Principal Component and Multiple Correspondence Analyses in Dimensionality Reduction: A Study on Aflatoxin Contamination of Peanuts in Kenya

Henry Mokua Momanyi

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science in Research Methods in the Jomo Kenyatta University of Agriculture and Technology

2013

DECLARATION

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

Signature ……. Date……..

Henry Mokua Momanyi

This dissertation has been submitted with our approval as university supervisors:

Signature ……. Date……..

Prof. Christine Onyango

Department of Food Science and Technology, Jomo Kenyatta University of Agriculture and Technology

Signature ……. Date……

MS. Caroline Mugo

Department of Statistics, Jomo Kenyatta University of Agriculture and Technology

Signature ……. Date……

Dr. Maina Wagacha

University of Nairobi / International Crops Research Institute for Semi-Arid Tropics

Signature ……. Date……

Dr. Charity Mutegi

International Institute of Tropical Agriculture / Kenya Agricultural Research Institute

DEDICATION

This dissertation is dedicated to my cousins Caroline Ondieki and Mariba Kegancha, my mum Pasikaria Burasi, my late dad Simeon Momanyi and my siblings: thanks for your encouragement, endurance and giving me a reason to carry on even during the hard times and most of all thanks to God the Divine who continues to make the impossible possible.

ACKNOWLEDGEMENT

I owe my greatest gratitude to God for all achievements in life even in the academic endeavor: His mercy and grace are true. I am indebted to my supervisors Prof. Christine Onyango for the academic support and for being an exempler of academic excellence and career focus; Ms. Caroline Mugo for the academic support and being a mentor in the field of statistics; Dr. Maina Wagacha for being an academic inspiration, moral support and guidance in the development of the thesis and Dr. Charity Mutegi for the academic support and willingness to incorporate me into the Peanut Collaborative Research Support Program. I am also very grateful for the scholarship award from RUFORUM for the MSc. Research Methods programme and JKUAT for providing logistical, technical, physical facilities and a serene learning environment for both students and lecturers. Moreover, thanks to the Peanut Collaborative Research Support Program (Peanut CRSP) funded by USAID under cooperative agreement USAID ECG-A-00-07-00001-00 for research support on aflatoxin contamination of peanuts in Kenya.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS AND DEFINITIONS

CDC Center for Disease Control and Prevention

Communality Refers to the percent of variance in an observed variable that is accounted for by the retained components (or factors). A given variable will display a large communality if it loads heavily on at least one of the study's retained components.

EC European Commission

EU European Union

IARC International Agency for Research on Cancer

ICRISAT International Crops Research Institute for the Semi Arid Tropics

KEBS Kenya Bureau of Standards

MCA Multiple correspondence analysis

NIEHS National Institute of Environmental Health Sciences

PCA/P.C.S Principal component analysis

Peanut CRSP-Peanut Collaborative Research Support Program

Rotation Is a linear transformation that is performed on the factor solution for making the solution easier to interpret.

WHO World Health Organization

ABSTRACT

In a study about the factors that contributed to the risk of aflatoxin contamination of peanuts in the Peanut CRSP project in Kenya, contingency table analysis (Pearson's chi-square) was used to analyze a large mixed data set from a survey. The data was collected between March and July 2009 from three provinces in Kenya namely Nairobi, Western and Nyanza. Data analysis with contingency tables has limitations since it cannot allow for testing of statistical significance, variables with many categories produce large tables that were difficult to read and the Chi-square test cannot provide predicted values and can only be used to analyze the effect of a single categorical variable on the response. This study was intended to identify more sensitive statistical methods that could overcome the above limitations by analyzing the data using multiple regression analysis, analysis of variance (ANOVA), Principal component analysis (PCA) and Multiple correspondence analysis (MCA). With such methods, 12 factors were identified as having played a significant role in enhancing aflatoxin contamination of peanuts. Principal component analysis was useful in reduction of the large data set of 37 variables into a lower dimension of six variables and in constructing data composites for MCA. Multiple correspondence analysis was applicable in the interpretation of aflatoxin contamination of peanuts by establishing associations for more than two categorical variables in a low-Euclidean dimensional space and was an excellent heuristic for getting into complex multi-factorial data than contingency tables. There is need for further studies on some of the variables that were identified as having played a significant role in aflatoxin contamination of the peanuts, especially those to do with peanut storage and housing conditions in order to qualify the findings.

CHAPTER ONE

INTRODUCTION

1.1 Background information

Peanuts are highly prone to aflatoxin contamination due to their susceptibility to *Aspergillus* molds that produce aflatoxin under favourable conditions. The threat of aflatoxin contamination to human and livestock health has caused a serious challenge in the international peanut markets and has seriously hampered the export business in developing countries (Nautiyal, 2003; Wagacha and Muthomi, 2008). Developed countries, which import peanuts, have set stringent total aflatoxin contamination limits for foodstuffs ranging from 4 to 10 μg kg^{-1} (compared to WHO's 15 μg kg^{-1} for total aflatoxin). The major aflatoxin-producing fungi, *Aspergillus flavus* and *A. parasiticus* whose native habitat is the soil, can invade peanut seed in the field before harvest or during postharvest, including drying and curing, in storage and transportation (Horn, 2003; Abbas et al., 2009).

In the statistical analysis of aflatoxin contamination of peanuts, contingency table analysis (Pearson's Chi-squared test) and t-test have been used (Mutegi et al., 2012). Other statistical techniques commonly used include Fisher's Exact probability test, G-Statistics and Z-test. However, exploitation of these tests depends on some conditions. Even though these conditions are met, there are still problems in interpretation of the results because obtained data are general and limited (Akrurk et al., 2007). The t-test and contingency tables (χ^2 test) analyze the effects of a single variable at a time and are part of univariate and bivariate methods of data analysis.

In real situations, several factors act simultaneously towards aflatoxin contamination in peanuts.

Multivariate analysis methods have the advantage of bringing in more information to bear on specific outcome and they take into account the continuing relationship among several variables (Anon, n.d; Shiker, 2012). Additionally, they allow easier visualization and interpretation of the data and more data can be analyzed simultaneously thereby providing greater statistical power. Regression models give more insight into relationships between variables and the focus is normally on relationships rather than on isolated factors. Multivariate statistics have been utilized in the statistical analysis of aflatoxin contamination and other mycotoxins in maize and other agricultural products. They are widely used to solve practical problems in an effective way in geology, meteorology, hydrology, medicine, industry, agriculture and economics (Alonso et al., 2011; Khatoon et al., 2012; Shiker, 2012). However, they are particularly important in social science research because social researchers are generally unable to use randomized laboratory experiments like those used in medicine and natural sciences (Shiker, 2012).

1.2 Problem statement

In many occasions but particularly the year 2004, several hundred Kenyans became severely ill and 125 died of acute aflatoxicosis: a disease of liver failure associated with consuming extremely high levels of aflatoxin in food. Aflatoxin-contaminated home-grown maize was the source of the outbreak (Lewis et al., 2005). While this severe outbreak was devastating, far more individuals suffer from diseases associated with lower chronic levels of aflatoxin consumption in maize and peanuts. The primary disease associated with

aflatoxin intake is hepatocellular carcinoma (liver cancer). This disease is the third-leading cause of cancer death globally according to WHO with about 600000 new cases each year and eighty-three percent of these deaths occur in Sub-Saharan Africa and East Asia. Some of the factors which had been implicated in aflatoxin contamination of maize and peanuts in Kenya included agro ecological zones, poor harvesting and storage of produce and susceptible cultivars (Barrett, 2005; Mutegi et al., 2009).

In a study investigating factors that contributed to the risk of aflatoxin contamination of peanuts in Kenya under the Peanut CRSP project, contingency table analysis (Pearson's chi-square) was used to analyze a large mixed data set from a survey (Mutegi et al., 2012). Contingency tables are easy to set up, easy to understand, are useful because little or no understanding of statistical concepts is necessary for interpretation and readers can easily observe patterns of association and can see if the pattern is weaker across some rows. However, they have several disadvantages including: not precisely measuring the nature of association between two variables and variables with many categories requires large tables that are difficult to read. Again, categories with few observations obfuscate the bivariate association and the Chi-square test cannot provide predicted values. Above all contingency tables can only be used to analyze the effect of a single categorical variable on the response**.** Therefore the current study was intended to overcome the above limitations by identifying suitable and better statistical methods that can be applied when mixed large data sets are encountered in the study of aflatoxin contamination of peanuts.

1.3 General objective

The main objective for this research was to identify suitable statistical method(s) for analyzing large and complex mixed data sets from surveys for aflatoxin contamination of

peanuts. The suitable statistical methods were those that were more sensitive in producing accurate results than categorical data analysis by means of contingency tables.

1.4 Specific Objectives

The specific objectives of this study were to:

a) Determine variables that play a significant role in aflatoxin contamination of peanuts by use of multiple linear regression and analysis of variance.
b) Evaluate the applicability of Principal component and Multiple correspondence analysis in interpretation of aflatoxin contamination of peanuts.
c) Evaluate the applicability of Multiple correspondence analysis in interpretation of identified significant variables compared to contingency table analysis (Pearson's Chi-square).

1.5 Justification and significance of the study

The study will increase and promote the adoption of less familiar statistical analysis options available to researchers in the field of aflatoxin contamination in peanuts and help to increase the validity of presentation of research findings. The identified statistical methods will contribute significantly to enhanced prediction of the risk of aflatoxin contamination in peanuts in relation to the sources of contamination. This will aggrandize future research and contribute to the development of suitable policies for handling of peanuts. Ultimately, this will lead to minimized health risk to humans and animals, enhanced peanut trade as well as contribute to better methods of awareness creation.

1.6 Hypothesis

H_0: $\beta_1 = \beta_2 = ... \beta_k = 0$

H_A: At least one β is not zero

CHAPTER TWO

LITERATURE REVIEW

2.1.Peanuts production

Peanut is the sixth most important oil seed crop in the world. The botanical name for peanut, *Arachis hypogaea Linnaeus*, is derived from two Greek words, *Arachis* meaning a legume and *hypogaea* meaning below ground, referring to the formation of pods in the soil. Peanut is an upright or prostrate annual plant. It is generally distributed in the tropical, sub-tropical and warm temperate zones (Nautiyal, 2003). It contains 48-50% oil, 26-28% protein and is a rich source of dietary fibre, minerals and vitamins (Rachier et al., 2010). China and India are the world's leading peanut producers accounting for nearly 60% of the production and 52% of the crop area. India cultivates about 7.74 million hectares and produces 7.61 million tonnes of peanut with the productivity level of 991.8 kg ha^{-1}. Nigeria is the major peanut producer in Africa, while in Latin America almost one half of the total peanut produced in that region may be credited to Argentina (Nautiyal, 2003).

Peanut production in Kenya is common in Western and Nyanza provinces. It is however produced in smaller amounts in other parts of the country such as Eastern, Rift valley and pockets of Coast province. Common varieties grown include ICGV 99568, ICGV 90704, Homa bay local, Valencia Red, ICGV 12988, ICGV 12991, JL24 and CG7, the latter four being improved varieties introduced by ICRISAT (Mutegi et al., 2013; Okoko et al., 2009). Other groups of peanuts such as Virginia and Spanish types are being evaluated at various KARI centers (Rachier et al., 2010).

According to Rachier et al. (2010), the crop is used for subsistence, cash-income and provides raw materials for agro-based industries. As food, peanut is used for human

consumption in the form of raw, boiled or roasted nuts. It is also pounded and used as vegetable oil for cooking or made into paste and eaten with sweet potatoes, cassava and bananas. As a cash crop, peanut is sold in the local market as boiled unshelled, raw unshelled, raw shelled and shelled roasted nuts while some is sold in the confectionery trade as peanut butter (Fig. 2.1), peanut sugar, peanut candy and peanut brittles among other products (Mutegi et al., 2013).

