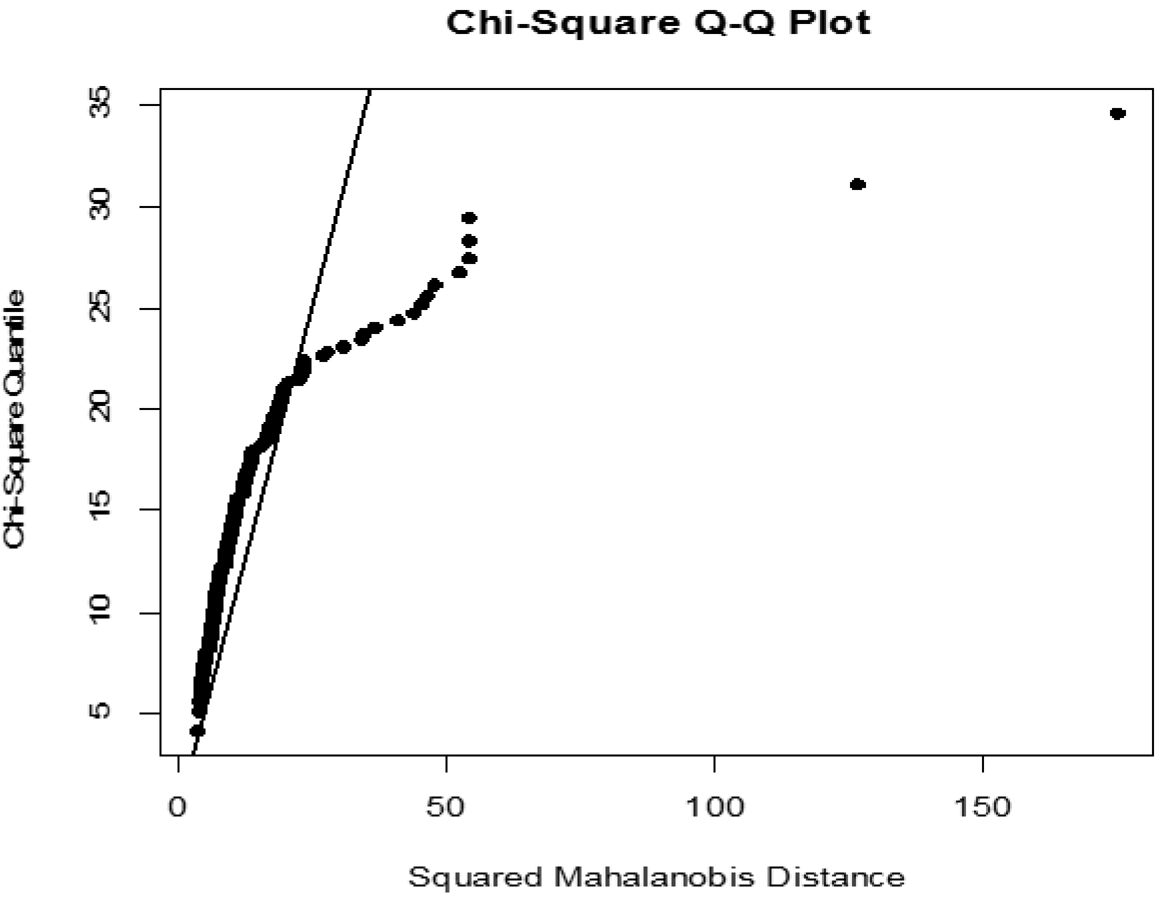

Figure 4.1: Q-Q plot for the multivariate normal test

4.3.2.2 Eigen Values, Proportion Of Variance Explained By Principal Components And Loadings.

The table below has the principal components from the PCA procedure.

