

Longitudinal Analysis of Cereal Yields in Ghana

Alfred Kwabena Amoah

Ghana National Ambulance Service, Brong Ahafo Regional Administration, Sunyani, Ghana

Email address:

alfredkwabenaamoah@gmail.com

To cite this article:

Alfred Kwabena Amoah. Longitudinal Analysis of Cereal Yields in Ghana. *International Journal of Statistical Distributions and Applications*. Vol. 4, No. 2, 2018, pp. 38-50. doi: 10.11648/j.ijstd.20180402.12

Received: September 13, 2018; **Accepted:** October 6, 2018; **Published:** November 14, 2018

Abstract: The aim of this study was to investigate cereal crop yields in Ghana. It was set out to specifically determine whether there is a significant difference in the yields across the ten regions in Ghana and also find out an evolution in the yields among the regions. A multivariate data of two major cereal crops (Maize and Rice) produced and consumed in Ghana from 2005 to 2014 was obtained from Statistical Research and Information Department (SRID) of Ministry of Food and Agriculture (MOFA). Multivariate Analysis of Variance (MANOVA) model as summarized by Casella & Berger (2002) and Linear Mixed Model (LMM) by Faraway (2007) were employed for the study. Diagnostic plots for the fitted LMM revealed a valid model for the analysis. The study revealed that significant differences exist in the yields of the two major cereal crops in all the regions in Ghana. Further analysis by LMM indicated that the yields of maize and rice varied between and within the regions of Ghana with maize yields having much less variability than rice yields. It also indicated that there is consistent decelerating trend in maize yields and gradual increasing trend in rice yields across all the regions in Ghana. Based on these findings, we recommend that intensive support must be given to farmers who engage in cereal crops production in all the regions in Ghana to help reduce this variability in the two major cereal crop yields (maize and rice).

Keywords: Longitudinal Analysis, Multivariate Analysis of Variance, Linear Mixed Model, Yields, Rice, Maize

1. Introduction

The concern of agricultural production and its role in ensuring food security in the last two decades seems to have moved away from global focus. There has been a decline in investment in research and infrastructure and advancements in technology have decelerated [1]. Agricultural development has been neglected and is now set to face its biggest challenges. The demands facing food production have come at a time when crop yields are stagnant and food reserves are the lowest they have been in 35 years [1-3].

In 2008 the FAO, World Bank and G8 all called for the renewed investment in agriculture directed at improving and expanding agricultural productivity and production [2]. The World Bank report (2009) highlights the global availability of agricultural land and the potential for improving production, particularly in areas with low agricultural development, such as Africa [4]. A consensus appears to be forming that “underutilized” land in developing countries can be the key to long-term global food security, whether this is through the introduction of new farmland or the closing of yield gaps [5-8].

Investing in the farming sectors of developing countries will increase productivity to a larger degree than possible in Europe or the US, where agricultural potential is very close to maximized and soil quality is now degrading [9]. This investment in policies, infrastructure, science and technology is required to access the potential productivity from Africa’s land and resources [1, 5]. The logic is that investment in developing nations, which are currently only achieving a small amount of their agricultural potential, could contribute to future global food security by increasing crop yields. This was reemphasized by UN General Assembly on 11th August, 2015 on the need to improved agriculture productivity by ensuring the attainment of Sustainable Development Goal 2 (SDGs 2) on hunger, achieves food security and improved nutrition and promote sustainable agriculture [10].

In developing country like Ghana, food crops such as maize, rice; plantain, cassava, yam and other vegetables are grown both on subsistence and commercial level. Annual production over the years continues to decline in relative term in spite of several programme interventions by the Municipal and District offices of the Ministry of Food and Agriculture [11]. This is due to increasing cost of farm inputs

and the low soil fertility. Crop production is largely rain-fed and traditional technology of production continues to dominate the sector with peasant farmers using simple tools such as hoes and cutlasses. The average land holding per farmer is relatively low and is about 0.5 hectares.

Maize is Ghana's number one staple cereal crop followed by rice, and domestic demand for both is growing. Between 2010 and 2015, rice demand is projected to have grown at a compound annual growth of 11.8 percent and maize at 2.6 percent [12]. However, the country is not self-sufficient in either of its two most important cereal crops, as Ghana has experienced average shortfalls in domestic maize supplies of 12 percent and domestic rice supplies of 69 percent in recent years [11].

In this paper we used multivariate analysis of variance (MANOVA) and linear mixed model (LMM) to analyze the yields of the two most important cereal crops (Maize and Rice) to come out with significant differences and evolution among them across all the ten (10) regions in Ghana.

2. Literature Survey

This section reveal related topics on how other researchers have handle group yields on other crops. It includes topics such as; mixed effects model of crop yields, longitudinal study on crop yields and modeling for crop yields in Ghana.

2.1. A Mixed Effects Model of Crop Yield

A study conducted by Verma, Piepho, Hartung, Ogutu, Connell and Goyal [13] of department of mathematics & statistics, Haryana Agriculture University India, developed methodology for pre-harvest crop yield prediction of major mustard growing districts in Haryana (India). They use linear mixed effects models with random time effects at district, zone and state level, to fit crop yield estimate. For mixed modeling, the hierarchical model structure of yield they used was represented as.

$$y_{ijt} = s_t + z_{it} + d_{ijt}$$

where,

y_{ijt} = yield in the j^{th} district within i^{th} zone in the t^{th} year

s_t = general state effect in the t^{th} year

d_{ijt} = effect of the j^{th} district within i^{th} zone in the t^{th} year.

For each of the three effects (state: s_t , zone: z_{it} , district: d_{ijt}), they set up a time-series model with three components:

Regression + Time Trend + White noise. Regression was a fixed part comprising regression on time as well as on the meteorological covariates. Time Trend comprises a random part for serial correlation with covariance structure and regression splines. White noise was an additional independently distributed random error term.

The purpose of their study was to show the usefulness of the mixed model framework for pre-harvest crop yield forecasting. The findings indicated that there was improvement in the

predictive accuracy of the zonal yield models using linear mixed modeling. The linear mixed models substantially improved the predictive accuracy and produced what they considered to be satisfactory district-level yield(s) estimation. They concluded by recommending the use of linear mixed models for pre-harvest yield forecasting of crop to enhance the predictive accuracy of the zonal models.

Roberts and Tack [14] also presented a paper on a mixed effects model of crop yields for purposes of premium determination at the Agricultural & Applied Economics Association at Pittsburgh, Pennsylvania, USA. The goal of their research was to determine empirical estimation of farm-level yield distributions, calculation of actuarially fair risk premiums, and prediction of potential efficiency gains using empirical mixed model for premium determination given below.

$$y_{it} = a_i + b_c t + c_c ins + \mu_{ct} + \varepsilon_{it}$$

where y_{it} represents yield on farm i in year t , a_i is a farm-specific intercept, b_c is a county-specific trend, c_c is a county-specific insurance effect, μ_{ct} is a county-specific random shock in year t , and ε_{it} is an idiosyncratic random shock on farm i in year t . The farm-specific intercept according to them can be written as a county-specific mean plus a farm-specific shock, $a_i = a_c + \mu_i$, and the county-specific slope parameters can be written as a population mean plus a county shock, $b_c = b + \mu_{c1}$ and $c_c = c + \mu_{c2}$.

These modifications generated the model:

$$y_{it} = a_c + bt + cins + \mu_i + \mu_{ct} + \mu_{c1}t + \mu_{c2}ins + \varepsilon_{it}$$

They indicated two types of effects in this model. The first part $a_c + bt + cins$, represents the fixed effects of crop yields, and the second part, $\mu_i + \mu_{ct} + \mu_{c1}t + \mu_{c2}ins + \varepsilon_{it}$, represents random effects. They cited two random intercepts, farm-specific $\{\mu_i\}$ and county/time-specific $\{\mu_{ct}\}$; and, there were two random slopes at the county level, for the trend parameter (μ_{c1}) and for the insurance parameter.