Figure 2.1: peanut butter

2.2 History of Aflatoxin

Aflatoxins were discovered about 40 years ago after an outbreak of Turkey X disease in England (ICRISAT, 2000; Yu, 2012). The disease was caused by toxins in Brazil nut meal infected with *Aspergillus flavus* and the toxins were named as 'aflatoxins'. Aflatoxin is mainly produced by *Aspergillus flavus* and *A. parasiticus* that can grow on different

substrates but particularly on poorly managed agricultural crops like maize and peanuts (Fig.2.2). It belongs to a group of toxic substances called mycotoxins (Sweets and Wrather, 2009). At least 14 mycotoxins are known carcinogens, with the aflatoxins having assumed economic importance because of their influence on the health of humans, livestock and on the marketability of agricultural products (ICRISAT, 2000; Wild and Turner, 2002; Klich, 2007; Wagacha and Muthomi, 2008; Wild and Gong, 2010).

Figure 2.2: peanuts contaminated with mould

2.3 Aflatoxin types

According to Yu et al.(2012), among the 16 structurally related aflatoxins that have been characterized, there are only four major aflatoxins, B_1, B_2, G_1, and G_2 (also named as AFB_1, AFB_2, AFG_1 and AFG_2 respectively), that contaminate agricultural commodities and pose a potential risk to human and livestock health. *Aspergillus flavus* produces AFB_1 and AFB_2.

Aspergillus parasiticus produces AFB_1, AFB_2, AFG_1 and AFG_2. Of the four aflatoxins, aflatoxin B_1 is the most potent hepatocarcinogenic compound (IARC, 2002; Yu, 2012).

Other significant members of the aflatoxin family, M_1 and M_2, are oxidative forms of aflatoxin B_1 modified in the digestive tract of some animals and isolated from milk, urine and feces. Aflatoxin B2A, G2A which may be produced in minor amounts have been isolated from cultures of *A. flavus* and *A. parasiticus* (Varga et al., 2009). A number of closely related compounds namely aflatoxin GM1, parasiticol and aflatoxicol are also produced by *A. flavus* (ICRISAT, 2000; IARC, 2002).

Some other species that produce aflatoxin are *Aspergillus nomius, Aspergillus pseudotamarii, Aspergillus bombycis, Aspergillus ochraceoroseus, Emericella venezuelensis, Aspergillus parvisclerotigenus, Aspergillus rambellii* and *Emericella astellata* (Klich, 2007; Yu, 2012).

2.4 Factors favouring aflatoxin contamination

Pre-harvest factors which contribute to aflatoxin contamination in peanuts include the presence of *A. flavus* in soil and air, use of susceptible cultivars, end-of-season moisture stress to the crop for more than 20 days, mean soil temperatures of 28-31^{o}C in the pod zone, growth cracks and mechanical injury to the pod (Fig.2.3), insect damage to pods by termites or pod borers, disease attack (stem, root and pod rots) at pod maturity stage and nematode damage to the pod (ICRISAT, 2000; Williams et al., 2004; Liang, 2006; Wang et al., 2010; Wu and Khlangwiset, 2010).

Figure 2.3: Some mechanically damaged peanut pods

The postharvest factors which contribute to aflatoxin contamination of peanuts include harvesting an over mature crop, mechanical damage to the pod at the time of harvest, stacking the harvest when pod moisture is more than 10% or under high humidity conditions, damage to the pod by insects during storage, storing haulms with immature or small pods which they tend to contain more aflatoxins, gleaning pods from the soil after harvest and rewetting stored pods due to factors like ground-moisture or roof leakage (ICRISAT, 2000; Cornell University, 2008; Nigam et al., 2009).

2.5 Natural occurrence of aflatoxin

Two fungi, Aspergillus flavus and Aspergillus parasiticus mainly produce aflatoxin. It is most frequently reported in the field in oilseed crops including maize, cotton, peanuts, tree

nuts and rarely in other crops. The reason for this may be partly biogeographical: these crops are grown in the latitudes where *A. flavus* is most frequently reported. Another possible reason may be the carbon utilization pattern of *A. flavus*. In cottonseed and maize, *A. flavus* first utilizes free saccharides and then oil before using starch (Klich, 2007).

The *Aspergillus flavus* and *A. parasiticus* can invade peanut seed in the field before harvest, during postharvest, drying, curing and in storage and transportation (Wagacha and Muthomi, 2008; CDC, 2012). Pre-harvest infection is significant in the semi-arid tropics, especially when end-of-season drought occurs (Rustom, 1997; Klich, 2007). In peanuts, experiments with drought stress and controlled soil temperatures (85–100 days after planting) demonstrated that drought stress and temperatures of 29°C yielded the greatest number of colonized edible grade peanuts and high aflatoxin levels (Klich, 2007). High temperatures and drought stress affect the physiology of plants, and therefore stressed plants may be more susceptible to infection or aflatoxin production. For instance, drought stress induces a great increase in proline production in plants and proline has been reported to enhance aflatoxin production (Reddy et al., 2003). Formation of some phytoalexins which are antimicrobial compounds produced by some plants is inhibited by drought stress.

Another possibility according to Klich (2007) is that the fungi that normally compete with *A. flavus* in the soil do not grow as readily under these conditions, giving *A. flavus* a competitive advantage. Even among other *Aspergillus* species, the temperature range for growth of *A. flavus* (25–42°C) is higher than for many other species and *A. flavus* is fairly xerotolerant.

Poor post-harvest conditions in warm humid areas, bad harvesting and storage practices lead to rapid development of the fungi and higher levels of toxins (Wagacha and Muthomi, 2008). This is especially true in developing countries where preventive measures are frequently ignored. Other food products contaminated with aflatoxins include cereals (maize, sorghum, pearl millet, rice, and wheat), oilseeds (soybean, sunflower, and cotton), spices (chili, black pepper, coriander, turmeric and zinger), tree nuts (almonds, pistachio, walnuts and coconut) and milk (ICRISAT, 2000; IARC, 2002).

Diet is the major way through which humans and animals are exposed to aflatoxin. Apart from this, exposure to aflatoxin can be through ingestion of contaminated milk containing aflatoxin M1 (metabolite of AFB1). Other reported avenues of exposure include aflatoxin inhalation and absorption through skin (Wagacha and Muthomi, 2008). Occupational exposure to aflatoxins in agricultural workers, people working in oil mills and granaries has been reported (ICRISAT, 2000; IARC, 2002; Wild and Gong, 2010; CDC, 2012).

2.6 Effects of aflatoxin contamination on peanut trade

According to FAO estimates, 25% of the world food crops are affected by mycotoxins each year. Crop loss due to aflatoxin contamination costs US producers more than $100 million per year on average including $ 26 million to peanuts (Klich, 2007). Production of aflatoxin due to the invasion of aflatoxin-producing fungi to peanut pod/kernel is a serious problem in the trade of peanuts in the international market. This has seriously hampered the export business of developing countries especially where the crop is grown under rain fed conditions (Mejia and Lewis, 2002; Wagacha and Muthomi, 2008).The aflatoxin contamination does not affect crop productivity but it makes produce unfit for consumption

as toxins are injurious to health. The marketability of contaminated produce, particularly in international trade is diminished to nil due to stringent standards of permissible limits on aflatoxin contamination set by the importing countries (Klich, 2007; Coulibaly, 2008). Developed countries that import peanuts have set total aflatoxin contamination limits for foodstuff ranging from 4 to 10µg kg^{-1} and this has resulted in import restrictions on aflatoxin-contaminated produce. As a result, many developing countries have been unable to export their peanuts and peanut products (Otsuki et al., 2001).

In developing countries, food safety is a major problem where detection and decontamination policies are impractical. Due to food shortage in those countries, routine consumption of aflatoxin-contaminated food is widespread (Guo et al., 2009).Outbreaks of acute aflatoxicosis from contaminated food in humans has been documented in Kenya, India, Malaysia and Thailand. The largest and most severe outbreak of acute aflatoxicosis occurred in Kenya during 2004 and involved 317 cases and 125 deaths, mainly among children due to consumption of aflatoxin-contaminated corn (CDC, 2004; Lewis et al., 2005; Guo et al., 2009).

2.7 Effects of aflatoxin contamination on health

Aflatoxin contamination in grain poses a great threat to human and livestock health (Wagacha and Muthomi, 2008). Epidemiological, clinical, and experimental studies reveal that exposure to large doses (above 6000mg) of aflatoxin may cause acute toxicity with lethal effects whereas exposure to small doses for prolonged periods is carcinogenic (Groopmann et al., 1988; Wild and Turner, 2002; Klich, 2007; Wild and Gong, 2010; Liu and Wu, 2010; CDC, 2012). After wide experimentation on many animal species like rats

and rainbow trouts, aflatoxin especially aflatoxin B1 is confirmed as a potent carcinogen (IARC, 1993; NIEHS, 2007; Klich, 2007; Cornell University, 2008).

During the 16th Aspen cancer conference, aflatoxicosis was reported in several countries such as India, China, Thailand and a number of African countries (ICRISAT, 2000). Studies on aflatoxin exposure and incidence of liver cancer by Groopmann and Wild (1996) in places like China and West Africa showed that the situation was alarming. Aflatoxin acts synergistically in the presence of confounding factors that include malnutrition, malaria, HIV/AIDS, kwashiorkor, alcoholic liver disease and Hepatitis B and C in the etiology of cancer (Rensburg et al., 1985; Debrah and Waliyar, 1996; NIEHS, 2007; Wagacha and Muthomi, 2008; Wild and Gong, 2010).

Foetal and childhood environment, including the nutritional status of the pregnant mother and the infant are considered critical for growth and risk of disease in earlier life. Apart from this, children are also exposed to high levels of mycotoxins of which aflatoxins are a key concern. Aflatoxins are immunogenic, teratogenic and retard growth among humans and experimental animals (Hendrickse, 1984; Klich, 2007; Hell et al., 2008). A study in West Africa showed a significant correlation between aflatoxin exposure and stunted growth in children who are exposed to aflatoxin right from neonatal stages while other studies have shown that aflatoxins have the capacity to cross the placental barrier and can cause genetic defects at foetal stages (Maxwell et al., 1998; Gong et al., 2002).

2.8 Control of aflatoxin contamination in peanuts

Since *Aspergillus flavus* and *A. parasiticus* can invade peanut seed in the field before harvest, during postharvest, drying, in storage and transportation, then it is possible that

aflatoxin contamination can be minimized by adopting certain improved cultural, produce handling and storage practices (Wagacha and Muthomi, 2008). However, these practices are not widely adopted particularly by the small farmers in developing countries, who contribute about 60% to the world peanut production (ICRISAT, 2000).

Pre-harvest strategies for controlling aflatoxin contamination include use of aflatoxin-resistant peanut varieties, selecting sound seed and treating them with Manganese ethylene bisdithiocarbante (Diathane M45) at 3g/kg before planting. Applying farm yard manure at 5-10 tons/ha, *Trichoderma harzianum* at 1kg/ha and anhydrous calcium sulphate (gypsum) at 400-500 kg/ha during flowering stage are effective measures for controlling aflatoxin contamination in peanuts in the field. Other strategies include maintaining optimal plant population in the field at $33m^2$ (Figure 2.4), avoiding end-of-season drought with irrigation if possible, controlling foliar diseases using chlorothalonil (Kavach) with 1-2 sprays, removing dead plants from the field and harvesting the crop at right maturity (ICRISAT, 2000; Liang, 2006; Klich, 2007;Wagacha and Muthomi, 2008; Wang et al., 2010).

Figure 2.4: maintaining optimal peanut population in the field

Postharvest strategies for controlling aflatoxin contamination include avoiding mechanical damage to the pods by inserting the plough below the pod zone at harvest and drying the harvested produce for 3-5 days using the inverted windrow method until the pod moisture is below 8% (Fig.2.5).

Figure 2.5: Inverted windrow drying method for peanuts in the field

Other postharvest strategies include threshing the pods immediately after drying, avoiding stacking when using mechanical threshers, using appropriate sieves based on pod size so that immature pods are blown off, removing mechanical and insect damaged pods and separating the fully mature large pods for raw consumption from the remaining produce that are used for oil extraction. Good practices for controlling aflatoxin contamination also entails not mixing the gleaned pods with the main produce, stacking the pod-filled gunny bags on a wooden plank and storing them in well aerated waterproof storage, preventing insect damage to the pods in storage and removing all immature pods attached to the

haulms (ICRISAT, 2000; Williams et al, 2004; Klich, 2007; Wagacha and Muthomi, 2008; Wu and Khlangwiset, 2010).