Table 4.5: Eigen values and proportion of variance explained by principal components

Principal Component	Standard Deviation	Prop of Variance	Cumulative Variance	Eigen Value
1	1.466	0.135	0.135	2.149
2	1.270	0.101	0.237	1.614
3	1.202	0.091	0.327	1.445
4	1.145	0.082	0.410	1.312
5	1.106	0.077	0.487	1.224
6	1.081	0.073	0.560	1.168
7	1.019	0.065	0.626	1.039
8	1.002	0.063	0.689	1.005
9	0.898	0.051	0.739	0.807
10	0.882	0.049	0.788	0.779
11	0.852	0.046	0.834	0.726
12	0.803	0.041	0.874	0.645
13	0.782	0.038	0.913	0.611
14	0.747	0.035	0.948	0.558
15	0.702	0.031	0.979	0.493
16	0.578	0.021	1.000	0.334

The total number of principal components returned was 16, equal to the total number of variables used in the principal component procedure. The principal components are arranged in order of size from the largest to the smallest, with the largest principal component contributing the largest proportion of variance and the smallest principal component contributing the smallest proportion of variance. The total variance explained by the components is the sum of the variances of the components which is unity (1). Using the amount of explained variance criterion (70%-80%), we select the first eight principal components from the table above which account for almost 70% of total variance. This is affirmed by the eigenvalue one rule in which we select the eigenvalues that are above value 1.

