Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.7, No.5, 2017

Modeling Cassava Yield In Western Kenya: Optimal Scaling Integrated With Principal Component Regression Approach

Vincent Alulu Harry[1*], George Orwa[1], Henry Athiany[1]
Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya
E-mail: harryalulu@gmail.com

Abstract

Cassava is a major food crop grown in the tropical and subtropical parts of the world. In this research work, we sought to develop a model for predicting cassava yield using the PCR model integrated with optimal scaling. Moreover, establishing relationship between the different factors of production, estimate the yield based on the key components adduced to the factors of production in trial data in Western region, Kenya. Principal component analysis and optimal scaling were 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, VIF 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%), we selected the first eight principal components which accounted for almost 70% of total model variance. Eight (8) key components were 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 indicated non-correlation among the eight principal components with the VIF attributed to the overall PCR model being 2.564, (R^2=0.610 (Adj R^2=0.590). The model offers an efficient alternative to existing models for crop yield prediction when the number of factors to be included in the model is high.

Keywords: PCR, PCA, VIF

Introduction

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. Although socioeconomic factors, market conditions and abiotic constraints negatively affect cassava yield, 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. Therefore, all these factors need to be considered while coming up with a model for yield prediction.

Fisher (1925) suggested a linear regression technique which requires small number of parameters to be estimated while taking care of distribution pattern of weather over the crop season. Models using spectral data have also been used in crop prediction. In 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).

Integrated models using data on plant characters along with agricultural inputs were found to be better than models based on plant characters alone in jowar and apple Jain et al. (1985). However there has been insuffiency in efficient models that incorporate all factors of production of cassava.

The objective of this study was to develop a model for predicting cassava yield using the PCR model integrated with optimal scaling.

Materials and methods

Data were obtained from six cassava breeding sites in Western Kenya namely Alupe, Kenya Agriculture and Livestock Organization-Kakamega, Kenya Agriculture and Livestock Organization-Kibos, Oyani, Sangalo and Siaya for the year 2016. Data was collected from 10 plots in each of the 3 replications in each site leading to 180 cases (n=180) of data. Complete responses were from 176 plots, that is a response rate of 98\%. The varibles collected were SITE (location where the trial was planted), REP (Replications), ENTRY (genotype), 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 in kg), YLD (Yield in ton/ha), CYN (Cyanide content of the storage roots on a scores scale of 1-9), RDM (Root dry matter content in %), 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) and CMBS (Cassava mealy bugs severity score, scale of 1-5) .

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

Before the PCA procedure, we used optimal scoring to assign numeric values to the observations on diseases and pests (on scale 1-5) 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). The elements of y (yield) had a one-to-one correspondence with the elements of x; that is, x_1 corresponded to y_1, x_2 corresponded to y_2, and so on.

Based on the transformed data set, preliminary diagnosis of bivariate correlation was done using Pearson correlation. Further analsysis using multiple linear regression (MLR) model $\mathbf{Y = XB + e}$ and output of variance inflation factor (VIF) on each factor of production was used to confirm the existence of multi-colinearity in the model**.**

PCA was used to reduce the dimensionality of the data set that contained cassava genotype,soil, pest and disease factors . This was done by identifying variances and correlations in the data set. We met the goal of reducing the dimensionality by maximizing the variance of a linear combination of the variables, Rencher (2002). The principal components retained were 8 from a possible maximum of 16 corresponding to the 16 factors of production. PC1 being the first principal component associated with the highest eigen value Λ_1, PC2 the second principal component associated with the second highest eigen value Λ_2 and so on. PCR model $Y=a_1PC1+a_2PC2+\ldots+a_8PC8$ was fitted on the 8 PCs obtained in the PCA procedure. The PCR co-effiecients were then transformed back to the linear scale using the transfrormation B=PA where:

$$A=(Z'Z)^{-1}Z'Y=D^{-1}Z'Y$$

P being the eigenvector matrix of factors of production extracted from the eigen values. Post-PCA correlation diagnostic was done by flagging the variance inflation factor associated with regression co-efficient of each component.

Results and discussion

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 $p<0.05$. Multiple linear regression and variance inflation factor analysis showed most variables in the data set had variance inflation factor, VIF>1 , implying existence of multicollinearity as shown in table 1 below. Moreover, most of the factors had higher values of standard error and this added 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 indicated 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 would be unstable and therefore results for yield prediction would be unreliable and invalid. This justified dimension reduction through principal component analysis.