2.9 A review of the selected statistical methods for the data analysis

Contingency table analysis is more often used with non-metric data which is nominal or ordinal. The advantages of contingency tables is that they are easy to set up and easy to understand. They are useful because little or no understanding of statistical concepts is necessary for interpretation and little technical know-how is necessary to build tables. Readers can easily observe patterns of association and can see if the pattern is weak across some rows (Namuth-Covert, Merk and Haines, 2012). However, they have the disadvantages of not allowing for testing of statistical significance or precisely measuring the association between two variables. Again variables with many categories require large tables that are difficult to read and categories with few observations can obfuscate the bivariate association (Clark, 1976; Dallal, 2000).

Regardless of the level of scaling, contingency tables are conventionally analyzed with chi-squared test. However, for this test to be useful the cell counts must be greater than or equal to some number (usually 5), otherwise this leads to the collapsing of the table and results in lost information (Namuth-Covert, Merk and Haines 2012). To counteract this effect, Fisher's Exact Test is used (Lowry, 1999; Routledge, 2005). Chi-square test again cannot provide predicted values and can only be used to analyze the effect of a single categorical variable on the response**.** These statistical limitations of contingency tables analysis can be overcome by fitting the data with regression model to provide predicted values and the application of multiple correspondence analysis (MCA) when the categorical data set is large (Greenacre, 2006; Greenacre and Blasius, 2006).

Principal component analysis (PCA) is a powerful statistical tool for analyzing data of high dimension by reducing the number of dimensions without much loss of information (Smith**,** 2002). By reducing the dimensionality of original data, PCA can often simplify many analyses. The disadvantage of this statistical technique is that interpretation can be more difficult since it is no longer possible to work with the original variables and the principal components are heavily affected by the scaling of variables (Anon, 1996). To overcome the above limitation of PCA, MCA can be utilized in the data analysis.

The MCA aims to identify a reduced set of synthetic dimensions maximizing the explained variability of the categorical data sets in question. The advantage in using MCA to study associations of categorical data are then to obtain a simplified representation of multiple associations characterizing attributes as to remove noise and redundancies in data. The exploratory and visualization based approach characterizing MCA provides immediate interpretation of the results.

However, the applicability of MCA on very large categorical data streams is limited due to the required Singular Value Decomposition (SVD). The applicability of SVD to large and high dimensional data is unfeasible since it requires a computational time that is quadratic in the data size. Furthermore, the SVD input matrix must be complete and stored in memory. This problem can be overcome by stratifying the data into different subgroups according to an external criterion related to time or another identified characteristic (Glynn, 2012; D'Enza, 2012). The other disadvantage of MCA is that it reconstructs a small part of the data, is sensitive to outliers and may yield solutions that display objects and categories in 2 dimensions in a horse-shoe shaped form known as Guttman effect (Groenen et al., 1998). To a great extent, these limitations can be overcome by fitting the data to a

regression model or data analyzed using Detrended correspondence analysis (Greenacre, 1984).

Principal Component analysis assumes a normal distribution and hence continuous variables. Multiple correspondence analysis on the other hand makes few assumptions on the nature of the distribution of individual variables and is more appropriate in the context of discrete and categorical variables (Shimeles and Thoenen., 2005).

2.9.1 Contingency table (cross tabulation)

The contingency table was first used by Karl Pearson in 1904. This is a type of table in a matrix format that displays the (multivariate) frequency distribution of variables. A cross tabulation is a joint frequency distribution of cases based on two or more categorical variables (Michael, 2001). The joint frequency distribution can be analyzed with the chi-square to determine whether the variables are statistically independent or if they are associated. If a dependency between variables does exist, then other indicators of association, such as Cramer's V and gamma; Sommer's d, and so forth, can be used to describe the degree which the values of one variable predict or vary with those of the other variable. More advanced techniques such as log-linear models and multinomial regression can be used to clarify the relationships contained in contingency tables (Michael, 2001; Howell, n.d).

2.9.2 Pearson's Chi-square test

The test serves both as a "goodness of- t" test, where the data are categorized along one dimension and as a test for the contingency table, in which categorization is across two or more dimensions (Howell, n.d).

It is calculated as:

$$X^2=\sum_{i=1}^{n}(O_i-E_i)^2/E_i$$

where X^2=Pearson's cumulative test statistic which asymptotically approaches a χ^2 distribution

O_i=an observed frequency

E_i=an expected frequency asserted by the null hypothesis

n=the number of cells in the table.

The Chi-squared statistic can then be used to calculate a P-value by comparing the value of the statistic to a Chi-squared distribution. The number of degrees of freedom is equal to the number of cells (n), minus the reduction in degrees of freedom **(**Shepard, 2008).

The chi-square test has four assumptions whereby the sample data is taken as a random sampling from a fixed distribution or population where each member of the population has an equal probability of selection. A sample with a sufficiently large size is also assumed otherwise Type II error will be committed with small samples (Michael, 2001). An adequate expected cell count of 5 or more in all cells of a 2-by-2 table, and 5 or more in 80% of cells in larger tables with no cells with zero expected count is assumed. When this assumption is not met, Yates's Correction is applied (Key, 1997). Lastly the observations are always assumed to be independent of each other and if not, McNemar's test is applied (Smith, 1996).

2.9.3 Pearson product-moment correlation coefficient

The strength of the linear association between two variables is quantified by the *correlation coefficient(r)*, which is also known as *Pearson product moment correlation coefficient*. Given a set of observations (x_1, y_1), (x_2, y_2), ...(x_n, y_n), the formula for computing the correlation coefficient is given by:

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

The correlation coefficient takes a value between -1 and 1, with 1 or -1 indicating perfect correlation. A positive correlation indicates a positive association between the variables in which increasing values in one variable corresponds to increasing values in the other variable. A negative correlation indicates a negative association between the variables. A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak. These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data. A correlation value close to 0 indicates no association between the variables.

Since the formula for calculating the correlation coefficient standardizes the variables, changes in scale or units of measurement will not affect its value. For this reason, the correlation coefficient is often more useful than a graphical depiction in determining the strength of the association between two variables (Meng, Rosenthal and Rubin, 1992).

2.9.4 Multiple regression analysis

Multiple regression is a flexible method of data analysis that may be appropriate whenever a dependent quantitative variable is to be examined in relationship to any other independent or predictor variables. Relationships may be nonlinear, independent variables may be quantitative or qualitative, and one can examine the effects of a single variable or multiple variables with or without the effects of other variables taken into account (Cohen et al., 2003).

In general, the multiple regression equation of Y on X_1, X_2, …, X_k is given by:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

Here b_0 is the intercept and b_1, b_2, b_3, … b_k are analogous to the slope in linear regression equation and are also called regression coefficients. They can be interpreted the same way as slope. Multiple regression analysis is used when one is interested in predicting a continuous dependent variable from a number of independent variables. If dependent variable is dichotomous, then logistic regression should be used (Choudhury, 2009).

According to Jason and Waters (2002) multiple regression has got the following assumptions:

- **Variables are normally distributed**

Regression assumes that dependent variables have normal distributions. Non-normally distributed variables (highly skewed or kurtotic variables, or variables with substantial outliers) can distort relationships and significance tests.

- **Relationship between the dependent and independent variable(s) is linear**

Standard multiple regression can only accurately estimate the relationship between dependent and independent variables if the relationships are linear in nature. If the relationship between independent variables and the dependent variable is not linear, the results of the regression analysis will under-estimate the true relationship. This under-estimation carries two risks: increased chance of a Type II error for that independent variables, and in the case of multiple regression, an increased risk of Type I errors (over-estimation) for other independent variables that share variance with that independent variables.

- **Variables are measured without error (reliably)**

In simple correlation and regression, unreliable measurement causes relationships to be under-estimated increasing the risk of Type II errors. In the case of multiple regression or partial correlation, effect sizes of other variables can be over-estimated if the covariate is not reliably measured, as the full effect of the covariate(s) would not be removed.

With each independent variable added to the regression equation, the effects of less than perfect reliability on the strength of the relationship becomes more complex and the results of the analysis more questionable. With the addition of one independent variable with less than perfect reliability each succeeding variable entered has the opportunity to claim part of the error variance left over by the unreliable variable(s). The apportionment of the explained variance among the independent variables will thus be incorrect. The more independent variables added to the equation with low levels of reliability the greater the likelihood that the variance accounted for is not apportioned correctly. This can lead to erroneous findings and increased potential for Type II errors for the variables with poor reliability, and Type I errors for the other variables in the equation.

- **Constant homoscedasticity**

Homoscedasticity means that the variance of errors is the same across all levels of the independent variables. When the variance of errors differs at different values of the independent variables, heteroscedasticity is indicated. Slight heteroscedasticity has little effect on significance tests; however, when heteroscedasticity is marked it can lead to serious distortion of findings and seriously weaken the analysis thus increasing the possibility of a Type I error.

2.9.5 Fisher's Exact Test

This is a test of independence in a 2×2 contingency table. It is more useful when the total sample size and the expected values are small (Routledge, 2005).The test holds the marginal total fixed and computes the hypergeometric probability that n_{11} is at least as large as the observed value. It is useful when the cell counts are less than 5 (Fu and Arnold, 1992; Bower, 2003; Shepard, 2008).

2.9.6 Multivariate statistics

Multivariate statistics are a form of statistics encompassing the simultaneous observation and statistical analysis of more than one response variable. The application of multivariate statistics is multivariate analysis and essentially models reality where each situation, product or decision involves more than a single variable (Kessler, 2007). It concerns understanding the different aims and background of each of the different forms of multivariate analysis and how they relate to each other.

The practical implementation of multivariate statistics to a particular problem may involve several types of univariate and multivariate analyses in order to understand the relationships

between variables and their relevance to the actual problem being studied. In addition, multivariate statistics are concerned with multivariate probability distributions, in terms of both how these can be used to represent the distributions of observed data and how they can be used as part of statistical inference, particularly where several different quantities are of interest to the same analysis.

According to Abeyasekera (2003) and Obuchowski (2005), multivariate methods in a strict statistical sense concern the collective study of a group of outcome variables, thus taking account of the correlation structure of variables within the group. Many researchers however, also use the term "multivariate" in the application of multiple regression techniques because this involves several explanatory (predictor) variables along with the main outcome variable (Abeyasekera, 2003). The benefit of exploring several variables together is that it allows for inter-correlations to be assessed.

With multivariate analysis the following can be achieved:

i. A summary or an overview of a table can be obtained. This analysis is often called principal component analysis or Factor Analysis. In the overview, it is possible to identify the dominant patterns in the data, such as groups, outliers, trends, and so on. The patterns are displayed as two plots (Kessler, 2007).

ii. Analysis of groups in the table, how these groups differ and to which group individual table rows belong. This type of analysis is called Classification and Discriminant Analysis (Fernandez, 2002).

iii. Relationships between columns in data tables can be established. For instance, relationships between process operation conditions and product quality whereby the objective is to use one set of variables (columns) to predict another, for the purpose

of optimization and to find out which columns are important in the relationship. The corresponding analysis is called Multiple Regression Analysis or Partial Least Squares (PLS), depending on the size of the data (Cramer, 1993; Stolzenberg, 2004).

2.9.7 Principal component analysis

This statistical methodology originated with Karl Pearson (1901) as a means of fitting planes by orthogonal least squares, but was later proposed by Hotelling (1933) for the particular purpose of analyzing correlation structures. It is used abundantly in all forms of analysis from neuroscience to computer graphics because it is a simple, non-parametric method of extracting relevant information from confusing data sets (Manly, 1986; Morrison, 1990).

With minimal additional effort Principal component analysis (PCA) provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified structures that often underlie it (Shlens, 2005). It is useful when you have obtained data on a large number of variables and believe that there is some redundancy in those variables. In this case, redundancy means that some of the variables are correlated with one another, possibly because they are measuring the same construct. Because of this redundancy, it is believed that it should be possible to reduce the observed variables into a smaller number of principal components which are artificial variables that will account for most of the variance in the observed variables (Hatcher, 1994).

According to Abeyasekera (2003) the technique is strictly applicable to a set of measurements which are either quantitative or have an ordinal scale. However, being largely a descriptive technique, the inclusion of binary variables and/or a small number of nominal categorical variables is unlikely to be of practical consequence. Principal component analysis is a large-sample procedure. To obtain reliable results, the minimum number of subjects providing usable data for the analysis should be the larger of 100 subjects or five times the number of variables being analyzed (Hatcher, 1994; Mead et al., 2003).