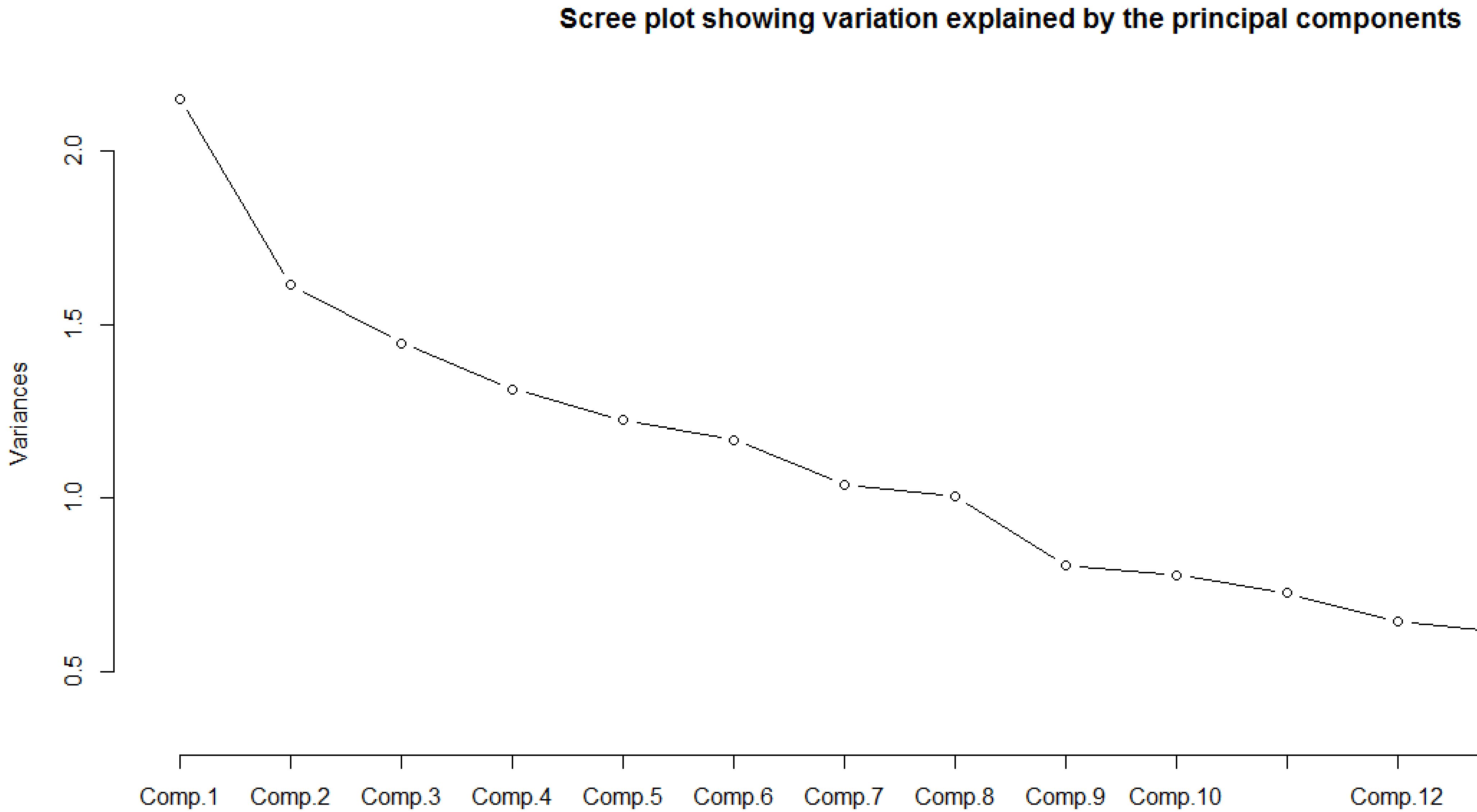

Figure 4.2: Scree plot for principal component importance

From the scree plot, there's a sharp decline in variance around PC 8. This indicates a a sharp reduction in the importance of the principal components. The components that follow from this point contribute very little to the overall variance.

Table 4.6: Loadings and importance of variables

Var/Comp	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
SITE	0.24	0.22	0.51	0.187	-0.153			
REP		0.23	−0.22		-0.366	0.518		-0.207
ENTRY		−0.3		0.554	0.13	0.102	-0.179	-0.352
SAH	−0.11		−0.26	0.519	-		0.221	0.269
BHT		−0.49		0.22	-0.375			
PHT	0.45	−0.24	−0.22		-0.24	-0.134		
NTOTAL	0.52		−0.1	0.107				-0.118
WTOTAL	0.49				0.198	0.256	0.111	-
RDM	0.16	−0.14	0.23		0.203	-0.14	-0.517	0.509
CYN	−0.19	−0.31	0.25	-0.263	-	0.259	0.187	-0.182
CADS	0.17	0.16	0.21	-0.138	0.281	0.57		
CBBS		0.41	0.3	0.417	-0.135		0.236	
CBSDS		−0.16	0.51		-	-0.205	0.31	-0.278
CMVS	0.23	0.19		-0.232	-0.563	-0.111	-	0.206
CGMS	0.12	0.36	−0.13	0	0.18	-0.318	-0.217	-0.539
CMBS	−0.17		0.2		-0.289	0.21	-0.606	-0.145

* N/B: The sum of squares of values in the loading column equal to unity. This implies that for uniformity each variable has to load/contribute 0.25 to each principal component. A value>0.25 shows the importance of the variable on that principal component.

From the table 4.6 above, PHT, N_TOTAL and W_TOTAL load highly on principal component 1 and among these three, only W_TOTAL loads on another component (comp. 6). BHT, CYN, CBBS and CMBS load highly on principal component 2. SITE, SAH, CYN, CBBS and CBSDS all contribute strongly on component 3. SAH, CYN and CBBS load highly on component 4. REP, BHT, CADS, CMVS and CMBS load highly on component 5 while REP, W_TOTAL, CYN, CADS and CMBS have high loadings on component 6. RDM, CBSDS and CMBS load highly on component 7 whereas SAH, RDM, CBSDS and CGMS load highly on component 8. Overall, CYN (cyanide content) loads highly on most components (4). Other factors that load highly on more components as compared to others include ENTRY (Genotype), SAH (plant population in the plot), CBBS (Cassava Bacterial Blight Disease Severity) and CBSDS (Cassava brown streak disease severity) which all load highly in 3 of the components. These high loadings in more than one component indicate the importance of these factors in yield prediction in cassava yield prediction. It is also worth mentioning that even though RDM (Root Dry Matter content) loads highly in only two components, these two loadings, -0.517 and 0.509 respectively, are very high as compared to the expected average loading of 0.25. These fewer set of important factors in cassava production could not be arrived at easily if not for the PCA procedure. Since the principal components were obtained on standardized data for the factors, the columns of the loadings are the eigenvectors. Therefore the transpose of this matrix generates the required eigenvector matrix. These eigenvectors will be vital together with the coefficients of principal component regression in the generation of MLR coefficients through the transformation stage. To generate the new scores (projection of the initial data on the new plane of PCs), we simply multiply the loadings and the standardized X variables. This projection of the original 16 variables data points in a 16 dimensional plane by the 8 best fitting planes as opposed to the original 16 correlated lines is the major advantage of the PCA procedure. This enables easy visualization of the data. This is in consistent with Pearson (1901) results in which he observed that in many physical, statistical and biological investigations it is desirable to represent a system of points in a plane, three or higher (p) dimensional space by the best fitting straight line or plane. This is also in consistent with Raychaudhuri et al. (2000) who used the PCA to analyze DNA microarray data set and they observed that PCA can find a reduced set of variables that are useful for understanding the experiments and aiding data visualization.The following table gives an excerpt of the scores produced on the first 8 principal components.

