Application of Response Surface Methodology in Modelling a Farm Production Process in the Presence of Random-Effects

Joel Cheruiyot Chelule

A thesis submitted in fulfilment for the Degree of Doctor of Philosophy in Applied Statistics in the Jomo Kenyatta University of Agriculture and Technology

2012

DECLARATION

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

Signature: ··························· Date: ··················

Joel Cheruiyot Chelule

This thesis has been submitted for examination with our approval as University Supervisors.

Signature: ··························· Date: ··················

Dr. George Otieno Orwa

JKUAT, Kenya

Signature: ······························ Date: ·················

Dr. Ronald Waweru Mwangi

JKUAT, Kenya

DEDICATION

To my dear mother Mary, beloved wife Divinah, cherished daughter Victoria and magical son Felix for their support, care, love and encouragement.

ACKNOWLEDGEMENT

The beginning of this work was difficult and the end looked uncertain. However, I did not despair and after a long struggle, the enormous task is finally accomplished. Admittedly, it was not possible to do this work all alone. I therefore wish to acknowledge the contributions of some personalities.

First and foremost, I thank the Almighty God for offering me good health, wisdom and strength during the entire period of this work.

My profound gratitude goes to my Principal Supervisor, Dr. George Otieno Orwa, for his able guidance and great humility displayed throughout the research. He offered me high quality, professional and academic guidance, and also made available, a number of reference materials. He always counseled and encouraged me whenever the journey seemed uncertain. He was always available for consultations, even at odd hours. His comments have immensely contributed to the development of this final document. I lack enough words to appreciate his assistance and no gift can repay what he has offered me. But one thing I can do is to request my Father in heaven to bless him, together with his family, abundantly.

My sincere appreciation extends to my Second Supervisor, Dr. Ronald Waweru Mwangi, who, despite his busy office, was able to go through my work and provide very constructive comments that led to the improvement of this document.

I cannot forget to acknowledge the efforts made by my colleague and friend, Dr. Joseph Kyalo Mung'atu, who trained me in data analysis. May Almighty God bless

him and his family.

I am greatly indebted to JKUAT Training Committee and the University Management for the financial assistance, and the Department of Statistics and Actuarial Science for according me humble time to carry out my studies. In particular, I wish to salute Prof. Romanus Odhiambo, for tirelessly encouraging me throughout my studies. Whenever we meet or talk on phone, he would always ask for progress in my studies.

Special appreciation goes to my beloved wife, Divinah, for her prayers, moral support, care and a lot of understanding during my research. My mother, Mary, has always played a crucial role in facilitating my education, particularly at the early stages. Without her positive efforts, I would not have received any meaningful education. I salute her!

Last, I appreciate the maize farmers in Eldoret East District, Kenya. They generously gave me the data I needed when I, together with my research assistants, visited them at their homes. Thanks too to my research assistants.

MAY GOD BLESS YOU ALL!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS

RSM	Response Surface Modelling
RRSM	Random-effects Response Surface Model
MCMC	Monte Carlo Markov Chain
NCPB	National Cereals and Produce Board
M-H	Metropolis-Hastings algorithm
KNBS	Kenya National Bureau of Statistics

ABSTRACT

Response Surface Methodology (RSM) for several explanatory variables and one response variable in the presence of random-effects was considered. In past studies, an assumption of non-randomness for explanatory variables was made. However, emerging situations reveal that randomness of explanatory variables is an aspect worth consideration (Kipchumba, 2008). In this thesis, a Random-effects Response Surface Model (RRSM) which is applicable to such situations is developed. The Bayesian approach is used to estimate the RRSM. A simulation study was undertaken to test the practicability of the RRSM and later, real data on maize farming in Eldoret East District of Kenya was used. WinBUGS and R Statistical Programming Packages were used in analysis. The simulation results showed that RRSM clearly reveals the randomness of explanatory variables when modeling a farm production process in the presence of random effects. When real data was used, RRSM similarly revealed the randomness in the real data.

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Response Surface Methodology (RSM) is an important subject in the statistical design of experiments. It is widely used in many fields such as Industrial, Biological, Clinical, Social, Food, Engineering, Agricultural sciences, amongst others. It is a tool in statistical analysis of experiments in cases where the yield is believed to be influenced by one or more controllable factors. The method explores relationships between several explanatory variables and one or more response variable(s). The main idea underlying RSM is to use a sequence of designed experiments to obtain an optimal response.

RSM was introduced by Box and Wilson (1951). They were motivated by the need to run experiments efficiently through a proper choice of design, and to determine operating conditions on a set of controllable variables that give rise to an optimal response.

The most extensive applications of RSM are in particular situations where several input variables potentially influence some performance measure or the quality characteristics of the process. RSM consists of the experimental strategy for exploring the space of the process or independent variables, empirical statistical modeling to develop an appropriate relationship between the yield and the process variables, and optimization methods for finding the values of the process variables that produce desirable responses.

As an illustration, consider a case whereby the growth of a plant is affected by certain amounts of water and sunshine. Denote these two treatments by X_1 and X_2 respectively. The plant can grow under any combination of these treatments which vary continuously.

When treatments are from a continuous range of values, the true relationship between the response and the explanatory variables may be difficult to know. However, the RSM provides a good approximation of such relationships. In the foregoing illustration, if we denote the plant growth by Y and consider it as the response variable, then it is a function of X_1 and X_2 . This relationship can be expressed mathematically as;

$$y = f\left(x_1,\, x_2\right) + e \tag{1.1}$$

where e in equation (1.1) is the experimental error which represents any measurement error on the response as well as other types of variations not explained by the variation in the explanatory variables. This implies that there are cases in which variation in the response variable Y is not fully explained by the total variation in the explanatory variables.

Since RSM finds its base in statistical models, which are usually approximations to reality, both the RSM and its parameters are subject to uncertainty. This means that an estimated point that is taken as optimum may not necessarily be optimal in reality due to estimation errors.

The explanatory variables may usually include main effects only or sometimes main effects plus their interactions if they exist. Quadratic and possibly cubic terms may also exist in the main effects to account for any curvature. The implication of this is that the response variable may be a function of the main effects only or main effects and their interactions.

When dealing with continuous treatments, the first goal for RSM is usually to find the optimum response. In a case of more than one response, we find the compromise optimum. If there are some prevailing constraints on the design data, the experimental design has to satisfy those constraints. The second goal is to understand how the response changes in a given direction by adjusting the design variables.

In most RSM problems, the true response function $f(.)$ is unknown and has to be approximated. In the approximation of $f(.)$, it is best to start with a low-order polynomial in some small region and proceed to a subsequent higher order polynomials if the problem has not been solved. This is usually the case when there is a curvature in the response surface. In general, all RSM problems use either one or a mixture of these models. In any model, the levels of each factor are independent of the levels of other factors.

Further, in order to get the most efficient results in the approximation of polynomials, proper experimental designs must be used to collect data. Once the data are collected, an appropriate technique of estimation must be used to estimate the parameters in the polynomial. Otherwise, fitting of the RSM can be tedious.

RSM has an effective track-record of helping researchers improve products and services. For example, Box and Wilson (1951) original RSM enabled chemical engineers to improve a process that had been stuck at a saddle point. Their design reduced the cost of experimentation and enabled a quadratic model to be fitted.

There are several applications of RSM in real life situations. In this thesis, we present an application of the RSM to model a farm production process in the presence of random effects. In particular, we consider maize farming in Eldoret East District, Kenya. The motivation behind this consideration is that maize is considered a staple food by many people yet there has been a general decline in production in the recent past. As a result, food insecurity has become a major concern to majority of people and the government. Indeed, one of the key aspects in the Kenya 2030 vision is food security.

1.2 Statement of the Problem

Response Surface Methodology (RSM) has been widely used in optimizing processes of designed experiments. The main objective of RSM is to summarize relationships between several explanatory variables and one or more response variable(s) through a mathematical model and thereafter, optimize the response variable. In the resulting models, some assumptions are usually made. A main assumption is that of non-randomness of the variables. However, emerging situations reveal that randomness is an aspect that should not be ignored (Kipchumba, 2008). Therefore, formulating a Random-effects Response Surface Model (RRSM) is inevitable.

1.3 Objectives of the Study

1.3.1 General Objective

To apply RSM to farm production processes in the presence of random effects.

1.3.2 Specific Objectives

1. To propose a model which incorporates randomness of independent variables, using RSM.
2. To estimate the proposed model using Bayesian approach.
3. To simulate the proposed model.
4. To apply the proposed model in solving a current problem of Maize production in Eldoret East District of Kenya.

1.4 Significance of the Study

There are many emerging real life situations, particularly in the area of production in agriculture, that require application of RSM in finding optimality of the yield. In most of these situations, the explanatory variables are practically random in nature. Therefore, in order to improve optimization of the processes, it is important to take into consideration the randomness, whenever it exists, of the explanatory variables.

In the application of this research, we use RSM considering the random nature of inputs in a maize farming process. In doing so, we develop a procedure for maximizing output per unit of land, while minimizing the costs involved. An optimal amount of each input to be used, within specified limits, is determined.

Maize is a staple food crop in Eldoret East District. This district is a major maize producing zone of Kenya. There is however a gap between production and consumption of this commodity. Implementation of this study is therefore in line with Vision 2030 of the Kenyan government, in which agriculture is listed as one of the key sectors of her economy that requires improvement.

1.5 Organization of the Thesis

The rest of this thesis is organized as follows; Chapter two presents a review of literature relating to the objectives. In chapter three, we discuss our methodology in which we first review RSM and thereafter, we propose a random-effects response surface model and discuss its estimation. In chapter four, we perform simulations of the proposed model and discuss the results. Chapter five has an empirical study in which we have data collection including choice of a sampling technique, data collection tools, and strategy of determining the sample size. The statistical analysis of the primary data collected and comprehensive discussion of the results is also undertaken in the same chapter. Finally, chapter six has the conclusion and recommendations for further research.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In this chapter, a review of literature related to our objectives is presented. First a review on advances related to each objective is provided, followed by some selected information on maize farming.

2.2 Response Surface Methodology (RSM)

RSM is a collection of mathematical and statistical techniques useful for modeling and analysis of problems in which a response of interest is influenced by several explanatory or design variables. It can also be defined as a collection of statistical and mathematical techniques useful for developing, improving, and optimizing processes. See Montgomery (2005) and Arap Koskei (2010).

The methodology was formally introduced and described by Box and Wilson (1951) who suggested the use of a first-degree polynomial to approximate a response variable, Hill and Hunter (1966), Arap Koskei (2010). They acknowledged that this methodology is an approximation procedure which is easy to use even when little is known about the process. However, available literature shows that some work had been done related to RSM prior to Box and Wilson (1951). For instance, see Mead and Pike (1975) in which several pre-Box era deliberations dating back to 1930s are provided.

During the pre-Box era, some important contributions were made to three main fields of applications. Foremost, response curves were extensively used as growth curves for studies on animals and plants. Empirical approaches to choosing a model were illustrated, Wishart (1938)and Wishart (1939). In this illustration, orthogonal polynomials are used for studying growth rates in nutrition studies of pigs. The functional model approach is also illustrated using the Gombertz curve, Winsor (1932). The logistic curve for growth studies is also proposed by, Reed and Berkson (1929).

Secondly, response curves were used in Probit Analysis. These led to modern probit analysis, Gaddum (1933), Bliss (935a). The third area where response curves were used and in which, for the first time, response surfaces were considered, was the agronomic study of the response of crop yield to fertilizer levels and crop spacing. In this study, an asymptotic relationship between plant yield and the supply of an essential growth factor was found to be biologically reasonable, Mitscherlich (1930). Later, Crowther and Yates (1941)improved the work of Mitscherlich (1930)using the proposed response equations to illustrate the response of arable crops to several different fertilizers.

Important developments of Optimal Design Theory in the field of experimental designs emerged following World War II, Myers et al. (1989), Chernoff (1953), Kiefer (1959), Kiefer and Wolfowitz (1952),Kiefer and Wolfowitz (1959),andKiefer and Wolfowitz (1960).

Box and Draper (1963) and later Peterson (1990) considered the second order response surfaces. Arap Koskei (1984)worked out fourth order rotatable designs. Njui (1985)extended this work to fifth order rotatable designs. Kosgei (2006) explored some important aspects of response surface designs with emphasis on those having the property of rotatability. This is a desirable quality of an experimental design.

In all the above studies, an important concern is whether the system contains a maximum or a minimum or a saddle point points which are of great interest to industry. RSM is therefore becoming increasingly useful. In recent years, more emphasis has been placed on RSM by the chemical and processing fields. Consequently, application and development of RSM continues to find use in many areas of research.

2.3 Bayesian Model Estimation

Bayesian estimation is based on Bayes' theorem. The theorem, also known as Bayes' law or Bayes' rule, is named in honour of Reverend Thomas Bayes (1701-1761), who first suggested using it to update prior beliefs in light of emerging evidence. However, his work did not gain popularity until the ideas were independently rediscovered and further developed by Laplace (1812), who first published the modern formulation.

Until late 20th century, the Bayesian approach was not accepted by Mathematicians who generally held frequentist views claiming that Bayesian is an unscientific method. Interestingly, it is currently widely accepted due to many examples of successful applications. As a result, many developments regarding Bayesian estimation have taken place. Monte Carlo Markov Chain (MCMC) sampling algorithms; the Metropolis, Metropolis-Hastings and the Gibbs sampler, have been developed, in that order. These algorithms are good especially for complex posterior distributions. Gilks et al (1993) show how to use Gibbs sampling to estimate a model. They illustrate the methodology with an analysis of long-term response to hepaptitis B vaccination, and demonstrate that the methodology can be easily and effectively extended to deal with censoring in the dependent variable.

2.4 Simulation Studies on RSM

Simulation is imitation of some real thing, or state of affairs, or a process. It entails representing certain key characteristics or behaviors of a selected physical or abstract system. Many simulation studies have been undertaken in RSM and decision makers are increasingly using simulations in their analysis where the aim is to determine the optimum combination of factors and/or to investigate relationship between response variable and explanatory variables, Hossein and Thornton (1984).

Baysal (2008) considered a situation in which one may wish to evaluate the distribution of profit and loss resulting from a dynamic trading strategy. In this case, a straight forward method is to simulate thousands of paths (i.e. time series) of relevant financial variables and track the profit and loss at every time at which the trading strategy rebalances its portfolio. In many cases, this requires numerical computation of portfolio weights at every rebalancing time on every path resulting in millions of simulations to compute portfolio weights, which is expensive. They show that RSM enables an efficient simulation procedure with reduced number of simulations, by modelling portfolio weights as a function of underlying financial variables.

There seems to be no established code of practice for the automated application of RSM in the field of simulation optimization, Nicolai and Dekker (2009). They aim to find the best settings for an automated RSM procedure especially when there is little information about the objective function. They present a framework of the RSM procedure for finding optimal solutions in the presence of noise, and compare various versions of the RSM algorithms on a number of test functions which include a simulation for cancer screening.

However, despite many efforts to encourage the application of RSM to simulation, this is yet to receive much attention and respect from practitioners and academicians, Hossein and Thornton (1984). They attempt to stimulate greater awareness on RSM and its associated experimental designs, as they relate to simulations.

2.5 Maize Farming and Causes of General Decline in Production

Since maize is a staple food crop for many people, extensive research has been done towards establishing the factors that affect its production, how to optimize the effect of such factors and hence optimize its yield. These efforts have been motivated by the need to find a lasting solution to food insecurity.

Kipchumba (2008) observed that an increasing number of food crop farmers in Uasin Gishu District of Kenya are abandoning maize farming in favour of small scale businesses and horticulture. He views this as a move that is likely to threaten food security because the district is one of those regions that constitute the bread basket of the country Kenya. He further established that the shift is due to relatively low returns mainly because of high cost of inputs and delays in payment from marketing bodies like the National Cereals and Produce Board (NCPB).

Uasin Gishu District is a major food-producing district of Kenya but over the recent years there has been a general decline of crop yields in the district, Cleopas et al. (2007). In their research, they observed that one of the factors affecting agricultural output is the level of mechanization. They found that there was stagnation in the level of agricultural mechanization in the district, which has contributed to the decline in crop yields.

Over the years, it has been established that mechanization is a main factor that contributes to increase in agricultural production. For example, in the United States, in 1950 one farm-worker produced enough food to support three other people while in 1970, one farm-worker supported 11 people, Wennblom, (1974). Additionally, improvements in crop varieties and use of chemicals/fertilizers and pesticides also contributed to increased production, Cleopas et al. (2007).

Any machinery will require repair at some stage of its life. Since manufacturers are usually far from farms, the farmers take the machines to the nearest workshop for repair. Most of these workshops are poorly-equipped and hence poor workmanship, Cleopas et al. (2007).

Increased agricultural output has also been attributed to timely weeding which can be achieved by use of agrochemicals using accurately calibrated machinery. Timely planting and application of fertilizers also contribute to increased production, Rider and Dickey (1982), Ksiazek (1985). Timely harvesting reduces crop losses as well. Moreover, if the whole crop can be harvested, losses can be reduced. Machinery known as combines, which harvest the whole crop, thresh and clean the grain, have been developed, Metianu et al. (1983). However, most farmers in Uasin Gishu District lack such machines, and this contributes to the reduced agricultural output. Others avoid using the machinery and instead opt to use human labour in a bid to reduce the cost of production, Kipchumba (2008), DAO (2001).

Soil acidity is one of the factors limiting maize production in some parts of Kenya notably in Uasin Gishu Plateau, Mwangi, et al. In their research, they found that several techniques of correcting the problem such as liming, use of organic farmyard manure alone or in combination with inorganic fertilizers and use of non-acidifying fertilizers have been suggested. Several formulations of fertilizers that can be used as alternatives to Diamonium Phosphate (DAP) fertilizer which is said to aggravate the acidity problem have been produced. Results of their research indicated that soil acidity can be improved with the application of agricultural lime in some seasons. It also indicated that farmyard manure also improves soil PH but the change is not as instant as is for lime. However, soil acidity improvement through manure is more sustainable with time than use of lime.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

Response Surface Methodology (RSM) consists of experimental strategies for exploring the space of independent variables, empirical statistical modelling used in developing an appropriate relationship between the yield and the process variables, and optimization algorithms for finding the values of independent variables that produce desirable values of the response. In this set of techniques, the performance measure or quality characteristic is called the response or the dependent variable while the input variables are known as the independent or explanatory or predictor variables.

We focus our study on statistical modelling to develop an approximating model between the response and the independent variables. In general, the relationship is;

$$Y = f\left(\xi_1, \xi_2, ..., \xi_k\right) + \varepsilon \tag{3.1}$$

where Y in equation (3.1) is the response, $\xi_1, \xi_2, ..., \xi_k$ are the independent variables and ε is a term that represents other sources of variability not accounted for in the function f . This may include effects such as measurement errors on the response, background noises and even effects of other variables. It is treated as a statistical error that is normally distributed with mean 0 and variance σ^2 i.e. $N\left(0, \sigma^2\right)$.

Consequently;

$$E\left[Y\left|\xi_1,\xi_2,...,\xi_k\right.\right]=E\left[f\left(\xi_1,\xi_2,...,\xi_k\right)\right]+E\left[\varepsilon|\xi_1,\xi_2,...,\xi_k\right]=E\left[f\left(\xi_1,\xi_2,...,\xi_k\right)\right] \tag{3.2}$$

The variables $\xi_1,\xi_2,...,\xi_k$ in equation(3.2) are called natural variables because they are expressed in the natural units in which the measurements being studied were made. In RSM, it is convenient to transform these natural variables into coded variables, say $X_1,X_2,...,X_k$, which are dimensionless with mean zero and same standard deviation. Accordingly, in terms of the coded variables, the response function can be written as;

$$Y=f\left(x_1,x_2,...,x_k\right)+\varepsilon \tag{3.3}$$

Where ε in equation (3.3)are random variables called error terms which are assumed to be identically and independently distributed, independent of X and normally distributed with zero mathematical expectation i.e. $E\left(\varepsilon\right)=0$, and constant and finite variance i.e.$Var\left(\varepsilon\right)=\sigma^2<\infty$. The explanatory variables X are assumed to be non-random.

Since the true response function f is unknown, it is approximated. In its approximation, the efficiency of the estimation procedure depends on the ability to develop a suitable approximation for this function. This tenability of an efficient approximation is usually the focus of RSM.

3.2 Types of RSM

There are basically three types of RSMs; the first-order, the second-order and the fractional factorial-order.

3.2.1 First-Order RSM

Let the response be defined by a linear function of independent variables. Then, the approximating response function is known as a first-order RSM. This is appropriate when one is interested in estimating the true response from a small region of an independent variable space where there is little curvature in the response function f .

A first-order RSM with two explanatory variables in terms of the coded variables can be expressed as;

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{3.4}$$

where β_0 is the intercept, and β_1 and β_2 are the regression coefficients for the independent variables X_1 and X_2 , respectively. The form of the first-order model in is also referred to as the main effects model, because it includes only the main effects of the two independent variables X_1 and X_2 . If there is an interaction between these variables, the model takes the form;

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon \tag{3.5}$$

The interaction term introduces curvature into the response function.

A first-order RSM with N experimental runs carried out on q input variables and a single response Y therefore resulting in N observations can be expressed as;

Since in this case, the response variable Y is a function of the design variables $X_1, X_2, X_3, ..., X_q$ and the experimental error ε , this model is a Multiple Regression Model with $\beta_{j's}$ being the regression coefficients.

In equations (3.4), (3.5) and (**??**) , the error terms ε are assumed to be identically and independently distributed, independent of X and normally distributed with zero mathematical expectation i.e. $E\left(\varepsilon\right) = 0$, and constant and finite variance i.e. $Var\left(\varepsilon\right) = \sigma^2 < \infty$. The explanatory variables X are assumed to be non-random.

First-order RSMs are used to describe flat surfaces which may or may not be tilted. A shortcoming of these models is that they are not suitable for analyzing maximum, minimum and ridge lines. Approximation of the response function in them is also only reasonable when the response function itself is neither too curved nor too big. First-order models are assumed to be adequate approximations of the true surfaces in a small region of the design variables $X's$, Montgomery (2005).

In view of the fact that it is important to design an efficient model, estimation of variances is considered. The orthogonal first-order RSM minimizes the variance of the regression coefficients β_j . A first-order RSM is orthogonal if the off-diagonal elements of the information matrix are all zero.

3.2.2 Second-Order RSM

If the curvature in the true response surface is too strong, the first-order RSM will be inadequate. A second-order RSM is then considered.

For the case of two input variables X_1 and X_2 , the second-order RSM is;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \tag{3.6}$$

Accordingly, a second-order response surface with $q-$input variables and which involves all the possible terms i.e. main effects, interaction of the main effects and quadratic term, is represented by the polynomial equation,

$$Y = \beta_0 + \sum_{i=1}^{q} \beta_i X_i + \sum_{i=1}^{q} \beta_i X_i^2 + \sum_{i=1}^{q} \sum_{i'=1}^{q} \beta_{ii'} X_i X_i' + \varepsilon\,; \quad i \neq i' \tag{3.7}$$

In the polynomial equation , the variable Y represents the dependent or response variable whereas$X_i's$ represent the explanatory or the independent variables and ε represents the error term. β_0, β_i and $\beta_{ii'}$ are unknown constants also known as coefficients of the factors or regressors. They represent proportions at which the respective terms contribute to the response variable and are tested to show the significance of the respective terms in estimating the response variable. Since they are unknown, they are estimated using an appropriate statistical estimation technique.

For Nobservations, the second order response surface model for $q-$input variables takes the form;

$$Y_j = \beta_0 + \sum_{i=1}^{q} \beta_i X_i + \sum_{i=1}^{q} \beta_i X_i^2 + \sum_{i=1}^{q} \sum_{i'=1}^{q} \beta_{ii'} X_i X_i' + \varepsilon\,; \quad i \neq i' \tag{3.8}$$

where $j = 1, 2, ..., N$. In matrix notation, this equation can be expressed as;

$$Y_j\left(x\right) = X'\beta + \varepsilon_j \tag{3.9}$$

Equation (3.9) is known as a quadratic RSM. The quadratic RSMs are always sufficient for industrial applications almost surely.

In equation (3.9), the error terms ε_j are assumed to be identically and independently distributed, independent of X and normally distributed with zero mathematical expectation i.e. $E\left(\varepsilon\right) = 0$, and constant and finite variance σ^2 i.e. $Var\left(\varepsilon\right) = \sigma^2 < \infty$. The explanatory variables X are assumed to be non-random. Further, it is assumed that there exists interaction effect between two inputs and outputs are independent.

In some special circumstances, a model involving only main effects and interactions may be appropriate to describe a response surface when analysis of results reveals no evidence of pure quadratic curvature in the response variable. In other circumstances, a complete description of the process might require at least a quadratic model. However, it is rare that all of the terms are needed in an application.