Because principal component analysis is performed on a matrix of Pearson correlation coefficients, it is assumed that all analyzed variables should be measured on an interval or ratio level. The relationship between all observed variables should be linear and each observed variable should be normally distributed. As such, variables that demonstrate marked skewness or kurtosis may be transformed to better approximate normality. Each pair of observed variables should display a bivariate normal distribution. However, the Pearson correlation coefficient is robust against violations of this assumption when the sample size is greater than 25. Also since each subject is expected to contribute one score on each observed variable, these sets of scores should represent a random sample drawn from the population of interest (Hatcher, 1994).

Technically, a principal component can be defined as a linear combination of optimally-weighted observed variables. The words “linear combination” refer to the fact that scores on a component are created by adding together scores on the observed variables being analyzed. “Optimally weighted” refers to the observed variables being weighted in such a

way that the resulting components account for a maximal amount of variance in the data set (Hatcher, 1994).

In PCA a new set of variables is created as linear combinations of the original set. If x_1, $x_2 \ldots$, x_p are the original set of p variables, then a variable Y formed from a linear combination of these takes the form $Y=a_1x_1+a_2x_2+\ldots+a_px_p$ where the a_1s (i=1, 2…, p) are numbers or principal component coefficients and x_p's are the subjects' score on observed variable *p*.

The linear combination that explains the maximum amount of variation is called the first principal component. A second principal component which is another linear combination is then found, independent of the first, so that it explains as much as possible of the remaining variability. Further components are then created sequentially, each new component being independent of the previous ones. If the first few components, say the first 3, explain a substantial amount, say 90 per cent of the variability amongst the original set of 15 variables, then essentially, the number of variables to be analyzed has been reduced from 15 to 3.

According to Fieller (2010), if the first few principal components (P.C.S) explain most of the variation in the data, then the later P.C.S are redundant and little information is lost if they are discarded or ignored. The number of components extracted is equal to the number of observed variables being analyzed. However, in most analyses, only the first few components account for meaningful amounts of variance, so only these first few

components are retained, interpreted, and used in subsequent analyses such as in multiple regression analyses.

The first component extracted in a principal component analysis accounts for a maximal amount of total variance in the observed variables. Under typical conditions, this means that the first component will be correlated with at least some of the observed variables. The second component extracted will account for a maximal amount of variance in the data set that was not accounted for by the first component. Again it will be correlated with some of the observed variables that did not display strong correlations with component 1 but it will be uncorrelated with the first component. The remaining components are extracted such that each component accounts for a maximal amount of variance in the observed variables that was not accounted for by the preceding components, and is uncorrelated with all of the preceding components.

A principal component analysis proceeds in this fashion, with each new component accounting for progressively smaller and smaller amounts of variance. When the analysis is complete, the resulting components will display varying degrees of correlation with the observed variables, but are completely uncorrelated with one another.

When a variable is given a great deal of weight in constructing a principal component, it is said that the variable loads on that component. It is highly desirable to have at least three and preferably more variables loading on each retained component when the principal component analysis is complete. Because some of the items may be dropped during the course of the analysis, it is generally good practice to write at least five items for each

construct that is to be measured. In this way, chances are increased such that at least three items per component will survive the analysis (Hatcher, 1994).

Mathematically principal component analysis entails 4 procedural steps which can be stated as: Starting by coding the variables $x_1,x_2,\ldots,x_p$ to have zero means and unit variances; Calculating the covariance matrix *C* which is a correlation matrix if step 1 has been done; Finding the eigenvalues $\lambda_1, \lambda_2, \ldots \lambda_p$ and the corresponding eigenvectors $a_1, a_2,\ldots a_p$. The coefficients of the *i*th principal components are then given by a_i while λ_i is its variance; Discarding any components that only account for a small proportion of the variation in the data (Manly, 1986).

According to Fieller (2010) and Hatcher (1994), the above steps can be summarized as follows when applied to given complex data set.

Step 1: Initial extraction of the components

The number of components extracted is equal to the number of variables being analyzed. Although a large number of components may be extracted, only the first few components will be important enough to be retained for interpretation.

Step 2: Determining the number of "meaningful" components to retain

In general, it is expected that only the first few components will account for meaningful amounts of variance, and that the later components will tend to account for only trivial variance. The next step of the analysis, therefore, is to determine how many meaningful components should be retained for interpretation. There are four criteria that may be used in making this decision: the eigenvalue-one criterion, the scree test, the proportion of variance accounted for, and the interpretability criterion.

With the eigenvalue-one criterion approach, any component with an eigenvalue greater than 1 is retained and interpreted. Since each observed variable contributes one unit of variance to the total variance in the data set, any component that displays an eigenvalue greater than 1 is accounting for a greater amount of variance than had been contributed by one variable. Such a component is therefore accounting for a meaningful amount of variance and is worthy of being retained.

With the scree test, eigenvalues associated with each component are plotted and a "break" between the components with relatively large eigenvalues and those with small eigenvalues identified. The components that appear before the break are assumed to be meaningful and are retained for rotation; those appearing after the break are assumed to be unimportant and are not retained.

The proportion of variance accounted for criterion entails retaining a component if it accounts for a specified proportion or percentage of variance in the data set. This proportion can be calculated with a simple formula:

$$\text{Proportion} = \frac{\text{Eigenvalue for the component of interest}}{\text{Total eigenvalues of the correlation matrix}}$$

The total eigenvalues of the correlation matrix is equal to the total number of variables being analyzed because each variable contributes one unit of variance to the analysis.

The most important criterion for solving the number of components to be retained is the interpretability criterion. This involves interpreting the substantive meaning of the retained components and verifying that this interpretation makes sense in terms of what is known about the constructs under investigation.

Step 3: Rotation to a final solution

After extracting the initial components, there will be created an unrotated factor pattern matrix. The rows of this matrix represent the variables being analyzed, and the columns represent the retained components. The entries in the matrix are factor loadings. A factor loading is a general term for a coefficient that appears in a factor pattern matrix or a factor structure matrix. A rotation is a linear transformation that is performed on the factor solution for the purpose of making the solution easier to interpret when more than one component has been retained in an analysis. A varimax rotation is an orthogonal rotation which results in uncorrelated components and tends to maximize the variance of a column of the factor pattern matrix as opposed to a row of the matrix.

Step 4: Interpreting the rotated solution

This means determining what each of the retained components measures. This involves identifying the variables that demonstrate high loadings for a given component and determining what these variables have in common.

Step 5: Creating factor scores or factor-based scores

Once the analysis is complete, it is often desirable to assign scores to each subject to indicate where that subject stands on the retained components. These component scores could be used either as predictor variables or as criterion variables in subsequent analyses. A separate equation, with different weights, is developed for each retained component.

2.9.8 Multiple correspondence analyses

Multiple correspondence analysis (MCA) is a useful technique for the structural analysis of multivariate categorical data (Glynn, 2012; Takane and Heungsun, 2006; Greenacre and

Nenadic, 2010). It gives insight into the complex dependence structure of such data sets by making plots. MCA has proved to be an important and useful tool for analyzing the association that is present in data sets with many variables (Schriever, 1986; Greenacre, 2006; Greenacre and Blasius, 2006). MCA assigns scores to rows (representing the subjects) and columns (representing the response categories) of a data matrix, yielding a graphical display of the rows and the columns of the data matrix. The graphical display facilitates the intuitive understanding of the relationships among the categories of the variables (Greenacre, 2010).

Let $x_1, x_2, \ldots x_k$ be categorical random variables. The technique MCA seeks *k* real valued functions $\Phi_{11}, \Phi_{21}, \ldots \Phi_{k1}$, defined on the categories (possible values) of $x_1, x_2, \ldots x_k$ respectively, such that the first principal component of the correlation matrix of $\Phi_{11}(X_1)$, $\Phi_{21}(X_2), \ldots \Phi_{kl}(X_k)$ has maximal variance. This principal component is called the first MCA component. It describes the most informative part of the variation between the categorical variables. Clearly, it is no restriction to assume that the derived variables $\Phi_{11}(X_1)$ have expectation zero and variance unity, for $i=1,\ldots k$. Subsequently, MCA seeks a second component which has maximal variance but which is uncorrelated with the first. This procedure is continued with a third component, a fourth component or until no new component that is uncorrelated with the previous components can be found.

The *t*-th MCA component is the linear combination of transformed variables.

$$Y_t = \sum_{l=1}^{k} \alpha_{lt} \Phi_{lt}(X_t)$$

for which $\mu_t = var\ (Y_t)$ is maximal subject to

$$E\ \Phi_{1t}(x_1)=0,\ \ var(\Phi_{lt}(X_l)) = 1 \text{ for } l=1,\ldots, k,$$

and the normalization constraint

$$\sum_{t=1}^{k} \alpha^2 \mathrm{lt} = 1,$$

Corr (Y_t, Y_s) = 0 for s =l,… *t*-l.

The MCA solution consists of all *k*+l tuples (μ_t,(μ,$\alpha_{lt}\Phi_{lt}(X_1)$, ..., $\alpha_{kt}\Phi_{kt}(X_k)$) for *t*=1,2,…The value $\alpha_{lt}\Phi_{lt}(x)$ is called the category score on the *t*-th MCA component of the category *x* of X_l; *l* =1,…, *k* ; *t* =1, 2,…

It follows directly from the definition that

$$var(Y_t) = \sum_{j=1}^{k} \sum_{l=1}^{k} \alpha_{jt}\alpha_{lt}\, Corr(\Phi_{jt}(X_j), \Phi_{lt}(X_l))$$

which means that MCA only considers the bivariate marginals of the *k*-dimensional probability distribution of $x_1,\ldots x_k$. It is well known (Schriever et al (n,d) that a MCA solution always exists and can be obtained by solving a generalized eigenvalue problem of the super matrix containing all bivariate marginal probability distributions. MCA can be seen as a generalization of the principal component analysis to nominal variables. Moreover, when X_1, X_2,…X_k are all dichotomous, e.g. 0-1 variables, then by the normalization Φ_{1t} (1) = $(\frac{1-\pi_l}{\pi_l})^{\frac{1}{2}}$ and $\Phi_{1t}(0)$ =- $(\frac{\pi_l}{1-\pi_l})^{\frac{1}{2}}$ where π_l = p{X_l =1}=1- p{ X_l = 0} for *l* =1, 2,…, *k* and *t* =1, 2,.. Hence the variance of Y_t is only maximized with respect to the variable weights α_{1t}, α_{2t},…α_{kt} for *t* =1, 2,…Therefore, MCA in the dichotomous case is equivalent to finding the principal components of the covariance matrix of $\Phi_{11}(X_1)$, $\Phi_{21}(X_2)$,… $\Phi_{k1}(X_k)$, that is, of the correlation matrix of $X_1, X_2,\ldots, X_k$.

Put differently, Everitt and Graham (2001) have explained that the starting point for MCA is the indicator matrix, Z. Each row of this matrix will have *k* values of unity and C-*k* zero values, where *k* is the number of categorical variables in the data set and C is the total

number of categories, so that C= $\sum_{i=1}^{k} C_i$ where C_i is the number of categories of the i^{th} variable. For a *k*-way contingency table, the indicator matrix can be written as

$$Z=[Z_1,\ldots Z_k],$$

where Z_1 is the n× C_i matrix containing the C_i indicator variables for the i^{th} way of the table.

The matrix given by:

$$B=Z'Z\ ,$$

is called the *Burt* matrix and contains the sub matrices $Z_i'Z_i$ the two–way contingency tables based on variables *i* and *j*.So

$$B=\begin{Bmatrix} Z_1Z_1 & Z_1Z_2 & .... & Z_1Z_k \\ Z_2Z_1 & Z_2Z_2 & ... & Z_2Z_k \\ ... & ... & ... & ... \\ Z_kZ_1 & Z_kZ_2 & ... & Z_kZ_k \end{Bmatrix}$$

B has a "block" structure, with the sub matrices $Z_i'Z_i$ on the diagonal being simply diagonal matrices of column sums, and every off-diagonal block being a two–way table of marginal totals for the i^{th} and j^{th} variable. Hence according to Everitt and Graham (2001), MCA involves essentially the extraction of the eigenvalues and eigenvectors of a Burt matrix.