Table 4.7: Excerpt of principal component scores for the first five observations on each of the first 8 PCs

No.	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
1	-4.199	-2.035	-2.756	1.741	-2.58	-6.354	0.105	-0.501
3	-1.271	0.217	-0.543	-0.423	0.734	-0.357	0.823	1.216
4	-0.283	-0.969	-0.33	1.382	1.838	-0.021	0.018	-0.035
5	-3.575	0.102	0.468	-0.122	2.328	-0.4	-0.037	0.384

4.3.3 Principal Component Regression And PCR Statistics

Fitting a principal component regression model for Yield on the 8 principal components produced the following PCR statistics.

Table 4.8: Principal Component Regression Statistics

Component	Coeff	Std Error	P-Value
Comp.1	3.494	0.24	0.000**
Comp.2	0.524	0.277	0.06
Comp.3	-0.018	0.292	0.95
Comp.4	-0.281	0.307	0.36
Comp.5	1.216	0.318	0.000**
Comp.6	1.63	0.325	0.000**
Comp.7	0.783	0.345	0.024**
Comp.8	0.354	0.351	0.314

[*] N/B: ** indicate significance at 0.05 and 0.01 respectively. F=32.850 with 168 degrees of freedom.

From the above table, the principal component regression equation is given by:

$$YLD = 3.494Comp.1 + 0.524Comp.2 - 0.018Comp.3 - 0.281Comp.4+$$

$$1.216Comp.5 + 1.630Comp.6 + 0.783Comp.7 + 0.354Comp.8.$$

The model has an F= 32.85 with a p-value=0.000 (DF=168). This implies the model consisting of the first 8 PCs is significant in prediction of yield. The models R^2=0.610 (Adj R^2=0.590) with the VIF attributed to the overall model being 2.564. The multicollinearity therefore is no longer a problem. Transforming these PCR statistics using the eigenvectors attributed to the eigenvalues of the standardized X variables results in the MLR coefficients for the cassava data results in the following PCR statistics.

Table 4.9: Transformed Multiple Linear Regression Coefficients

variable	Linear Reg Coefficient
SITE	2.197
REP	-0.091
ENTRY	1.288
SAH	0.848
BHT	-1.506
PHT	0.142
NTOTAL	-0.02
WTOTAL	-0.27
RDM	0.171
CYN	0.991
CADS	-0.187
CBBS	-1.611
CBSDS	1.608
CMVS	0.71
CGMS	0.405
CMBS	0.665

Transforming to a multiple linear regression model yields the above coefficients. The MLR equation therefore becomes:

$$YLD = \beta_0 + 2.197SITE - 0.091REP + 1.288ENTRY + 0.848SAH - 1.506BHT + 0.142PHT -$$

$$0.020NTOTAL - 0.270WTOTAL + 0.171RDM + 0.991CYN - 0.187CADS -$$

$$1.611CBBS + 1.608CBSDS + 0710CMVS + 0.405CGMS + 0.665CMBS$$

In this equation there are many regression coefficients and some are close to zero and this increases the likelihood of some being insignificant and unstable and hence leading to unreliable prediction: therefore the above 8 principal components make it easier, convenient and efficient for forecasting yield. This stability achievement is another major aspect that this research project found out and it is recommended that more research be done in this area to add on existing inconclusive findings.