They followed conventional methods for estimating linear mixed models by assuming that the random components follow a multivariate normal distribution and employed maximum likelihood estimation. With a separate model for each state-crop combination in their dataset, their empirical results suggested that there were several important sources of heterogeneity for crop yield distribution.

For example, for Arkansas rice, the farm-specific random intercept accounted for 42% of the random variation of their model, the county-specific intercept accounted for 7.5%, the random slope for the time trend accounted for 0.1% and the random slope accounted for 1% [14]. These findings confirm their objectives that there were significant farm-level crop distribution shifters across space and time.

2.2. Longitudinal Study on Crop Yields

Another study conducted on longitudinal and spatial analyses applied to corn yield data from a long-term rotation trial by Brownie, Larry and Tina [15] investigated a number of analyses of rotation trial of corn yields. The study used several types of split plot, repeated measures and spatial analyses, each with and without soil covariates and restricted attention to models that could be implemented using standard software. Important features of the yield data from these authors trial include considerable unbalance, possible correlations across time and space, possible heterogeneity in the error variances across years, and the availability of pretreatment measurements of soil properties.

Ignoring the soil covariates, a linear mixed model for the corn yields in their methodology was given as

$$Y_{ijk} = \mu + \beta_i + (RY)_{jk} + \delta_{ij} + \varepsilon_{ijk}$$

where

Y_{ijk} is the yield in year k for rotation j in block i , $k = 1, \dots, 6, 8, 9, 10$, $j = 1, \dots, 27$, $i = 1, \dots, 4$, β_i is a random effect for the i^{th} block or rep,

$(RY)_{jk}$ is a fixed effect for rotation j in year k , with

$$\sum_{jk} (RY)_{jk} = 0,$$

δ_{ij} is a random effect for the ij^{th} plot (the plot in block i containing rotation j), and

ε_{ijk} is a random error associated with the ij^{th} plot in year k .

The corn yield data were highly unbalanced because of the different crop sequences with corn planted in 155 of the possible $9 \times 27 = 243$ rotation-year combinations [15]. There were also several missing yields, thus instead of including main and interaction effects for rotation and year, they fitted an effect for each observed rotation-year combination, represented as $(RY)_{jk}$ in the model above. The Authors diagnostic graphs and AIC values indicated that the model should allow for heterogeneity across years in the error variance. The precision of contrasts in years with high error

variance according to their recommendation was underestimated in the analysis that assumed constant variance. Allowing spatial correlations resulted in a small reduction in estimates of precision, and including soil covariates which were important in two of the nine years used. The authors indicated that the best models were repeated measures with banded covariance, as well as autoregressive repeated measures and isotropic spatial models, both modified to allow heterogeneity of the error variance across years.

Also a study on climate variability and crop production in Tanzania conducted by Rowhani, Lobell, B, Linderman, Ramankutty and Navin [16] revealed the relationship between seasonal climate and crop yields in Tanzania, focusing on maize, sorghum and rice. The study which makes use of linear mixed model outline below revealed an interesting result.

In order to determine the effects of climate on agricultural yields, and to exploit the cross-sectional and temporal attributes of their dataset, they develop linear mixed models for each of the three crops (maize, rice, and sorghum). The method was appropriate for longitudinal [17, 18] where observations within a group are often more similar than would be predicted on a pooled-data basis. The model was given as

$$y_{ij} = \beta_0 + \beta_1 T_{i,j} + \beta_2 P_{i,j} + \beta_3 P_{i,j}^2 + \beta_4 CW_{T-i,j} + \beta_5 CV_{P-i,j} + a_i + \varepsilon_{ij}$$

where

y_{ij} is yield, i represents the regions and j the observations within a region, β_{0-5} represent model parameters, a_i represents the random intercept term, T represent temperature, P represents precipitation and ε_{ij} is an error term. Also the model include a fixed part comprised of P and T (and their interaction term), CV_T and CV_P (and their interaction term), as well as P^2 , and random intercepts. For comparison, the coefficients resulting from their mixed models were compared to those obtained from simple linear regression models that included the different regions as a dummy variable to account for fixed effects. Again a time variable (year from 1992) was also used in the linear models to capture yield changes related to non-climatic factors and other technological development:

$$y_j = \beta_0 + \beta_1 T_j + \beta_2 P_j + \beta_3 P_j^2 + \beta_4 CV_{T-j} + \beta_5 CV_{P-j} + \beta_6 \text{Region}_j + \beta_7 \text{Year}_j + \varepsilon_j.$$

These linear models were developed using stepwise model selection based on the AIC. In order to compare climate data effects on yield estimates, they used both sets of models, the mixed and linear models, were also developed using climate data they extracted from the CRU dataset. The results of their study indicated that both intra and inter seasonal changes in temperature and precipitation influenced cereal yields in Tanzania. They projected that by their studies, in Tanzania, by 2050, projected seasonal temperature increases by 2°C will reduce average maize, sorghum and rice yields by 13%, 8.8% and 7.6% respectively [16].

2.3. Modeling for Crop yield in Ghana

Some researchers in Ghana have also conducted studies on crop yields using mixed modeling and other methods but usually concentrated on a single crop. For instance a study conducted by Isaac [19] of KNUST uses multiple comparison and random Effect model to analyzed Cocoa production in Ghana from 1969/70 to 2010/11 production years.

In his studies he set the linear mixed effects models as

$$Y_i = X_i \beta + Z_i b_i + \varepsilon_i \quad i = 1, 2, \dots, N$$

where Y_i is a vector of observations with mean $E(Y) = X\beta$.

X_i and Z_i are the design matrices corresponding to the fixed and random effects respectively, β is fixed effects vector, b_i is a vector of independent and identically distributed (iid) random effects with mean $E(b) = 0$ and variance-covariance matrix variable $(b) = D$.

ε is a vector of iid random errors with mean $E(\varepsilon) = 0$ and variance $\text{var}(\varepsilon) = R$. It was assumed that $b \sim N(0, D)$ and $\varepsilon \sim N(0, \Lambda)$, with b independent of ε . His analysis using mixed effect model revealed that from 1969/70 production, all the six regions used in the study experienced increasing cocoa production trend with exception of Volta region. Western and Ashanti regions had the highest production over the years.

Again, a study was carried out by Azinu [20] of department of Crop Science University of Ghana, on the evaluation of hybrid maize varieties in three agro-ecological zones (transition forest, Guinea savannah and coastal savannah) in Ghana. The study which was undertaken to assess the relative yielding abilities and stability of 20 hybrids selected from the breeding programme of the West Africa Centre for Crop improvement of maize uses analysis of variance (ANOVA) model. ANOVA per location and across location or environment for agronomic traits were carried out using Genstat 12th edition. Genotypes were considered as fixed effects, whilst environments and replication were considered as random effects. For each agronomic and morphological trait, an individual ANOVA was conducted by the researcher to determine the statistical significance of the genotypes at each environment and across environment. According to the author, the significant genotype-environment interaction revealed by additive main effect and multiplicative interaction (AMMI) analysis of variance for grain yield suggested that the relative performance of the genotypes changed for grain yield across all environments.

The study also identified Wenchi as the location for the best grain yield performance and Tamale as an environment yielded low grain yield in both seasons

3. Methodology

Multivariate analysis of variance model (MANOVA) and Linear mixed (LM) model enable us to study longitudinal data of cereal crop yields. Since its introduction by Wilks in 1932, multivariate generalization of the ANOVA-model has become well established and widely used in many research areas ranging from Agriculture to psychology [21-23].

Our study make use of two-way MANOVA model as summarized by Casella and Berger [21, 24].