In data analysis three types of multiple correspondence analysis are encountered: Indicator multiple correspondence analysis (Indicator MCA), Burt multiple correspondence analysis (Burt MCA) and Joint multiple correspondence analysis-Joint MCA (Glynn, 2012). Indicator MCA that is also called homogeneity analysis uses a binary matrix of indicators to combine the binary correspondence analyses. Results obtained are similar to Burt MCA

but according to Greenacre (2007), Burt MCA produces more optimistic percentages of inertia. However, in multiple correspondence analysis the percentage of explained inertia is not very important in interpretation since it severely underestimates the representative quality of the biplot map (Glynn, 2012). Joint MCA is based on Burt MCA and according to Greenacre (2006; 2007) it is superior both in terms of explained inertia and in the accuracy of visualization. It works by restricting the analysis to the cross tabulations that typically contain the correlations of interest that explain the inertia.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Data source

In order to identify suitable statistical method for analyzing large and complex mixed data sets from surveys for statistical analysis of aflatoxin contamination in peanuts, data from ICRISAT collected under the Peanut CRSP project was used for this study. The data was collected between March and July 2009 from three provinces of Kenya namely Nairobi, Western (Busia district) and Nyanza (Homa bay, Rachuonyo, Kisii Central and Kisumu East districts).

3.2 Sampling, data collection and aflatoxin analysis

A survey exercise was conducted in the three provinces and purposeful sampling was utilized in identifying vendors that were trading in peanuts. The sampling method targeted areas where peanuts were majorly produced or traded. Nairobi is a major market outlet of peanuts and peanut products sourced from within Kenya and other countries. It has both large and small scale peanut processing enterprises. Busia district is a major peanut produce, has several market outlets for peanuts and has a border point with Uganda which is another major peanut producer characterized by a thriving cross-border trade. Nyanza province is also a leading producer of peanuts and has several peanut processors as well as a high demand for peanut products.

The data was collected using a questionnaire and issues addressed were those practices that were related to either mould or aflatoxin contamination in peanuts. Some of the variables collected(factors considered) included: gender, age and educational level of trader; the type

of peanut products in the market which included podded raw kernels, shelled raw kernels, roasted kernels, peanut butter, boiled kernels, fried kernels, or spoilt kernels; packaging material used for peanut products whether it was jute bags, propylene bags, metal tins, PVC bags, paper, plastic jars, plastic basins, or reeded baskets; source of peanuts whether from own harvest, bought locally or imported from neighboring countries; mode of peanut product transaction whether it was direct or through middlemen; nature of market outlets whether the peanut products were sold through hawking, informal market structures, formal market structures, stockists, or supermarkets; mode of transporting peanut products to the markets whether it was through the use of bicycles, vehicles, boats, carts and donkeys.

Other aspects (variables or factors) detailed included the duration that peanut products took before being sold, the state of marketing structures by describing the condition of the roofing materials, walls, floors and ventilation. Post harvest pest and disease control measures that were done were determined and varieties of peanuts grown. A total of 1260 vendors were interviewed and a peanut sample taken from each interviewee was analyzed for aflatoxin contamination with an indirect competitive ELISA method by preparing an aflatoxin-bovine serum albumin conjugate in carbonate coating buffer at 100 ng/ml concentration and dispensing 150 µl in each well of the Nunc-Maxisorp ELISA plates. Absorbance was then measured at 405 nm in an ELISA plate reader (appendix 2) as described in Mutegi et al. (2013).

3.3 Categorization of peanut samples according to aflatoxin content

Peanut samples were grouped into three categories based on their aflatoxin content: samples with ≤4µg/kg (Category 1), >4-10µg/kg (category 2) and >10µg/kg (category 3). Aflatoxin category ≤4µg/kg represented the European Union (EU) regulatory limit for total

aflatoxin for peanuts (EC, 2006). The category >4-10µg/kg represented peanuts which could be rejected in the EU countries but could be accepted in Kenya under the Kenya Bureau of Standards (KEBS) regulations (KEBS, 2007) while category >10µg/kg aflatoxin contaminated peanuts could be rejected under the KEBS standards. The dependent categorical variable (aflatoxin category) was to be analyzed in relation to predictor variables with utilization of multiple correspondence analysis.

3.4 Statistical analysis approach

The data was cleaned, validated and coded for nominal categorical variables. It was then analyzed for Normality test (Shapiro-Wilk test) and the response variable (aflatoxin level) was not normally distributed ($T= 0.0563$; $p< 0.001$) and hence in subsequent analysis it was to be analyzed through generalized linear model (GLM). This was done to assess whether some of the assumptions for subsequent analyses such as multiple regression and principal component analysis could hold when the dependent variable was continuous and when fitted in a model. The data was then analyzed with contingency tables analysis (Pearson chi-square and Fisher's Exact test methods) as the benchmark statistical method against which results from other statistical methods could be evaluated. Multiple correspondence analysis (MCA) was used to analyze the large categorical variables in a low-dimensional Euclidean space. Principal component analysis (PCA) was applied to reduce the large data set into a lower dimension of few but significant variables and multiple regression as the statistical method for handling mixed predictor variables when the response variable was continuous. The data was also analyzed by analysis of variance (ANOVA) for categorical predictor variables when the response was continuous through generalized linear model (GLM). All data was analyzed at 5% level of significance where applicable and the

statistical analysis was done using Genstat 14th edition, STATA version 11 and SPSS 20th edition.

3.4.1 Contingency tables analysis (Pearson chi-square and Fisher's Exact Test methods)

The data was analyzed by contingency tables (Pearson's Chi-square) and when the cell counts were below 5 they were analyzed by Fisher's Exact Test method in establishing an association between any two variables.

3.4.2 Multiple correspondence analysis (MCA) in categorical data analysis

The data was subjected to Burt matrix MCA to make biplots to show the dependence structure in the data set. The data to be analyzed had to be stratified into subgroups since the data to be analyzed was very large. This was due to the fact that MCA implementation consists of Singular Value Decomposition (SVD) or the related Eigenvalue Decomposition (EVD) of the data (D'Enza and Greenace, 2012). Multiple correspondence analysis then aimed to identify a reduced set of synthetic dimensions maximizing the explained variability of the categorical data set. MCA assigns scores to rows (representing the subjects) and columns (representing the response categories) of a data matrix, yielding a graphical display of the rows and the columns of the data matrix. The graphical display facilitated the intuitive understanding of the relationships among the categories of the variables.

3.4.3 Principal component analysis in variable reduction

To statistically reduce the number of variables encountered in the study, the data was subjected to PCA and 37 principal components were extracted according to the number of variables being analyzed. Only variables with significant factor loadings per principal component were retained for further analysis.

3.4.4 Multiple regression analysis for categorical, discrete and continuous variables

In determining variables that played a significant role in the aflatoxin contamination of peanuts, the data was also analyzed by multiple regression analysis and also in developing a model for aflatoxin contamination in peanuts from the variables analyzed. The data was analyzed through GLM (Generalized linear model) when aflatoxin level was used as the dependent variable. In the determination of the most significant variables that could explain aflatoxin contamination in peanuts, Wald Test was used in model fitting by forward and backward selection of the terms (variables).

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Determination of variables that played a significant role in aflatoxin contamination of peanuts by use of multiple linear regression and analysis of variance

Results obtained from analysis of variance (ANOVA) for the data indicated that 10 variables were statistically significant in having played a significant role in aflatoxin contamination of the peanuts (Table 4.1A). They included the province where peanuts were sampled from, the education level of peanut vendors, type of peanut varieties, storage period before selling of peanuts in the market, the month when the peanuts were harvested, the mode of transacting the peanuts either purchasing them directly from farmers or from middlemen and non-application of any peanut protection methods. Leaking of the roof and the type of materials used on the wall structure and how the peanut house was used, either full time or part time were also significant.

Western province had the highest mean for aflatoxin contamination at 678675.4 µg/kg followed by Nyanza (7735.1 µg/kg) and finally Nairobi (933.5 µg/kg).The mean values were significantly different for Western and Nairobi provinces (lsd=219693.58). Respondents with tertiary level of education had the highest mean for aflatoxin (430646.3 µg/kg) followed by those with secondary education (177068.4 µg/kg), primary education (57594.4 µg/kg and last by those without formal education (1166.0 µg/kg). The mean value for those with tertiary education was significantly different from those with primary education and those without formal education (lsd=315887.71).

The aflatoxin mean value for not applying any peanut protection method (291591.3µg/kg) was significantly higher than when applied (54754.9µg/kg; lsd=182476.5). Practices such as drying, sorting and proper storage of peanuts have been documented in reducing aflatoxin contamination of peanuts significantly (N'dede et al., 2012). When the peanut housing structure was used part time as opposed to full time, it led to less aflatoxin contamination of the peanuts (115020.6 µg/kg part time, 127491.6 µg/kg fulltime) and when roof was leaking there was more contamination (295462.2 µg/kg) than when not leaking(134518.1 µg/kg).

Among the peanut varieties under study, the most susceptible varieties with the highest mean value for aflatoxin contamination included Red mixed (7928993µg/kg), Red small(593262.73µg/kg), Uganda red(163334.23µg/kg) and Tatu tatu (35464.63µg/kg).The peanut varieties with the lowest mean value for aflatoxin contamination included Brown medium(420.7 µg/kg), Homabay local(1443.8 µg/kg) and Brown kubwa (2033.4 µg/kg).Materials for peanut wall structures with the highest mean level for aflatoxin contamination included blocks(2.00E+07 µg/kg) and bricks(4.00E+05 µg/kg). Studies conducted by Mutegi et al.(2009) in western Kenya observed that planting improved cultivars would lower the odds of aflatoxin contamination to a half those for local landraces.

Compared to the rest of the harvesting months (January and June, mean of 1.00E+03 µg/kg) for the peanuts, peanuts harvested in October had the highest mean for aflatoxin contamination (2.00E+06 µg/kg). This is the month when a recent incident of aflatoxin food contamination occurred in the year 2011 when Proctor and Allan East Africa, a cereal

manufacturer, recalled 25 tons of contaminated Unimix (a high-protein mix containing corn flour) destined for relief efforts in drought-affected areas of Kenya(Grohe et al., 2011).

Table 4.1A: Determination of significant variables in aflatoxin contamination of peanuts

Source of variation	Seq. SS	df	MS	F	Prob > F
Model	1.02E+15	132	7.71E+12	5.84	**0.0000**
Province	7.02E+13	2	3.51E+13	26.59	**0.0000**
District	2.08E+10	9	2.31E+09	0.00	1.0000
Peanut variety	6.59E+13	17	3.88E+12	2.94	**0.0001**
Peanut sample type	3.21E+12	8	4.02E+11	0.30	0.9645
Packaging material	1.29E+13	8	1.61E+12	1.22	0.2833
Mode of transportation	3.36E+12	6	5.61E+11	0.42	0.8629
Where samples were sourced from	3.89E+12	8	4.86E+11	0.37	0.9375
Duration before storage of peanuts	6.57E+12	2	3.29E+12	2.49	0.0837
Storage period before selling of peanuts	1.75E+14	8	2.19E+13	16.61	**0.0000**
Year of harvest	2.06E+10	1	2.06E+10	0.02	0.9007
Month of harvest	5.80E+13	12	4.83E+12	3.66	**0.0000**
Type of vendor	1.09E+13	5	2.17E+12	1.65	0.1453
Mode of transaction	1.12E+13	1	1.12E+13	8.46	**0.0037**
Gender of respondent	7.99E+11	1	7.99E+11	0.61	0.4368
Age of respondent	5.41E+12	6	9.02E+11	0.68	0.6637
Education level of respondents	6.39E+13	3	2.13E+13	16.13	**0.0000**
Sieving as protection method	1.05E+12	1	1.05E+12	0.80	0.3719
Sorting as protection method	2.77E+12	1	2.77E+12	2.10	0.1482
Tumbling as protection method	2.23E+12	1	2.23E+12	1.69	0.1938
Drying as protection method	1.91E+12	1	1.91E+12	1.44	0.2299
Non-use of protection methods	7.78E+12	1	7.78E+12	5.89	**0.0154**
Roofing materials	7.36E+12	6	1.23E+12	0.93	0.4733
Leaking of roof	1.06E+13	1	1.06E+13	8.01	**0.0048**
Materials used for walls	4.66E+14	9	5.18E+13	39.19	**0.0000**
Presence of crevices in house	3.46E+11	2	1.73E+11	0.13	0.8774
Use of pallets	4.47E+11	1	4.47E+11	0.34	0.5609
Hygiene of pallets	1.53E+11	2	7.65E+10	0.06	0.9438
Hygiene of floor	3.23E+12	1	3.23E+12	2.45	0.1182
Insects in house	4.09E+12	1	4.09E+12	3.10	0.0788
Floor cracked	4.08E+12	1	4.08E+12	3.09	0.0793
Type of floor	7.41E+12	2	3.70E+12	2.81	0.0611
Enough lighting	1.26E+11	1	1.26E+11	0.10	0.7571
Windows present	3.41E+10	1	3.41E+10	0.03	0.8723
House used fulltime/part time	5.56E+12	1	5.56E+12	4.21	**0.0405**
Musty smell in house	6.31E+11	1	6.31E+11	0.48	0.4896