4.3.4 Model accuracy assessment and forecasting

The model accuracy and forecasting results were as below:

Table 4.10: Model validation results

PC	Crossvalidation	variance-explained-X	ForecastYLD
1	0.94	13.51	49.31
2	0.94	23.65	50.14
3	0.94	32.74	50.14
4	0.958	40.98	50.34
5	0.93	48.68	53.74
6	0.906	56.02	59.57
7	0.879	62.55	60.77
8	0.888	68.87	61
9	0.887	73.94	63.49
10	0.856	78.83	66.37
11	0.811	83.39	72.9
12	0.699	87.45	81.43
13	0.633	91.29	82.44
14	0.645	94.8	83.64
15	0.62	97.9	86.46
16	0.675	100	97.72

In the analysis results in the table above,8 principal components were enough to explain nearly 70% of the variability in the data although the CV score is a little higher than with more than 8 components. Finally,we note that the 16 components explain total variability as expected.

Chapter 5

SUMMARY OF FINDINGS,CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

In this chapter, summary of key findings, conclusions and recommendations are made.

5.2 Summary Of Key Findings

The major objective of this study was to develop a model for predicting cassava yield using the PCR model integrated with optimal scaling. The model is to be used for prediction of response variable yield using few principals. Optimal scaling was used in transformation of categorical variables while PCA was used in dimension reduction. Preliminary analyses on all the factors of yield indicated a high amount of correlation among the factors of production, with most of the bivariate combinations resulting in a $p<0.05$ as shown in table 4.2. Moreover, regressing yield on all factors of production showed that most of the co-efficients were statistically insignificant, $p>0.05$. Regressing each factor of production on the remaining k-1 factors of production resulted in VIF>10 with some factors resulting to as high as VIF>100 as evidenced in table 4.3. This indicated existence of multicollinearity. The PCA technique applied in the analysis had the shrinkage capability on the data set dimension, from 16 variables to 8 principal components that best model the cassava yield. Nonetheless, the variance inflation factor for the full model at 1666.667 reduced to 2.565<10 , therefore providing a more stable and reliable model. However the variability explained by the PCR model dropped to 61% from 99% as expected, however the multicolliarity problem had been solved. Model validation indicated a high validation error when one component was used for forecasting, explaining only 13.51% of the variation in yield but the accuracy of the model optimized at PCs<=8 with the PCR regression co-efficients being statistically significant, p<0.05 and increasing model reliability for prediction.

5.3 Conclusions

From the foregoing summarized results, the following are conclusions that can be drawn. From the results of objective 1, factors of cassava production are correlated and therefore requires dimension reduction before being used in yield prediction models.

From the results of objective 2, eight key principals are sufficient in predicting cassava yield and offer optimal and accurate results. From the results of objective 3, the model for cassava prediction is;

$$YLD = 3.494Comp.1 + 0.524Comp.2 - 0.018Comp.3 - 0.281Comp.4+$$

$$1.216Comp.5 + 1.630Comp.6 + 0.783Comp.7 + 0.354Comp.8.$$

This model therefore not only offers an alternative to existing models but also an efficient solution when the number of factors of production is high.

5.4 Recommendations For Further Studies

These empirical results show model stability achievement, nonetheless more models that incorporate pests and disease data that is optimally scaled should be developed and their stability, validity and reliability assessed. These models could go a long way in informing policy making in the crop breeding field.

More research should be done in forecasting using the principal component regression in quadratic or higher polynomial in order to assess model validity, reliability and stability as this work solely focused on linear combinations of the principal components.

The government and other key stakeholders should allocate resources and support researchers attempting to develop incorporated statistical models for cassava and other crops prediction.

REFERENCES

D. M. Aberi. *Innovative post-harvest drying technology for small-scale production of quality starch from kenyan cassava cultivars.* PhD thesis, M. Sc Thesis, University of Nairobi, 2012.

R. Agrawal and R. Jain. Forecast of sugarcane yield using eye estimate alongwith plant characters. *Biometrical journal*, 38(6):731–739, 1996.

R. Agrawal, R. Jain, M. Jha, and D. Singh. Forecasting of rice yield using climatic variables [india]. *Indian Journal of Agricultural Sciences*, 1980.