The second-order RSM is widely used for several reasons; it is flexible in the sense that it can take on a wide range of functional forms, it is easy to estimate its parameters, and there is considerable practical experience indicating that it works well in solving real response surface problems.

3.2.3 Fractional Factorial RSM

A fractional factorial RSM is a factorial design with no run to completion of the full factorial design. Factorial designs are usually denoted by p^q where p refers to the number of levels for which the variates are observed with q being the number of factors being considered. For instance, the 3^q factorial RSM is a factorial arrangement with q factors, each at three levels.

The levels of a chosen factor are usually referred to as low, intermediate and high, represented by digits 0, 1 and 2 respectively. In a 3^3 factorial design, 0, 2, 1 indicates the treatment combination corresponding to, say, factor A at the low level, factor$B$ at the high level and factor$C$ at the intermediate level. When the measurements on the response variable contain all possible combinations of the levels of the factors, this type of experimental design is called a complete factorial design.

In general, a factorial design require many runs, therefore it is unlikely that the runs can be carried out under homogenous conditions. As a result, the confounding in blocks is unavoidable. A complete factorial experiment can be placed in a block of units, where units in the same block are homogenous. This type of design technique is called Confounding.

The complete blocks include every treatment in every block. On the contrary, the incomplete blocks do not include all the treatments or treatment combinations in each block. The incomplete blocks are less efficient than complete blocks due to the loss of some information, usually the high order interactions. However, confounded factorials will tolerate main effects and low-order interactions.

The 3^q design, for instance, can be confounded in 3^s blocks, each with $3^{(q-s)}$ units, where $q > s$. If say, $q = 3$ and $s = 2$, then,3^3 factorial design is confounded in $3^3 = 9$ incomplete blocks, each with $3^{3-2} = 3^1$ units. We then define a contrast by choosing a factorial effect to confound with blocks. The general defining contrast is given by;

$$Y_j = \beta_0 + \sum_{i=1}^{q} \beta_i X_i + \sum_{i=1}^{q} \beta_i X_i^2 + \sum_{i=1}^{q} \sum_{i'=1}^{q} \beta_{ii'} X_i X_i' + \varepsilon \, ; \quad i \neq i' \tag{3.10}$$

where α_i represents the exponents on the i^{th} factor in the effect to be confounded and X_i is the level of the i^{th} factor in a particular treatment combination (Montgomery, 2005). Therefore,X_i takes the values of 0 (low level), 1 (intermediate level), or 2 (high level), where α_i is 0, 1, or 2.

One important concern about 3^q design is that it can require a large number of runs even for moderate values of q . For example, a 3^9 design with a single replicate will have 19,683 observations. If the design is confounded in $3^{9-6} = 27$ incomplete blocks, then each block will require 27 observations. In this case, the fractional factorial design might be an alternative approach when dealing with a large number of factors.

Remark 1. In each of the types of RSM described above, aspects of multiple regression feature.

3.3 Multiple Regression Model

The relationship between a set of independent variables and the response variable Y is determined by a mathematical model called regression model. When there are more than two independent variables, the regression model is called multiple-regression model. In general, a first-order multiple linear regression model with q independent variables and N experimental runs or observations takes the form;

$$y_j = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q + \varepsilon_j; \quad j = 1, 2, ..., N \tag{3.11}$$

$$= \beta_0 + \sum_{i=1}^{q} \beta_i X_i + \varepsilon_i;\ i = 1, 2, ..., q; i = 1, 2, ..., q \tag{3.12}$$

While a second-order multiple-regression model with q input variables and N experimental runs that result in N corresponding observations takes the form;

$$Y_j = \beta_0 + \sum_{i=1}^{q} \beta_i X_i + \sum_{i=1}^{q} \beta_i X_i^2 + \sum_{i=1}^{q} \sum_{i'=1}^{q} \beta_{ii'} X_i X_i' + \varepsilon\ ; \quad i \neq i'; i = 1, 2, ..., q; j = 1, 2, ..., N \tag{3.13}$$

The parameter β_i measures the expected change in response Y per unit increase in X_i when the other independent variables are held constant.

A multiple-regression model can be written in matrix form as;

$$Y_j\left(x\right) = X'\beta + \varepsilon_j \tag{3.14}$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}_{nx1} \quad X = \begin{bmatrix} x_{11}\ x_{12} \ldots x_{1q} \\ x_{21}\ x_{22} \ldots x_{2q} \\ .\quad .\quad . \\ .\quad .\quad . \\ .\quad .\quad . \\ x_{n1}x_{n2} \ldots\ x_{nn} \end{bmatrix}_{nx1} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_q \end{bmatrix}_{kx1} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}_{nx1}$$

Y is an $(nx1)$ vector of observations, X is an (nxk) matrix of levels of independent variables, β is a $(kx1)$ vector of regression coefficients, and ε is an $(nx1)$ vector of random errors. If X is a (kxk) non-singular matrix, then the linear system $Y_j\left(x\right) = X'\beta + \varepsilon_j$ has a unique least squares solution given by $\widehat{\beta} = \left(X'X\right)^{-1} X'Y$. The data structure for the multiple-regression model can be displayed as shown in table 3.1 below

3.4 The Random-effects Response Surface Model (RRSM)

3.4.1 Introduction

In this section, we propose a Random-effects Response Surface Model (RRSM), in which we consider random effects that may come with the independent variables.

This means that the proposed model tackles both fixed and random effects. Accordingly, the proposed RRSM takes the form of a mixed effects model. Traditionally, mixed effects models are used in analysis of multilevel data structures. Our contribution is that the proposed model not only acts on multilevel data structures, but also on response surfaces. Notably, estimation of variance in traditional mixed effects models is done considering the multilevel structure in the population. In doing this, one computes the variation at the different levels of the exhibited hierarchy. In our case however, the variation is based on the response surface only.

3.4.2 Notations Used

Let Y denote the response variable of interest. We suppose that the population of our study is partitioned into strata with the assumption that each stratum has same inherent characteristics. Therefore, we let Y_{ijt} to represent the response from respondent i selected from stratum j at time t in years where $i = 1, 2, ..., n_j$, $j = 1, 2, ..., m$, $t = 1, 2, ..., T$. Let $Y^*_{i,j,t-1} = (Y_{i,j,t-1}, Y_{i,j,t-2}, ..., Y_{i,j,1})$ represent past responses from respondent i in stratum j for time $t = t - 1, t - 2, ..., 1$.

We suppose that there are p quantitative input factors denoted by

$X_{1,i,j} = (X_{1,1,i,j}, X_{1,2,i,j}, ..., X_{1,p,i,j})$ and q qualitative input factors denoted by $X_{2,i,j} = (X_{2,p+1,i,j}, X_{2,p+2,i,j}, ..., X_{2,p+q,i,j})$, that significantly contribute to the response variable so that $X_{i,j} = (X_{1,i,j}, X_{2,i,j})$ constitute a set of explanatory variables; both quantitative and qualitative. In matrix form, this can be written as;

$$X_{i,j} = (X_{1,i,j}, X_{2,i,j})^T = \begin{pmatrix} X_{1,1,1,1}\ X_{1,2,1,2}\ \ldots\ X_{1,p,1,m}\ X_{2,p+1,1,1}, ..., X_{2,p+q,1,m} \\ X_{1,1,2,1}, X_{1,2,2,2}, ..., X_{1,p,2,m}, X_{2,p+1,2,1}, ..., X_{2,p+q,2,m} \\ . \\ . \\ . \\ X_{1,1,n_1,1}, X_{1,2,n_2,2}, ..., X_{1,p,n_m,m}, X_{2,p+1,n_1,1}, ..., X_{2,p+q,n_m,m} \end{pmatrix}$$

For example,$X_{1,1,1,1}$ denotes the first quantitative factor applied by the first respondent in the first stratum, $X_{1,1,2,1}$ denotes the first quantitative factor applied by the second respondent in the first stratum, $X_{1,1,n_1,1}$ denotes the first quantitative factor applied by the n_1^{th} respondent in the first stratum, where n_1 is the total number of respondents interviewed in the first stratum (i.e. the sample size of stratum 1). Further, $X_{2,p+1,1,1}$ denotes the first qualitative factor applied by the first respondent in the first stratum, $X_{2,p+1,2,1}$ denotes the first qualitative factor applied by the second respondent in the first stratum, and so on.

Further, let Z_{ijt} and W_{ijt} represent population-specific and stratum-specific design vectors, respectively, whereby W_{ijt} is a sub vector of Z_{ijt}. The design vector Z_{ijt} and therefore W_{ijt} may depend on deterministic or stochastic covariates and on past responses, such that;

$$Z_{ijt} = Z\left(X_{ijt}, Y_{ijt-1}^{*}\right) \tag{3.15}$$

3.4.3 The Model

We develop our model in two phases as follows;

Phase 1: Fixed-effects Component

This model specifies effects that are constant across the whole population under study. Such effects may also be referred to as population-specific effects and the corresponding model is sometimes called population-averaged model. For each selected respondent i , we have;

$$Y_{ijt} = \tau + \beta Z_{ijt} + e_{ijt} \tag{3.16}$$

where β denotes a vector of unknown regression coefficients for the fixed-effect explanatory variables. Here, the intercept and the slope are considered to be fixed.

The error terms e_{ijt} are assumed to be identically and independently distributed, independent of Z_{ijt} and normally distributed with zero mathematical expectation i.e. $E\left(e_{ijt}\right) = 0$, and constant and finite variance i.e. $Var\left(e_{ijt}\right) = \sigma^2 < \infty$. The explanatory variables are assumed to be non-random.

Phase 2: Random-effects Component

In this phase, we foremost assume that the normal responses depend linearly on unknown population-specific effects β and on unknown stratum-specific intercepts τ_{ij} such that we have the model;

$$Y_{ijt} = \tau_{ij} + \beta Z_{ijt} + e_{ijt} \tag{3.17}$$

where e_{ijt} are the error terms assumed to be uncorrelated normal random variables such that $e_{ijt}\ N\left(0,\sigma^2\right)$. Y_{ijt} represents the response (yield) realized by respondent i in stratum j at time t; $t = 1, 2, ..., T_{ij}$; . $Z_{ijt} = \left(X_{ijt}, Y^*_{ijt-1}\right)$ are population-specific explanatory variables which depend on deterministic or stochastic covariatesX_{ijt} and past responses Y^*_{t-1} . The stratum-specific intercepts τ_{ij} are also assumed to vary amongst respondents within same stratum.

Next, we assume that the normal responses depend linearly on unknown stratum-specific intercepts τ_{ij} , population-specific effects β , stratum-specific effects α_{ijt} and the error term e_{ijt} such that we have the model;

$$Y_{ijt} = \tau_{ijt} + \beta Z_{ijt} + \alpha_{ijt} W_{ijt} + e_{ijt} \tag{3.18}$$

Here, the effects α_{ijt} are assumed to vary independently from one stratum to another according to a mixing distribution with zero mean. Since the errors are assumed to be Gaussian, a normal mixing density with unknown covariance matrix $\operatorname{cov}\left(\alpha_{ijt}\right) = Q$ is chosen such that;

$$\alpha_{ijt}\left(0,\, Q\right), \quad Q > 0 \tag{3.19}$$

and the sequences $\{\varepsilon_{ijt}\}$ and $\{\alpha_{ijt}\}$ are assumed to be independent.

Model (3.18) is now our proposed Random-effects Response Surface Model (RRSM).

3.4.4 Assumptions

In the proposed model (3.18), the following assumptions are made;

3.4.4.1 Assumptions on the Error Term

A_1 : e_{ijt} are independent and identically distributed (iid) random variables.

A_2: e_{ijt} , $e_{ijt} \neq e_{ijt'}$ are independent of the quantitative factors and the qualitative factors

A_3: The covariance between error terms in any two different observations

equals to zero i.e. $\text{cov}\,(e_{ijt}, e_{ijt'}) = 0$, $e_{ijt} \neq e_{ijt'}$,

A_4: e_{ijt} are normally distributed with mean 0 and variance σ^2 .

A_5: There is no curvature in the explanatory variables. This therefore implies that our RSM model is a multiple linear regression model with experimental variables comprising of the main effects and their interactions where they exist.

3.4.4.2 Assumptions on the Population under Study

A_1: The population is organized into non-overlapping strata of varying sizes $n_j; j = 1, 2, ..., m$

A_2: Each stratum has homogenous inherent characteristics and the units in

one stratum are independent of those in all the other strata.

3.4.5 Remarks on the Model

1. Using assumption we can rewrite our model as a multivariate heteroscedastic linear regression model;

$$Y_{ijt} = \tau_{ijt} + \beta Z_{ijt} + e^*_{ijt} \tag{3.20}$$

where the multivariate errors $e^*_{ijt} = \alpha_{ijt} w_{ijt} + e_{ijt}$ and;

$$e^*_{ijt}\ iid\ N\left(0, V_{ijt}\right); V_{ijt} = I\sigma^2_e + W_{ijt} Q W'_{ijt} \tag{3.21}$$

2. The model (3.19) is a general version of our proposed model. From it, special cases may be derived and used in analyzing varying intercepts and varying slopes, or stratum-specific effects. Below are examples of some of these cases.

Random Intercepts Model

In some empirical studies, stratum-specific intercepts that may significantly influence the response variable, in addition to the observed covariates, may exist. Stratum-specific effects do not cut across the entire population under study but are only specific to a portion of a population. To account for such effects, a Random-effects Model with stratum-specific intercept is appropriate. Such a model can be written as;

$$Y_{ijt} = \tau_{ijt} + \beta Z_{ijt} + e_{ijt} \tag{3.22}$$

where the slope coefficients β are taken to be constant and the intercepts τ_{ijt} are random and assumed to be *iid* with unknown parameters $E\left(\tau_{ijt}\right) = \tau$ and $Var\left(\tau_{ijt}\right) = \sigma^2$. The unobservable deviations between the population mean τ and the stratum-specific realizations τ_{ijt} may be interpreted as effects of omitted covariates.

Random Slopes Model

The Random-intercept Model (3.22) do not alleviate the restrictive assumption that the slope coefficients are equal for each observation. In most longitudinal studies, varying slope coefficients usually arise, where intercept and slope coefficients are specific to each stratum. Therefore, to take into account such parameter heterogeneity, the Random- intercept model (3.22) can be extended so that we treat, not only the intercept as random, but also the slope coefficients. Doing this, the corresponding model will have the form;

$$Y_{ijt} = \tau_{ijt} + \beta_{ijt} Z_{ijt} + e_{ijt} \tag{3.23}$$

Sometimes, however, the assumption that some coefficients are stratum-specific may be less realistic than the assumption that some coefficients are constant across strata (Ludwig and Gerhard, 1994). For instance, suppose β_{i1} denotes the stratum-specific coefficients and β_{i2} denotes the remaining coefficients which are constant across strata, then, the parameter vector β_i may be partitioned into $\beta_i = (\beta_{i1}, \beta_{i2})$ with $\beta_{i2} = \beta_2$ for all i . The design vector $Z_{ijt} = (1, Z_{ijt})$ also has to be rearranged according to the structure $Z_{ijt} = (Z_{ijt1}, Z_{ijt2})$. Then the probability model for β_i is a multivariate normal density with singular covariance matrix expressed as;

$$\beta_i = \begin{bmatrix} \beta_{i1} \\ \beta_{i2} \end{bmatrix} \left\{ \begin{bmatrix} \beta_{i1} \\ \beta_{i2} \end{bmatrix}, \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} \right\} \tag{3.24}$$

where the sub matrix Q is assumed to be positive definite. Due to the mixing of ‘fixed effects’ and ‘random effects’, models of this type are called Random-effects models.

3.4.6 Model Estimation

Estimation of our proposed RRSM (3.18) implies estimation of the parameters $\tau,\ \beta, \alpha$ and σ_e^2 of the model. Their estimation is often based on a frequentist or classical approach, where $\tau,\ \beta, \alpha$ and σ_e^2 are treated as ‘fixed’ parameters. However, we can also use Bayesian estimation, which is fundamentally different from the classical approach in the sense that the parameters to be estimated are treated as ‘random’ variables with suitable prior probability distribution. For this reason, we estimate our model using the Bayesian approach.

3.4.6.1 Review of the Bayesian Estimation

Let $Y_1, Y_2, ..., Y_n$ be a random sample taken from a population indexed by the parameter, say θ, and the prior distribution is updated or modified using the information from the sample. The resulting distribution (modified prior) is called the posterior distribution, and is assumed to contain all the relevant information about θ. The posterior Bayes estimator of θ is the mean of the posterior distribution.

Suppose $f(\theta)$ is the prior distribution of θ . Then, $f(\theta)$ expresses what is known about θ prior to observing $Y = Y_1, Y_2, ..., Y_n$. The first step in Bayesian estimation procedure is usually to decide on this prior. The second step involves deciding on the likelihood. Let $f(\theta)$ be the likelihood function of Y given θ. The likelihood describes the process giving rise to the data in terms of unknown θ. Accordingly,

$$f(y|\theta) = \frac{f(y,\theta)}{f(\theta)} \tag{3.25}$$

Let $f(\theta|y)$ be the posterior distribution. This expresses what is known about θ after observing Y. Thus,

$$f(\theta|y) = \frac{f(\theta,y)}{f(y)} \tag{3.26}$$

Deriving the posterior by applying the Bayes theorem, is usually the third step in Bayesian estimation. The fourth and the last step is deriving inference from the posterior.

From equation (3.25), we have that $f(y,\theta) = f(y|\theta) f(\theta)$. Substituting this in equation (3.26), leads to;

$$f(\theta|y) = \frac{f(y|\theta) f(\theta)}{f(y)} \tag{3.27}$$

But $f(y) = \int f(y,\theta) \, d\theta$. Therefore, equation (3.27) may be written as;

$$f\left(\theta|y\right)=\frac{f\left(y|\theta\right)f\left(\theta\right)}{\int f\left(y,\theta\right)d\theta}=\frac{f\left(y|\theta\right)f\left(\theta\right)}{\int f\left(y|\theta\right)f\left(\theta\right)d\theta} \tag{3.28}$$

Equation (3.28) represents the posterior distribution when no sampled observations are available. Suppose a random sample $Y_1, Y_2, ..., Y_n$ is taken from the random variable Y, then, we have;

$$f\left(\theta|y\right)=\frac{f\left(y|\theta\right)f\left(\theta\right)}{\int f\left(y|\theta\right)f\left(\theta\right)d\theta} \tag{3.29}$$

Assuming $Y_1, Y_2, ..., Y_n$ are independent observations on Y, then, equation (3.29) can be re-written as;

$$f\left(\theta|y\right)=\frac{\prod_{i=1}^{n}f\left(y_i|\theta\right)f\left(\theta\right)}{\int\prod_{i=1}^{n}f\left(y|\theta\right)f\left(\theta\right)d\theta} \tag{3.30}$$

Equation (3.30) now represents the posterior distribution when sampled observations are available. The posterior Bayes estimator of θ is the mean of equation (3.30). That is;

$$\hat{\theta}=E\left(\theta\right)=\int\theta f\left(\theta|x\right)d\theta \tag{3.31}$$

As an alternative to equation (3.31), we may use MCMC sampling algorithms. This class of algorithms is good especially for complex posterior distributions. It involves sampling from the posterior distribution using an appropriate MCMC sampling algorithm.

3.4.6.2 Review of Monte Carlo Markov Chain (MCMC) Sampling Methods

MCMC sampling methods are a class of algorithms for sampling the posterior distribution based on constructing a Markov chain that has the desired density as its limiting distribution. The idea behind MCMC sampling is to simulate a random walk in the space of parameters of interest, $\theta = (\theta_1, \theta_2, ..., \theta_d)'$, which converges to the joint posterior distribution $p(\theta|y)$. The samples are drawn sequentially and the distribution of the sampled draws depends on the last value drawn. The state of the chain after a large number of iterations is then used as a sample from the desired posterior distribution.

The Metropolis Algorithm

Given a target posterior distribution $p(\theta|y)$, the metropolis algorithm creates a sequence of random vectors $\left(\theta^{(1)}, \theta^{(2)}, ...\right)$ whose distribution converges to the target distribution . Each sequence can be considered as a random walk whose stationary distribution is $p(\theta|y)$. The algorithm proceeds as follows;

Start with some initial value θ^0 . For $t = 1, 2, ...$, obtain $\theta^{(t)}$ from $\theta^{(t-1)}$ using the following steps:

1. Sample a candidate point θ^* from a proposal distribution at time t , $q\left(\theta^*|\theta^{(t-1)}\right)$ .The proposal distribution must be symmetric; that is $q(\theta_a|\theta_b) = q(\theta_b|\theta_a)$, for all θ_a and θ_b .

2. Calculate the ratio of the densities;

$$r = \frac{p\left(\theta^*|y\right)}{p\left(\theta^{(t-1)}|y\right)} \tag{3.32}$$

3. Set;

$$\theta^{(t)} = \begin{cases} \theta^* \; with\; probability \; \min\left(r,1\right) \\ \theta^{(t-1)} \; otherwise \end{cases} \tag{3.33}$$

It is important to note that the algorithm requires the ability to draw θ^* from the proposal distribution $q\left(\theta^*|\theta\right)$ for all θ .

The Metropolis-Hastings (M-H) Algorithm

The M-H algorithm generalizes the metropolis algorithm in two ways. First, the proposal distribution q needs no longer to be symmetric. That is, there is no requirement that $q\left(\theta_a|\theta_b\right) = q\left(\theta_b|\theta_a\right)$. Secondly, to correct for the asymmetry in the proposal density, the acceptance ratio is now (Tierney, 1994; Chib and Greenberg, 1995; Gelman et al, 1995)

$$r = \frac{p\left(\theta^*|y\right) q\left(\theta^{(t-1)}|\theta^*\right)}{p\left(\theta^{(t-1)}|y\right) q\left(\theta^*|\theta^{(t-1)}\right)} \tag{3.34}$$

Gibbs Sampler

The Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990; Gilks, 1996) is a MCMC algorithm that has been found to be very useful in multidimensional problems. It is defined in terms of sub vectors of θ . At each iteration t, the

Gibbs sampler cycles through the sub vectors θ of θ_j , drawing from the conditional distribution given all the remaining components of θ:

$$p_j\left(\theta_j|\theta_{-j}^{(t-1)},y\right),$$

Where θ_{-j} represents all the components of θ ,

except for θ_j , i.e. $\theta_{-j} = (\theta_1, \theta_2, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_d)'$ This suggests the following MCMC scheme;

1. Generate $\theta_1^{(t)}$ from $p_1\left(\theta_1|\theta_2^{(t-1)},\theta_3^{(t-1)},...,\theta_d^{(t-1)},y\right)$
2. Generate $\theta_2^{(t)}$ from $p_2\left(\theta_2|\theta_1^{(t-1)},\theta_3^{(t-1)},...,\theta_d^{(t-1)},y\right)$

   .

   .

   .

3. Generate $\theta_d^{(t)}$ from $p_d\left(\theta_d|\theta_1^{(t-1)},\theta_3^{(t-1)},...,\theta_{d-1}^{(t-1)},y\right)$

At the completion of these steps, the vector $\theta^{(t)} = \left(\theta_1^{(t)},\theta_2^{(t)},...,\theta_d^{(t)}\right)'$ provides the simulated value of θ at the t^{th} iteration of sampling. The d steps of this Gibbs sampling scheme completes one iteration of the simulated method.

After a large number,T , of iterations, we obtain $\theta^{(T)}$. Geman and Geman(1984) show that under mild conditions, the joint distribution of $\theta^{(T)}$ converges at an exponential rate to $p(\theta|y)$ as $T \to \infty$. The desired joint posterior distribution, $p(\theta|y)$, can be approximated by the empirical distribution of Mvalues of $\theta^{(t)}$ for $t = T+1, T+2, ..., T+M$, where $T$is large enough so that the Gibbs sampler has converged and M is chosen to give sufficient precision to the empirical distribution of interest.

3.4.6.3 The Random-effects Response Surface Model in Bayesian Framework

In Bayesian framework and under assumptions $A_1 - A_5$ on the error term, and assumptions $A_1 - A_2$ on the population under study, our proposed Random-effects Response Surface Model (3.18) can be rewritten as;

$$\left[Y|\tau, \alpha, \beta, \sigma_e^2\right]\ \left(\tau + \beta Z + \alpha W, \sigma_e^2\right) \tag{3.35}$$

independently for each respondent i, where Z is a matrix of population-specific regressors; β is a vector of population-specific (fixed effects) parameters linking Z to Y; W is a matrix of stratum-specific (random effects) regressors with corresponding vector of parameters α linking W to Y ; and σ_e^2 is the error variance. In general, W is a sub set of Z i.e. $W \subseteq Z$.