Note: Figure in bold indicate significant variables. R squared=0.504, Adjusted R squared=0.4177

Results from multiple regression analysis indicated that 7 variables were statistically significant in having played a significant role in aflatoxin contamination of the peanuts (Table 4.2B; Appendix 1). Compared with respondents without formal education, those with tertiary education were significant in contributing to aflatoxin contamination of peanuts. They had higher mean for aflatoxin contamination than those without formal education (Table 4.1A) and this could be attributed to the use of plastic jars as the preference peanut packaging material(Fig 4.6) which had been associated with aflatoxin contamination category >10 µg/kg(Table 4.6). In reference to mud floor, cemented floors were significantly contributing to aflatoxin contamination of peanuts. Peanuts harvested in the month of October were more contaminated than those harvested in January. When the peanut housing structure was used part time as opposed to full time it led to significantly less aflatoxin contamination. Compared to peanut seeds that were in pods, peanut samples that were taken from spoilt peanuts were significantly more aflatoxin contaminated. This observation was in agreement with Mutegi et al.(2013) who found out that the most aflatoxin contaminated peanut products in Kenyan peanut market were peanut butter and spoilt peanut products.

The use of blocks as wall materials as compared when there is no wall in the peanut house structure led to significantly more aflatoxin levels in the peanuts. Storing peanuts for a period of 6 months before selling them significantly enhanced aflatoxin contamination as compared to storage period of 1 month.

Table 4. 1B: Parameter estimates from multiple regression for significant variables

Parameter	Estimate	s.e.	t(757)	t pr.	Factor reference level
Constant(Intercept)	-1E+06	1811903	-0.74	0.457	
Respondents with primary education	72069	148771	0.48	0.628	*No formal education*
Respondents with secondary education	225094	171813	1.31	0.191	
Respondents with tertiary education	969996	318562	3.04	**0.002**	
Cemented floor	447687	199922	2.24	**0.025**	*Mud floor*
Wooded floor	370732	724309	0.51	0.609	
February harvest	156506	1013169	0.15	0.877	*January harvest*
March harvest	-289327	747854	-0.39	0.699	
April harvest	-61718	579947	-0.11	0.915	
May harvest	153201	596828	0.26	0.797	
June harvest	-10611	558305	-0.02	0.985	
July harvest	-115364	554210	-0.21	0.835	
August harvest	-15394	573153	-0.03	0.979	
September harvest	-24323	622313	-0.04	0.969	
October harvest	2146804	628746	3.41	**<.001**	
November harvest	-3338	560019	-0.01	0.995	
December harvest	172096	569389	0.3	0.763	
Unknown harvest month	-335300	1360364	-0.25	0.805	
House used part time	248258	123268	2.01	**0.044**	*House used fulltime*
Whole seed(shelled) peanuts	173918	215880	0.81	0.421	*Peanut seeds in pods*
Roasted peanuts	252114	278775	0.9	0.366	
Peanut butter	93756	408245	0.23	0.818	
Boilled peanuts	-91643	409342	-0.22	0.823	
Podded peanuts	48601	823939	0.06	0.953	
Fried peanuts	461541	354235	1.3	0.193	
Spoilt peanuts	800811	325926	2.46	**0.014**	
Other peanut products	-51830	738878	-0.07	0.944	
Concrete wall	-263924	731266	-0.36	0.718	*No wall*
Iron sheets as wall	-188875	724332	-0.26	0.794	
Timber walled	-90193	873474	-0.1	0.918	
Brick walled	30685	752396	0.04	0.967	
Cement & sand walled	-363747	1024948	-0.35	0.723	
Reeded mats walled	161496	1053443	0.15	0.878	
Mud walled	-234583	1123245	-0.21	0.835	
Blocks as wall	1.2E+07	967007	12.15	**<.001**	
Sticks as wall	-375273	1517932	-0.25	0.805	
2 months in storage before selling	9785	100526	0.1	0.922	*1 months in storage before selling*
3 months in storage before selling	123336	137477	0.9	0.37	
4 months in storage before selling	131859	265094	0.5	0.619	
5 months in storage before selling	16087	915200	0.02	0.986	
6 months in storage before selling	3813293	516356	7.39	**<.001**	
7 months in storage before selling	-114279	1312224	-0.09	0.931	
10 months in storage before selling	315689	1194599	0.26	0.792	
20 months in storage before selling	196546	1230035	0.16	0.873	

Figures in bold indicate significant variables in reference to given factor reference level ($p<0.05$)

4.2 Evaluation of applicability of Multiple correspondence (MCA) and Principal component analyses (PCA) in interpretation of aflatoxin contamination of peanuts

In trying to establish if there could be any association for more than two categorical variables, multiple correspondence analysis was applied and the following information was obtained. In Nairobi males dominated the peanut trade while in Nyanza it was the females who dominated it (Figure 4.1). The peanut varieties traded in Nairobi included White small, Red kubwa, Unknown, Red medium and Mixed brown. Varieties in Nyanza included Brown medium, Mixed white, White medium, Brown kubwa and Homabay local while in Western province the varieties included Uganda red, Mbilimbili, Red small, Mixed variety, Red mixed and Brown small.

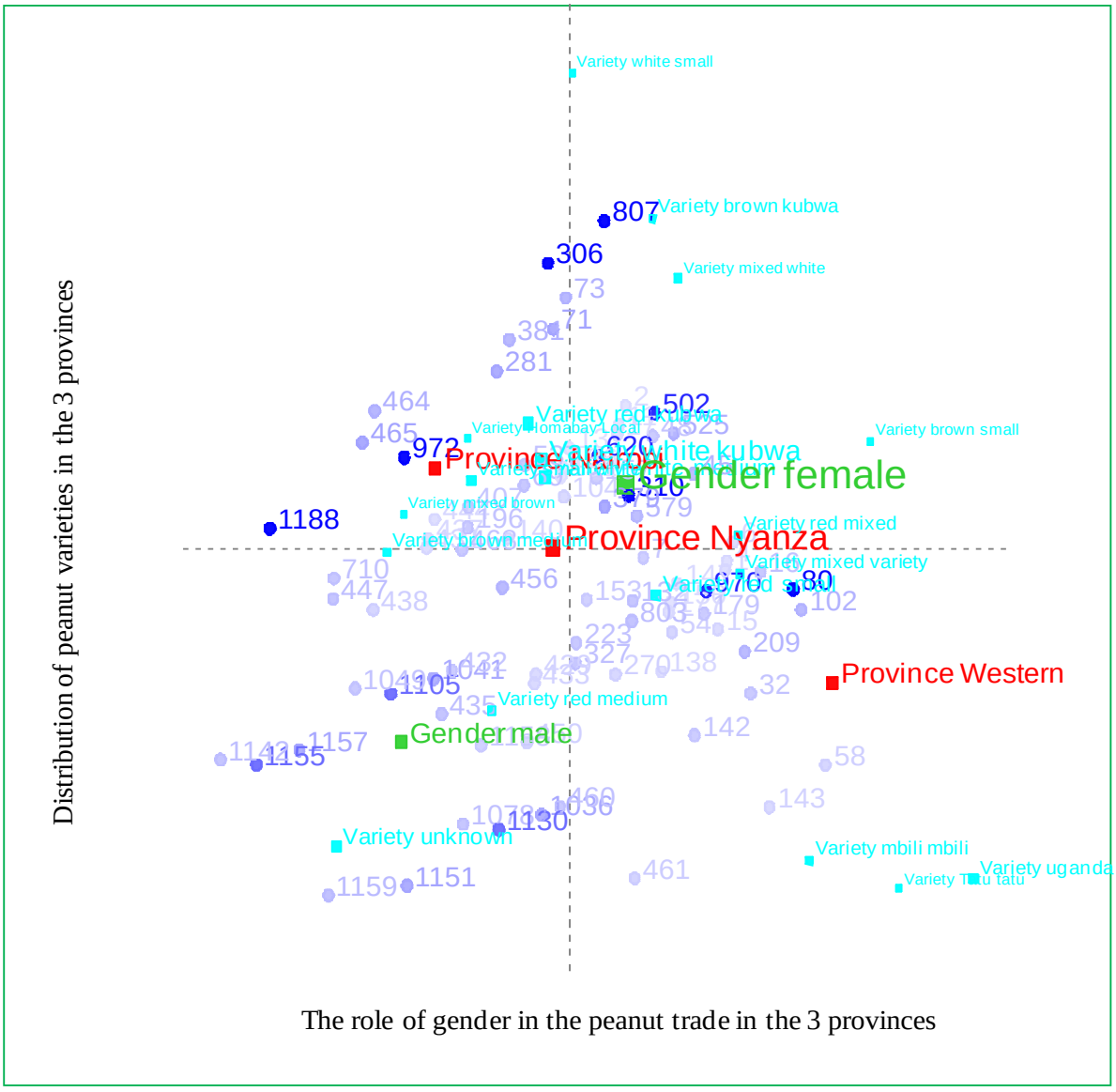

Figure 4.1: The role of gender in the peanut trade in the provinces and distribution of peanut varietiesq

Direct transaction in the peanut trade was associated with aflatoxin contamination category ≤4µg/kg (Fig. 4.2).When middlemen were involved, then aflatoxin contamination was associated with the category >10 µg/kg. Peanut samples obtained from supermarkets, formal open-air market and stockists were associated with aflatoxin category ≤4µg/kg while those obtained from hawkers and informal open-air market were associated with aflatoxin category >10 µg/kg. Formal open-air markets differed from informal open-air

markets in terms of available social amenities such as waste disposal facilities, water, public toilets etc that the former had while the later did not have. Peanuts obtained from own harvest and those purchased from neighbours were associated with aflatoxin category ≤4µg/kg. Those peanuts did not pass through middlemen and therefore had few post-harvest sources for aflatoxin contamination. As the number of middlemen increased in the peanut trade, so were the potential sources for aflatoxin contamination. And as such peanuts imported from Uganda, Tanzania, Zambia, Malawi or obtained from processing company, local eastern region & local western region of Kenya were associated with aflatoxin category >10 µg/kg .