The two-way MANOVA model with interactions is expressed as:

$$y_{ij\ell} = \mu + \alpha_\ell + \beta_j + \alpha\beta_{\ell j} + \varepsilon_{ij\ell} \quad (1)$$

where

$y_{\ell i} : p \times I$ is vector of p response variables for the ℓ^{th} replicate on the i^{th} level of factor A, and the j^{th} level of factor B, $\ell = 1, 2, \dots, g$, $j = 1, 2, \dots, b$, $i = 1, 2, \dots, n$. In the two-way MANOVA-model, vectors α_ℓ , β_j and $\alpha\beta_{\ell j}$ represent main and interaction effects respectively. Also, it is assumed that $\varepsilon_{ij\ell} \sim N_p^{iid}(0, \Sigma)$ so that:

$$E(y_{ij\ell}) = \mu + \alpha_\ell + \beta_j + \alpha\beta_{\ell j}, \quad V(y_{ij\ell}) = V(\varepsilon_{ij\ell}) = \Sigma \quad (2)$$

For the MANOVA-model, the testing of hypotheses based on the partitioning of sums of squares becomes more complex because of the interrelationships between the p include ANOVA-models. Unlike the univariate models, we considered sums of squares but also cross products for the factors in the MANOVA-model. In the resulting matrices, called sums of squares and cross products (SSCP), diagonal elements corresponds to the usual sums of squares for each of the P response variables whereas the off-diagonal elements correspond to the cross products for each response variable pair [22].

When data is balanced, the partitioning of SSCP matrices is independent in analogy with the ANOVA-models described earlier. For instance, in MANOVA-model:

$$T = H + E \quad (3)$$

where $T: p \times p$ is the total SSCP matrix, $H: p \times p$ is the hypothesis SSCP matrix and $E: p \times p$ is the error SSCP matrix. The general hypothesis under the MANOVA can be written as:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_g \quad (4)$$

H_1 : at least one group centroid is different.

With p responses, 1 factor and g groups. The assumptions are given as:

Independent groups, independent observations

Responses are multivariate normal with each group

Population covariance matrices are equal across groups

According to [24]. The observations can be decomposed as:

$$X_{\ell i} = \bar{x} + (\bar{x} - \bar{x}) + (x_{\ell i} - \bar{x}_\ell) \quad (5)$$

That is, observation is equal to overall mean + treatment effect (or group effect) + residuals

This implies

$$\begin{aligned} (x_{\ell i} - \bar{x})(x_{\ell i} - \bar{x})' &= [(\bar{x}_\ell - \bar{x}) + (x_{\ell i} - \bar{x}_\ell)][(\bar{x} - \bar{x}) + (x_{\ell i} - \bar{x}_\ell)]' \\ &= (\bar{x}_\ell - \bar{x})(\bar{x}_\ell - \bar{x})' + (\bar{x}_\ell - \bar{x})(x_{\ell i} - \bar{x}_\ell)' + \\ &\quad (x_{\ell i} - \bar{x}_\ell)(\bar{x}_\ell - \bar{x})' + (x_{\ell i} - \bar{x}_\ell)(x_{\ell i} - \bar{x}_\ell)' \end{aligned}$$

Summing over i , we obtain the following:

$$\begin{aligned}\sum_{i=1}^{n_\ell} (x_{\ell i} - \bar{x})(x_{\ell i} - \bar{x})' &= \sum_{i=1}^{n_\ell} (\bar{x}_\ell - \bar{x})(\bar{x}_\ell - \bar{x})' + (\bar{x}_\ell - \bar{x}) \sum_{i=1}^{n_\ell} (x_{\ell i} - \bar{x}_\ell)' + \sum_{i=1}^{n_\ell} (x_{\ell i} - \bar{x}_\ell)(\bar{x}_\ell - \bar{x})' + \sum_{i=1}^{n_\ell} (x_{\ell i} - \bar{x}_\ell)(x_{\ell i} - \bar{x}_\ell)' \\ &= n_\ell (\bar{x}_\ell - \bar{x})(\bar{x}_\ell - \bar{x})' + \sum_{i=1}^{n_\ell} (x_{\ell i} - \bar{x}_\ell)(x_{\ell i} - \bar{x}_\ell)' \text{ Since } \sum_{i=1}^{n_\ell} (x_{\ell i} - \bar{x}_\ell) = 0\end{aligned}$$

Next, we sum over ℓ and obtain;

$$\sum_{\ell=1}^g \sum_{i=1}^{n_\ell} (x_{\ell i} - \bar{x})(x_{\ell i} - \bar{x})' = \sum_{\ell=1}^g n_\ell (\bar{x}_\ell - \bar{x})(\bar{x}_\ell - \bar{x})' + \sum_{\ell=1}^g \sum_{i=1}^{n_\ell} (x_{\ell i} - \bar{x}_\ell)(x_{\ell i} - \bar{x}_\ell)' \quad (6)$$

$$\left(\begin{array}{c} \text{Total (corrected) Sum} \\ \text{of Squares and Cross} \\ \text{Products} \end{array} \right) = \left(\begin{array}{c} \text{Treatment (Between)} \\ \text{Sum of Squares and} \\ \text{Cross Products} \end{array} \right) + \left(\begin{array}{c} \text{Residual (Within) Sums} \\ \text{of Squares and Cross} \\ \text{Products} \end{array} \right)$$

$$T = B + W$$

where $T: p \times p$ is the total SSCP matrix, B is the “between” matrix which is denoted by H , the “hypothesis” as indicated in (4) and “within” matrix W is often denoted as E , “error” matrix.

3.1. Statistics Used in MANOVA Model

From [24] Johnson and Wichern [25] we outline four (4)

$$\Lambda = \frac{|E|}{|E + H|} = \det(E(H + E)^{-1}) = \prod_{j=1}^p \lambda_j = \prod_{j=1}^p \theta_j = \prod_{j=1}^p \frac{1}{1 + \phi_j} \quad (8)$$

is generally known as Wilks' Λ after Wilks (1932) and cited in [23]. The null hypothesis is rejected for small values of Λ , showing that E is small compared to the total SSCP matrix $E + H$.

2. Hotelling-Lawley Trace

The test statistic is given by:

$$U = \text{tr}(HE^{-1}) = \sum_{j=1}^p \phi_j = \sum_{j=1}^p \frac{\theta_j}{1 - \theta_j} \quad (9)$$

is often referred to as Hotelling-Lawley Trace after and Hotelling who took part in developing the statistic and cited in [26, 27]. Naturally, a large H relative to E would indicate a larger support for H and a larger trace. Hence the null hypothesis in equation (4) of no effects is rejected for large values of U .

3. Pillai's Trace

Pillai (1955) developed the following statistic:

$$V = \text{tr}((E + H)^{-1}H) = \sum_{j=1}^p \theta_j, \quad (10)$$

which is commonly known as Pillai's Trace. As with Hotelling-Lawley's Trace, the null hypothesis is rejected for large values of V , indicating a large H relative to E Crowder [26].

statistics commonly use in MANOVA-model and considered in this study as:

1. Wilks' Lambda (Λ)

Under the null hypothesis of no factor γ effects:

$$H_0: \gamma = 0 \quad (7)$$

the likelihood ratio test statistic is given as,

4. Roy's Greatest Root

Roy [28] also developed the Statistic: The largest root of the equation $|H - \phi E| = 0$

$$\phi_{\max} = \frac{\theta_{\max}}{1 - \theta_{\max}} \quad (11)$$

3.2. Linear Mixed Model (LMM)

Another model used in the study is Linear mixed model (LMM) which is an extension of Linear model for data that were collected and summarized in groups (profile) over a period [24, 29, 30]. The model is used in outcome variables in which residuals is normally distributed but may not be independent or have constant variance. Example is longitudinal or repeated-measures in which subjects are measured repeatedly over time. LMM may include both fixed effects and random effects parameters.