Further, we assume that for each factor K, the random effects vector α is independently normally distributed with mean zero and variance σ_α^2 . That is;

$$\alpha\left(0,\, \sigma_\alpha^2\right) \tag{3.36}$$

We also assume that random effects are independent between factors.

3.4.6.4 Prior Distributions of Parameters

In line with Bayesian Philosophy, we must complete the Bayesian formulation of the RRSM (3.18) by specifying the prior distributions of all parameters to be estimated. Under assumptions $(A_1 - A_5)$, therefore, we specify conjugate prior distributions for τ, β, α and σ^2 , respectively, as;

$$\tau\left(0, \sigma_\tau^2\right) \tag{3.37}$$

$$\beta\left(0, \sigma_\beta^2\right) \tag{3.38}$$

$$\alpha\left(0, \sigma_\alpha^2\right) \tag{3.39}$$

$$\sigma^2\left(0, \sigma_e^2\right) \tag{3.40}$$

Where $\sigma_\tau^2, \sigma_\beta^2, \sigma_\alpha^2$ and σ_e^2 are the variances associated with τ, β, α and σ^2 , respectively.

3.4.6.5 Posterior Distributions

To run Gibbs Sampler in estimating τ, β, α and σ^2 in RRSM (3.35) , we require the posterior distribution for each parameter (Gilks et al, 1993). The joint posterior distribution of τ, β, α and σ^2 is expressed as;

$$f\left(\tau, \beta, \alpha | z, w\right) = \frac{\prod\limits_{i=1}^{n} f\left(z_i, w_i | \tau, \beta, \alpha\right) g\left(\tau, \beta, \alpha\right)}{\int\int\int \prod\limits_{i=1}^{n} f\left(z_i, w_i | \tau, \beta, \alpha\right) g\left(\tau, \beta, \alpha\right) d\tau d\beta d\alpha} \tag{3.41}$$

But $g\left(\tau, \beta, \alpha\right) = g\left(\tau\right) g\left(\beta\right) g\left(\alpha\right)$ assuming independency between τ, β and α. Therefore, equation (3.41) becomes;

$$f\left(\tau,\beta,\alpha|z,w\right)=\frac{\prod_{i=1}^{n}f\left(z_i,w_i|\tau,\beta,\alpha\right)g\left(\tau\right)g\left(\beta\right)g\left(\alpha\right)}{\int\int\int\prod_{i=1}^{n}f\left(z_i,w_i|\tau,\beta,\alpha\right)g\left(\tau,\beta,\alpha\right)d\tau d\beta d\alpha} \tag{3.42}$$

But the posterior distribution (3.42) depends on the random effects W. We need to infer from the marginal posterior that does not depend on the random effects. Thus, we integrate equation (3.42) with respect to W. That is;

$$f\left(\tau,\beta,\alpha|z\right)=\frac{\int\prod_{i=1}^{n}f\left(z_i,w_i|\beta,\alpha\right)g\left(\tau\right)g\left(\beta\right)g\left(\alpha\right)dw_i}{\int\left(\int\int\int\prod_{i=1}^{n}f\left(z_i,w_i|\tau,\beta,\alpha\right)g\left(\tau,\beta,\alpha\right)d\tau d\beta d\alpha\right)dw_i} \tag{3.43}$$

One important trick that is usually used when deriving the posterior distribution in Bayesian analysis is to ignore terms that are constant with respect to the unknown parameters. Since the expression in the denominator of equation (3.43) results to a constant, we can rewrite it as;

$$f\left(\tau,\beta,\alpha|z\right)\ \alpha\ \int\prod_{i=1}^{n}f\left(z_i,w_i|\tau,\beta,\alpha\right)g\left(\tau\right)g\left(\beta\right)g\left(\alpha\right)dw_i \tag{3.44}$$

It can be seen that the functions $g\left(\tau\right)$, $g\left(\beta\right)$ and $g\left(\alpha\right)$ in expression (3.44) do not depend on $W$and therefore this expression can be rewritten as;

$$f\left(\tau,\beta,\alpha|z\right)\ \alpha\ \int\prod_{i=1}^{n}f\left(z_i,w_i|\tau,\beta,\alpha\right)dw_ig\left(\tau\right)g\left(\beta\right)g\left(\alpha\right) \tag{3.45}$$

3.4.6.6 Gibbs Sampling

We now run Gibbs sampler on the posterior distribution (3.45) together with the conjugate prior distributions (3.37), (3.38),(3.39) and (3.40).

3.5 Data Collection

This section presents the sampling technique, data collection tools and the strategy of determining the sample size.

3.5.1 Sampling Technique

Eldoret East District consists of twenty administrative locations, namely; Moiben, Sergoit, Ainabkoi, Tembelio, Meibeki, Mumetet, Karuna, Kaplolo, Koitoror, Kapsoya, Kaptagat, Plateau, Kipsinende, Kipkabus, Kapngetuny, Olare, Chepngoror, Chepkero, Chepkoilel, Kimoning'.

In view of this geographical description of the district, we apply stratified sampling method. First, stratified sampling is convenient to use administratively. Often, one of the typical variables used in stratification are administrative regions e.g. counties, divisions, e.t.c. In this research, therefore, we use the existing administrative locations as strata. Secondly, the administrative locations are non-overlapping. It is a key assumption in stratified sampling that the strata should be non-overlapping. Finally, the population in an administrative location is assumed homogeneous in terms of soil inherent characteristics, farming practices, inputs applied and the weather pattern. This is another key aspect of stratified sampling whereby a stratum is considered homogenous unlike in cluster sampling where the cluster is heterogeneous almost like the whole population. This last point was also crucial in developing our model in chapter four of this thesis whereby we considered stratum-specific characteristics and assumed that each stratum possessed similar characteristics that contribute to the same intercept.

The population of maize farmers in each location is considered as a stratum which makes a subpopulation of the whole population of maize farmers in the district. Thus, we can denote the totals of sub populations as $N_1, N_2, ..., N_{20}$ where $N_i \neq N_j; i, j = 1, 2, ..., 20$ and their sum $N_1 + N_2 + ... + N_{20} = \sum\limits_{i=1}^{20} N_i = N$ where N is the total population in the district.

The samples drawn from the strata are of sizes denoted by $n_1, n_2, .., n_{20}$, respectively such that $n_1 + n_2 + ... + n_{20} = n$ where $\sum\limits_{i=1}^{20} n_i = n$, the size of the sample selected from the whole district.

The sample selected using stratified sampling method can either be a simple random sample or systematic sample. In this research, we decided to select a simple random sample from each stratum (location). This is because in simple random sampling, each maize farmer will have equal opportunity of being selected into the sample.

3.5.2 Sample Size

There are four commonly used strategies of determining a sample size. These include using census for small populations, imitating a sample size of similar study, using published tables and using formulas. In our study, we use a formula. Consequently, there are many formulas which have been tested to work well. In our case, we use the formula due to Yamane (1967). We adopted this formula because of its simplicity.

Also, this formula has been used to calculate the sample sizes in some published tables that have been reliably used for many years in determining sample sizes in various studies.

The Yamane (1967) formula is given by;

$$n = \frac{N}{1 + N\left(e\right)^{2}} \tag{3.46}$$

where n is the sample size, e is the desired level of precision and N is the population size in the whole district. We got the estimated value of N from the Kenya National Bureau of Statistics (KNBS) census report of 2009.

3.5.3 Data Collection Tools

When collecting data, it is important to consider whether they are primary or secondary data. Our data is Primary data. Some of the main sources of data are census where data is obtained from every member of a population, sample survey where data is obtained from a subset of a population in order to estimate population attributes, experiment and observational study. The source of our data was a sample survey. This was done by trained enumerators through investigation whereby they contacted individuals and filled in questionnaires (see appendix 1) after asking the required information. The data collected this way is usually accurate and reliable.

3.6 Conclusion

Chapter has provided our model and the estimation procedure using Gibss sampling technique. It has also explained sampling procedure to be used in the data collection in the emperical study. The subsequent chapter provides simulation study of the proposed RRSM.

CHAPTER FOUR

SIMULATION STUDY

4.1 Introduction

This chapter presents simulation of the proposed RRSM. First, we assume some range of the response variable and fix some values for the regression coefficients of the explanatory variables and the intercept. We then proceed to simulate from the proposed RRSM under three different settings: Fixed effects only and then random effects. Under the random effects setting, we have the random effects model whereby the intercept alone is random (assuming explanatory variables are constant) and then we have the random effects model in which both the intercept and the explanatory variables are random.

4.2 Simulation Study Set Up

We suppose that there are 6 strata depicting the administrative locations. From each stratum, we simulate ten (10) yields representing the harvest from ten farmers. We assume the harvest is measured in 90Kg bags with the yield ranging between 10-40 bags of maize per acreage of land. Each of the 60 farmers in the assumed sample is observed for a period of 8 years.

Let Y be the yield per acre of land measured in 90Kg bags. The acreage then becomes an offset variable. The vector of covariates Z consists of past responses and quantity of fertilizer applied. Let r, s , and t index the administrative location (stratum), the sampled farmer and year of measurement, respectively.

We simulate from the proposed RRSM under three different settings as follows;

Case 1: A fixed effects model only

$$Y_{rst} = \tau + \beta Z_{rs} + e_{rst} \qquad r = 1, 2, \dots, 6; s = 1, 2, ...10; t = 1, 2, \dots\ , 8$$

Case 2: A random effects model with random intercepts only

$$Y_{rst} = \tau_r + \beta Z_{rs} + e_{rst} \qquad r = 1, 2, \dots, 6; s = 1, 2, ...10; t = 1, 2, \dots\ , 8$$

Case 3: A random effects model with both random intercepts and random slope

$$Y_{rst} = \tau_{rst} + \beta Z_{rst} + \alpha_r W_{rst} + e_{rst} \qquad r = 1, 2, \dots, 6; s = 1, 2, ...10; t = 1, 2, \dots\ , 8$$

Where the parameters and data are as defined in the previous Chapter.

4.3 Simulation Results

We now present simulation results and explanation from the three cases using tables and box plots.

Table 4.1: Regression Coefficient Estimates and their Precision Measures from a Fixed-effects model

```
           true  estimate  SE    LCI    UCI     coverage  bias   rmse
Intercept 10.00    9.971  4.331  1.482  18.459     94.6   0.029  4.331
Fert       0.10    0.100  0.047  0.007   0.193     94.4   0.000  0.047
Educ2      2.50    2.530  1.121  0.332   4.728     94.9  -0.030  1.122
Educ3      5.00    5.017  1.580  1.920   8.113     94.8  -0.017  1.580
Early      2.00    2.041  1.047 -0.011   4.094     95.3  -0.041  1.048
Time      -0.5    -0.497  0.150 -0.790  -0.203     94.3   0.003  0.031
```

In Table (4.1) , we have presented a few explanatory variables which possibly have some effect on the response variable. These include the fertilizer quantity, the education level, ploughing time and time of harvesting. We simulated their estimated contribution to the response variable and the confidence interval at 5% level of significance.

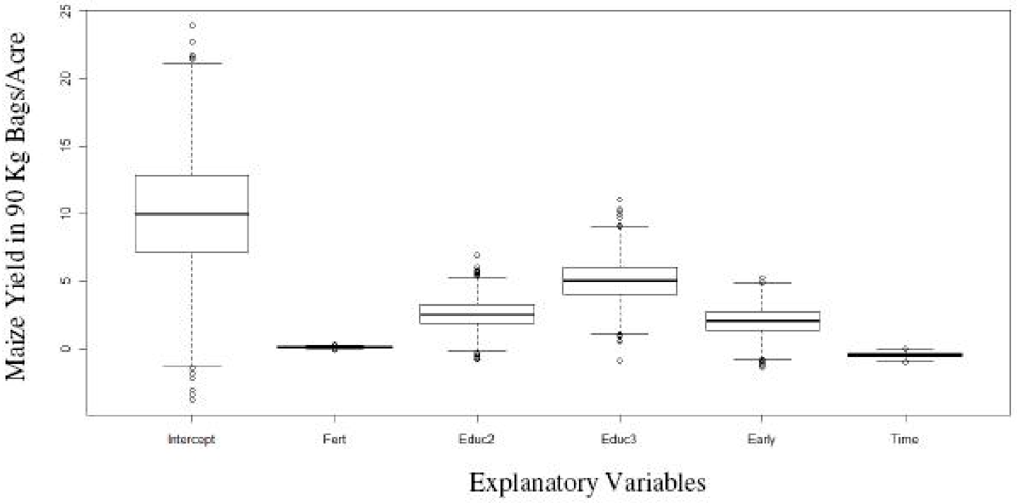


Figure 4.1: A Box Plot of the distribution of the Regression Coefficients of Fixed-effects Model

In this case, we considered a few explanatory variables which possibly have some effect on the response variable. These include the fertilizer quantity, the education level, ploughing time and time of harvesting. We consider these factors to be constant across the whole population under study. The intercept is also considered to be fixed. Figure (4.3.1) above illustrates how each of these factors contributes to the response variable in addition to the intercept. For instance, higher education level of a maize farmer contributes most while harvesting time contributes the least to the response variable.

Table 4.2: Regression Coefficient Estimates of RRSM with Random Intercept

```
          True     estimate   SE      LCI    UCI   coverage bias  rmse
Intercept 10.00     9.982  4.357  1.443 18.521    95.3  0.018  4.357
Educ2      2.50     2.491  1.102  0.331  4.651    94.9  0.009  1.102
Educ3      5.00     5.014  1.597  1.883  8.145    95.2 -0.014  1.597
Early      2.00     1.998  1.047 -0.055  4.051    94.9  0.002  1.047
Time      -0.50    -0.500  0.148 -0.791 -0.209    95.1  0.000  0.148
```

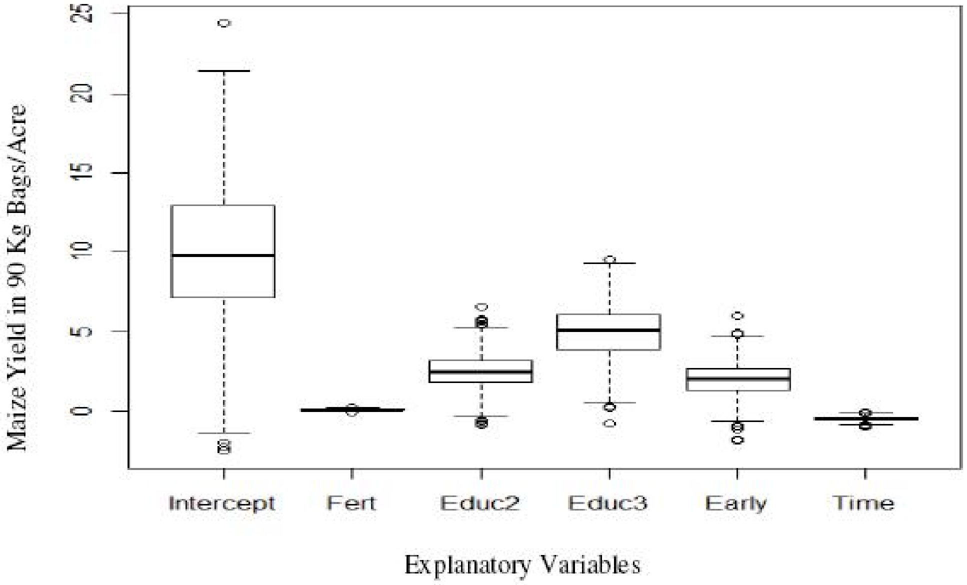


Figure 4.2: A Box Plot showing the distribution of the Regression Coefficients of Random-intercepts Model

This simulation considered the intercept alone to be random. That is, its values are randomly distributed amongst the strata. The explanatory variables are assumed to be constant across the entire population. It is not a common characteristic that the intercept is random while the explanatory variables are constant. Indeed, it has been observed that in most longitudinal studies, varying slope coefficients usually arise, where both the intercept and the slope coefficients are specific to each stratum. Figure (4.3.2) shows how each of the explanatory variables contributes to the response variable in addition to the random intercept, in such a set up.

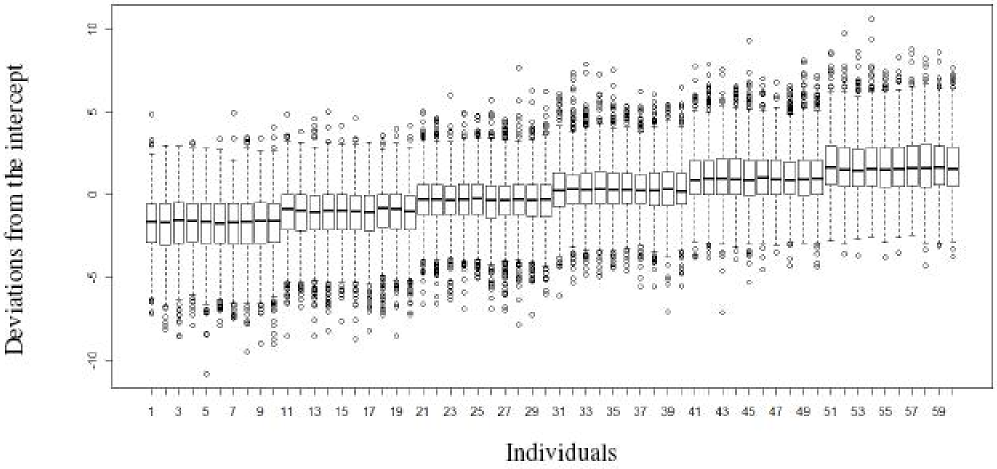


Figure 4.3: Individual Random Intercepts from the Random-intercepts Regression Model

Figure (4.3.3) illustrates simulation from a random intercepts model. In this particular case, we consider six strata, and from each stratum, we consider ten responses. Homogeneity within a stratum is clearly evident from the figure in the sense that the randomly selected responses within each stratum are almost the same. It suggests that each stratum has some inherent characteristics which are almost uniform across the entire stratum, hence yielding almost the same intercept for each response within the same stratum. It may also suggest that farming practices and inputs applied amongst respondents within a stratum are almost similar, hence yielding almost the same response from each respondent in the same stratum. This explanation is in line with assumption on the assumptions on the population under study. Further, it is evident from figure 3 that neighbouring strata may possess similar inherent characteristics and almost the same farming practices since the responses from the randomly selected respondents are almost the same.

Figure (4.3.4) below illustrates distribution of random intercepts on average as per stratum. It shows that yields vary from one location to another. This variation in yield is informed by similar explanation in figure 3. Further, this variation justifies the use of random-effects model in modelling maize farming and other related studies, in a given place.

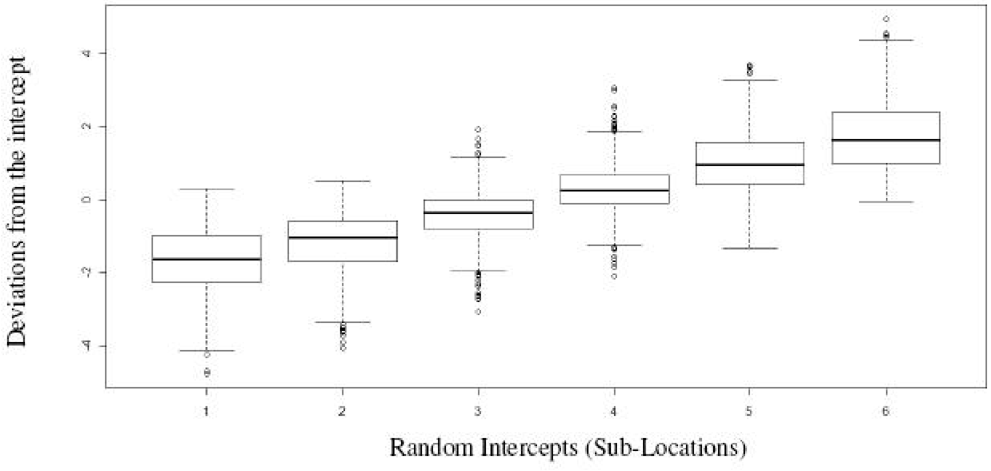


Figure 4.4: Box Plot of Random Intercepts by Sub Locations

Table 4.3: Regression Coefficient Estimates and their Precision Measures from a Random Effects Model with both Intercept and Slope being Random

	True	estimate	SE	LCI	UCI	coverage	bias	rmse
Intercept	10.0	10.098	4.137	1.989	18.206	95.3	-0.098	4.138
Fert	0.1	0.099	0.045	0.011	0.188	95.0	0.001	0.045
Educ2	2.5	2.485	1.108	0.313	4.656	94.3	0.015	1.108
Educ3	5.0	5.028	1.567	1.957	8.099	94.7	-0.028	1.567
Early	2.0	1.982	1.066	-0.107	4.072	95.4	0.018	1.066
Time	-0.5	-0.510	0.146	-0.795	-0.224	95.1	0.010	0.146

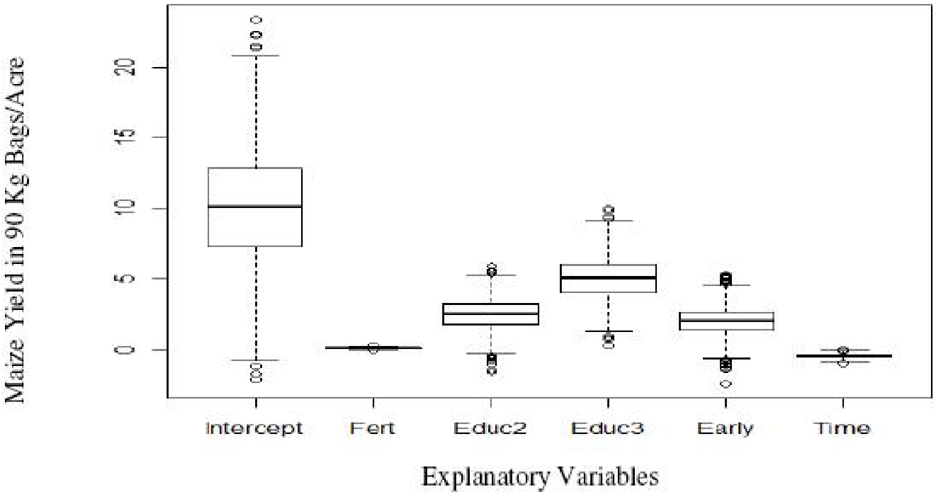


Figure 4.5: A Box Plot Showing Distribution of Parameters from a Random-effects Model with both Intercept and Slope being Random

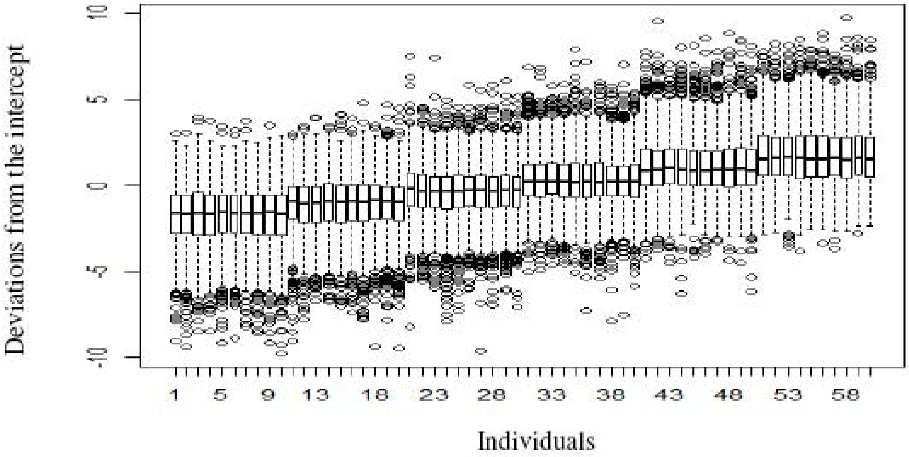


Figure 4.6: Individual Random Intercepts Coefficients from a Random Effects Model with both Intercept and Slope being Random

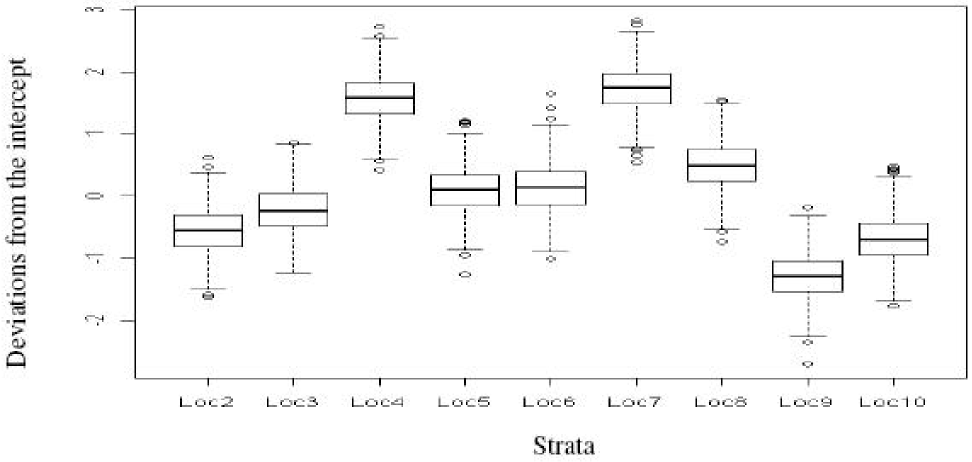


Figure 4.7: Random Intercepts from a Random Effects Model with both Intercept and Slope being Random

The explanation of figure (4.3.7) is in many ways similar to the explanation of figure 4. The only main difference is that figure 4 considers only the intercept as random whereas figure 7 considers both the intercept and the explanatory variables as random. Indeed, it is evident from figure 7 that variation between locations is clearer.