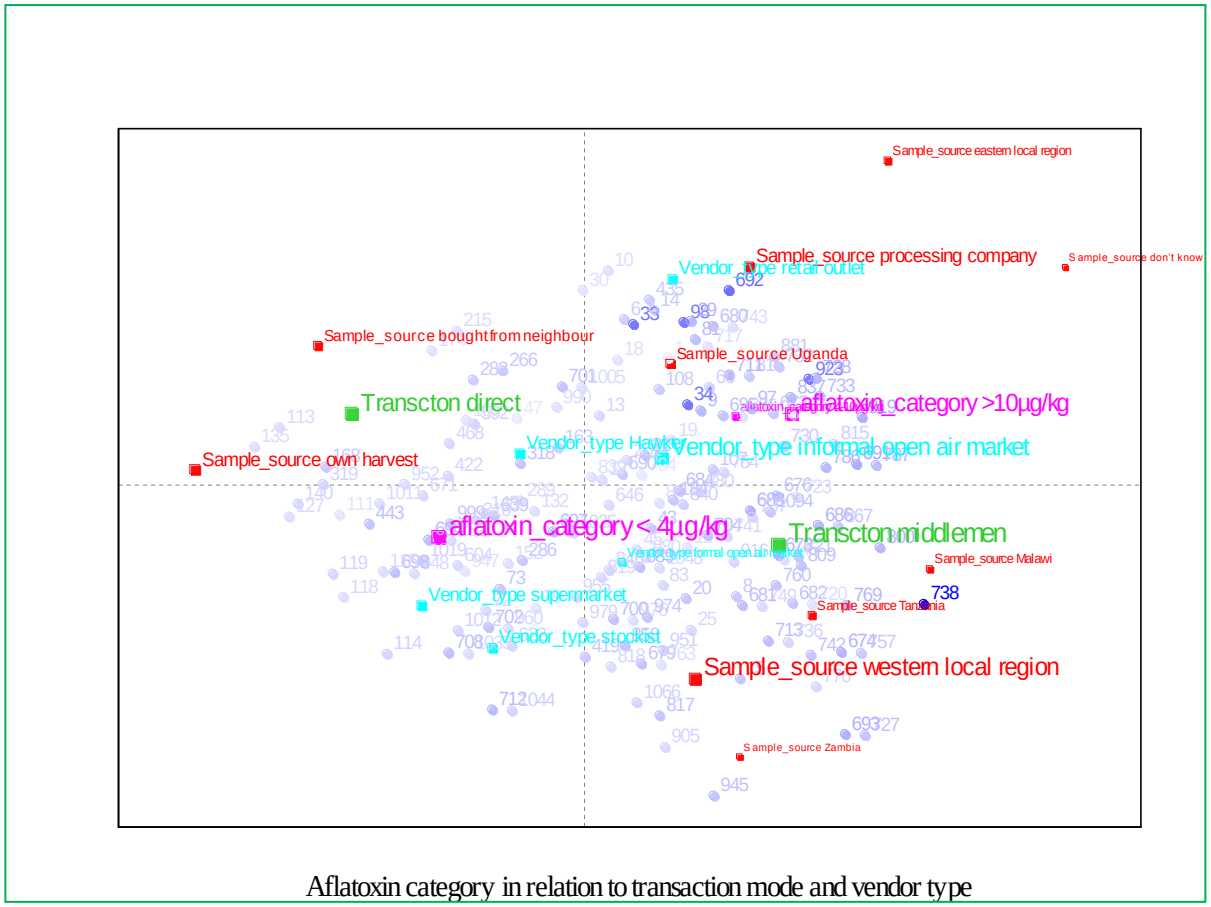

Figure 4.2: Multiple correspondence analysis plot on aflatoxin contamination category, vendor type, peanut sample source and peanut transaction mode.

Musty odour was associated with the presence of insects on the floor of the peanut storage structure (Fig. 4.3). Observations repeatedly showed that absence of insects was associated with the absence of musty odour in peanut storage structure. Perceived musty odour and visible mould are some of the indicators of microbial growth and proxy indicators for aflatoxin production in agricultural storage structures (Ayanbimpe et al., 2012). The research findings were also in agreement with some other studies that had indicated that high insect activity in peanuts and corn were associated with mould growth and aflatoxin production in peanuts and corn (Widstrom, 1979; Diener et al., 1987; Saad, 2004).

Peanut varieties Red medium, Mixed white, Brown medium, White medium and Red mixed were associated with aflatoxin category ≤4µg/kg. The peanut varieties Red kubwa, White kubwa, Red small, Mixed variety and Mbilimbili were associated with aflatoxin category >4-10µg/kg. The varieties Small white, Uganda red, Unknown variety and Brown small were associated with aflatoxin category >10 µg/kg (Fig.4.3).

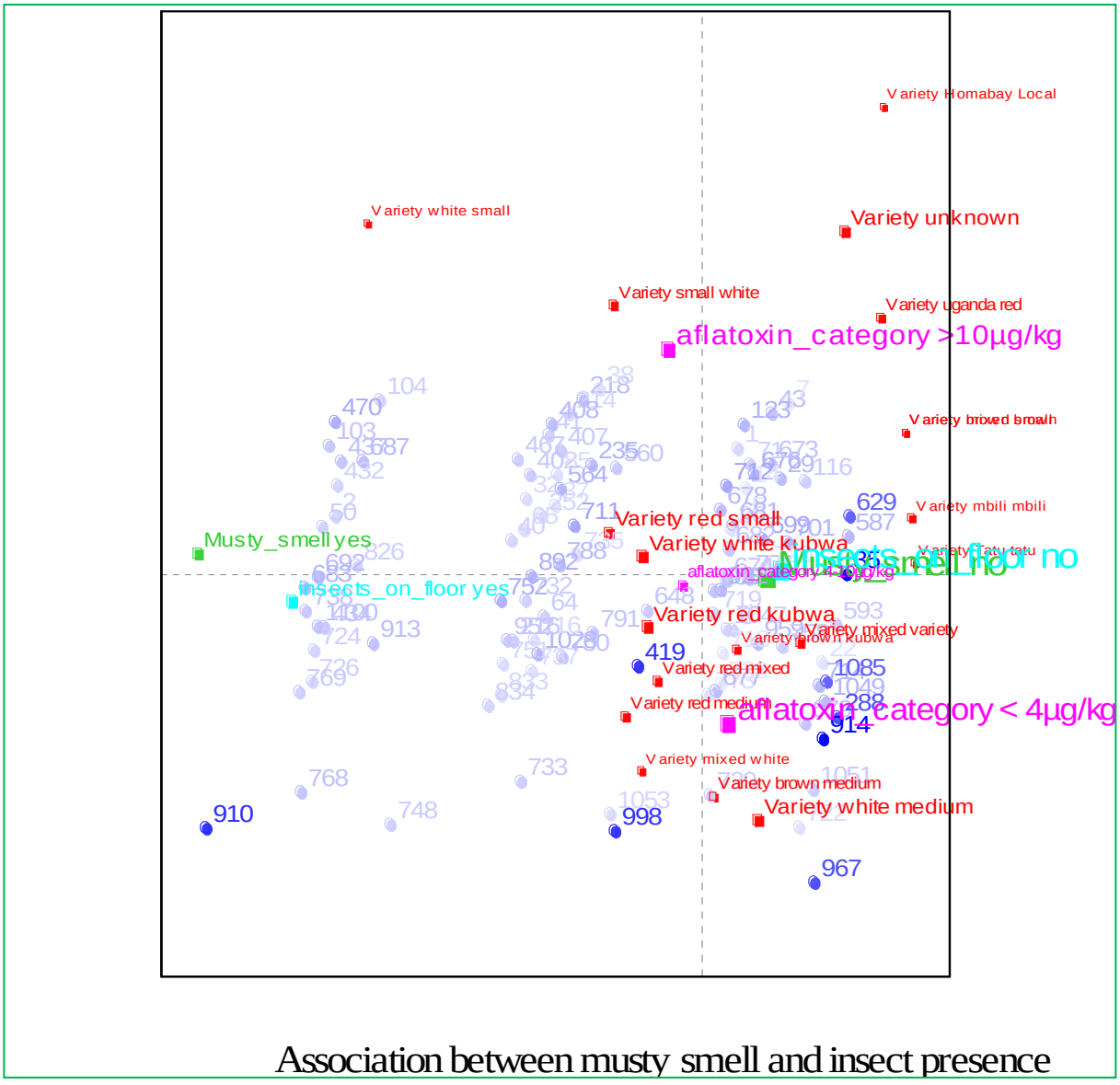

Figure 4.3: Multiple correspondence analysis plot on aflatoxin contamination category, implication of musty smell, peanut varieties and insect attack on peanuts.

Pallets were mainly used in Nairobi and Nyanza while they were not utilized in Western province (Fig. 4.4). Almost half of the pallets used in Nairobi were clean while the rest were dirty. The pallets had been considered clean on the basis of being kept dry, not being stored outdoors unprotected, keeping of pallets separated by having special pallets for hygienic zones, use of pallet inverters and pallets being pasteurized by any applicable method including use of high pressure sprinkling water, high temperature treatment and microwave technology.

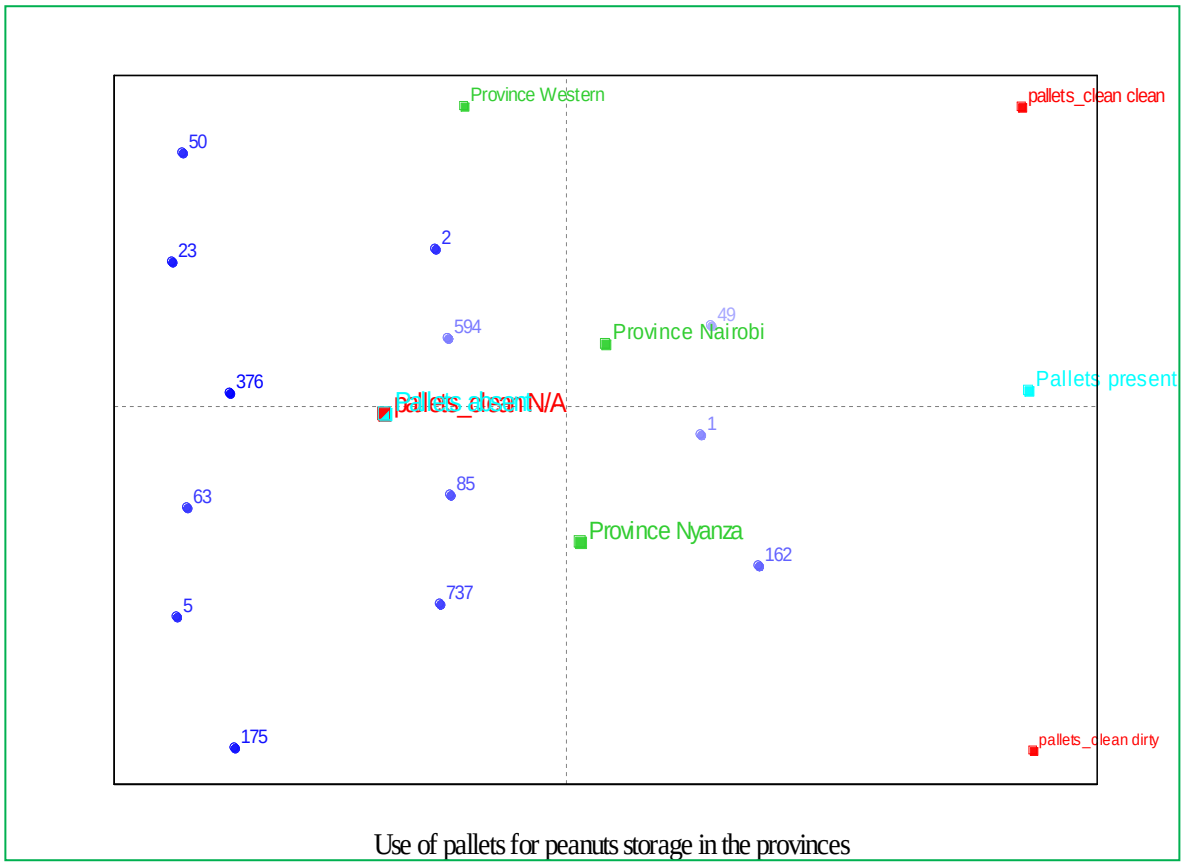

Figure 4.4: Multiple correspondence analysis plot on the use of pallets in the 3 provinces and determination of their hygiene status.

In Nyanza Province, the retailers purchased the peanuts directly from peanut farmers while middlemen were involved in Nairobi and Western provinces (Fig. 4.5). In Nyanza province,

the peanuts traded were from the vendors' own harvest and the rest purchased from neighbours for retailing. In Nairobi, peanuts in the market were obtained from the eastern region of Kenya, processing companies and the rest imported from Tanzania. In Western Province, peanuts in the market had originated from the western region of Kenya with the rest having been imported from Uganda.