R. Agrawal, R. Jain, and M. Jha. Joint effects of weather variables on rice yield. *Mausam*, 1983.

R. Agrawal, R. Jain, and M. Jha. Models for studying rice crop-weather relationship. *Mausam*, 37(1):67–70, 1986.

R. Agrawal, R. Jain, and S. Mehta. Yield forecast based on weather variables and agricultural inputs on agro-climatic zone basis. *The Indian Journal of Agricultural Sciences*, 71(7), 2001.

R. Agrawal, S. Mehta, A. Kumar, and L. Bhar. Development of weather based forewarning system for crop pests and diseases. *Project Report, IASRI, New Delhi Mission mode project under NATP, PI, Dr. YS Ramakrishna, CRIDA, Hyderabad*, 2004.

A. Amadei, A. Linssen, and H. J. Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics*, 17(4):412–425, 1993.

A. Amadei, M. A. Ceruso, and A. Di Nola. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, 36(4):419–424, 1999.

P. K. Anderson. Whitefly and whitefly-borne viruses in the tropics: Building a knowledge base for global action. introduction. 2005.

F. Aryee, I. Oduro, W. Ellis, and J. Afuakwa. The physicochemical properties of flour samples from the roots of 31 varieties of cassava. *Food control*, 17(11):916–922, 2006.

W. Baier. *Crop-weather models and their use in yield assessments.* 1977.

I. R. M. Benesi. Characterisation of malawian cassava germplasm for diversity, starch extraction and its native and modified properties. 2005.

J. Cadima and I. T. Jolliffe. Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics*, 22(2):203–214, 1995.

CFSAM. Service validation report. Technical report, Internal Report, 2006.

T. R. Chandrahas and T. Rai. Pilot study for developing bayesian probability forecast model based on farmers appraisal data on wheat crop. *IASRI, New Delhi publication*, 2001.

W. W. Cooley and P. R. Lohnes. *Multivariate data analysis.* J. Wiley, 1971.

O. Coulibaly, V. Manyong, S. Yaninek, R. Hanna, P. Sanginga, D. Endamana, A. Adesina, M. Toko, and P. Neuenschwander. Economic impact assessment of classical biological control of cassava green mite in west africa. *IITA, Cotonou, Benin Republic*, 2004.

E. De Wolf and L. Francl. Neural network classification of tan spot and stagonospora blotch infection periods in a wheat field environment. *Phytopathology*, 90(2):108–113, 2000.

E. De Wolf and L. Franel. Neural networks that distinguish infection periods of wheat tan spot in an outdoor environment. *Phytopathology*, 87(1):83–87, 1997.

FAO. Cassava production statistics, November 2007. URL `http://www.fao.org.2007/`.

F. FAOstat. agriculture organization of the united nations. *Statistical database*, 2009.

C. M. Fauquet, M. A. Mayo, J. Maniloff, U. Desselberger, and L. A. Ball. *Virus taxonomy: VIIIth report of the International Committee on Taxonomy of Viruses.* Academic Press, 2005.

A. M. Fermont et al. *Cassava and soil fertility in intensifying smallholder farming systems of East Africa.* publisher not identified, 2009.

R. A. Fisher. The influence of rainfall on the yield of wheat at rothamsted. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:89–142, 1925.

V. Fondong, J. Thresh, and C. Fauquet. Field experiments in cameroon on cassava mosaic virus disease and the reversion phenomenon in susceptible and resistant cassava cultivars. *International Journal of Pest Management*, 46(3):211–217, 2000.

M. Frere and G. Popov. *Agrometeorological crop monitoring and forecasting.* FAO, 1979.

C. Githunguri, S. Mwiti, Y. Migwa, et al. Cyanogenic potentials of early bulking cassava planted at katumani, a semi-arid area of eastern kenya. In *African Crop Science Conference Proceedings*, volume 8, pages 925–927, 2007.

J. W. Hansen, A. Mishra, K. Rao, M. Indeje, and R. K. Ngugi. Potential value of gcm-based seasonal rainfall forecasts for maize management in semi-arid kenya. *Agricultural Systems*, 101(1):80–90, 2009.