4.4 Conclusion

This chapter has presented a simulation study under three settings; fixed effects model in which it was assumed that all effects are fixed across the entire population, random effects model in which only the intercept is considered to be random, and the random effects model in which both the intercept and the slopes are considered to be random. The outcomes are presented using tables and box plots, and explanation of each outcome is given. Most importantly, the use of a random-effects model in our problem and perhaps other related studies has been justified through our simulation. In the subsequent chapter, we will analyze empirical data and see if the results are related to the simulation results.

CHAPTER FIVE

EMPIRICAL STUDY

5.1 Introduction

This chapter presents a test of practicability of our proposed model to a real life problem; maize farming in Eldoret East District, Kenya. We fit the model using four-year data of maize production realized as a result of application of various combinations of inputs (treatments). First, we give an overview of maize farming in general and particularly maize farming in Eldoret East District. Secondly, we discuss data collection. This entails suitable sampling technique, data collection tools, and the strategy for determining the sample size i.e. the number of maize farmers interviewed. Finally, we perform data analysis, presentation and discussion of the results.

5.2 Overview of Maize Farming

Mostly, output production per acre of land in a maize farming system varies year after year. For instance, the following table shows a case of a farmer in which production kept varying within a period of 10 years (1999-2008).

Table 5.1: A Case of Production Per Acre in a Period of 10 years (1999-2008)

Year	Production/acre(No. of 90 kgs Bags)
1999	15
2000	17
2001	20
2002	25
2003	28
2004	24
2005	35
2006	30
2007	30
2008	33

This scenario is not unique to this particular farmer. Indeed, it is a common experience to many other maize farmers, even within the same area. The varying production is as a result of using different inputs and is a clear indication that there are some dynamics in the maize farming system that are not clear to maize farmers. Notably, this subjects many of them to making decisions that lead to sub-optimal production or sometimes, they may realize relatively good production by chance.

It therefore follows that there exist an optimal production guided by many factors as identified in section 5.3 of this thesis. It is evitable from the varying yearly production that these factors must be contributing proportionately to maize production depending on their amounts used. Otherwise, if they were contributing in the same proportions, then, the output production would portray Uniform distribution.

Our case study revealed that majority of the maize farmers within Eldoret East District make trials each year by trying various combinations of inputs. The result of this is realization of the ever varying production. They do this while seeking optimal production.

In our study, we consider maize farming as a process that combines a variety of inputs. From each unit of a particular input, the process derives a certain amount of the output, with the amount of the output produced being proportional to the amount of each input consumed. The inputs also have cost per unit hence the total cost of inputs is also proportional to the amount of inputs consumed. Our decision variables are the amount of inputs consumed. Our goal is to determine an amount of each input to be consumed, within specified limits, so that the output thereby produced meet specified requirements at the least cost.

5.3 Factors that affect Maize Output

This section presents the categories of factors or explanatory variables that affect maize output, giving some examples of each category.

5.3.1 Quantitative Variables

These are variables which can be measured or weighed using appropriate instruments. They comprise both dependent and independent variables. In a maize farming system, for example, the independent variables may include such variables as quantity of fertilizer applied per unit of land, Quantity of seeds applied per unit of land, soil chemical properties (carbon, nitrogen, potassium, iron, ph level etc), soil physical characteristics (clay content, sand content etc) and the dependent variable is the yield. Measurement of these variables is usually done in their SI units.

To identify these variables, we shall use the symbol X as a random variable to refer to the entire set of the independent factors, and to refer to an individual factor in this set, we use the subscripts on the symbol X as; $X_1, X_{2,}..., X_n$, that is, supposing we have n factors.

5.3.2 Qualitative Variables

These are variables with no natural sense of ordering. They are therefore measured on a nominal scale. For instance, quality of maize seeds, System of cultivation, Texture of soil, etc. Qualitative variables can be coded to appear numeric but the numbers used for coding do not have effect.

Socioeconomic factors can also be treated as categorical variables. These are factors that characterize an individual or a group within a given social structure. They include income level, ethnicity, sense of community and many other such factors. Studies have shown that certain segments of society are exposed to environmental hazards, and may be more vulnerable to such hazards than other populations. In most cases, these socioeconomic factors characterize the kind of farming done by various individuals or groups in terms of inputs used as well as the corresponding outputs realized.

The quantitative and qualitative variables account for fixed effects only. However, there may exist some other factors which are random in their nature and hence cannot be taken into account under these sub sections. Such random factors can be well taken into consideration in the model by incorporating random intercepts and random slopes.

5.3.3 Random Intercepts

In many empirical studies, there may be some inherent characteristics that are only specific to a certain portion of the population and may significantly contribute to the performance of the response variable in addition to the applied treatments; quantitative or categorical. These may be referred to as cluster-specific or strata-specific factors, depending on how the population is partitioned. These characteristics contribute to the response variable in such a way that without application of any treatment (s), various portions of the population can still give different intercepts. In our application problem, for example, there could be some factors which are specific to a certain parcel of land or strata i.e. they do not cut across the whole district and their effect on the response variable may be significant.

To take into account such factors, a model with random intercepts, β_{it} , where $i = 1, 2, ..., n$ representing various portions of the population and $t = 1, 2, ..., T_i$ representing time, is appropriate. In this case, the slope coefficients are assumed to be constant and the random intercepts are assumed to be identically and independently distributed (iid) with unknown parameters $E(\beta_{it}) = \beta$ and $Var(\beta_{it}) = \sigma^2$. The unobservable deviations between the population intercept β and the cluster specific realizations β_{it} may be interpreted as effects of omitted covariates.

Random intercept models are also called error components or variance components (Hsiao, 1986). It is interesting to note that a random intercept model also takes into account intracluster correlation of the Gaussian outcomes.

5.3.4 Random Slopes

The random intercept models do not alleviate the restrictive assumption that the slope coefficients are equal for each observation. Varying slope coefficients arise in particular in longitudinal studies, where intercept and slope coefficients are specific to each time series. To take into account such parameter heterogeneity, the random intercept model can be extended in such a way that we treat, not only the intercept, but also all regression coefficients, as random.

Models where all coefficients are assumed to vary randomly over strata are also called random coefficient regression models (Hsiao, 1986). We can also note that in a longitudinal setting, the responses from i^{th} stratum form a time series, and the possible effects of past responses are also allowed to vary from time series to time series.

Sometimes, however, assuming that some coefficients are stratum-specific is less realistic than the assumption that some coefficients are constant across clusters. Suppose β_{i1} denote some stratum-specific coefficients and denote the remaining coefficients which are constant across the remaining strata, then, the parameter vector β_i can be partitioned into $\beta_i = (\beta_{i1}, \beta_{i2})$ with $\beta_{i2} = \beta_2$ for all i. The design vector will also have to be rearranged in the same manner. Due to the mixing of "fixed" and "random" coefficients, models of this type can also be called linear mixed models.

5.3.5 Error-Term Factor

In regression modeling, the variation in the response variable may not be totally explained by the explanatory variables. Practically, there is, in most cases, a small percentage of the total variation that may remain unexplained. The error term represents such unexplained variation. The error term is usually treated as a random variable. Later in our application problem, factors such as the natural calamities-hailstone, strong winds etc, can be taken as error terms. For purposes of analysis of the model, appropriate assumptions are usually specified regarding the error term depending on the nature of the study being undertaken.

5.4 Frequentist Empirical Results

This section presents empirical results on estimation of our model (3.18) using frequentist approach. We have used the R statistical package to perform the analysis.

5.4.1 Discussion of the Frequency Distributions

The frequency distribution tables generated from the questionnaires are displayed in Appendix 3. A total of 587 maize farmers were contacted and interviewed. The interpretation of these tables follows below.

The study established that maize is the staple food to all the people who were interviewed, and so to the entire population in the district. Maize is cultivated alone or in addition to another crop e.g. wheat, potatoes, etc grown for subsistence or commercial purpose. Being the staple food, the respondents put much emphasis on maize production throughout the entire process of cultivation. For instance, all the respondents practice mechanized farming i.e. use of tractors in ploughing of land and planting maize. Over 95% of the respondents practice crop rotation whereby they don't plant maize consequently on the same piece of land. Over 95% of the respondents also plough their land early; about two months prior to planting season, which is recommendable in sound agriculture. Majority also applies pest control and weed control two or three times per season. The type of soil in almost the entire district is loamy, which is favorable to farming.

However, there are some factors which probably work against the production. For instance, most parcels of land have been cultivated continuously for over 23 years. Also, most respondents practise burning as a way of clearing their lands in preparation for ploughing. This is a poor farm preparation method.

5.4.2 Estimation of Parameters

This section presents empirical estimation of the parameters (regression coefficients) in the model for purposes of fitting our model (3.18) using the maize production data in which maize production is considered as the response variable and the inputs like fertilizer quantity, fertilizer type, seed type, seed quantity, e.t.c. are the explanatory variables. Since the model comprises of fixed and random effects, we categorize the maize input variables into fixed effects and random effects. We consider strata (locations) and seed type as the random effects, and the fixed effects include top dresser quantity, weed control method, pest control frequency per season, cultivation period, farm preparation method and time of ploughing. Some input variables were found to be constant amongst all the interviewees and therefore could not be categorized into any of the two categories. These include factors like the quantity of fertilizer and quantity of seed.

Estimation was performed using code. The following were the results realized;

Table 5.2: Frequentist Estimation of Regression Coefficients in our proposed model

```
Random effects:
Groups       Name             Variance   Std.Dev.
Stratum    (Intercept)        0.25812   0.50806
Seeds Type (Intercept)        0.07200   0.26834
Residual                     15.03042   3.87691
Number of obs: 2348, groups: Stratum, 18; SeedsType, 8
```

```
Fixed effects:
                                           Estimate   Std. Error  t value
(Intercept)                              19.057445    0.982049   19.406
TopdresserQty50                           0.616733    0.408972    1.508
TopdresserQty75                           0.311713    0.403987    0.772
WeedcontrolMethod Spray.herbicides        0.558709    0.207289    2.695
PestcontrolFrequency Freq.2               0.209152    0.197881    1.057
PestcontrolFrequency Freq.3              -0.231011    0.228125   -1.013
PestcontrolFrequency Freq.4               0.746269    0.565382    1.320
CultivationPeriod                        -0.006517    0.011873   -0.549
FarmPreparationMethodHallowing           -0.167356    0.432418   -0.387
PloughingPeriod 2.months                  1.445772    0.826763    1.749

Correlation of Fixed Effects:
            (Intr) TpdQ50 TpdQ75 WdcMS. PFP..F.2 PFP..F.3 PFP..F.4 CltvtP
TpdrssrQt50 -0.355
TpdrssrQt75 -0.368  0.875
WdcntrlMtS. -0.104 -0.104 -0.108
PstcFP..F.2 -0.045 -0.074 -0.105  0.001
PstcFP..F.3 -0.041 -0.052 -0.074 -0.063  0.512
PstcFP..F.4 -0.045 -0.041 -0.015 -0.035  0.206    0.173
CultivtnPrd -0.341  0.034  0.014  0.031 -0.027   -0.050    0.090
FrmPrprtnMH  0.033 -0.042  0.000 -0.017 -0.056   -0.098    0.012   -0.021
PlghnPP..2. -0.837  0.012  0.030 -0.015 -0.009    0.007    0.001    0.017
```

5.4.3 Discussion of the Results

The intercept in any regression model represents the base line value of the response variable assuming no contribution of any explanatory variable. In our analysis under fixed effects, we got the intercept as 19.057445. This implies that even with no application of any input, the maize yield will be about 19 bags per acre.

The regression coefficient often referred to as the slope, measures the steepness of a regression line at a given point. It gives the approximate change in the response variable contributed by a unit change in a given explanatory variable. In our study, for instance, top dressing is one of the explanatory variables with regression coefficient 0.616733. This implies that application of 1 kg of top dresser (C.A.N) per acre of land results into a corresponding change in maize production by approximately 0.616733 kgs per acre of land. Further, this is a positive contribution (increase). The study also found out that application of 50 kgs per acre of top dresser contributes to better yield than application of 75 kgs per acre. Probably, much top dresser is unnecessary.

We also found out that some inputs contribute negatively to the maize yield. For example, the cultivation period in which we obtained the value of the regression coefficient as -0.006517. This implies that as land is cultivated continuously for many years, the level of yield will decline. Further, the box plots below clearly show how the various explanatory variables contribute to the response variable, including the intercept.

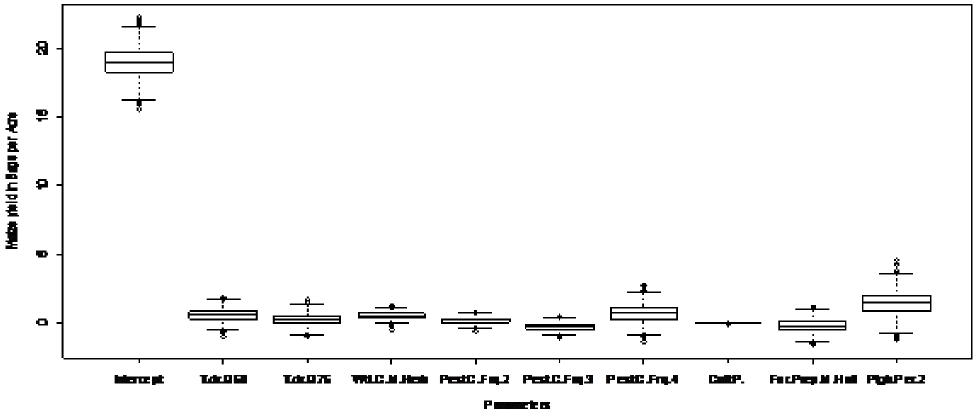


Figure 5.1: Box Plot of Maize Yield per acre (in 90 Kgs bags) versus the various Explanatory variables

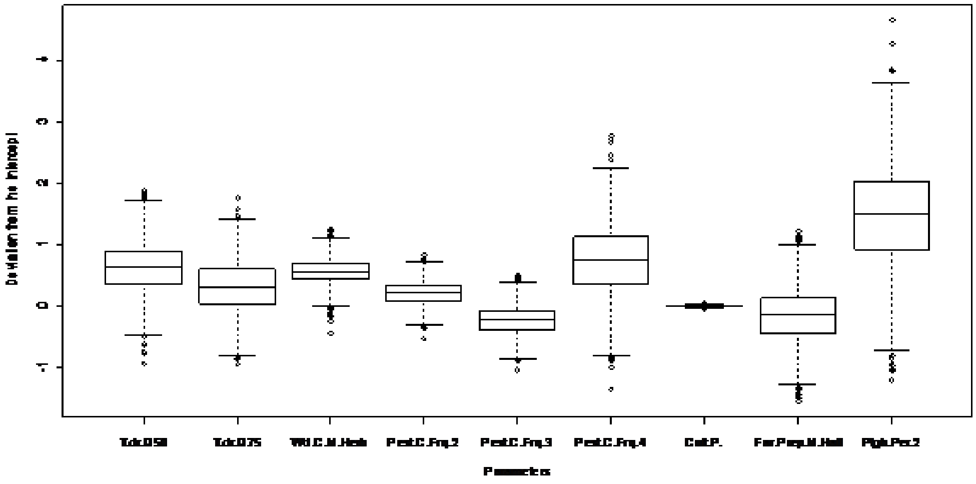


Figure 5.2: Box Plot of Deviation from the intercept versus the Explanatory Variables

From these two figures, it is clear that application of 50 Kgs of top dresser per acre results in better yield than application of 75 kgs per acre. The reason for this interesting result may be investigated in a different study. This analysis also reveals that application of a pesticide four (4) times in the entire season gives better yield than application of the pesticide three or two times. Further, ploughing the land at least two months before planting contributes to better yield.

A good estimated regression model has to explain variation of the response variable as a result of the various explanatory variables. This calls for test of significance of the explanatory variables. This requires that the error terms e_i be normally and identically distributed with mean 0 and variance σ^2; see assumptions A_1 and A_5. To check this assumption, we graphed the normal plots of the residuals as shown in figure (5.4.3) below.

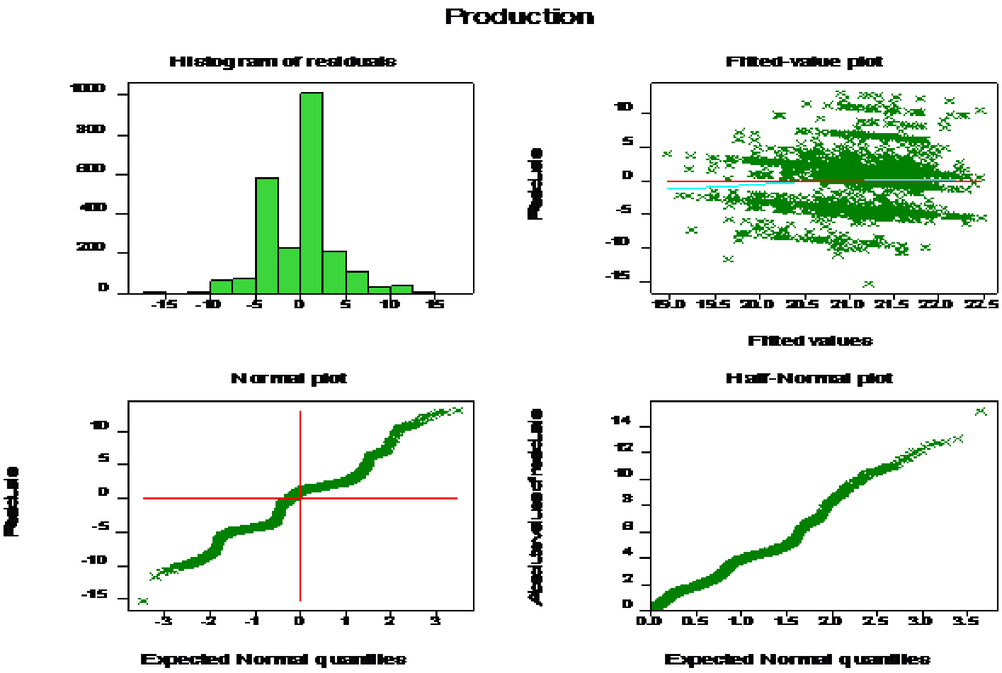


Figure 5.3: Diagnostic Plots for the Residuals

From the figures above, the histogram of residuals, normal plot and the half normal plot, all indicate approximately normality of the residuals, hence satisfying the normality assumption A_4 . In addition, the plot of the residuals versus the fitted values clearly indicates randomness of the residuals, which is a desirable feature of any good model. This satisfies assumption A_1 .

5.5 Bayesian Estimation

In this section, we present results of the Bayesian analysis of our proposed RRSM in which we employ Gibbs sampling estimation technique to the model using Win bugs for Bayesian Statistics. Table 7 below presents estimates of the regression coefficients of our model, together with their corresponding standard errors and t-values.

Table 5.3: Bayesian Estimation of Regression Coefficients of the Proposed Model

```
                              Fixed.effects  Estimate Std..Error t.value Estimate1
1                                 Intercept 19.057445   0.982049  19.406  19.05745
2                           TopdresserQty50  0.616733   0.408972   1.508  19.67418
3                           TopdresserQty75  0.311713   0.403987   0.772  19.36916
4         WeedcontrolMethodSpray.herbicides  0.558709   0.207289   2.695  19.61615
5   PestcontrolFrequencyPest.control.Freq.2  0.209152   0.197881   1.057  19.26660
6   PestcontrolFrequencyPest.control.Freq.3 -0.231011   0.228125  -1.013  18.82643
7   PestcontrolFrequencyPest.control.Freq.4  0.746269   0.565382   1.320  19.80371
8                         CultivationPeriod -0.006517   0.011873  -0.549  19.05093
9            FarmPreparationMethodHallowing -0.167356   0.432418  -0.387  18.89009
10 PloughingPeriodPloughing.period.2.months  1.445772   0.826763   1.749  20.50322
```

It can be observed from this table that application of pest control (three times) and farm preparation method by hallowing, are input factors that contribute negatively to the maize production. The other factors result in positive contribution but in various proportions, with early ploughing of land (at least two months prior to planting) yielding a larger proportion.

We then considered our response variable Y to have a prior normal with mean μ and variance σ^2 i.e. $Y(\mu, \sigma^2)$. We performed posterior analysis using Gibbs sampler (1000 iterations) and figures (5.5.1) and (5.5.2) below are the resulting plots of posterior mean and posterior standard deviation, respectively.

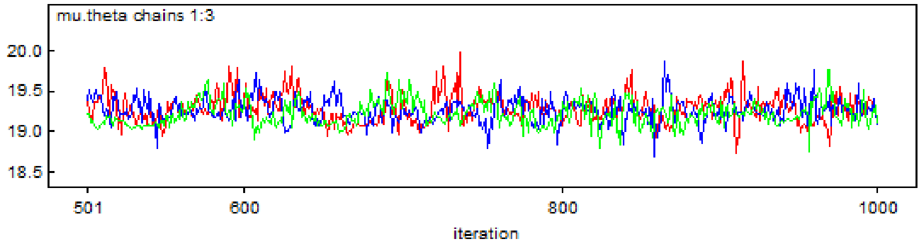


Figure 5.4: Trace Plot of Posterior mean (1000 iterations)

It can be seen from figure (5.5.1) that the posterior mean oscillates between 19.0 and 19.5 bags of maize, which is in line with the intercept of 19.2 bags.

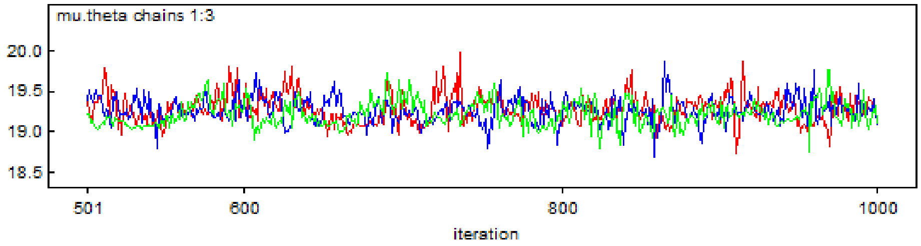


Figure 5.5: Trace Plot of Posterior Standard Deviation (1000 Iterations)

The posterior standard deviation as shown in figure (5.5.2) lies mainly between 5 bags on the negative side (signifying a reduction from the mean) and 10 bags on the positive (signifying an increase to the mean or intercept), and in some few cases, over 10 bags on the positive side (signifying some few extreme outliers), which is expected under any normal practice.

5.6 Conclusion

Chapter 5 has provided empirical estimation results of our model based on a four-year data of maize production. It has also explained the methodologies of how these data were collected, analyzed and presented, together with explanation of each result.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1 Introduction

We now present conclusions and recommendations for further research arising from our research.

6.2 Conclusion

We achieved the objectives set at the beginning. We considered randomness of the explanatory variables and therefore deviated from the tradition whereby, in most situations, assumption of non-randomness of the variables is considered, which is in most cases, not real in practice. In this respect, we considered a random-effects response surface model, as a suitable model to our problem. Our consideration was also informed by the fact that our case study problem i.e. maize production, is longitudinal in nature. In chapter three, we showed how our random-effects response surface model can be estimated straightforwardly using Gibbs sampling.

Since our study is applied statistics, we laid much emphasis on the application problem. We applied our random-effects response surface model to food production, and in particular, did a case study of maize production in Eldoret East District, Kenya. An overview of maize farming indicated suitability of our proposed model in modelling maize production as well as other related studies. The empirical results realized in chapter five closely resemble the simulation results in chapter four.

6.3 Suggestions for Further Research

There are a few issues emerging from this research work that could be taken up in future researches.