Table 1: 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	1.040	0.001	p<0.001	214.549
BHT	1.000	0.000	0.930	19.931
PHT	1.000	0.000	0.118	37.693
NTOTAL	1.000	0.000	0.798	11.815
WTOTAL	1.018	0.000	p<0.001	22.305
RDM	1.005	0.001	0.001	46.038
CYN	1.030	0.004	0.0039**	17.156
CADS	0.983	0.009	0.430	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	0.010	0.551	14.810
CMBS	1.479	0.034	p<0.001	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.

The total number of principal components returned was 16, equal to the total number of variables used in the principal component procedure. 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 selected the first eight principal components from the table above which account for almost 70% of total variance. This was affirmed by the eigenvalue one rule in which we select the eigenvalues that are above value 1.

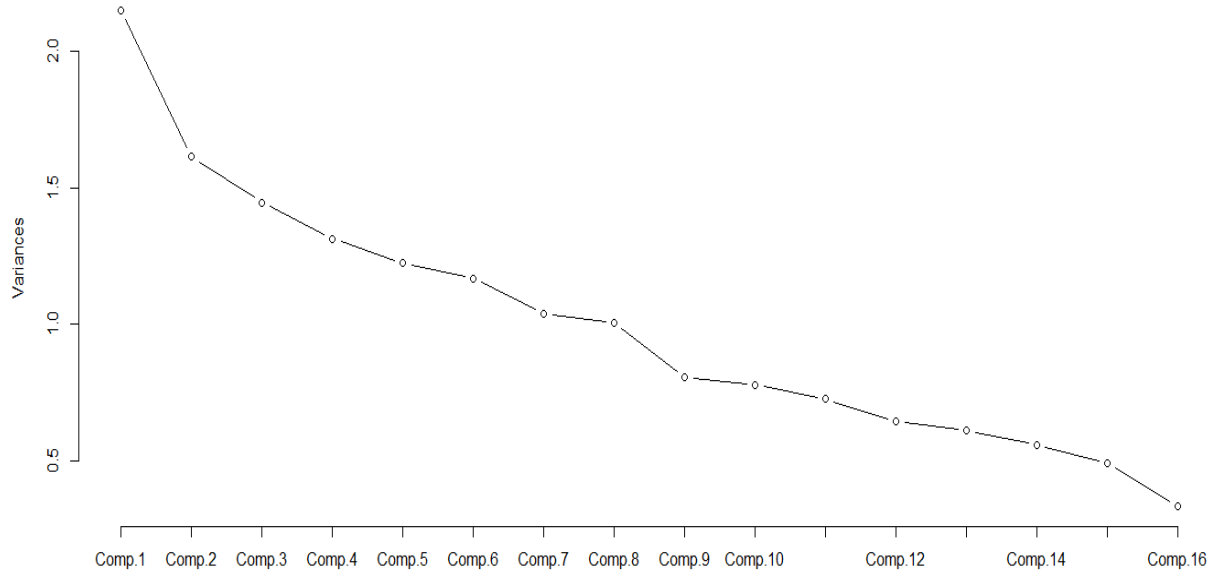
Eigen Values, Proportion Of Variance Explained By Principal Components And Loadings.

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

Table 2: 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	0.41	1.312
5	1.106	0.077	0.487	1.224
6	1.081	0.073	0.56	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

Figure 1: Scree plot for principal component importance



From the scree plot, a sharp decline in variance around PC 8 indicated a a sharp reduction in the importance of the principal components. The components that followed from this point contributed very little to the overall variance.