W. Hendrick and J. Scholl. Technique in measuring joint relationship. the joint effects of temperature and precipitation on crop yield. *N. Carolina Agric. Exp. Stat. Tech. Bull*, 74, 1943.

B. Hess. Convergence of sampling in protein simulations. *Physical Review E*, 65(3): 031910, 2002.

R. Hillocks and D. Jennings. Cassava brown streak disease: a review of present knowledge and research needs. *International Journal of Pest Management*, 49(3):225–234, 2003.

R. J. Hillocks, J. Thresh, and A. Bellotti. *Cassava: biology, production and utilization.* CABI, 2002.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

C. C. House. within-year growth model approach to forecasting corn yields. 1977.

A. Hyvarinen, J. Karhunen, and E. Oja. Independent components analysis, 2001.

W. G. Jacoby. Levels of measurement and political research: An optimistic view. *American Journal of Political Science*, pages 271–301, 1999.

R. Jain and R. Agrawal. Probability model for crop yield forecasting. *Biometrical journal*, 34(4):501–511, 1992.

R. Jain and V. Ramasubramanian. Forecasting of crop yields using second order markov chains. *Journal of the Ind. Soc. of Agril. Stats*, 51(1):61–72, 1998.

R. Jain, H. Sridharan, and R. Agrawal. Principal component technique for forecasting sorghum yield. *Indian journal of agricultural sciences*, 1984.

R. Jain, R. Agrawal, and M. Jha. Use of growth indices in yield forecast. *Biometrical journal*, 27(4):435–439, 1985.

A. Jamak, A. Savatić, and M. Can. Principal component analysis for authorship attribution. *Business Systems Research*, 3(2):49–56, 2012.

M. Jha, R. Jain, and D. Singh. Pre-harvest forecasting of sugarcane yield. *Indian Journal of Agricultural Sciences (India)*, 1981.

D. A. Johnson, J. R. Alldredge, and D. L. Vakoch. Potato late blight forecasting models for the semiarid environment of south-central washington. *Phytopathology*, 86(5):480–484, 1996.

I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

E. E. Karuri, S. K. Mbugua, J. Karugia, J. Wanda, and J. Jagwe. Marketing opportunities for cassava based products: An assessment of the industrial potential in kenya. *University of Nairobi, Department of Food science, technology and nutrition food net/international institute of tropical agriculture*, 2001.

J. Kiura, C. Mutegi, P. Kibet, and M. Danda. Cassava production, utilisation and marketing in coastal kenya. a report of a survey on cassava enterprise conducted between july and october 2003 in kwale, kilifi, mombasa and malindi districts. Technical report, Internal Report, 2005.

A. D. Lanterman. Bayesian inference of thermodynamic state incorporating schwarz-rissanen complexity for infrared target recognition. *Optical Engineering*, 39(5):1282–1292, 2000.

G. A. Larsen. Forecasting 1977 kansas wheat growth. 1978.

J. Legg, F. Ndjelassili, and G. Okao-Okuja. First report of cassava mosaic disease and cassava mosaic geminiviruses in gabon. *Plant Pathology*, 53(2):232–232, 2004.

J. S. Lohse, P. Giordano, M. C. Williams, F. A. Vogel, and P. Law. Illinois agricultural land productivity formula. *Public Law*, 95:87, 1985.

P. Mair, K. Hornik, and J. de Leeuw. Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32(5):1–24, 2009.

M. Maruthi, S. Seal, J. Colvin, R. Briddon, and S. Bull. East african cassava mosaic zanzibar virus–a recombinant begomovirus species with a mild phenotype. *Archives of virology*, 149(12):2365–2377, 2004.

J. Matis, T. Saito, W. Grant, W. Iwig, and J. Ritchie. A markov chain approach to crop yield forecasting. *Agricultural systems*, 18(3):171–187, 1985.

S. Mehta, R. Agrawal, and V. Singh. Strategies for composite forecast. *J. Ind. Soc. Agril. Statist*, 53(3):262–272, 2000.