3.3. The Model Formulation

The model for this study which is also longitudinal data collected over ten years period are given below.

The model without random intercept slope is given as:

$$y_{ij} = \beta_0 + \sum_{r=1}^9 \beta_r X_{ij} + \beta_{10} Year_i + \sum_{r=1}^9 \beta_r X_{ij} Year_i + Z_{ij} b_i + \xi_{ij} \quad (12)$$

where y_{ij} is a $(n \times 1)$ vector of the j^{th} region yields for the i^{th} year,

β_0 is the intercepts of the fixed effects, for the fixed effects we have

$$\sum_{r=1}^9 \beta_r X_{ij} + \beta_{10} Year_i$$

$\sum_{r=1}^9 \beta_r X_{ij} Year_i$ is the interaction between regions and years,

$Z_{ij} b_i$ is random effects,

X_{ij} is an $(n \times p)$ vector fixed-effects design or explanatory variable of region j at year i ,

β_r is a $(p \times 1)$ vector offixed effects parameters, that is region

Z_{ij} is $(n \times g)$ random effects design or regressor matrix, that is districts,

b_i is $(g \times 1)$ vector of random effects parameter that occur in the data vector y and

ξ_{ij} is an $(n \times 1)$ vector of model errors (also random effects) specific to j^{th} region at year i .

The model with random intercept and slope is given as

$$y_{ij} = \beta_0 + \sum_{r=1}^9 \beta_r X_{ij} + \beta_{10} Year_i + \sum_{r=1}^9 \beta_r X_{ij} Year_i + b_{0i} + Z_{ij} b_i + \xi_{ij} \quad (13)$$

where b_{0i} is the random intercept slope, all other variables retain their usual meaning.

Unlike linear model where $E(y) = X_i \beta$ with β as fixed effects; in LMM $X_i \beta$ is used for the fixed effects and b_i used as random effects. The model is specified conditionally as

$$E(y/b) = X_i \beta + Z_{ij} b_i \quad (14)$$

where b_i is the conditional mean and realization of the random variable. The distribution assumptions are:

$$b_i \sim N(0, D) \text{ and } \xi_i \sim N(0, \Sigma_i) \quad (15)$$

The random components b_i and ξ_{ij} are independent. Also

$$\text{var}(y/b) = Z D Z + \Sigma \text{ and } y \sim (X \beta, Z D Z + \Sigma) \quad (16)$$

The parameters of this model are:

Fixed –effects β and random effect b_i

All unknowns in the variance matrices D and Σ .

The unknown variance elements are also referred to as the covariance parameters and collected in the vector θ . The vector of covariance parameters,

$$\theta = \begin{bmatrix} \theta_D \\ \theta_\Sigma \end{bmatrix} \quad (17)$$

combines all parameters from the covariance matrices D and Σ_i in the vectors θ_D and θ_Σ respectively.

3.4. Information Criteria

The study used two types of information criteria often used to choose the best fitted model for the data, the Akaike information criteria and the Bayes information criteria developed by [31] and [32, 33] also known as Schwarz Criterion [34], respectively.

The Akaike information criteria, AIC , is defined by

$$AIC = -2l(\hat{\beta}, \hat{\theta}) + 2p \quad (18)$$

where $l(\hat{\beta}, \hat{\theta})$ can be either the ML or REML log-likelihood function and p represents the total number of parameters, both the fixed and random effects, being estimated in the model. The model with the lowest AIC value is assumed to be the best fit for the data.

The Bayes information criteria, BIC is defined by

$$BIC = -2l(\hat{\beta}, \hat{\theta}) + p \ln(n) \quad (19)$$

where $l(\hat{\beta}, \hat{\theta})$ is the ML log-likelihood function, p represent the total number of parameters, both the fixed and random effects, being estimated in the model and n is the total number of observations used in estimation of the model. That is

$$n = \sum_{m=1}^m n_i.$$

According to [35] we can calculate the REML version of the BIC by simply using REML log-likelihood function and replacing $\ln(n)$ by $\ln(n - p_{fixed})$, where p_{fixed} is the number of estimated fixed effect parameter in the model, in equation (19).

In other words, the BIC applies a greater penalty for models with more parameters than the AIC . And as such, the model with the lowest BIC value is assumed to be the best fit and preferable for the data. West, Welch and Galecki [29] cited that there is no information criterion which stands apart as the best criterion to be used when selecting linear mixed effects models.

3.5. Diagnostics

It is necessary after a linear mixed effects model is fitted to check whether the underlying distributional assumptions for the

random effects and the residuals appear valid for the data. Diagnostic methods for linear models are well established, but diagnostics for linear mixed effects models are however more difficult to perform and interpret due to the complexity of the model. The most useful method for diagnostics according to Bates and Pinheiro [35] are based on plots of the residuals, the fitted values and the estimated random effects. In this paper we will do diagnostic by using the functions *qqnorm.lme*, *qqline.lme*, *histogram* and *plot.lme* in [36]. Here the standardized, or residuals, defined as the raw residuals divided by the estimated corresponding standard deviation, are used.

4. Results and Discussion

4.1. Exploratory and Preliminary Analysis

This paper applies MANOVA and Linear Mixed models to cereal crop yields in Ghana. All the analysis in this paper is limited to the yields of the two major cereal crops (Maize and Rice) recorded over the period 2005 to 2014. The methodologies described in the previous section are applied to the data and the results are discussed in this section.

Exploratory analysis was first conducted on the assumptions of both models to verify the satisfactory level of

these assumptions. All the assumptions seem to be satisfied except the constant variance assumption which was violated. This happens because the yields were correlated to each other as results of presents of soil nutrients due to farm chemicals like fertilizers used in the successive years on same plot for production.

Table 1. Descriptive statistics of the two major cereal crops.

Yield	Minimum	1 st Quartile	3 rd Quartile	Mean	Maximum
Maize	0.130	1.252	1.900	1.605	9.560
Rice	0.680	1.265	2.665	2.098	6.720

The lowest maize yields for the period under study is 0.13 metric tons per hectare (Mt/Ha) and the maximum maize yields was 9.56 Mt/Ha. In the same period the lowest rice yield recorded was 0.68 Mt/Ha and the maximum rice yields for the ten years period was 6.72 Mt/Ha. We observed from the above table that the average yield of rice in the period is higher than the average maize yield. This indicates that rice yields in Ghana for the period under study were averagely higher than maize yields in most of the regions for the ten years period. Figure 1 below give graphical representation of this.

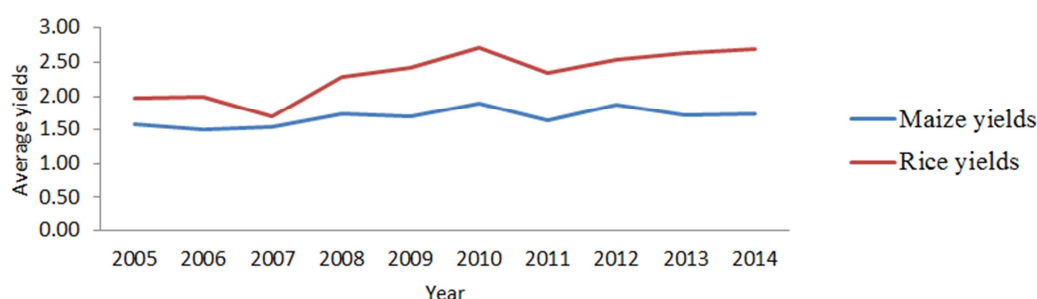


Figure 1. Line graph showing average yields by year.

Table 2. Pillai's Test for differences in crop yields.