In our research, we assumed that each stratum has homogenous inherent characteristics and that the units in one stratum are independent of those in the other strata. This might not be ideal. For instance, suppose we have two respondents in two different strata but each is close to the shared boundary, obviously, they cannot be independent. Such a scenario could form a good case for further research, and is recommended.

We also assumed no curvature in our model; this again might not be real. Some curvature might exist. This scenario can also form a case for future research.

REFERENCES

Arap Koskei, J. K. (1984). *Fourth Order Rotatable Designs. Ph. D Thesis.* University of Nairobi, Kenya.

Arap Koskei, J. K. (2010). *Rotatability and Response Surface Designs of Experiments.Inaugural Lecture 11.* Moi University, Kenya.

Baysal, R. E. (2008). *Advances in risk management simulation. Ph. D Thesis.* North Western University, UK.

Bliss, C. I. (1935a). The calculation of the dosage-mortality curve. *Ann Appl. Biol.*, 22:134–167.

Box, G. E. P. and Draper, N. R. (1963). The choice of second order rotatable design. *Biometrika*, 50:335–352.

Box, G. E. P. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *J. R. Statist. Soc*, B 13:1–45.

Chernoff, H. (1953). Local optimal designs for estimating parameters. *Ann. Math. Statist.*, 24:586–602.

Cleopas, K. L., Okemwa, P., Dimo, H., Lagat, K., and Korir, J. K. (2007). The state of agricultural mechanisation in uasin gishu district, kenya, and its impact on agricultural output. *Agricultural Engineering International: the CIGR Ejournal*, IX.

Crowther, E. M. and Yates (1941). Fertilizer policy in war-time. *Empire J. Exp. Agric.*, 9:77–97.

DAO (2001). Uasin gishu district agricultural annual report. *District Agricultural Office, Eldoret, Kenya.*

Gaddum, J. H. (1933). *Reports on Biological Standards III Methods of Biological Assay Depending on a Quantal Response.Special Report No. 183*. Medical Research Council, H.M. S. O. London.

Hill, W. J. and Hunter, W. G. (1966). A review of response surface methodology. a literature survey. *Techno Metric*, 8:571–590.

Hossein, M. S. and Thornton, B. (1984). Optimization in simulation experiments using response surface methodology. *Biometrics*, 8:11–27.

Kiefer, J. (1959). Optimum experimental designs (with discussion). *J. R. Statist. Soc*, 21:272–319.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function). *Ann. Math. Statist.*, 23:462–466.

Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems. *Ann. Math. Statist.*, 30:271–294.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Ann. Math. Statist.*, 12:363–366.

Kipchumba, S. (2008). Kenya: Farmers shift threatens bread basket. *Business Daily*.

Kosgei, K. M. e. a. (2006). On optimality of a second order rotatable design in three dimensions. *East African Journal of Statistics*, 1:123–128.

Ksiazek, T. (1985). Automatic controls for chemical application. *Agri-Matron Conference and Exposition*, pages 165–172.

Mead, R. and Pike, D. J. (1975). A review of response surface methodology from a biometric viewpoint. *Biometrics*, 31:803–851.

Metianu, A., Tinker, D. B., and Farrant, J. (1983). Evaluation of the prototype whole crop harvest under yield conditions in schag. *National Institute of Agricultural Egineering*.

Mitscherlich, E. A. (1930). *Dic Bestimmung Des Dungerbedurfnisses Des Bodens*. Paul Parey, Berlin.

Montgomery, D. C. (2005). *Design and Analysis of Experiments: Response Surface Method and Designs*. John Weley and Sons, Incl., New Jersey.

Myers, R. H., Khuri, A. I., and Carter, W. H. (1989). Response surface methodology: 1966-1988. *Techno Metric*, 31(2):137–153.

Nicolai, R. and Dekker, R. (2009). Automated response surface methodology for simulation optimization models with unknown variance. 6:325–352.

Njui, F. (1985). *Fifth Order Rotatable Designs. Ph. D Thesis*. University of Nairobi, Kenya.

Peterson, R. G. (1990). *Design and Analysis of Experiments*. Marcel Dekker, New York.

Reed, L. J. and Berkson, J. (1929). The application of the logistic function for experimental data. 33:760–799.

Rider, A. R. and Dickey (1982). Field evaluation of calibration accuracy for pesticide application equipment. *Transactions of the ASAE*, 25(2):258–260.

Winsor, C. P. (1932). The gompertz curve as a growth curve. *Proc. Natl. Acad. Sci.*, 18:1–8.

Wishart, J. (1938). Growth rate determination in nutrition studies with the bacon pig, and their analysis. *Biometrika*, 30:16–28.

Wishart, J. (1939). Statistical treatment of animal experiments. *J. R. Statist. Soc*, B6:1–22.

APPENDICES

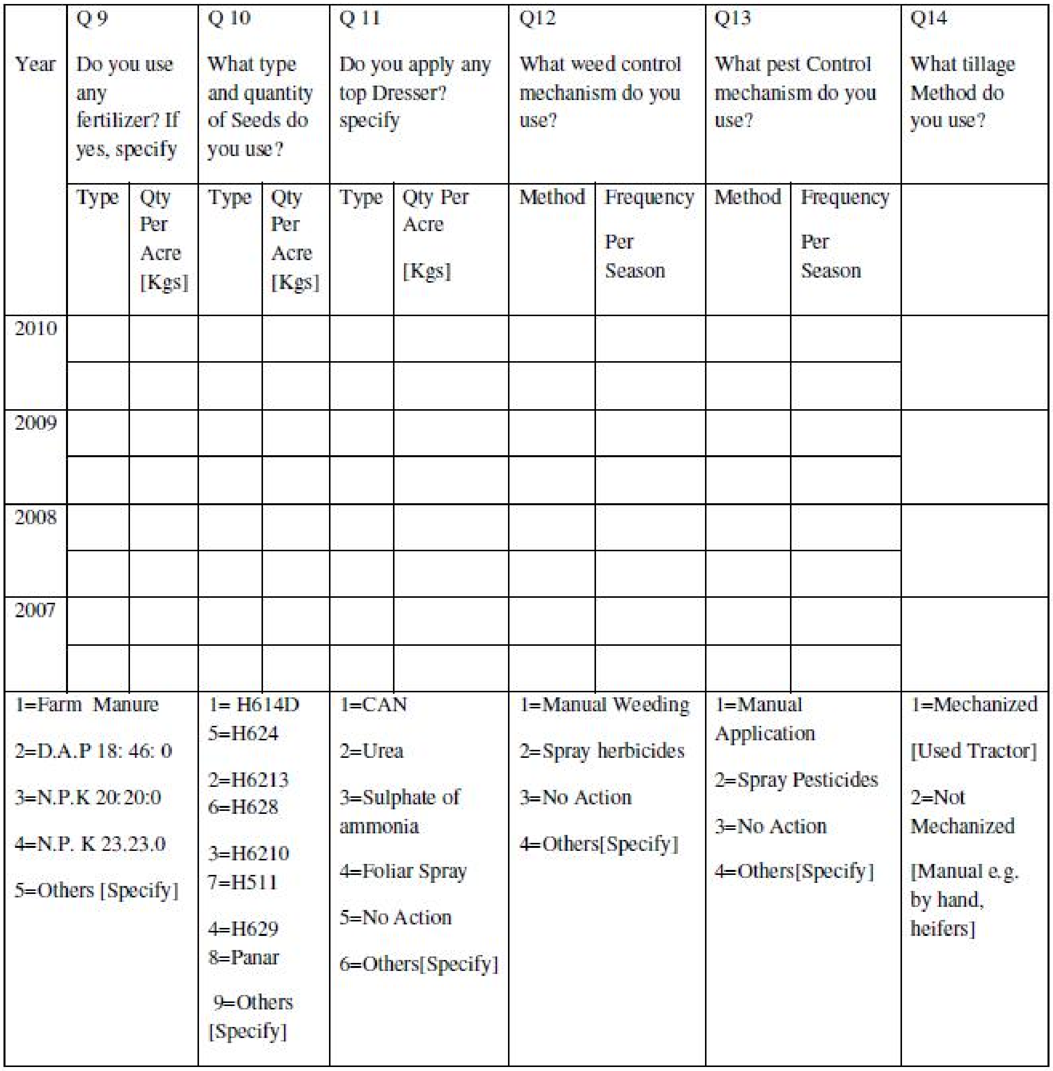
Appendix 1: Questionnaire

Part I: Biodata [The choices for each question are listed at the bottom of the table]

Location (Stratum): . . . .

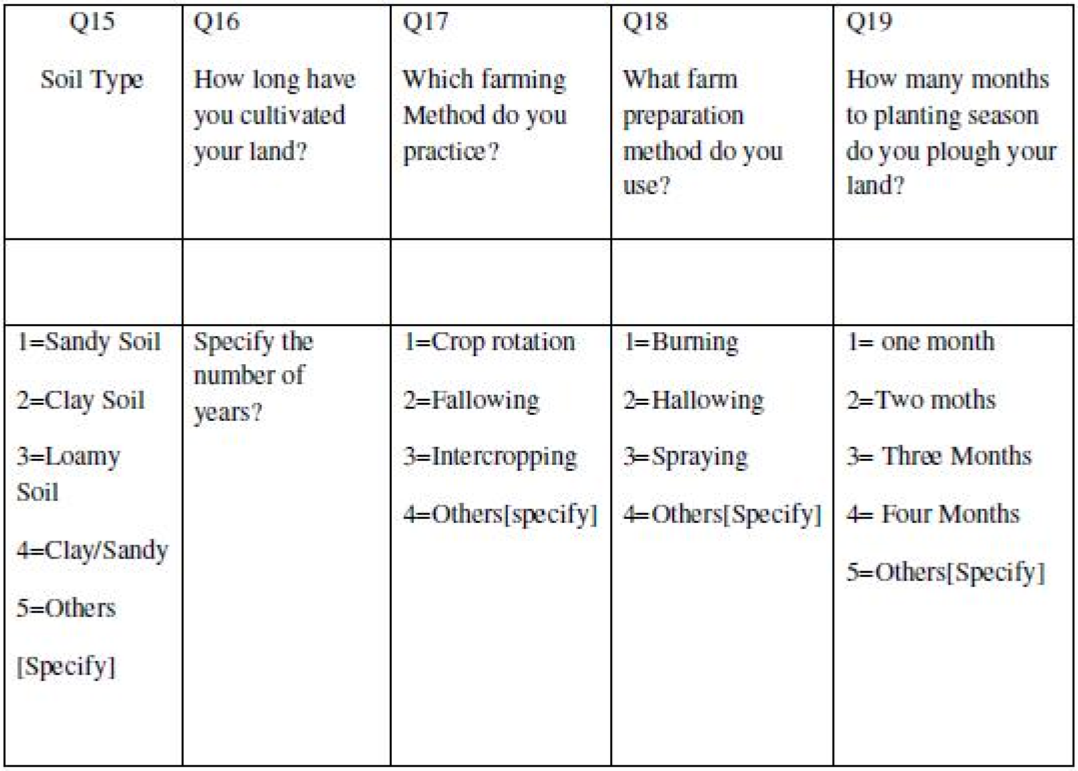
Q 1 Name of Respondent	Q 2 Gender	Q 3 Age	Q 4 What is your highest Education Level?	Q5 What is your marital Status?	Q6 What is the number of Members in your household?	Q7 Who is the household head? Who assists him/her?	Q8 What is your staple Food?
Kindly write your full names [Optional]	1=Male 2=Female	[No. of Years]	1=No Formal education 2=Primary 3=Secondary 4=Tertiary 5=University	1=Married 2=Single[Unmarried] 3=Widow/Widower 4=Divorced 5=Separated	House hold is defined as people leaving together and having a common eating arrangement [According to UN]	1=Father 2=Mother 3=Son(s) 4=Daughter(s) 5=Others [Specify]	1=Maize 2=Other(s) [Specify]

Part II: Inputs per Acre [The choices for each question are listed at the bottom of the table]



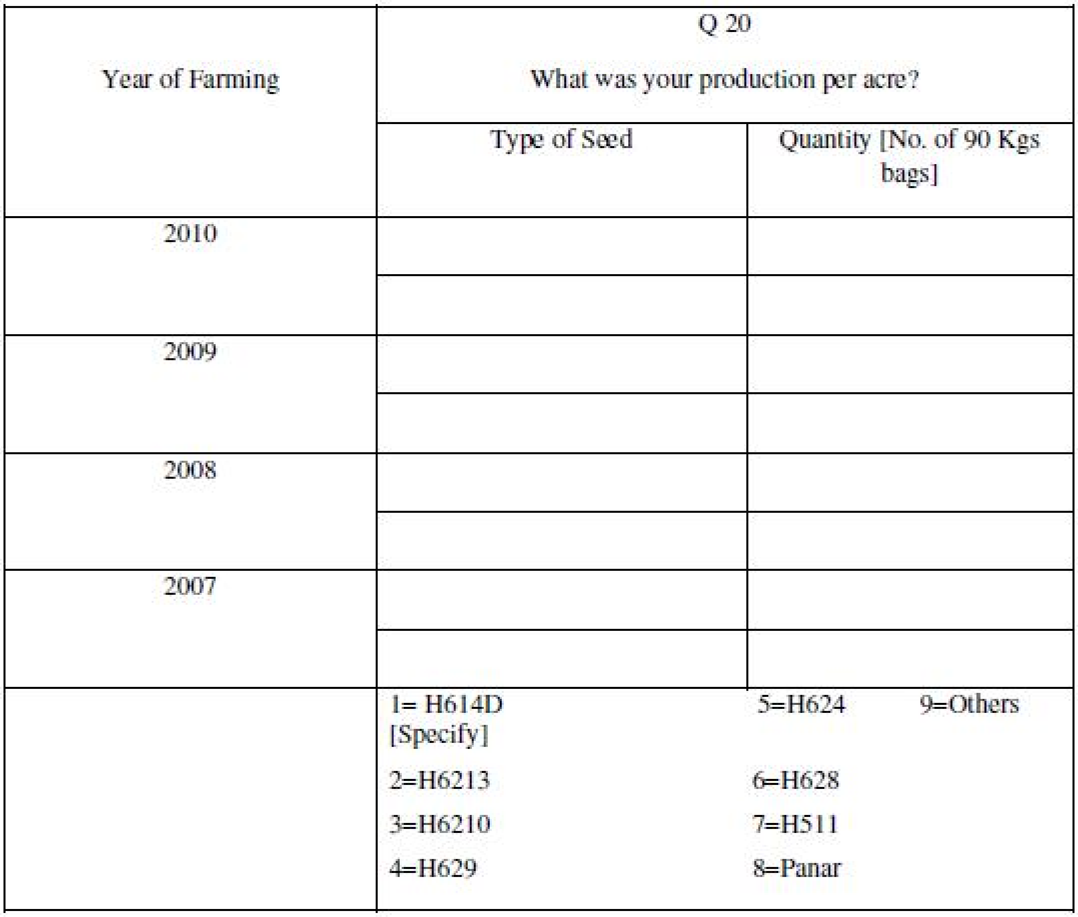
	Q 9		Q 10		Q 11		Q12		Q13		Q14
Year	Do you use any fertilizer? If yes, specify		What type and quantity of Seeds do you use?		Do you apply any top Dresser? specify		What weed control mechanism do you use?		What pest Control mechanism do you use?		What tillage Method do you use?
	Type	Qty Per Acre [Kgs]	Type	Qty Per Acre [Kgs]	Type	Qty Per Acre [Kgs]	Method	Frequency Per Season	Method	Frequency Per Season	
2010											
2009											
2008											
2007											
1=Farm Manure 2=D.A.P 18: 46: 0 3=N.P.K 20:20:0 4=N.P. K 23.23.0 5=Others [Specify]			1= H614D 5=H624 2=H6213 6=H628 3=H6210 7=H511 4=H629 8=Panar 9=Others [Specify]		1=CAN 2=Urea 3=Sulphate of ammonia 4=Foliar Spray 5=No Action 6=Others[Specify]		1=Manual Weeding 2=Spray herbicides 3=No Action 4=Others[Specify]		1=Manual Application 2=Spray Pesticides 3=No Action 4=Others[Specify]		1=Mechanized [Used Tractor] 2=Not Mechanized [Manual e.g. by hand, heifers]

Part III: Site Characterization [The choices for each of the items are listed at the bottom of the table]



Q15 Soil Type	Q16 How long have you cultivated your land?	Q17 Which farming Method do you practice?	Q18 What farm preparation method do you use?	Q19 How many months to planting season do you plough your land?
1=Sandy Soil 2=Clay Soil 3=Loamy Soil 4=Clay/Sandy 5=Others [Specify]	Specify the number of years?	1=Crop rotation 2=Fallowing 3=Intercropping 4=Others[specify]	1=Burning 2=Hallowing 3=Spraying 4=Others[Specify]	1= one month 2=Two moths 3= Three Months 4= Four Months 5=Others[Specify]

Part IV: Output per Acre



Year of Farming	Q 20 What was your production per acre?	
	Type of Seed	Quantity [No. of 90 Kgs bags]
2010		
2009		
2008		
2007		
	1= H614D [Specify] 2=H6213 3=H6210 4=H629	5=H624 9=Others 6=H628 7=H511 8=Panar

Appendix 2: Programming Codes

```
####################################################################
setwd("C:\\S-Disk\\Research\\JCC\\out")
nloc=10   # number of locations
ni =6     # farmers per location
n=nloc*ni # sample size
nt=12     # number of obs/times
N = n*nt  # total num of observations
loc=as.numeric(gl(nloc,ni))
idx=as.numeric(gl(n,nt))           # subject
# set.seed(123)          # keep seed for future replica
# loc.coeffs=round(c(0,rnorm(nloc-1,0,1)),2) # location coeff; allow for
above/below zero
loc.coeffs = c(0.00, -0.56, -0.23,  1.56,  0.07,  0.13,  1.72 , 0.46
,-1.27, -0.69)
alpha =  10    # intercept
fert.coeff= 0.15       # coefficient of fertilizer
time.coeffs = c(0.75,0.05) # allow for quadratic time effect
sigma  =  sqrt(5) # standard devition of the error, yield to vary
between 20-30
loc2=loc[idx] # locations for all
tt = rep(1:nt,ni*nloc) # a vector of time
tt2=tt^2               # time square
#output to file
cat(file="out-fixed-1.txt","Iteration Intercept Fert Loc2 Loc3
Loc4 Loc5 Loc6 Loc7 Loc8 Loc9 Loc10 Time Time2\n",append=F)
nsim = 1000
for(iter in 1:nsim)
{
Fert=round(runif(n,75,100)) # allow to vary 75 to 100
Fert2=ceiling(rnorm(N,Fert[idx],5)) # fixed per farmer but
with random varitions over time
error = rnorm(N,0,sigma) # error
Yield=alpha+ fert.coeff*Fert2+loc.coeffs[loc2]+
time.coeffs[1]*tt+time.coeffs[2]*tt2 + error
fit.fixed=lm(Yield~Fert2+as.factor(loc2)+tt+tt2)
cat(file="out-fixed-1.txt",iter,fit.fixed$coeff,"\n",append=T)
cat("\tIteration = ",iter,"\n")
}
#################################################################
setwd("C:\\S-Disk\\Research\\JCC\\out")
setwd("C:\\Users\\Joseph\\Documents\\Chelule\\out")
nloc=10   # number of locations
ni =6     # farmers per location
n=nloc*ni # sample size
nt=12     # number of obs/times
N = n*nt  # total num of observations
loc=as.numeric(gl(nloc,ni))
idx=as.numeric(gl(n,nt))           # subject
# set.seed(123)          # keep seed for future replica
# loc.coeffs=round(c(0,rnorm(nloc-1,0,1)),2) # location
coeff; allow for above/below zero
loc.coeffs = c(0.00, -0.56, -0.23,  1.56,  0.07,  0.13,
 1.72 , 0.46 ,-1.27, -0.69)
alpha =  10    # intercept
fert.coeff= 0.15       # coefficient of fertilizer
time.coeffs = c(0.75,0.05) # allow for quadratic time effect
sigma  =  sqrt(5) # standard devition of the error, yield
to vary between 20-30
loc2=loc[idx] # locations for all
tt = rep(1:nt,ni*nloc) # a vector of time
tt2=tt^2               # time square
ran.ints = scan()
 -1.25 -0.51  3.49  0.16  0.29  3.84  1.03 -2.83 -1.54 -1.00  2.74  0.80
  0.90  0.25 -1.24  4.00  1.11 -4.40  1.57 -1.06 -2.39 -0.49 -2.29 -1.63
 -1.40 -3.77  1.87  0.34 -2.54  2.80  0.95 -0.66  2.00  1.96  1.84  1.54
  1.24 -0.14 -0.68 -0.85 -1.55 -0.46 -2.83  4.85  2.70 -2.51 -0.90 -1.04
  1.74 -0.19  0.57 -0.06 -0.10  3.06 -0.50  3.39 -3.46  1.31  0.28  0.48
#output to file
cat(file="out-random intercept.txt","Iteration Intercept Fert Loc2 Loc3
Loc4 Loc5 Loc6 Loc7 Loc8 Loc9 Loc10 Time Time2  ranint1  ranint2  ranint3
ranint4 ranint5  ranint6  ranint7  ranint8  ranint9  ranint10 ranint11
ranint12  ranint13 ranint14 ranint15 ranint16 ranint17 ranint18 ranint19
ranint20 ranint21 ranint22 ranint23 ranint24  ranint25 ranint26 ranint27
ranint28 ranint29 ranint30 ranint31 ranint32 ranint33 ranint34 ranint35
ranint36 ranint37 ranint38 ranint39 ranint40 ranint41 ranint42 ranint43
ranint44 ranint45 ranint46 ranint47 ranint48 ranint49 ranint50 ranint51
ranint52 ranint53 ranint54 ranint55 ranint56 ranint57 ranint58 ranint59
ranint60\n",append=F)
nsim = 1000
library(nlme)
for(iter in 1:nsim)
{
Fert=round(runif(n,75,100))       # allow to vary 75 to 100
Fert2=ceiling(rnorm(N,Fert[idx],5)) # fixed per farmer but with random
varitions over time
error = rnorm(N,0,sigma) # error
Yield=alpha+ fert.coeff*Fert2+loc.coeffs[loc2]+ time.coeffs[1]*tt+
time.coeffs[2]*tt2 + error + ran.ints[idx]
fit.mixed=lme(Yield~Fert2+as.factor(loc2)+tt+tt2,random=~1|idx)
cat(file="out-random intercept.txt",iter,
    as.vector(c(fit.mixed$coeff$fixed,fit.mixed$coeff$random$idx)),
"\n",append=T)
cat("\tIteration = ",iter,"\n")
}
################################################################

setwd("C:\\Users\\Joseph\\Documents\\Chelule\\out")
nloc=10   # number of locations
ni =6     # farmers per location
```

```
n=nloc*ni # sample size
nt=12     # number of obs/times
N = n*nt  # total num of observations
loc=as.numeric(gl(nloc,ni))
idx=as.numeric(gl(n,nt))           # subject


#set.seed(123)          # keep seed for future replica
#loc.coeffs=round(c(0,rnorm(nloc-1,0,1)),2) # location coeff;
allow for above/below zero
loc.coeffs = c(0.00, -0.56, -0.23,  1.56,  0.07,  0.13,  1.72
, 0.46 ,-1.27, -0.69)
alpha =  10    # intercept
fert.coeff= 0.15       # coefficient of fertilizer
time.coeffs = c(0.75,0.05) # allow for quadratic time effect
sigma  =  sqrt(5) # standard devition of the error, yield to
vary between 20-30
loc2=loc[idx] # locations for all
tt = rep(1:nt,ni*nloc) # a vector of time
tt2=tt^2               # time square

ran.ints = scan()
 -1.25 -0.51  3.49  0.16  0.29  3.84  1.03 -2.83 -1.54 -1.00  2.74  0.80
  0.90  0.25 -1.24  4.00  1.11 -4.40  1.57 -1.06 -2.39 -0.49 -2.29 -1.63
 -1.40 -3.77  1.87  0.34 -2.54  2.80  0.95 -0.66  2.00  1.96  1.84  1.54
```

```
  1.24 -0.14 -0.68 -0.85 -1.55 -0.46 -2.83  4.85  2.70 -2.51 -0.90 -1.04
  1.74 -0.19  0.57 -0.06 -0.10  3.06 -0.50  3.39 -3.46  1.31  0.28  0.48

#output to file
cat(file="out-random intercept.txt","Iteration Intercept Fert Loc2 Loc3
Loc4 Loc5 Loc6 Loc7 Loc8 Loc9 Loc10 Time Time2  ranint1  ranint2  ranint3
ranint4  ranint5  ranint6  ranint7  ranint8  ranint9  ranint10 ranint11
ranint12  ranint13 ranint14 ranint15 ranint16 ranint17 ranint18 ranint19
ranint20 ranint21 ranint22 ranint23 ranint24  ranint25 ranint26 ranint27
ranint28 ranint29 ranint30 ranint31 ranint32 ranint33 ranint34 ranint35
ranint36  ranint37 ranint38 ranint39 ranint40 ranint41 ranint42 ranint43
ranint44 ranint45 ranint46 ranint47 ranint48   ranint49 ranint50
ranint51 ranint52 ranint53 ranint54 ranint55 ranint56 ranint57 ranint58
ranint59 ranint60\n",append=F)
nsim = 100
library(nlme)

for(iter in 1:nsim)
{
Fert=round(runif(n,75,100))        # allow to vary 75 to 100
Fert2=ceiling(rnorm(N,Fert[idx],5)) # fixed per farmer but with random
varitions over time
error = rnorm(N,0,sigma) # error
Yield=alpha+ fert.coeff*Fert2+loc.coeffs[loc2]+ time.coeffs[1]*tt+
time.coeffs[2]*tt2 + error + ran.ints[idx]
fit.mixed=lme(Yield~Fert2+as.factor(loc2)+tt+tt2,random=~1|idx)
cat(file="out-random intercept.txt",iter,
    as.vector(c(fit.mixed$coeff$fixed,fit.mixed$coeff$random$idx)),
"\n",append=T)
cat("\tIteration = ",iter,"\n")
}


################################################################

setwd("C:\\Users\\Documents\\Chelule\\out")
simrani=read.table("out-random intercept.txt",h=T)
boxplot(simrani[,2:14])
boxplot(simrani[,15:74],axes=F)
axis(2)
axis(1,1:60,1:60)
box()



fixe=simrani[,2:14]
rani=simrani[,15:74]
CI = function(X) c(mean(X)-1.96*sd(X),mean(X)+1.96*sd(X))

outf=round(t(apply(fixe,2,function(x) rbind(mean(x),sd(x),CI(x)[1],
CI(x)[2]))),3)
```

```
outr=round(t(apply(rani,2,function(x) rbind(mean(x),sd(x),CI(x)[1],
CI(x)[2]))),3)

cover <- function(x,t)
{
ci=CI(x)
out= (x>ci[1]) * (x<ci[2])
n=length(x)
round((sum(out)/n)*100,1)
}
rmse <- function(x,t)
{
bias= t-mean(x)
vari=var(x)
mse=bias^2+vari
rmse=sqrt(mse)
round(c(bias,rmse),3)
}
loc.coeffs2 = c(-0.56, -0.23,  1.56,  0.07,  0.13,  1.72 ,
0.46 ,-1.27, -0.69)
alpha =  10    # intercept
fert.coeff= 0.15       # coefficient of fertilizer
time.coeffs = c(0.75,0.05) # allow for quadratic time effect
ran.ints = scan()
 -1.25 -0.51  3.49  0.16  0.29  3.84  1.03 -2.83 -1.54 -1.00  2.74  0.80
```