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

Table 3: Principal Component Regression Statistics

Component	Coeff	Std Error	P-Value
Comp.1	3.494	0.240	p<0.001
Comp.2	0.524	0.277	0.060
Comp.3	-0.018	0.292	0.950
Comp.4	-0.281	0.307	0.360
Comp.5	1.216	0.318	p<0.001
Comp.6	1.630	0.325	p<0.001
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

Principal component regression equation:

$$\mathit{YLD} = 3.494\mathit{Comp}.1 + 0.524\mathit{Comp}.2 - 0.018\mathit{Comp}.3 - 0.281\mathit{Comp}.4 +$$

$$1.216\mathit{Comp}.5 + 1.630\mathit{Comp}.6 + 0.783\mathit{Comp}.7 + 0.354\mathit{Comp}.8.$$

The model had an F-value, F= 32.85 with a p-value<0.001 (DF=168). This implied the model consisting of the first 8 PCs was significant in prediction of yield. The model's R^2=0.610 (Adj R^2=0.590) and the VIF attributed to the overall model being 2.564. Moreover, regressing yield on all factors of production showed that most of the co-efficients were statistically insignificant, p>0.05. 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 modelled 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

Conclusions

The PCR model solved the problem of multicolliarity and provided stability in regression co-efficients. Therefore reliability on the model was achieved even though the variability explained dropped. The model therefore not only offers an alternative to existing models but also an efficient solution when the number of factors is high.

References

Andrea Amadei, Antonius Linssen, and Herman JC Berendsen.(1993). Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics*, 17(4):412–425.

Andrea Amadei, Marc A Ceruso, and Alfredo Di Nola.(1999). 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.

Pamela K Anderson.(2005). Whitefly and whitefly-borne viruses in the tropics: Building a knowledge base for global action. Introduction.

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

Ibrahim Robeni Matete Benesi.(2005). Characterisation of Malawian cassava germplasm for diversity, starch extraction and its native and modified properties.

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

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

William W Cooley and Paul R Lohnes.(1971). *Multivariate data analysis*. J. Wiley.

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

FAO.(2007). Cassava production statistics, November. URLhttp://www.fao.org.2007/.

Claude M Fauquet, M A Mayo, Jack Maniloff, Ulrich Desselberger, and Laurence Andrew Ball.(2005). *Virus taxonomy: VIIIth report of the International Committee on Taxonomy of Viruses*. Academic Press.

Anneke Marijke Fermont et al.(2009). *Cassava and soil fertility in intensifying smallholder farming systems of East Africa*. publisher not identified.

VN Fondong, JM Thresh, and C Fauquet.(2000). 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.

CM Githunguri, S Mwiti, Y Migwa, et al.(2007). 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.

James W Hansen, Ashok Mishra, KPC Rao, Matayo Indeje, and Robinson Kinuthia Ngugi.(2009). Potential value of gcm-based seasonal rainfall forecasts for maize management in semi-arid kenya. *Agricultural Systems*, 101(1):80–90.

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

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

Rory J Hillocks, JM Thresh, and Anthony Bellotti.(2002). *Cassava: biology, production and utilization*. CABI.

Aapo Hyvarinen, Juha Karhunen, and Erkki Oja.(2001). Independent components analysis.

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

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

Ian Jolliffe.(2002). *Principal component analysis*. Wiley Online Library.

Edward E Karuri, Samuel K Mbugua, Joseph Karugia, J Wanda, and John Jagwe.(2001). 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*.

JN Kiura, CK Mutegi, P Kibet, and MK Danda.(2005). 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..

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

MN Maruthi, Susan Seal, John Colvin, RW Briddon, and SE Bull.(2004). East African cassava mosaic Zanzibar virus–a recombinant begomovirus species with a mild phenotype. *Archives of virology*, 149(12):2365–2377.

Soumya Raychaudhuri, Joshua M Stuart, and Russ B Altman.(2000). 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.

Alvin C Rencher.(2002). M éthods of multivariate analysis. a john wiley & sons. *Inc. Publication*.

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

Stephen R Waddington, Xiaoyun Li, John Dixon, Glenn Hyman, and M Carmen De Vicente.(2010). Getting the focus right: production constraints for six major food crops in asian and african farming systems. *Food security*, 2(1):27–48.

AP Walker, PK Mutuo, Meine van Noordwijk, Alain Albrecht, and Georg Cadisch.(2007). Modelling of planted legume fallows in western kenya using wanulcas.(i) model calibration and validation. *Agroforestry systems*, 70(3):197–209.

Qing-Song Xu and Yi-Zeng Liang.(2001). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1):1–11.

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

Wenjun Zhu, Lysa Porth, and Ken Seng Tan.(2012). Improving crop yields forecasting using weather data: A comprehensive approach combining principal component analysis and credibility model. *IARFIC publication*.