Year	Estimates	DF	Approx.F	Pr(>F)
2005	1.0741	9	13.148	2.201e-16
2006	1.1379	9	18.772	2.211e-16
2007	0.9189	9	12.088	2.234e-16
2008	1.0420	9	11.845	2.231e-16
2009	0.9501	9	12.870	2.220e-16
2010	0.9069	9	11.800	2.242e-16
2011	1.1942	9	21.075	2.247e-16
2012	0.5287	9	6.3077	2.874e-13
2013	0.8559	9	13.302	2.210e-16
2014	0.9018	9	14.598	2.112e-16

The results of multivariate analysis of variance conducted on the data set using the four most commonly used statistic discussed in the previous section revealed that, significant differences exist in the yields of two major cereal crops in the country. For instance, the results of Pillai's test in table 2 below indicate that the yields of the two cereal crops are significantly different in all the years across the various

regions. All the statistics has P-value less than 0.05, signifying significant evidence against the null hypothesis that, there are no significant differences in yields of the two major cereal crops across the regions. The other three statistics indicated similar results.

4.2. Model Analysis

This section of the analysis deals with the linear mixed model considered for the study. As stated earlier, the discussions were based on the earlier methodology. Two types of mixed models were considered for the study: model with random intercepts and model with random intercepts and slope. In order to know which model best fit the data, we used Akaike Information Criteria (AIC) and the Bayes Information Criteria (BIC) developed by [31]and [32] to select best fitted model to the data. Table 3 below present the two models and the values of AIC and BIC. It is important to state that for all analyses, P-value < 0.05 was considered to be statistically significant.

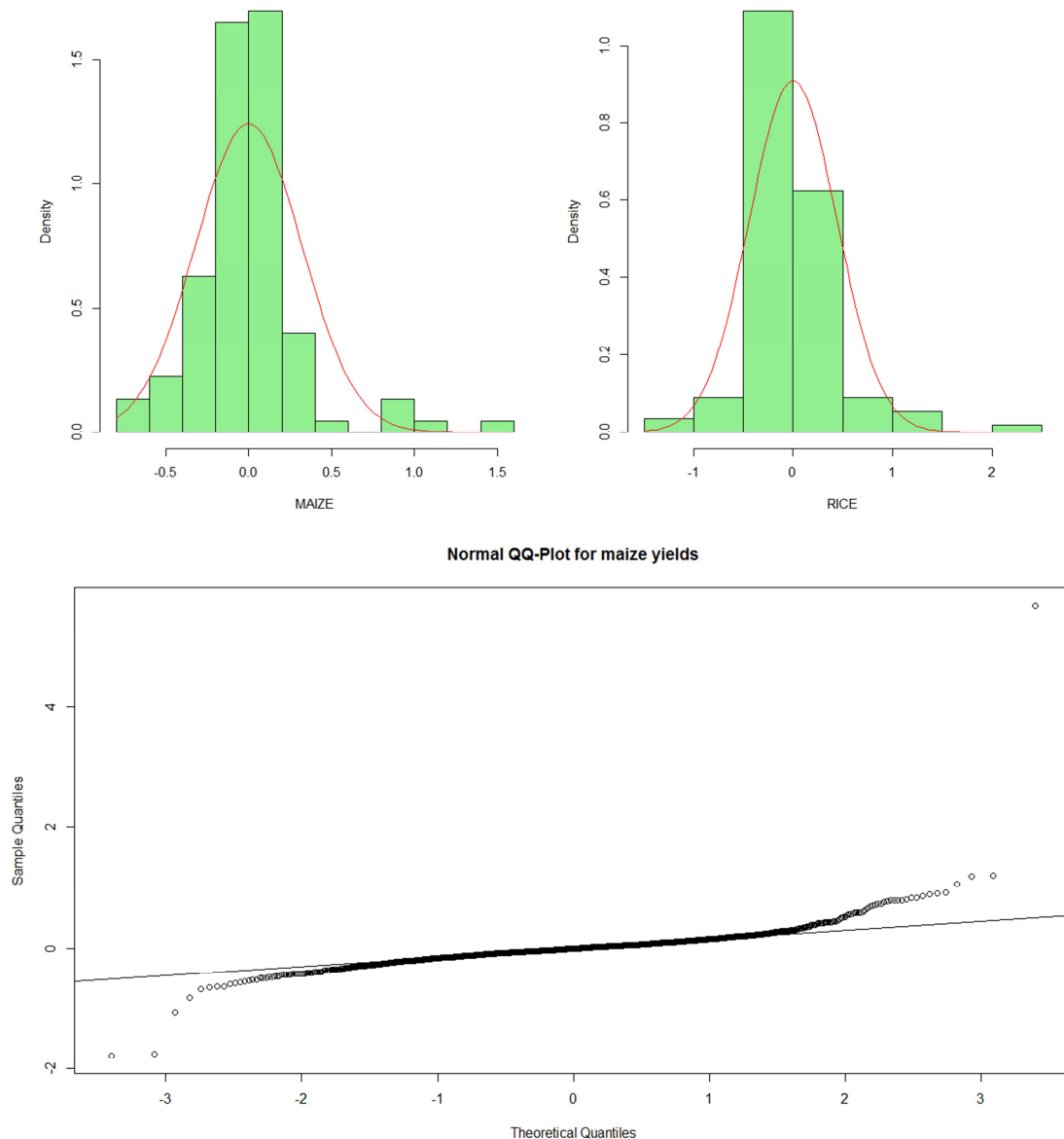
Table 3. Selection of best model.