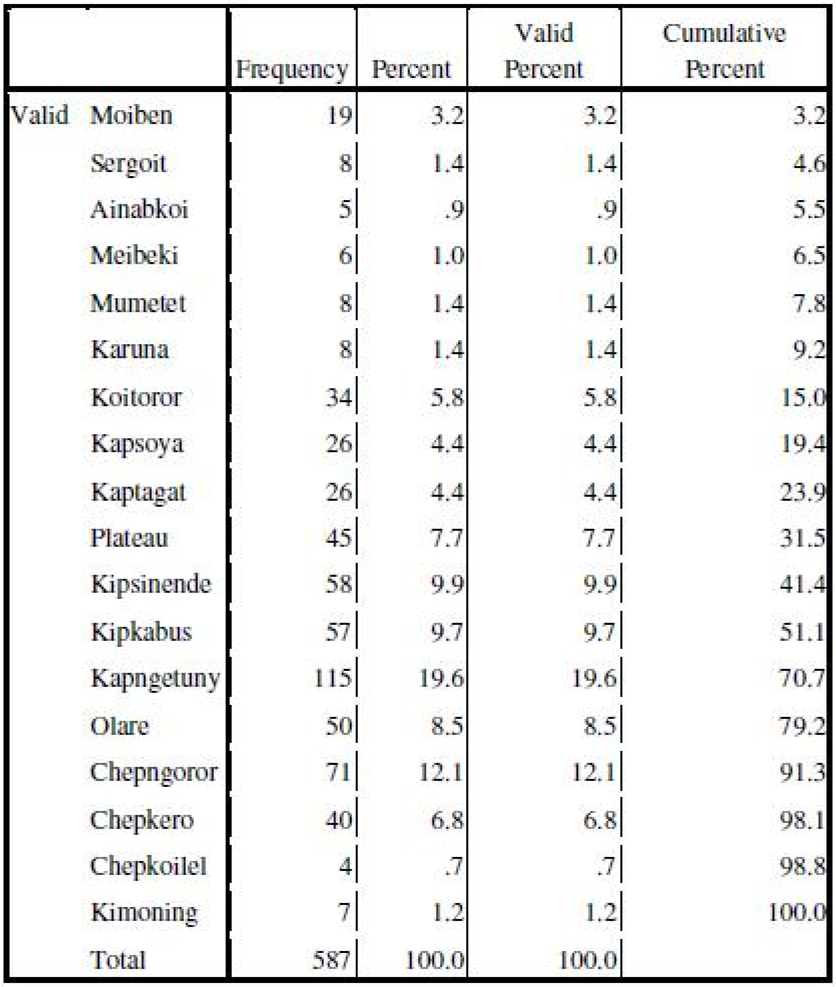
```
  0.90  0.25 -1.24  4.00  1.11 -4.40  1.57 -1.06 -2.39 -0.49 -2.29 -1.63
 -1.40 -3.77  1.87  0.34 -2.54  2.80  0.95 -0.66  2.00  1.96  1.84  1.54
  1.24 -0.14 -0.68 -0.85 -1.55 -0.46 -2.83  4.85  2.70 -2.51 -0.90 -1.04
  1.74 -0.19  0.57 -0.06 -0.10  3.06 -0.50  3.39 -3.46  1.31  0.28  0.48
true.coeffs = c(alpha,fert.coeff,loc.coeffs2,time.coeffs)
covrmse=NULL
for(j in 1:13)
 covrmse =rbind(covrmse, c(cover=cover(fixe[,j], true.coeffs[j]),
                         bias = rmse(fixe[,j], true.coeffs[j])[1],
                         rmse = rmse(fixe[,j], true.coeffs[j])[2]))
ans=cbind(true.coeffs,outf,covrmse)
colnames(ans)=c("true","estimate","SE","LCI","UCI","coverage","bias","rmse")
ans

covrmser=NULL
for(j in 1:60)
 covrmser =rbind(covrmser, c(cover=cover(rani[,j], ran.ints[j]),
                         bias = rmse(rani[,j], ran.ints[j])[1],
                         rmse = rmse(rani[,j], ran.ints[j])[2]))

ans2=cbind(ran.ints,outr,covrmser)
colnames(ans2)=c("true","estimate","SE","LCI","UCI","coverage","bias","rmse")
ans2
```

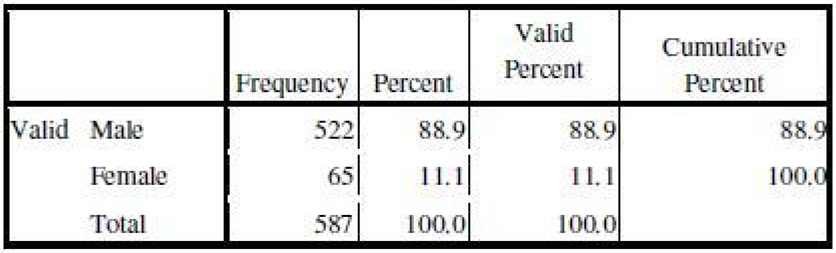
Appendix 3: Frequency Distribution Tables

Respondents as per Stratum (Location)

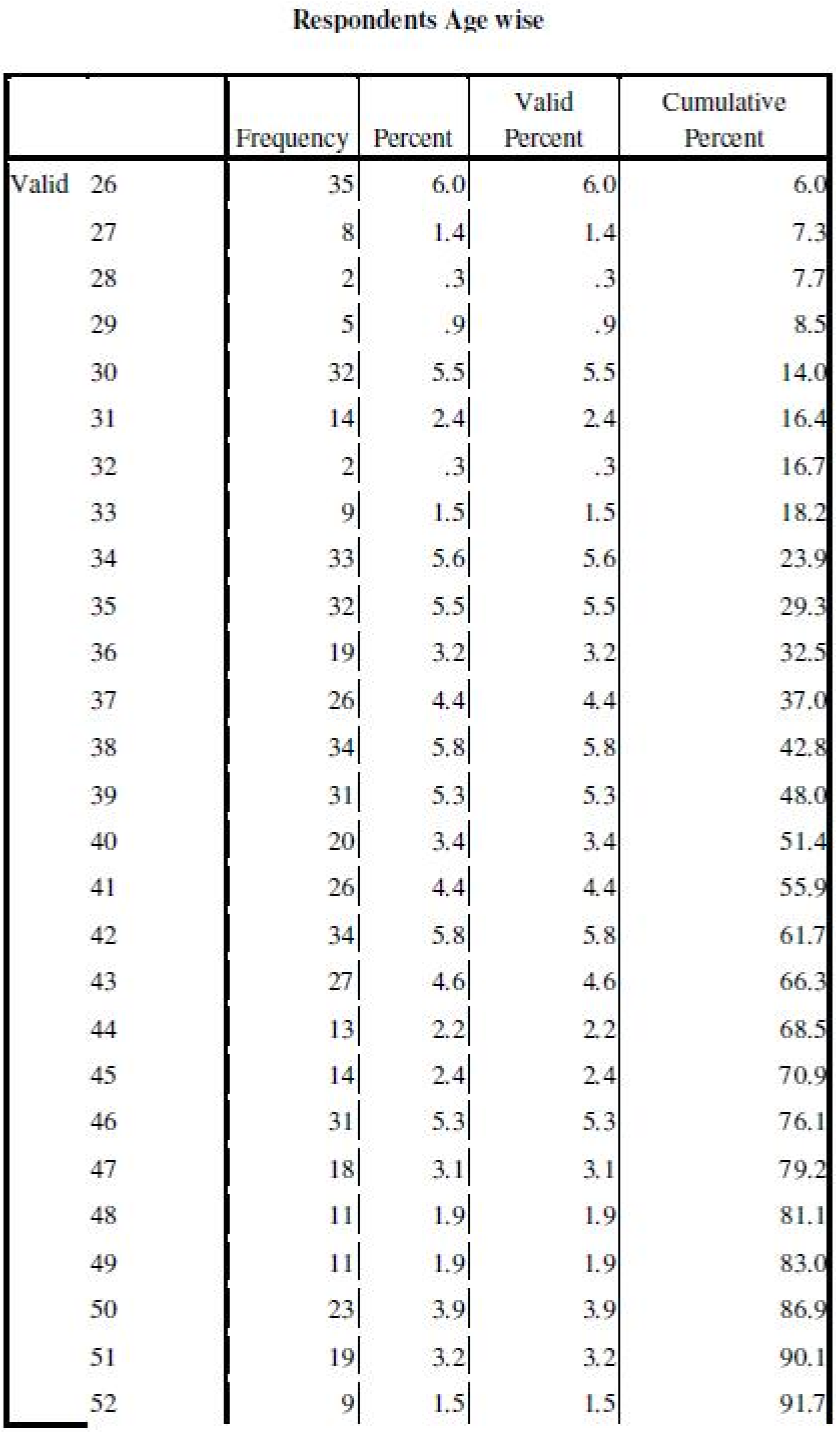


		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Moiben	19	3.2	3.2	3.2
	Sergoit	8	1.4	1.4	4.6
	Ainabkoi	5	.9	.9	5.5
	Meibeki	6	1.0	1.0	6.5
	Mumetet	8	1.4	1.4	7.8
	Karuna	8	1.4	1.4	9.2
	Koitoror	34	5.8	5.8	15.0
	Kapsoya	26	4.4	4.4	19.4
	Kaptagat	26	4.4	4.4	23.9
	Plateau	45	7.7	7.7	31.5
	Kipsinende	58	9.9	9.9	41.4
	Kipkabus	57	9.7	9.7	51.1
	Kapngetuny	115	19.6	19.6	70.7
	Olare	50	8.5	8.5	79.2
	Chepngoror	71	12.1	12.1	91.3
	Chepkero	40	6.8	6.8	98.1
	Chepkoilel	4	.7	.7	98.8
	Kimoning	7	1.2	1.2	100.0
	Total	587	100.0	100.0	

Respondents as per Gender

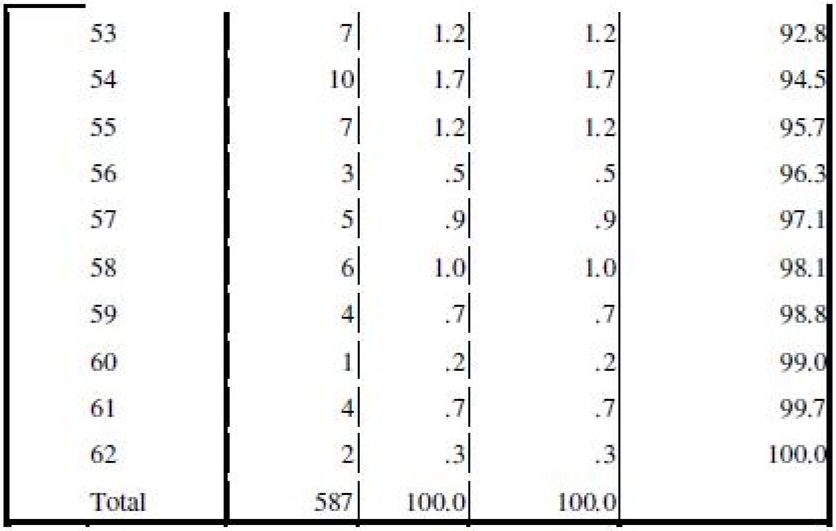


		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	522	88.9	88.9	88.9
	Female	65	11.1	11.1	100.0
	Total	587	100.0	100.0	



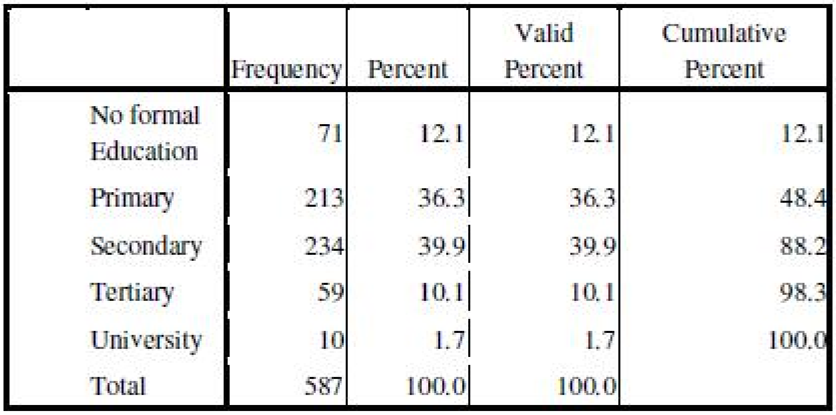
Respondents Age wise

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	26	35	6.0	6.0	6.0
	27	8	1.4	1.4	7.3
	28	2	.3	.3	7.7
	29	5	.9	.9	8.5
	30	32	5.5	5.5	14.0
	31	14	2.4	2.4	16.4
	32	2	.3	.3	16.7
	33	9	1.5	1.5	18.2
	34	33	5.6	5.6	23.9
	35	32	5.5	5.5	29.3
	36	19	3.2	3.2	32.5
	37	26	4.4	4.4	37.0
	38	34	5.8	5.8	42.8
	39	31	5.3	5.3	48.0
	40	20	3.4	3.4	51.4
	41	26	4.4	4.4	55.9
	42	34	5.8	5.8	61.7
	43	27	4.6	4.6	66.3
	44	13	2.2	2.2	68.5
	45	14	2.4	2.4	70.9
	46	31	5.3	5.3	76.1
	47	18	3.1	3.1	79.2
	48	11	1.9	1.9	81.1
	49	11	1.9	1.9	83.0
	50	23	3.9	3.9	86.9
	51	19	3.2	3.2	90.1
	52	9	1.5	1.5	91.7



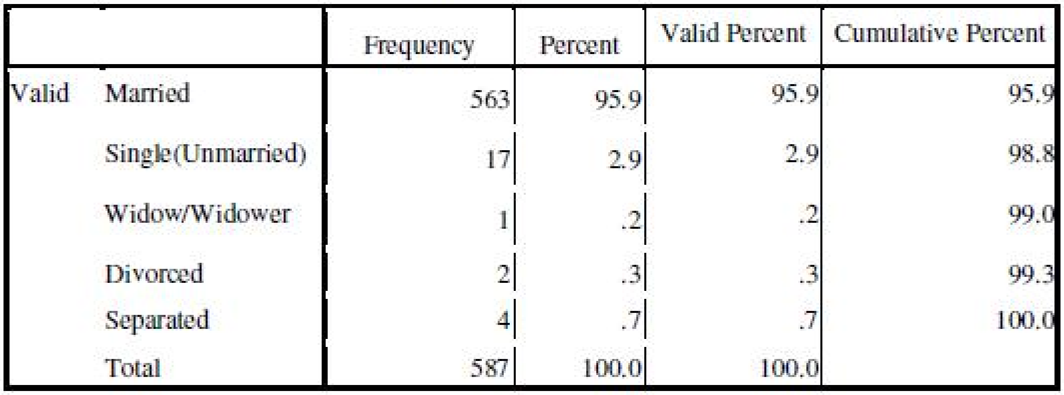
53	7	1.2	1.2	92.8
54	10	1.7	1.7	94.5
55	7	1.2	1.2	95.7
56	3	.5	.5	96.3
57	5	.9	.9	97.1
58	6	1.0	1.0	98.1
59	4	.7	.7	98.8
60	1	.2	.2	99.0
61	4	.7	.7	99.7
62	2	.3	.3	100.0
Total	587	100.0	100.0	

Respondents' Level of Education



	Frequency	Percent	Valid Percent	Cumulative Percent
No formal Education	71	12.1	12.1	12.1
Primary	213	36.3	36.3	48.4
Secondary	234	39.9	39.9	88.2
Tertiary	59	10.1	10.1	98.3
University	10	1.7	1.7	100.0
Total	587	100.0	100.0	

Respondents' Marital Status



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Married	563	95.9	95.9	95.9
	Single(Unmarried)	17	2.9	2.9	98.8
	Widow/Widower	1	.2	.2	99.0
	Divorced	2	.3	.3	99.3
	Separated	4	.7	.7	100.0
	Total	587	100.0	100.0	

Respondents' household number of Membership

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	27	4.6	4.6	4.6
	2	85	14.5	14.5	19.1
	3	108	18.4	18.4	37.5
	4	106	18.1	18.1	55.5
	5	70	11.9	11.9	67.5
	6	75	12.8	12.8	80.2
	7	61	10.4	10.4	90.6
	8	27	4.6	4.6	95.2
	9	20	3.4	3.4	98.6
	10	4	.7	.7	99.3
	11	4	.7	.7	100.0
	Total	587	100.0	100.0	

Respondents' Household Head

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Father	573	97.6	97.6	97.6
	Mother	14	2.4	2.4	100.0
	Total	587	100.0	100.0	

Respondents' Staple Food

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Maize	587	100.0	100.0	100.0

Type of Fertilizer Used in 2010

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	D.A.P 18:46:0	585	99.7	99.7	99.7
	N.P.K 23.23.0	1	.2	.2	99.8
	Others(Specify)	1	.2	.2	100.0
	Total	587	100.0	100.0	

Type of Fertilizer Used in 2009

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	D.A.P 18:46:0	586	99.8	99.8	99.8
	N.P.K 23.23.0	1	.2	.2	100.0
	Total	587	100.0	100.0	

Type of Fertilizer used in 2008

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	D.A.P 18:46:0	586	99.8	99.8	99.8
	N.P.K 23.23.0	1	.2	.2	100.0
	Total	587	100.0	100.0	

Type of Fertilizer used in 2007

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	D.A.P 18:46:0	585	99.7	99.7	99.7
	N.P.K 23.23.0	1	.2	.2	99.8
	Others(Specify)	1	.2	.2	100.0
	Total	587	100.0	100.0	

Quantity of Fertilizer Per Acre (in Kgs) Used in 2010

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	50	587	100.0	100.0	100.0

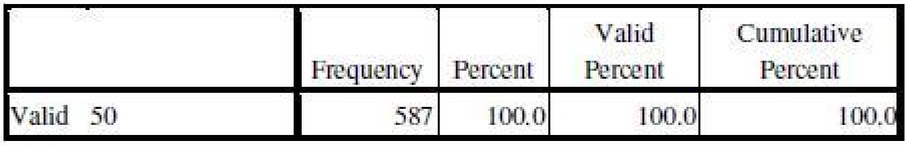
Quantity of Fertilizer Per Acre (in Kgs) Used in 2009

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	50	587	100.0	100.0	100.0

Quantity of Fertilizer Per Acre (in Kgs) Used in 2008

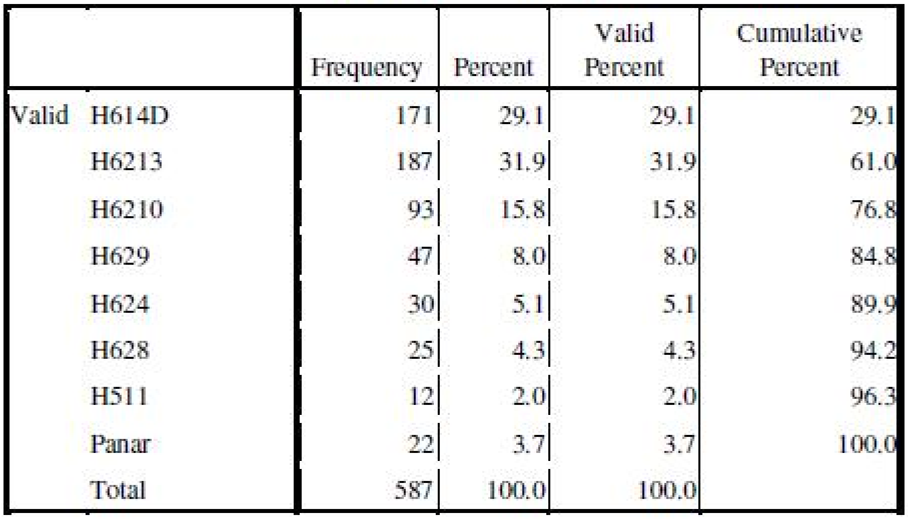
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	50	587	100.0	100.0	100.0

Quantity of Fertilizer Per Acre (in Kgs) Used in 2007



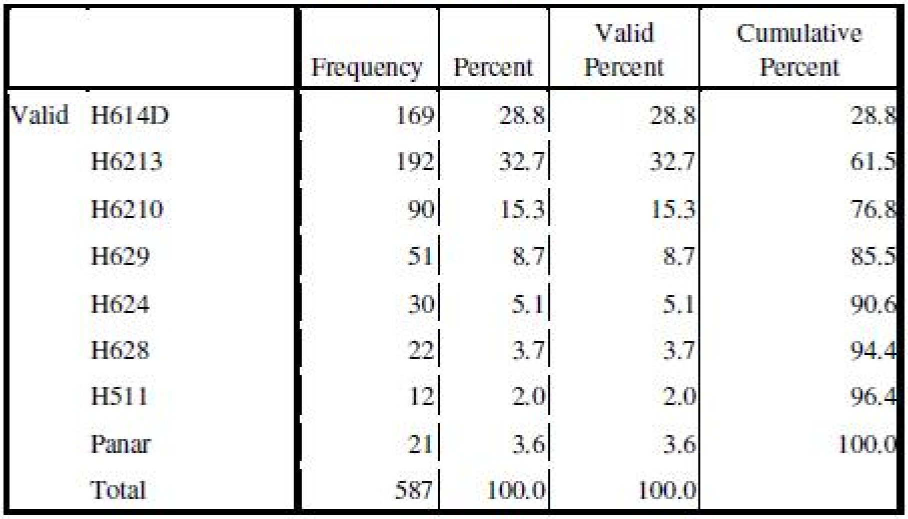
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	50	587	100.0	100.0	100.0