G. Michailidis and J. de Leeuw. The gifi system of descriptive multivariate analysis. *Statistical Science*, pages 307–336, 1998.

J. Nealon. development of within-year forecasting models for spring wheat (other than durum). 1976.

N. Patel, M. Chakraborty, S. Dutta, C. Patnaik, J. Parihar, S. Moharana, A. Das, B. Sarangi, and G. Behera. Multiple forecasts of kharif rice in orissa state-four year experience of fasal pilot study. *Journal of the Indian Society of Remote Sensing*, 32 (2):125–143, 2004.

K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901.

T. Rai. Use of discriminant function of weather parameters for developing forecast model for rice crop. *IASRI Publication*, 2000.

Y. Ramakrishna, H. Singh, and G. N. Rao. Weather based indices for forecasting foodgrain production in india. *Jounral of Agrometeorology*, 5(1):1–11, 2003.

V. Ramasubramanian and R. Jain. Use of growth indices in markov chain models for crop yield forecasting. *Biometrical journal*, 41(1):99–109, 1999.

V. Ramasubramanian, R. Agrawal, and L. M. Bhar. Forecasting sugarcane yield using multiple markov chains. *Project Report, IASRI, New Delhi*, 2004.

S. Raychaudhuri, J. M. Stuart, and R. B. Altman. Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 455. NIH Public Access, 2000.

A. C. Rencher. Méthods of multivariate analysis. a john wiley & sons. *Inc. Publication*, 2002.

M. Renkow and D. Byerlee. The impacts of cgiar research: A review of recent evidence. *Food policy*, 35(5):391–402, 2010.

A. Saksena, R. Jain, and R. Yadav. Development of early warning and yield assessment model for rainfed crops based on agrometeorological indices. *IASRI publication*, 2001.

M. Sardana, R. Khosla, N. Ohri, and P. Mitra. Preharvest forecasting of yield rate of jute. *Jute Bull*, 1972.

J. Sarkar. Forecasting rice and wheat yield over different meterological sub-divisions of india using statistical models. *Indian Society of Agricultural Statistics (India)*, 2003.

G. Sarwade. Meteorological yield models. In *Workshop on Crop Yield Modelling*, pages 1–8. Space Applications Centre India, 1988.

B. Singh and S. Bapat. Pre-harvest forecast models for prediction of sugarcane yield. *Indian Journal of Agricultural Sciences (India)*, 1988.

D. Singh, H. Singh, and P. Singh. Pre-harvest forecasting of wheat yield. *INDIAN JOURNAL OF AGRICULTURAL SCIENCES*, 46(10):445–450, 1976.

K. Singh. Preharvest forecast of yield–an overview. *IASRI publication*, 2010.

R. Singh, K. Singh, A. Kumar, and H. Chandra. Crop yield estimation and forecasting using remote sensing. *Forecasting Techniques in Agriculture*, pages 201–215, 2012.

A. Space Application Centre. Review document of rsam project : Crop acreage and production estimation. Technical report, Internal Report, 1990.

J. Stevens. Applied multivariate statistics for the social sciences.. mahwah, nj: Lawrence erlbaurn associates, 2002.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

S. R. Waddington, X. Li, J. Dixon, G. Hyman, and M. C. De Vicente. Getting the focus right: production constraints for six major food crops in asian and african farming systems. *Food security*, 2(1):27–48, 2010.

A. Walker, P. Mutuo, M. van Noordwijk, A. Albrecht, and G. Cadisch. Modelling of planted legume fallows in western kenya using wanulcas.(i) model calibration and validation. *Agroforestry systems*, 70(3):197–209, 2007.

Y. Wang and L. H. Staib. Boundary finding with prior shape and smoothness models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):738–743, 2000.

Q.-S. Xu and Y.-Z. Liang. Monte carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1):1–11, 2001.

F. W. Young. Quantitative analysis of qualitative data. *Psychometrika*, 46(4):357–388, 1981.

W. Zhu, L. Porth, and K. S. Tan. Improving crop yields forecasting using weather data: A comprehensive approach combining principal component analysis and credibility model. *IARFIC publication*, 2012.