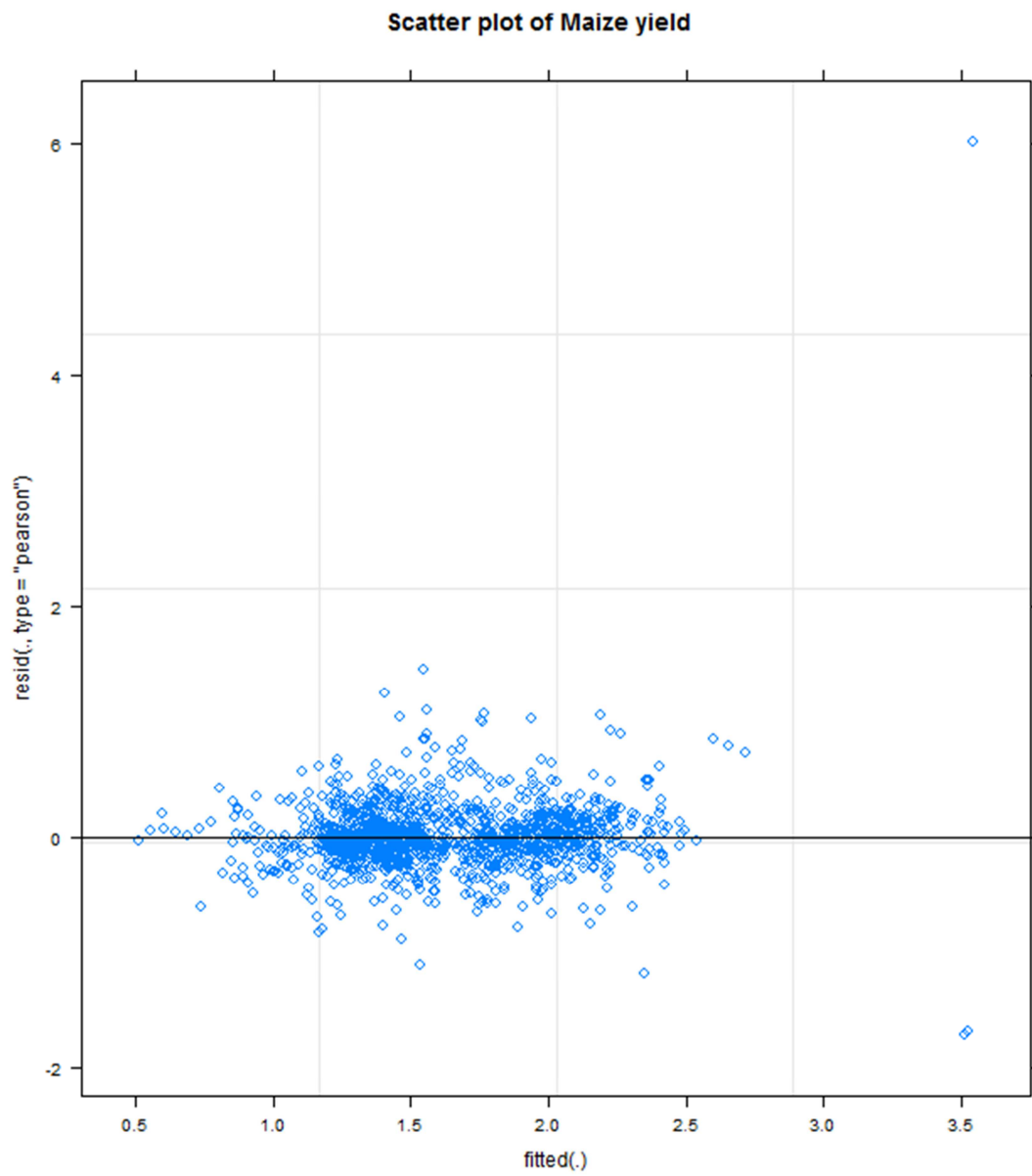
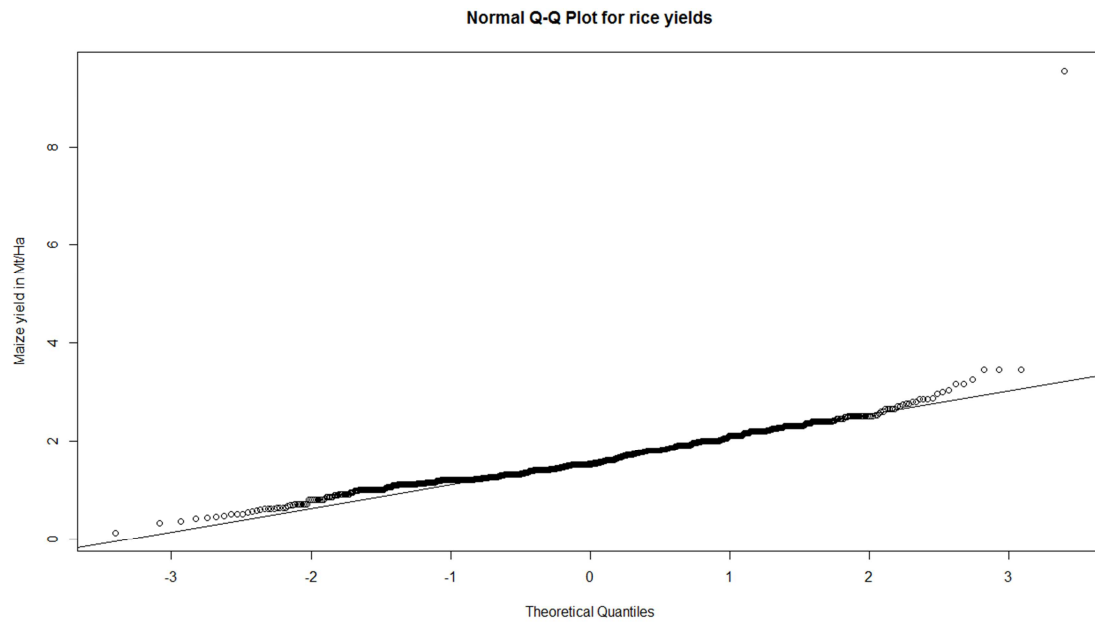
Maize	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
Random Intercept	1	22	1326.305	1442.632	-641.1527			
Ran. Int. and slope	2	24	1329.980	1456.882	-640.9902	1 vs 2	0.3250162	0.85
Rice	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
Random Intercept	1	22	3084.077	3200.403	-1520.039			
Ran.Int. and slope	2	24	3087.685	3214.587	-1519.843	1 vs 2	0.3917895	0.821

From the table, we observed that the P-value associated with Maize model is not statistically significant ($0.85 > 0.05$) which indicates Random Intercept model best fit the data than Random intercept and slope. It is also confirmed by the lower AIC and BIC values associated with the model (1326.305 and 1442.632).

The Rice model also revealed similar findings. It also indicated that the model with Random intercept best fit the data than Random intercept and slope. As it can be seen from the table, the P-value is not statistically significant ($0.821 > 0.05$) which was supported by lower AIC and BIC values given by the model (3084.077 and 3200.403).

We further assess the fitted LM Models using diagnostic plots for maize and rice yields. The plots were (1) QQ-plot, (2) scattered plot and (3) a histogram with a density curve of the fitted LMM. A good fit for QQ-plot should produce a straight one-to-one line of points, for the scatter plot, a good fit should not produce a pattern in the plot. Generally, QQ-plot is often desired to the scatter plot. The normality nature of the density curve on the histogram with the points and approximated linearity of the QQ-plot indicates that, the model is effective model. Hence we conclude that the diagnostic plots support the fitted model and so LMM is a good model for fitting maize and rice yields.