Seed Hybrid Used in 2010



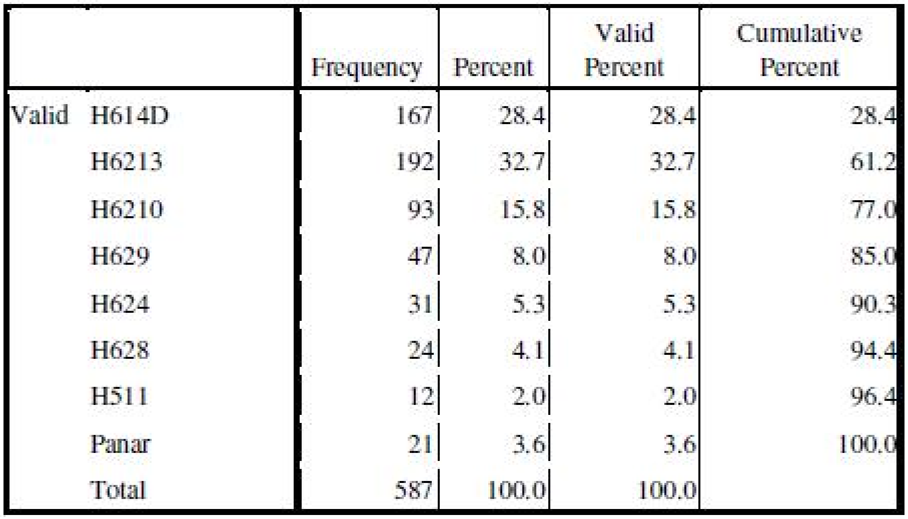
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	H614D	171	29.1	29.1	29.1
	H6213	187	31.9	31.9	61.0
	H6210	93	15.8	15.8	76.8
	H629	47	8.0	8.0	84.8
	H624	30	5.1	5.1	89.9
	H628	25	4.3	4.3	94.2
	H511	12	2.0	2.0	96.3
	Panar	22	3.7	3.7	100.0
	Total	587	100.0	100.0	

Seed Hybrid Used in 2009



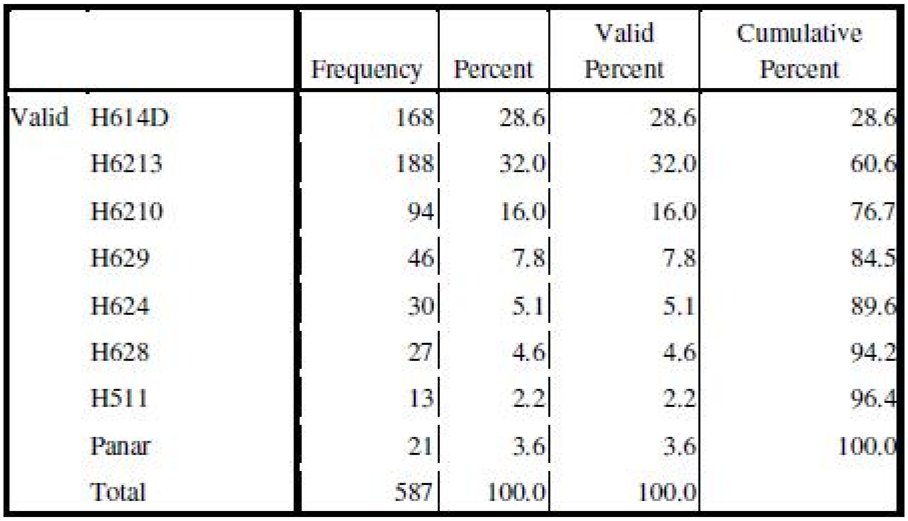
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	H614D	169	28.8	28.8	28.8
	H6213	192	32.7	32.7	61.5
	H6210	90	15.3	15.3	76.8
	H629	51	8.7	8.7	85.5
	H624	30	5.1	5.1	90.6
	H628	22	3.7	3.7	94.4
	H511	12	2.0	2.0	96.4
	Panar	21	3.6	3.6	100.0
	Total	587	100.0	100.0	

Seed Hybrid Used in 2008



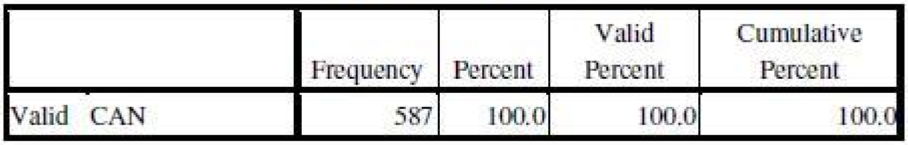
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	H614D	167	28.4	28.4	28.4
	H6213	192	32.7	32.7	61.2
	H6210	93	15.8	15.8	77.0
	H629	47	8.0	8.0	85.0
	H624	31	5.3	5.3	90.3
	H628	24	4.1	4.1	94.4
	H511	12	2.0	2.0	96.4
	Panar	21	3.6	3.6	100.0
	Total	587	100.0	100.0	

Seed Hybrid Used in 2007



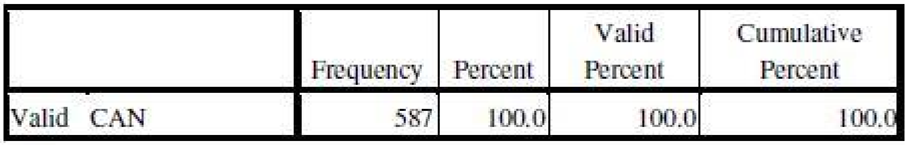
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	H614D	168	28.6	28.6	28.6
	H6213	188	32.0	32.0	60.6
	H6210	94	16.0	16.0	76.7
	H629	46	7.8	7.8	84.5
	H624	30	5.1	5.1	89.6
	H628	27	4.6	4.6	94.2
	H511	13	2.2	2.2	96.4
	Panar	21	3.6	3.6	100.0
	Total	587	100.0	100.0	

Top Dresser Used in 2010



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	CAN	587	100.0	100.0	100.0

Top Dresser Used in 2009



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	CAN	587	100.0	100.0	100.0

Top Dresser Used in 2008

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	CAN	587	100.0	100.0	100.0

Top Dresser Used in 2007

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	CAN	587	100.0	100.0	100.0

Quantity of Top Dresser Per Acre (in Kgs) Used in 2010

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	10	33	5.6	5.6	5.6
	50	224	38.2	38.2	43.8
	75	330	56.2	56.2	100.0
	Total	587	100.0	100.0	

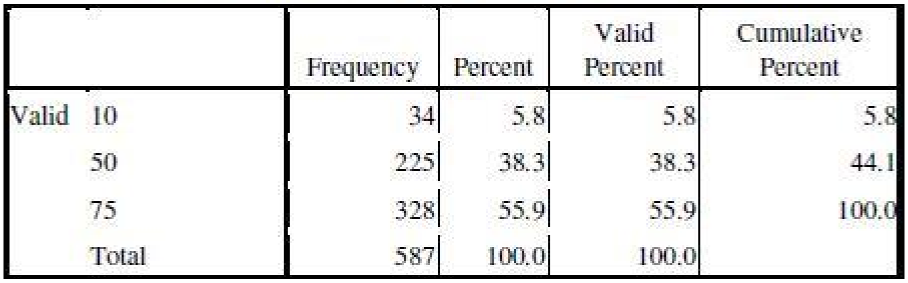
Quantity of Top Dresser Per Acre (in Kgs) Used in 2009

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	10	35	6.0	6.0	6.0
	50	223	38.0	38.0	44.0
	75	329	56.0	56.0	100.0
	Total	587	100.0	100.0	

Quantity of Top Dresser Per Acre (in Kgs) Used in 2008

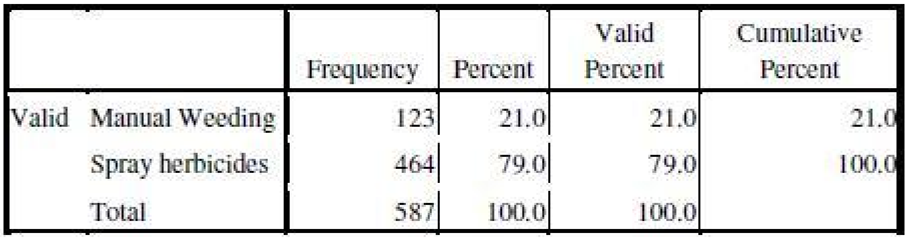
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	10	36	6.1	6.1	6.1
	50	223	38.0	38.0	44.1
	75	328	55.9	55.9	100.0
	Total	587	100.0	100.0	

Quantity of Top Dresser Per Acre (in Kgs) Used in 2007?



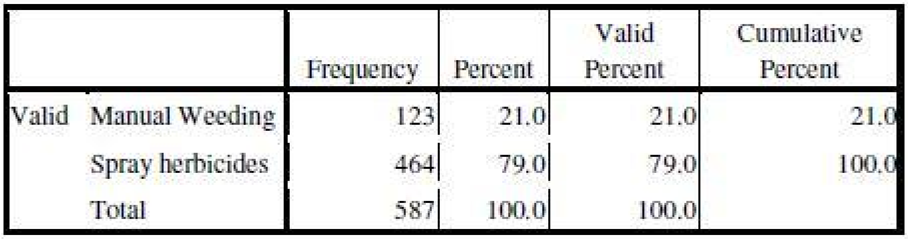
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	10	34	5.8	5.8	5.8
	50	225	38.3	38.3	44.1
	75	328	55.9	55.9	100.0
	Total	587	100.0	100.0	

Weed Control Mechanism Used in 2010



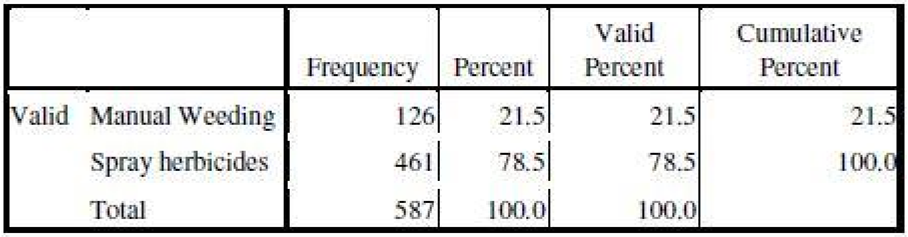
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Manual Weeding	123	21.0	21.0	21.0
	Spray herbicides	464	79.0	79.0	100.0
	Total	587	100.0	100.0	

Weed Control Mechanism Used in 2009



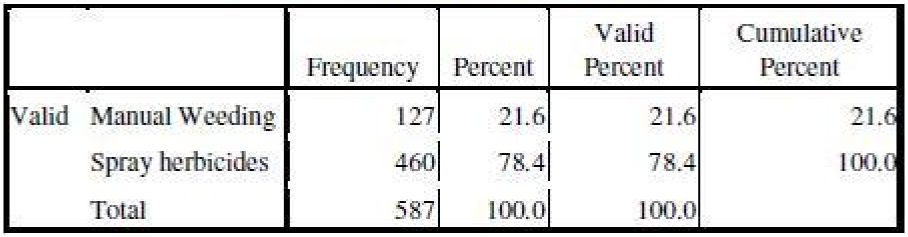
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Manual Weeding	123	21.0	21.0	21.0
	Spray herbicides	464	79.0	79.0	100.0
	Total	587	100.0	100.0	

Weed Control Mechanism Used in 2008



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Manual Weeding	126	21.5	21.5	21.5
	Spray herbicides	461	78.5	78.5	100.0
	Total	587	100.0	100.0	

Weed Control Mechanism Used in 2007



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Manual Weeding	127	21.6	21.6	21.6
	Spray herbicides	460	78.4	78.4	100.0
	Total	587	100.0	100.0	

Frequency of the Weed Control Mechanism in 2010

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	94	16.0	16.0	16.0
	2	493	84.0	84.0	100.0
	Total	587	100.0	100.0	

Frequency of the Weed Control Mechanism in 2009

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	98	16.7	16.7	16.7
	2	489	83.3	83.3	100.0
	Total	587	100.0	100.0	

Frequency of the Weed Control Mechanism in 2008

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	103	17.5	17.5	17.5
	2	484	82.5	82.5	100.0
	Total	587	100.0	100.0	

Frequency of the Weed Control Mechanism in 2007

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	105	17.9	17.9	17.9
	2	482	82.1	82.1	100.0
	Total	587	100.0	100.0	

Pest Control Mechanism Used in 2010

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Spray pesticides	587	100.0	100.0	100.0

Pest Control Mechanism Used in 2009

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Spray pestcides	587	100.0	100.0	100.0

Pest Control Mechanism Used in 2008

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Spray pestcides	587	100.0	100.0	100.0

Pest Control Mechanism Used in 2007

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Spray pestcides	587	100.0	100.0	100.0

Frequency of Application of Pest Control Mechanism in 2010

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	190	32.4	32.4	32.4
	2	244	41.6	41.6	73.9
	3	138	23.5	23.5	97.4
	4	15	2.6	2.6	100.0
	Total	587	100.0	100.0	

Frequency of Application of Pest Control Mechanism in 2009

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	194	33.0	33.0	33.0
	2	248	42.2	42.2	75.3
	3	133	22.7	22.7	98.0
	4	12	2.0	2.0	100.0
	Total	587	100.0	100.0	

Frequency of Application of Pest Control Mechanism in 2008

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	200	34.1	34.1	34.1
	2	242	41.2	41.2	75.3
	3	131	22.3	22.3	97.6
	4	14	2.4	2.4	100.0
	Total	587	100.0	100.0	

Frequency of Application of Pest Control Mechanism in 2007

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	193	32.9	32.9	32.9
	2	242	41.2	41.2	74.1
	3	140	23.9	23.9	98.0
	4	12	2.0	2.0	100.0
	Total	587	100.0	100.0	

Tillage Method Used in 2010

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Mechanized (Used Tractor)	587	100.0	100.0	100.0

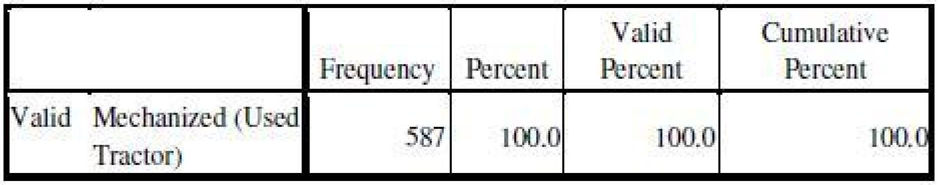
Tillage Method Used in 2009

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Mechanized (Used Tractor)	587	100.0	100.0	100.0

Tillage Method Used in 2008

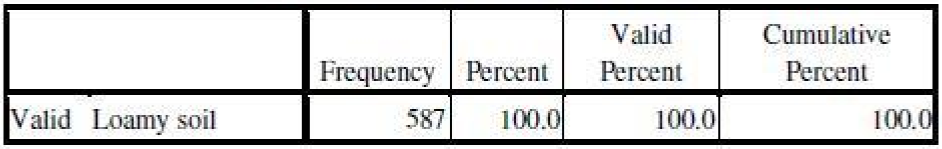
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Mechanized (Used Tractor)	587	100.0	100.0	100.0

Tillage Method Used in 2007

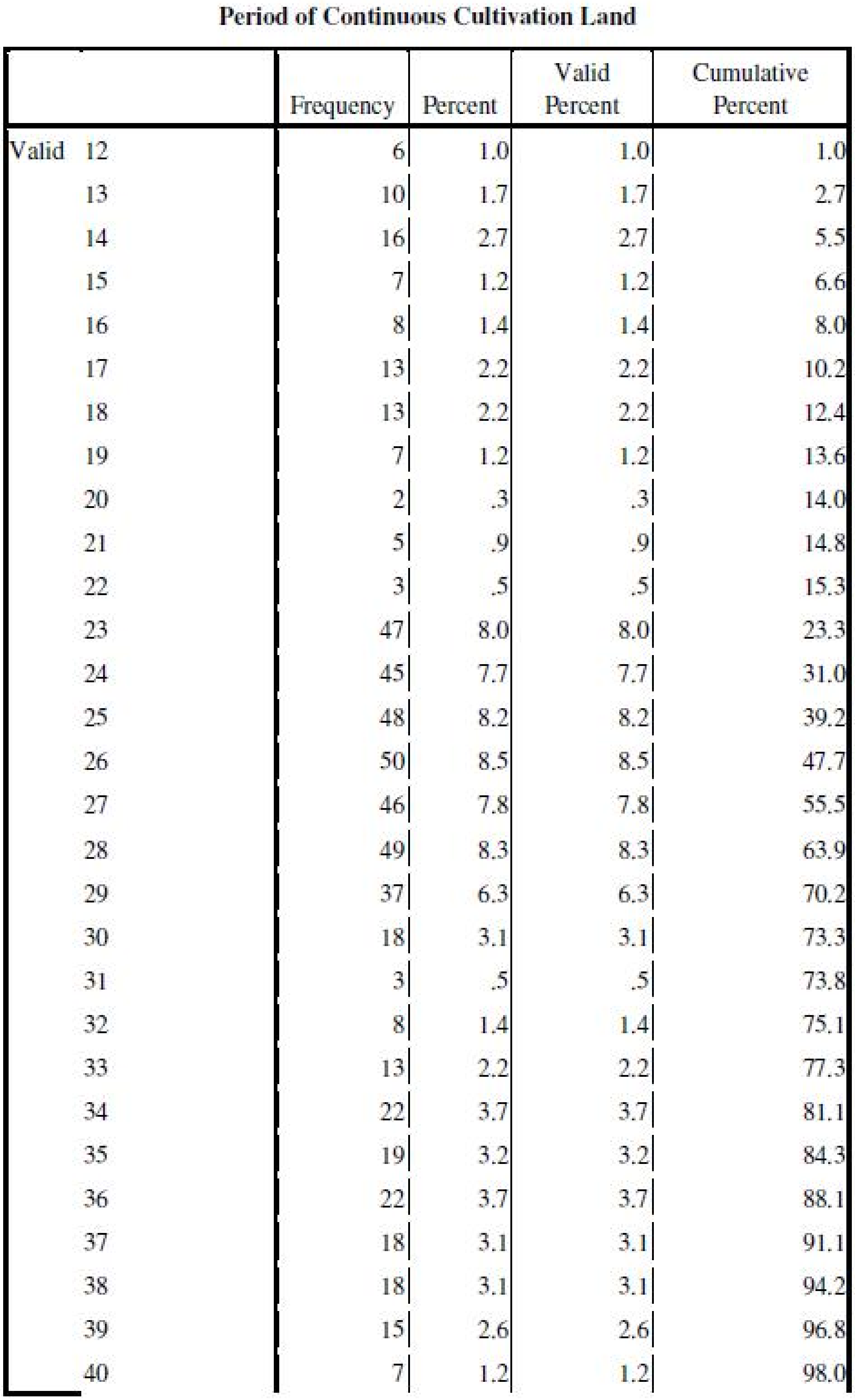


		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Mechanized (Used Tractor)	587	100.0	100.0	100.0

Type of Soil in the Respondents' Farm



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Loamy soil	587	100.0	100.0	100.0



Period of Continuous Cultivation Land

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	12	6	1.0	1.0	1.0
	13	10	1.7	1.7	2.7
	14	16	2.7	2.7	5.5
	15	7	1.2	1.2	6.6
	16	8	1.4	1.4	8.0
	17	13	2.2	2.2	10.2
	18	13	2.2	2.2	12.4
	19	7	1.2	1.2	13.6
	20	2	.3	.3	14.0
	21	5	.9	.9	14.8
	22	3	.5	.5	15.3
	23	47	8.0	8.0	23.3
	24	45	7.7	7.7	31.0
	25	48	8.2	8.2	39.2
	26	50	8.5	8.5	47.7
	27	46	7.8	7.8	55.5
	28	49	8.3	8.3	63.9
	29	37	6.3	6.3	70.2
	30	18	3.1	3.1	73.3
	31	3	.5	.5	73.8
	32	8	1.4	1.4	75.1
	33	13	2.2	2.2	77.3
	34	22	3.7	3.7	81.1
	35	19	3.2	3.2	84.3
	36	22	3.7	3.7	88.1
	37	18	3.1	3.1	91.1
	38	18	3.1	3.1	94.2
	39	15	2.6	2.6	96.8
	40	7	1.2	1.2	98.0

	41	2	.3	.3	98.3
	42	3	.5	.5	98.8
	43	3	.5	.5	99.3
	44	1	.2	.2	99.5
	45	2	.3	.3	99.8
	47	1	.2	.2	100.0
	Total	587	100.0	100.0	

Farming Method Used in 2010

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Crop rotation	587	100.0	100.0	100.0

Farming Method Used in 2009

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Crop rotation	587	100.0	100.0	100.0

Farming Method Used in 2008

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Crop rotation	587	100.0	100.0	100.0

Farming Method Used in 2007

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Crop rotation	586	99.8	99.8	99.8
	Fallowing	1	.2	.2	100.0
	Total	587	100.0	100.0	

Farm Preparation Method Used in 2010

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Burning	562	95.7	95.7	95.7
	Hallowing	25	4.3	4.3	100.0
	Total	587	100.0	100.0	

Farm Preparation Method Used in 2009

				Valid	Cumulative
		Frequency	Percent	Percent	Percent
Valid	Burning	563	95.9	95.9	95.9
	Hallowing	24	4.1	4.1	100.0
	Total	587	100.0	100.0	

Farm Preparation Method Used in 2008

				Valid	Cumulative
		Frequency	Percent	Percent	Percent
Valid	Burning	564	96.1	96.1	96.1
	Hallowing	23	3.9	3.9	100.0
	Total	587	100.0	100.0	

Farm Preparation Method Used in 2007

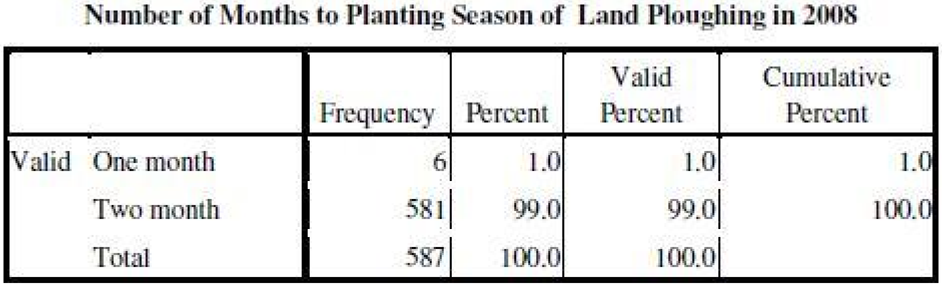
				Valid	Cumulative
		Frequency	Percent	Percent	Percent
Valid	Burning	565	96.3	96.3	96.3
	Hallowing	22	3.7	3.7	100.0
	Total	587	100.0	100.0	

Number of Months to Planting Season of Land Ploughing in 2010

				Valid	Cumulative
		Frequency	Percent	Percent	Percent
Valid	One month	6	1.0	1.0	1.0
	Two month	581	99.0	99.0	100.0
	Total	587	100.0	100.0	

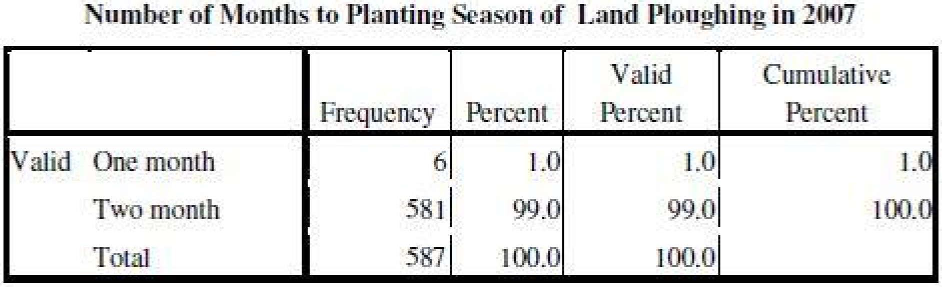
Number of Months to Planting Season of Land Ploughing in 2009

				Valid	Cumulative
		Frequency	Percent	Percent	Percent
Valid	One month	6	1.0	1.0	1.0
	Two month	581	99.0	99.0	100.0
	Total	587	100.0	100.0	