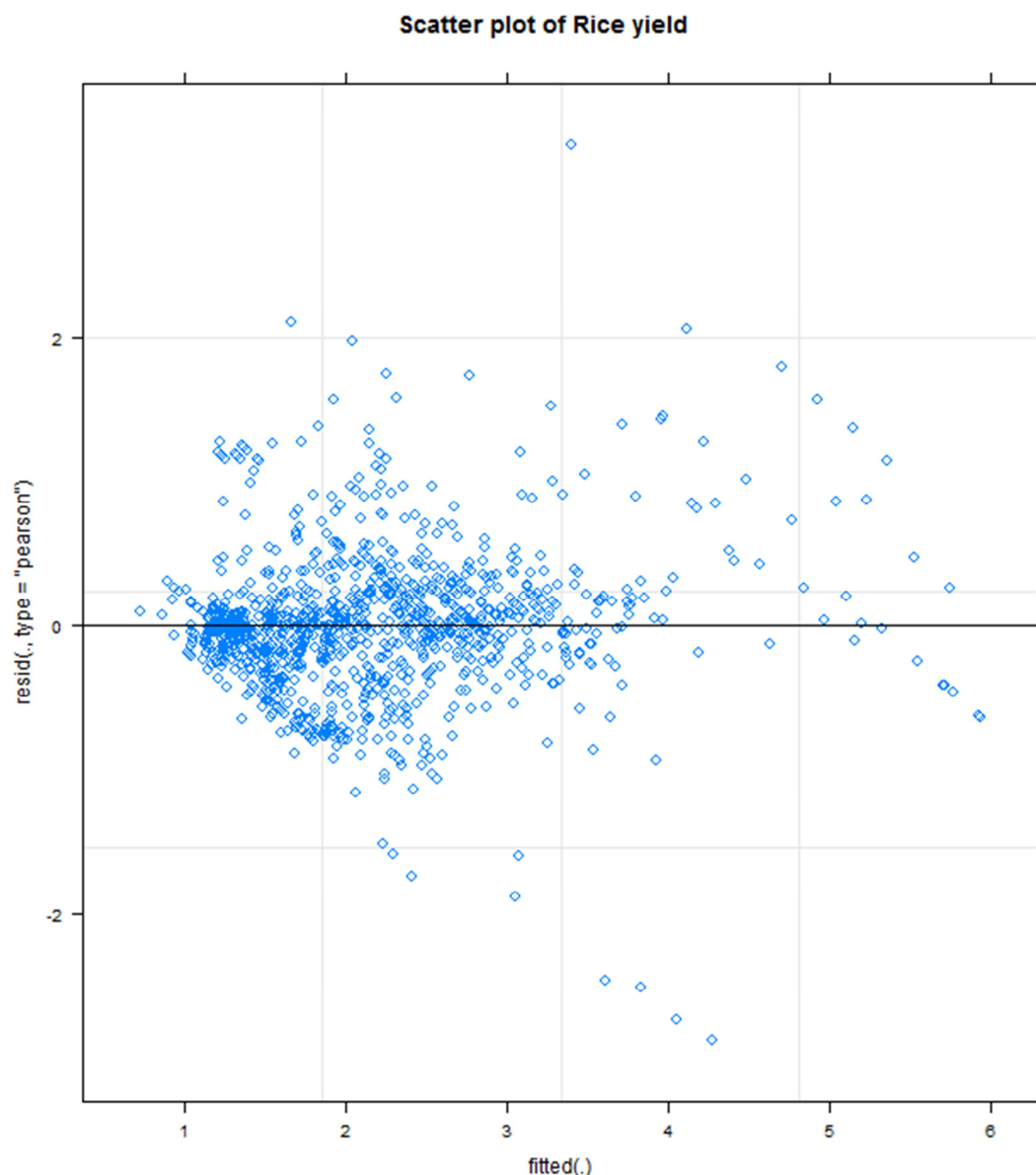


Figure 2. Diagnostic plots for fitted Linear Mixed model.

Since it is clear that the LM model is a good for modeling cereal crop yields data, we continue to estimates random and fixed effects parameters of the model.

4.3. Random Effects Estimates

The first part of the output that we would like to discuss is

the random effect estimates. The random effects (random intercepts) are random values associated with the levels of a random factor (District) in the LMM. These values which are specific to a given level of a random factor District represent random deviations from a given district from overall fixed effects.

Table 4. Random effectsestimates.

Random effects of maize				Random effects of rice			
Groups Name	Par.	Variance	Std.Dev.	Groups Name	Par.	Variance	Std.Dev.
DISTRICT (Intercept)	(b)	0.5885	0.2426	DISTRICT (Intercept)	(b)	0.1911	0.4371
Residual	(ξ)	0.1314	0.3212	Residual	(ξ)	0.2668	0.5165
95% Conf. Interval	Lower	Est.	Upper	95% Conf. Interval	Lower	Est.	Upper
	0.220	0.251	0.285		0.542	0.609	0.685
Number of obs: 1482, groups:DISTRICT, 241				Number of obs: 1195, groups:DISTRICT, 241			

The standard deviation column measures the variability for each random effect that we added to the model. As we can see from the table, maize has much less variability (24.26%) than rice (43.71%). The variance column measures the variability that is explained by our model. We observed that our model explains 58.85% of variability in maize yields and 19.11% of rice yields across various districts found in the ten regions of Ghana. This suggests that different regions have different variations in the yields of the two major cereal crops.

Of course, not all the variability was accounted for by our model and that is indicated in the residuals which stand for variability that are not due to District. This is our “ ξ ”, again, the “random” deviations from the predicted values that are not due to District. This value is 32.12% and 51.65% for maize and rice yields respectively. It is believed that there is other variability from unknown covariates which also contribute to deviations in yields of these cereal crops.

4.4. Fixed Effects Parameter Estimates

This show a portion of the model output that contain the fixed effect parameter estimates, their corresponding standard errors, the degrees of freedom, the t-test values and the corresponding p-values. From table 5 the intercept

(=1.37) represents the estimated average maize yields for 10 years period in the reference category of Ashanti Region. The value reported for Central Region represents estimated difference between the average values for 10 years maize yields in Central Region versus the reference category. In this case, the estimate is negative (-0.49), which means that the average maize yields for Central Region is 0.49 Mt/Ha lower than yields in the reference category. Similarly, the estimated average maize yields of Brong Ahafo Region is positive (0.71) which means that the average yield of maize for Brong Ahafo Region for the 10 years period is 0.71Mt/Ha higher than the yields in Ashanti Region, the reference category.

The average yields of six regions are negative and significant, which suggests a trend in maize yields of these regions that is consistently decelerating. Only two of the regions (Brong Ahafo and Western) aside the reference region (Ashanti) has positive average fixed effect values which indicate stable improvement in maize yields for the period under study. The value of the year is positive and not significant that indicates insignificant effects of years on the yields of maize in the regions. The interaction between the regions and year produced similar results.

Table 5. Fixed effects parameter estimates.

Maize	Parameter	Estimates	Std.Error	DF	t-value	p-value
(Intercept)	β_0	1.3717652	0.07036351	1231	19.495406	0.0000***
REGIONBRONG AH	β_1	0.7114422	0.10177370	231	6.990433	0.0000***
REGIONCENTRAL	β_2	-0.4903036	0.11528284	231	-4.253050	0.0000***
REGIONEASTERN	β_3	-0.3068082	0.10128959	231	-3.029020	0.0027***
REGIONGREATER A	β_4	-0.5163818	0.12797532	231	-4.035011	0.0001***
REGIONNORTHERN	β_5	0.0169314	0.10067520	231	0.168178	0.8666
REGIONUPPER EAST	β_6	-0.3625924	0.11875902	231	-3.053178	0.0025***
REGIONUPPER WEST	β_7	-0.1489881	0.12956015	231	-1.149953	0.2514
REGIONVOLTA	β_8	-0.0605070	0.10997136	231	-0.550207	0.5827
REGIONWESTERN	β_9	0.0261694	0.11341834	231	0.230733	0.8177
YEAR	β_{10}	0.0045390	0.00796872	1231	0.569605	0.5690

Source: SRID of MOFA.

Table 6 present fixed effects parameter of rice yields. It can be observed that the reference region (Ashanti) has estimated average value of rice (-1.03) lower than all other regions. This value is significant suggesting continue decreasing of the yields of rice for the region. Other regions have estimated average value of rice yields higher than the reference region and significant, which suggest a trend in rice

yields that, is reliably increasing. This confirms the earlier assessment in figure 2 which indicated stable improvement in rice yields for all the regions. The value of year is positive and significant indicating the effects of years on the yields of rice in all the regions. Interaction between region and the year also revealed similar results.

Table 6. Fixed effects Parameter estimates.

Rice	Parameter	Estimates	Std.Error	DF	t-value	p-value
(Intercept)	β_0	-1.0349446	0.14509882	1231	-7.132688	0.0000***
REGIONBRONG AH.	β_1	0.4862897	0.21185939	231	2.295342	0.0226***
REGIONCENTRAL	β_2	0.3202962	0.23532551	231	1.361077	0.0048***
REGIONEASTERN	β_3	0.5833880	0.20834345	231	2.800126	0.0055***
REGIONGREATER A.	β_4	0.8772819	0.26037494	231	3.369303	0.0009***
REGIONNORTHERN	β_5	1.0327226	0.20976650	231	4.923201	0.0000***

Rice	Parameter	Estimates	Std.Error	DF	t-value	p-value
REGIONUPPER EAST	β_6	1.5809686	0.24136448	231	6.550130	0.0000***
REGIONUPPER WEST	β_7	0.1551816	0.27012671	231	0.574477	0.0062***
REGIONVOLTA	β_8	0.7140924	0.22778364	231	3.134959	0.3763
REGIONWESTERN	β_9	0.2051626	0.23146262	231	0.886375	0.0019***
YEAR	β_{10}	0.1867735	0.01420502	1231	13.148423	0.0000***

YEAR β_{10} 0.18677350.01420502123113.1484230.0000Source: SRID of MOFA.

5. Conclusion

This study uses data from Statistical Research and Information Department of Ministry of Food and Agriculture (MOFA) to analyze major cereal crop yields in Ghana. The paper was aimed at investigating whether significant differences exist in the crop yields and to determine an evolution of crop yields between the regions in Ghana. MANOVA and Linear Mixed Models (LMM) were used to assess these objectives in two major cereal crops (Maize and Rice) yields produce and consume in Ghana. The Restricted Maximum Likelihood Method was employed to select best fitted LMM to estimates random and fixed effects parameters. The diagnostic plots indicate that, the LMM selected is good for determine evolution in cereal crop yields. It was found that significant differences exist in maize and rice yields across the ten (10) regions in Ghana. Descriptive statistics revealed that within the ten years period considered for the study, the highest average maize yields of 9.56 metric tons per hectare (Mt/Ha) was recorded and the lowest average maize yields recorded was 0.13 metric tons per hectare (Mt/Ha). Similarly, the highest rice yields recorded was 6.72 metric tons per hectare (Mt/Ha) and the lowest rice yields recorded was 0.68 metric tons per hectare (Mt/Ha).

The study also confirms that there is variability in cereal crop yields of maize and rice between and within the regions with maize yields having much less variability than rice yields. It reveals a trend that is decelerating in maize yields among majority of the regions and steadily increasing trend in rice yields in all the regions with exception of Ashanti region (reference category). Against the popular believe that maize is number one cereal crop produced and consumed in Ghana and hence may have higher average yields, the study generally indicated that, the yields of rice are averagely higher than the yields of maize from 2005 to 2014.

Based on these findings, we recommended that more support must be given to farmers who engage in production of cereal crops in Ghana to enable them improve upon their production across all the regions in the country. Maize production should be given more attention likewise rice production which was given a boost due to National Rice Development Strategy (NRDS) programme in 2009.

We therefore conclude by declaring that Joint Models that may consider more crops and other factors such as rainfall and climate data which may influence crop yields should be employed to investigate cereal crop yields in Ghana.

References

- [1] M. W. Rosegrant, and S. A. Cline, "Global food security: challenges and policies," *Science*, vol. 302, no. 5652, pp. 1917-1919, 2003.
- [2] OECD, and FAO, "OECD-FAO Agricultural Outlook 2009-2018, OECD," 2009.
- [3] J. Premanandh, "Factors affecting food security and contribution of modern technologies in food sustainability," *Journal of the Science of Food and Agriculture*, vol. 91, no. 15, pp. 2707-2714, 2011.
- [4] J. Foley, "Can We Feed the World and Sustain the Planet? An Overview of the Issues."
- [5] W. Bank, "Food Price Watch April 2011. Poverty Reduction and Equity Group, Poverty Reduction and Economic Management (PREM) Network," 2011.
- [6] J. Foley, "Can We Feed the World and Sustain the Planet? An Overview of the Issues." p. 01.
- [7] J. Kearney, "Food consumption trends and drivers," *Philosophical transactions of the royal society B: biological sciences*, vol. 365, no. 1554, pp. 2793-2807, 2010.
- [8] M. Tran, "Africa can feed the world? The Guardian [online]. London," 2011.
- [9] J. Beddington, "Food security: contributions from science to a new and greener revolution," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1537, pp. 61-71, 2010.
- [10] D. Griggs, M. Stafford-Smith, O. Gaffney, J. Rockström, M. C. Öhman, and I. Noble, "Policy: Sustainable development goals for people and planet," *Nature*, vol. 495, no. 7441, pp. 305-307, 2013.
- [11] Agyare, W. Agyei, I. K. Asare, Sogbedji, Jean Clottey, and V. Attuquaye, "Challenges to maize fertilization in the forest and transition zones of Ghana," *African Journal of Agricultural Research*, vol. 9, no. 6, pp. 593-602, 2014.
- [12] M. SRID, "National Crop production estimates 2002-2006," *Statistical Research and Information Department, Ministry of Food and Agriculture*, 2007.
- [13] U. Verma, H. Piepho, K. Hartung, J. Ogutu, C. Connell, and A. Goyal, "Linear Mixed Modeling for Mustard Yield Prediction in Haryana State (India)," *Issues*, vol. 1, no. 1, 2014.
- [14] M. J. Roberts, and J. B. Tack, "A Mixed Effects Model of Crop Yields for Purposes of Premium Determination."

- [15] C. Brownie, D. Larry, and J. Tina, "Longitudinal and Spatial Analyses Applied to Corn Yield Data from a Long-Term Rotation Trial-Institute of Statistics Mimeo Series No. 2559," 1993.
- [16] P. Rowhani, Lobell, D. B, M. Linderman, Ramankutty, and Navin, "Climate variability and crop production in Tanzania," *Agricultural and Forest Meteorology*, vol. 151, no. 4, pp. 449-460, 2011.
- [17] Pinheiro, Jose, D. Bates, S. DebRoy, Sarkar, and Deepayan, "Linear and nonlinear mixed effects models," *R package version*, vol. 3, pp. 57, 2007.
- [18] A. F. Zuur, "AED: data files used in mixed effects models and extensions in ecology with R," *R package version*, vol. 1, no. 9, 2010.
- [19] A. A. Isaac, "Multiple Comparison And Random Effect Model On Cocoa Production In Ghana (From 1969/70 To 2010/11 Production Years)," College of Science Department of Mathematics. A Thesis Submitted to the department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi, 2012.
- [20] A. R. Azinu, "Evaluation of Hybrid Maize Varieties in Three Agro-Ecological Zones in Ghana," University of Ghana, 2014.
- [21] G. Casella, and R. L. Berger, *Statistical inference*: Duxbury Pacific Grove, CA, 2002.
- [22] J.-T. Zhang, and S. Xiao, "A note on the modified two-way MANOVA tests," *Statistics & Probability Letters*, vol. 82, no. 3, pp. 519-527, 2012.
- [23] G. A. Marcoulides, and S. L. Hershberger, *Multivariate statistical methods: A first course*: Psychology Press, 2014.
- [24] R. A. Johnson, and D. Wichern, "Multivariate analysis," *Wiley StatsRef: Statistics Reference Online*, pp. 1-20, 2014.
- [25] R. A. Johnson, and D. W. Wichern, *Applied multivariate statistical analysis*: Prentice hall Englewood Cliffs, NJ, 2007.
- [26] M. Crowder, *Analysis of repeated measures*: Routledge, 2017.
- [27] K. ESKOG, "Analysis of Covariance Structures." p. 263.
- [28] S. N. Roy, "On a heuristic method of test construction and its use in multivariate analysis," *The Annals of Mathematical Statistics*, pp. 220-238, 1953.
- [29] B. T. West, K. B. Welch, and A. T. Galecki, *Linear mixed models: a practical guide using statistical software*: CRC Press, 2014.
- [30] J. J. Faraway, *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*: CRC press, 2016.
- [31] H. Akaike, "Likelihood of a model and information criteria," *Journal of econometrics*, vol. 16, no. 1, pp. 3-14, 1981.
- [32] K. P. Burnham, and D. R. Anderson, "Multimodel inference understanding AIC and BIC in model selection," *Sociological methods & research*, vol. 33, no. 2, pp. 261-304, 2004.
- [33] L. Davies, "Data Analysis and Approximate Models," *Monographs on Statistics and Applied Probability*, vol. 133, 2014.
- [34] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461-464, 1978.
- [35] D. Bates, and J. Pinheiro, *Mixed-effects models in S and S-PLUS*: Springer Science & Business Media, 2006.
- [36] Pinheiro, Céline, B. Sousa, A. Albergaria, J. Paredes, R. Duflath, D. Vieira, F. C. Schmitt, and F. Baltazar, "GLUT1 and CAIX expression profiles in breast cancer correlate with adverse prognostic factors and MCT1 overexpression," *Histology and histopathology*, vol. 26, no. 10, pp. 1279-1286, 2011.