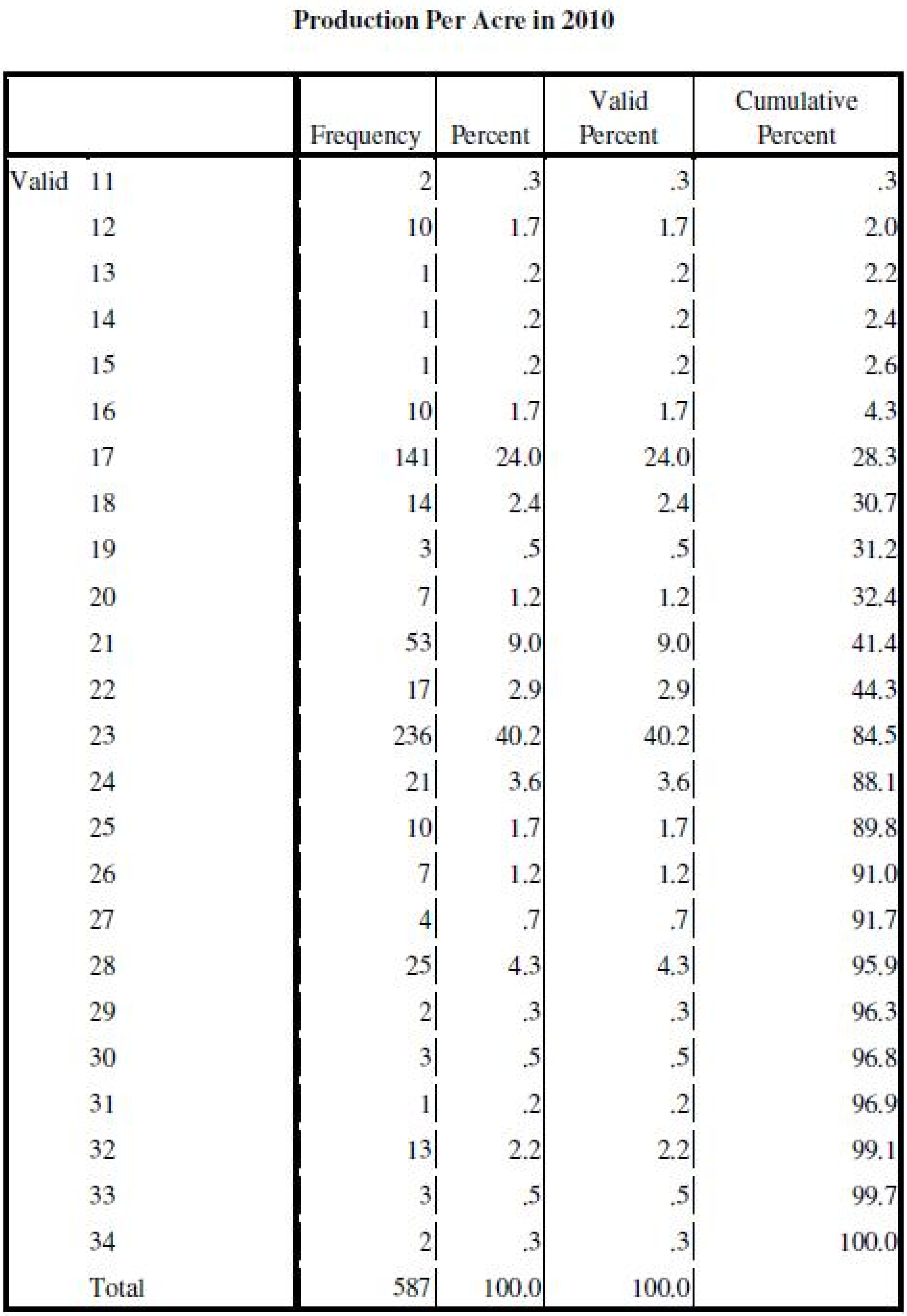
Number of Months to Planting Season of Land Ploughing in 2008

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	One month	6	1.0	1.0	1.0
	Two month	581	99.0	99.0	100.0
	Total	587	100.0	100.0	



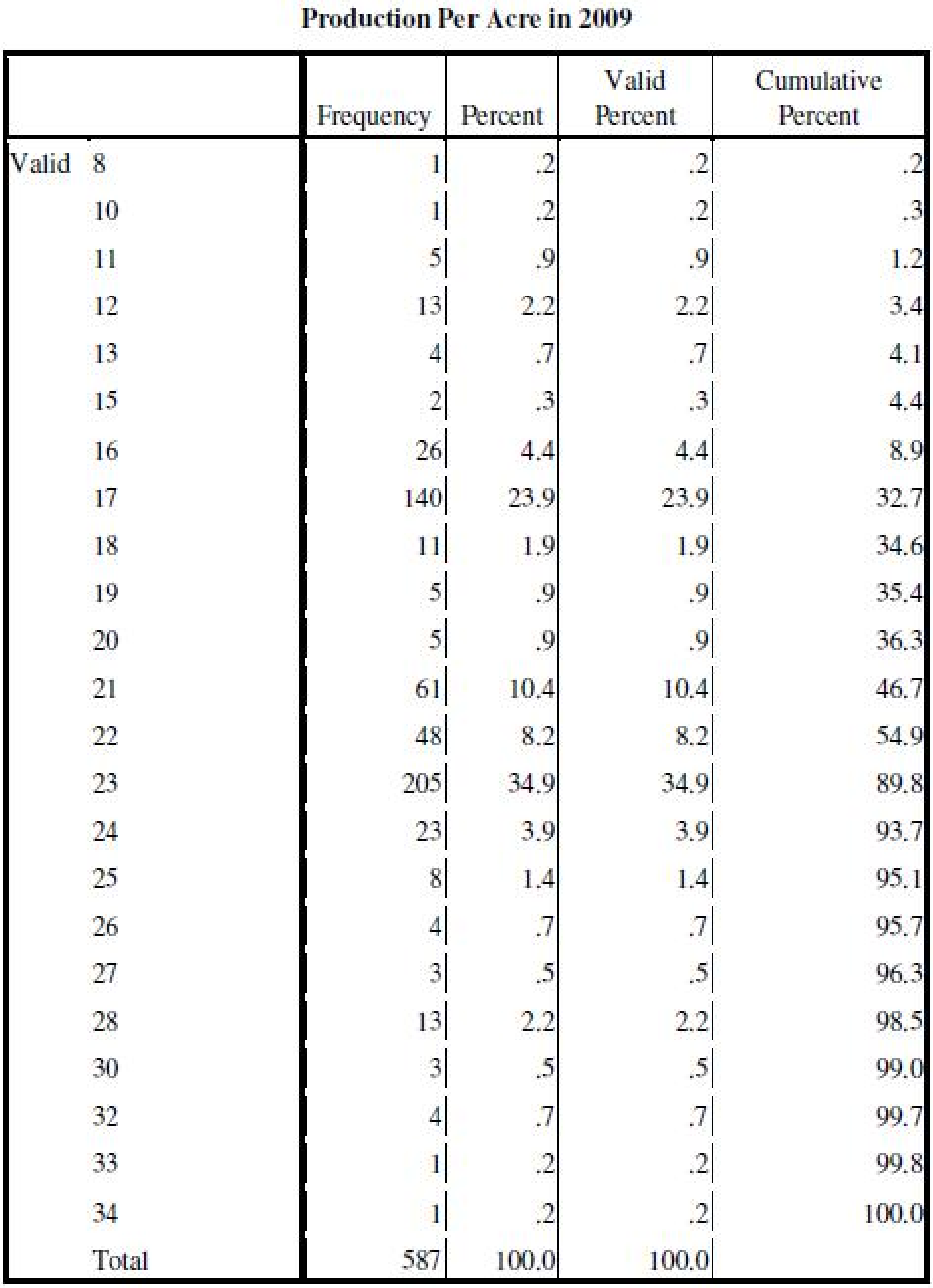
Number of Months to Planting Season of Land Ploughing in 2007

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	One month	6	1.0	1.0	1.0
	Two month	581	99.0	99.0	100.0
	Total	587	100.0	100.0	



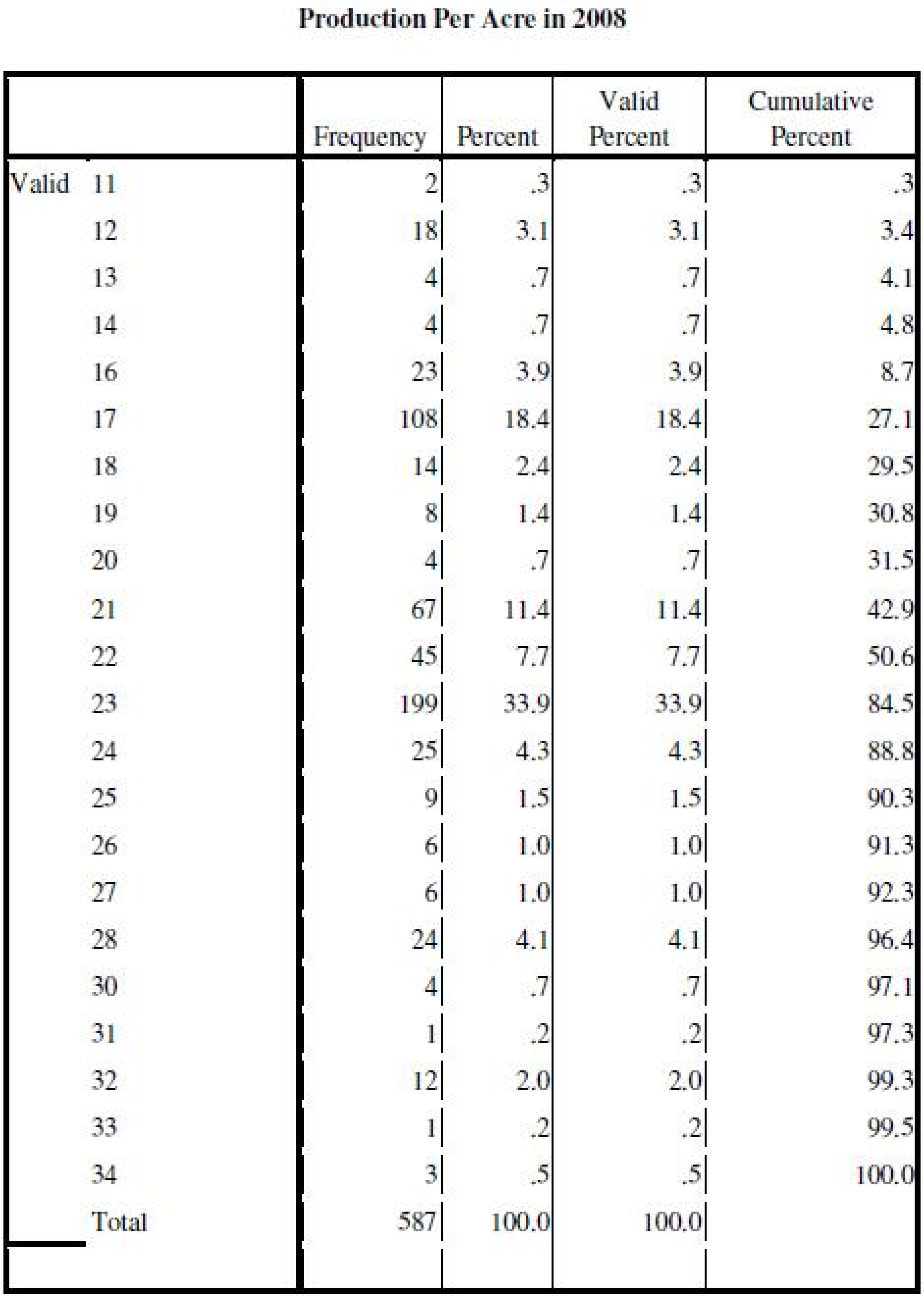
Production Per Acre in 2010

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	11	2	.3	.3	.3
	12	10	1.7	1.7	2.0
	13	1	.2	.2	2.2
	14	1	.2	.2	2.4
	15	1	.2	.2	2.6
	16	10	1.7	1.7	4.3
	17	141	24.0	24.0	28.3
	18	14	2.4	2.4	30.7
	19	3	.5	.5	31.2
	20	7	1.2	1.2	32.4
	21	53	9.0	9.0	41.4
	22	17	2.9	2.9	44.3
	23	236	40.2	40.2	84.5
	24	21	3.6	3.6	88.1
	25	10	1.7	1.7	89.8
	26	7	1.2	1.2	91.0
	27	4	.7	.7	91.7
	28	25	4.3	4.3	95.9
	29	2	.3	.3	96.3
	30	3	.5	.5	96.8
	31	1	.2	.2	96.9
	32	13	2.2	2.2	99.1
	33	3	.5	.5	99.7
	34	2	.3	.3	100.0
	Total	587	100.0	100.0	



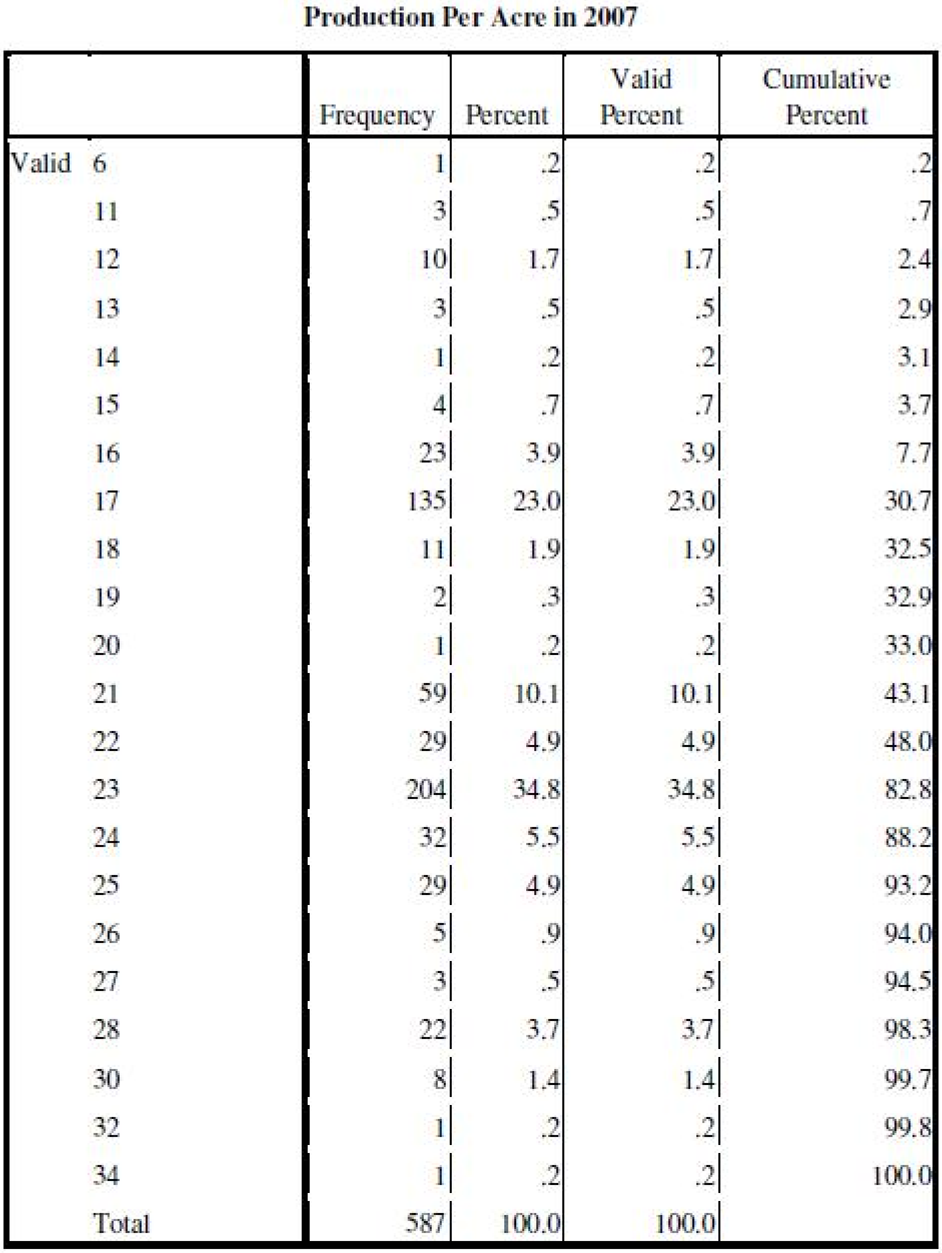
Production Per Acre in 2009

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	8	1	.2	.2	.2
	10	1	.2	.2	.3
	11	5	.9	.9	1.2
	12	13	2.2	2.2	3.4
	13	4	.7	.7	4.1
	15	2	.3	.3	4.4
	16	26	4.4	4.4	8.9
	17	140	23.9	23.9	32.7
	18	11	1.9	1.9	34.6
	19	5	.9	.9	35.4
	20	5	.9	.9	36.3
	21	61	10.4	10.4	46.7
	22	48	8.2	8.2	54.9
	23	205	34.9	34.9	89.8
	24	23	3.9	3.9	93.7
	25	8	1.4	1.4	95.1
	26	4	.7	.7	95.7
	27	3	.5	.5	96.3
	28	13	2.2	2.2	98.5
	30	3	.5	.5	99.0
	32	4	.7	.7	99.7
	33	1	.2	.2	99.8
	34	1	.2	.2	100.0
	Total	587	100.0	100.0	



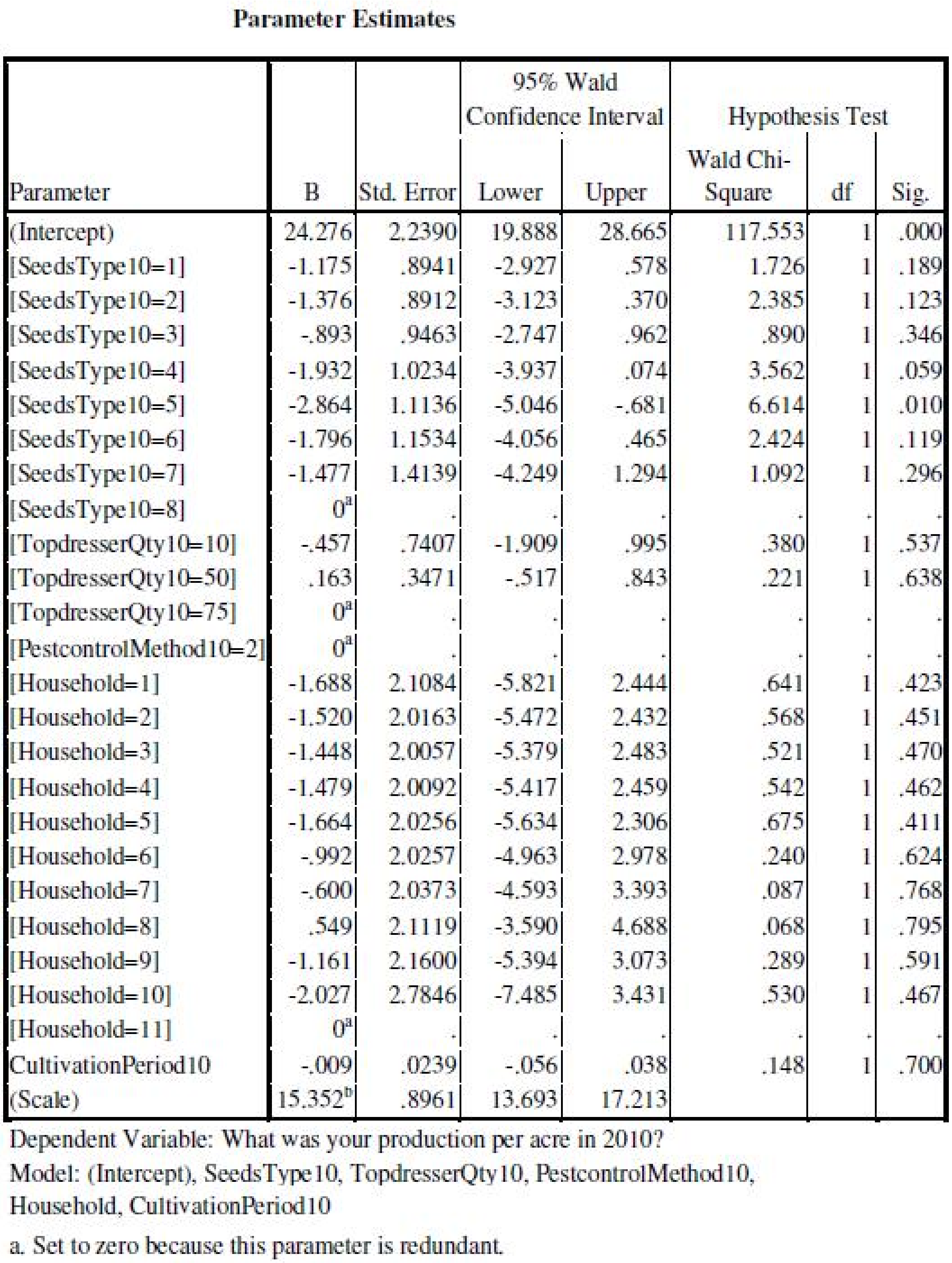
Production Per Acre in 2008

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	11	2	.3	.3	.3
	12	18	3.1	3.1	3.4
	13	4	.7	.7	4.1
	14	4	.7	.7	4.8
	16	23	3.9	3.9	8.7
	17	108	18.4	18.4	27.1
	18	14	2.4	2.4	29.5
	19	8	1.4	1.4	30.8
	20	4	.7	.7	31.5
	21	67	11.4	11.4	42.9
	22	45	7.7	7.7	50.6
	23	199	33.9	33.9	84.5
	24	25	4.3	4.3	88.8
	25	9	1.5	1.5	90.3
	26	6	1.0	1.0	91.3
	27	6	1.0	1.0	92.3
	28	24	4.1	4.1	96.4
	30	4	.7	.7	97.1
	31	1	.2	.2	97.3
	32	12	2.0	2.0	99.3
	33	1	.2	.2	99.5
	34	3	.5	.5	100.0
	Total	587	100.0	100.0	



Production Per Acre in 2007

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	6	1	.2	.2	.2
	11	3	.5	.5	.7
	12	10	1.7	1.7	2.4
	13	3	.5	.5	2.9
	14	1	.2	.2	3.1
	15	4	.7	.7	3.7
	16	23	3.9	3.9	7.7
	17	135	23.0	23.0	30.7
	18	11	1.9	1.9	32.5
	19	2	.3	.3	32.9
	20	1	.2	.2	33.0
	21	59	10.1	10.1	43.1
	22	29	4.9	4.9	48.0
	23	204	34.8	34.8	82.8
	24	32	5.5	5.5	88.2
	25	29	4.9	4.9	93.2
	26	5	.9	.9	94.0
	27	3	.5	.5	94.5
	28	22	3.7	3.7	98.3
	30	8	1.4	1.4	99.7
	32	1	.2	.2	99.8
	34	1	.2	.2	100.0
	Total	587	100.0	100.0	



Parameter Estimates

			95% Wald Confidence Interval		Hypothesis Test		
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	24.276	2.2390	19.888	28.665	117.553	1	.000
[SeedsType10=1]	-1.175	.8941	-2.927	.578	1.726	1	.189
[SeedsType10=2]	-1.376	.8912	-3.123	.370	2.385	1	.123
[SeedsType10=3]	-.893	.9463	-2.747	.962	.890	1	.346
[SeedsType10=4]	-1.932	1.0234	-3.937	.074	3.562	1	.059
[SeedsType10=5]	-2.864	1.1136	-5.046	-.681	6.614	1	.010
[SeedsType10=6]	-1.796	1.1534	-4.056	.465	2.424	1	.119
[SeedsType10=7]	-1.477	1.4139	-4.249	1.294	1.092	1	.296
[SeedsType10=8]	0[a]	.	.	.	.	.	.
[TopdresserQty10=10]	-.457	.7407	-1.909	.995	.380	1	.537
[TopdresserQty10=50]	.163	.3471	-.517	.843	.221	1	.638
[TopdresserQty10=75]	0[a]	.	.	.	.	.	.
[PestcontrolMethod10=2]	0[a]	.	.	.	.	.	.
[Household=1]	-1.688	2.1084	-5.821	2.444	.641	1	.423
[Household=2]	-1.520	2.0163	-5.472	2.432	.568	1	.451
[Household=3]	-1.448	2.0057	-5.379	2.483	.521	1	.470
[Household=4]	-1.479	2.0092	-5.417	2.459	.542	1	.462
[Household=5]	-1.664	2.0256	-5.634	2.306	.675	1	.411
[Household=6]	-.992	2.0257	-4.963	2.978	.240	1	.624
[Household=7]	-.600	2.0373	-4.593	3.393	.087	1	.768
[Household=8]	.549	2.1119	-3.590	4.688	.068	1	.795
[Household=9]	-1.161	2.1600	-5.394	3.073	.289	1	.591
[Household=10]	-2.027	2.7846	-7.485	3.431	.530	1	.467
[Household=11]	0[a]	.	.	.	.	.	.
CultivationPeriod10	-.009	.0239	-.056	.038	.148	1	.700
(Scale)	15.352[b]	.8961	13.693	17.213			

Dependent Variable: What was your production per acre in 2010?

Model: (Intercept), SeedsType10, TopdresserQty10, PestcontrolMethod10, Household, CultivationPeriod10

a. Set to zero because this parameter is redundant.