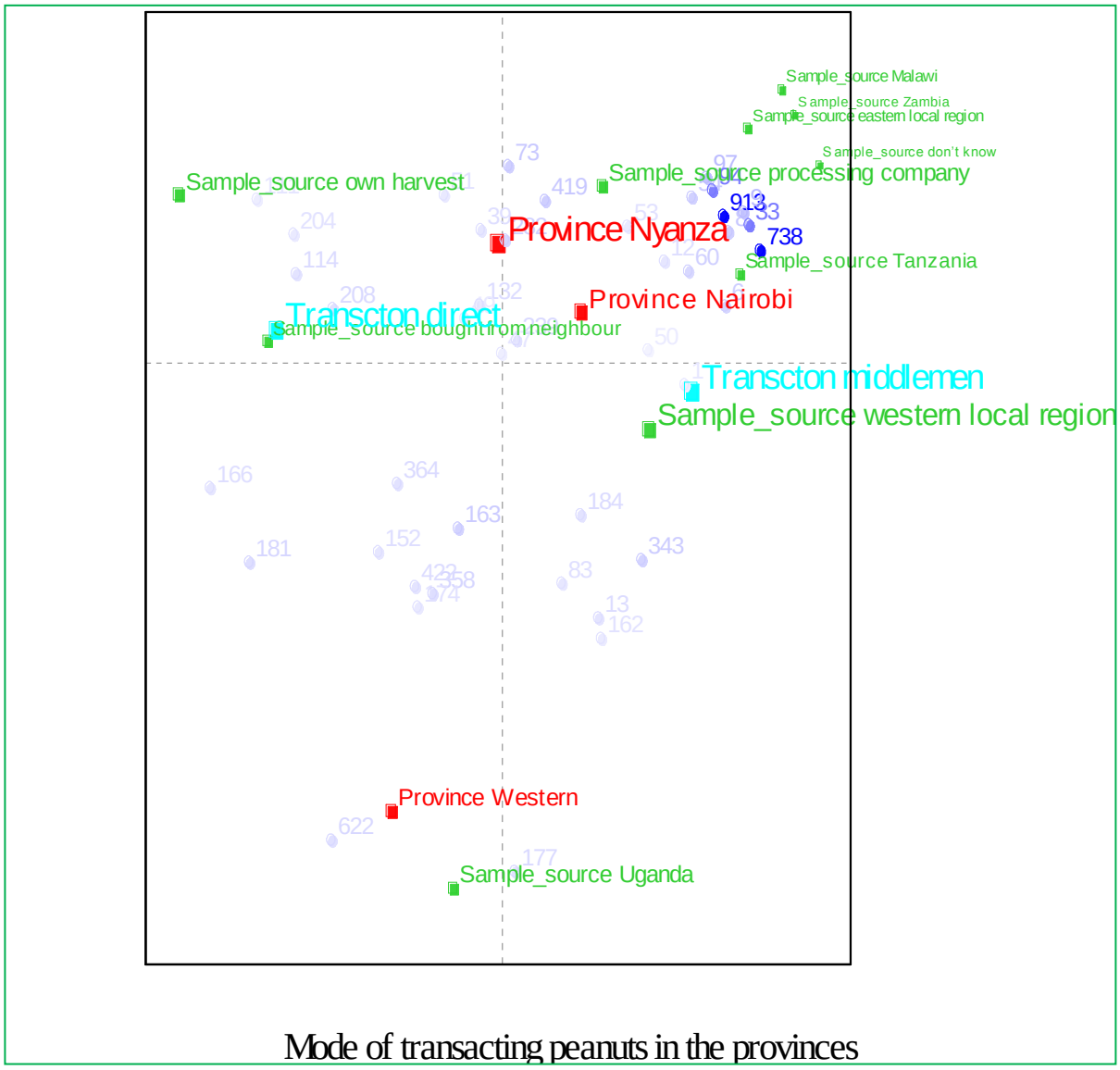

Figure 4.5: Multiple correspondence analysis plot on the mode of peanut transaction and the source of the peanuts sampled in the 3 provinces.

Insects were present where the floor of the housing structure was cracked (Fig. 4.6). Insects were not present where there were no cracks on the floor of the peanut housing structure. Respondents with no formal education or those with primary schooling were associated with the use of metal tins, plastic basins, reeded baskets and papers as the packaging materials for the peanuts. Respondents with secondary education were associated with the utilization of PVC bags, propylene bags and jute bags while those with tertiary education used mainly plastic jars as the package materials for the peanuts sampled in the survey. The findings that peanut vendors with different education levels preferring different peanut packaging materials could help explain why aflatoxin contamination levels increased with higher education attainment of the respondents (Fig.4.6; Table 4.2).

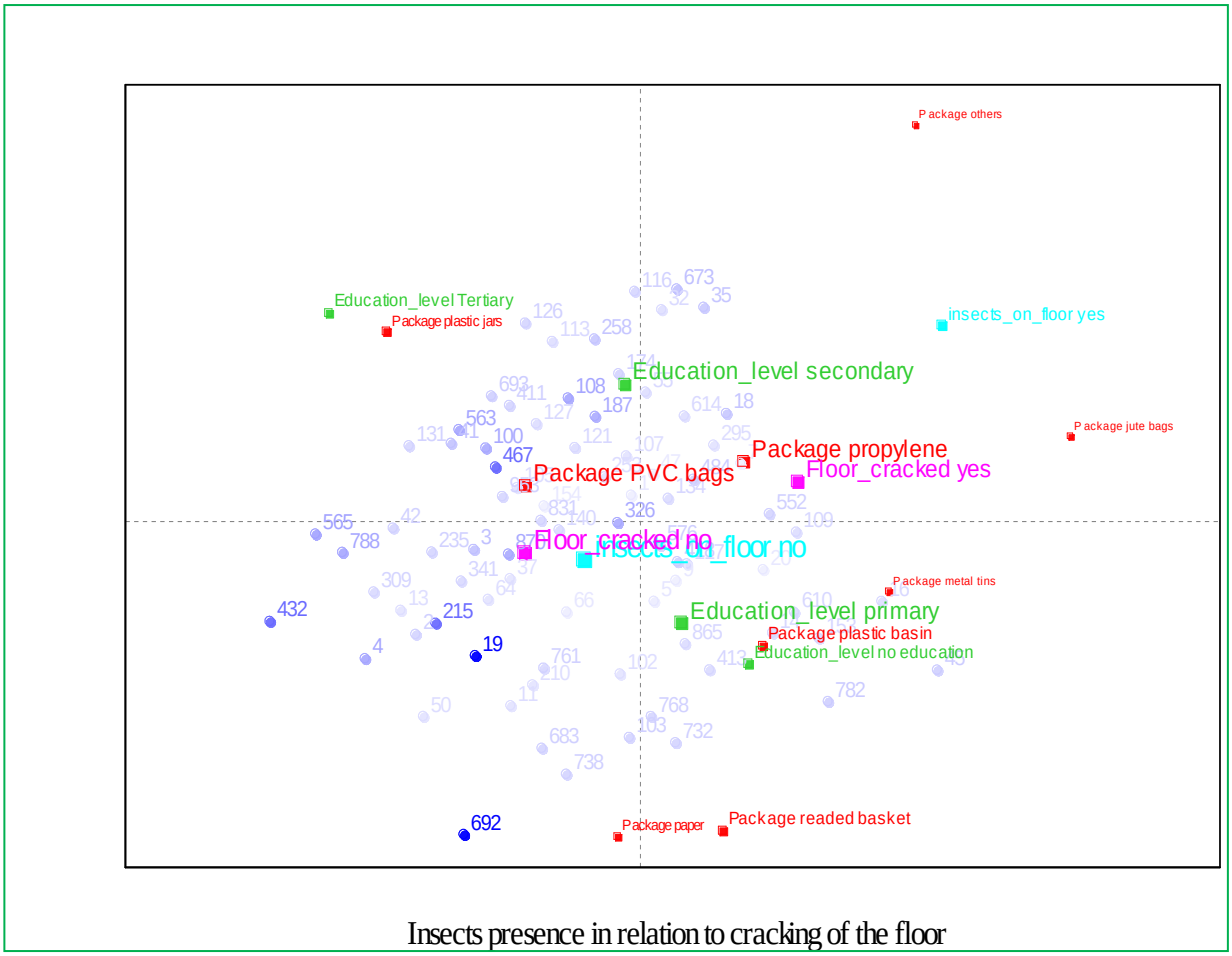

Figure 4.6: Multiple correspondence analysis plot on education level in relation to peanut packaging materials, presence of insects and cracking of the floor

The application of peanut processing methods i.e. drying, tumbling, sieving and sorting was associated with aflatoxin contamination category ≤4µg/kg in peanuts. Failure to apply crop protection methods was associated with aflatoxin contamination category >4-10µg/kg and category >10 µg/kg (Fig. 4.7). Hence these findings were in agreement with other studies that aflatoxin contamination of peanuts could be minimized by adopting certain cultural, produce handling and storage practices (ICRISAT, 2000; Liang, 2006; Klich, 2007;Wagacha and Muthomi, 2008; Wang et al., 2010; Wu, 2010).

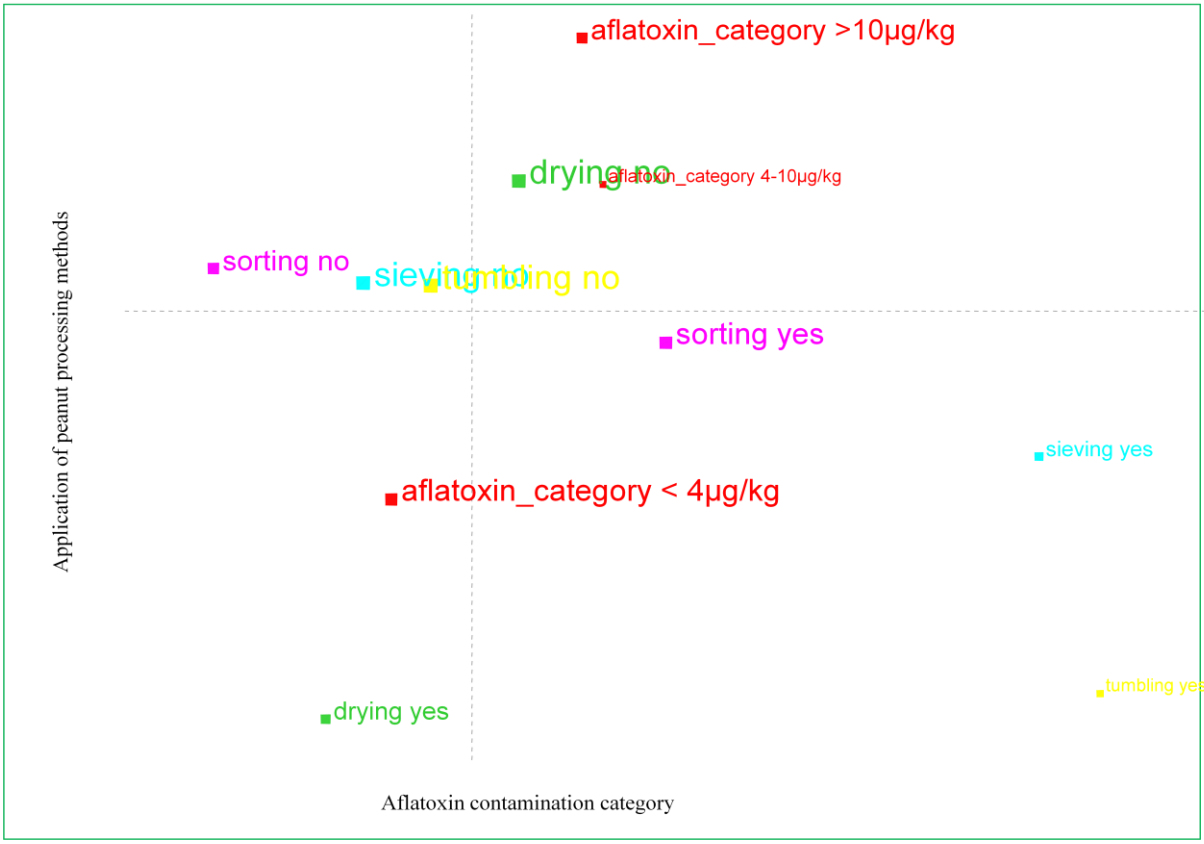

Figure 4.7: Multiple correspondence analysis plot on the effect of tumbling, sieving, sorting and drying in relation to aflatoxin level of peanut samples

When the data was subjected to principal component analysis, 37 principal components were extracted according to the number of variables being analyzed. However only 10 components could have been retained for interpretation according to Kaiser's eigenvalue criterion whereby any component displaying an eigenvalue greater than 1.00 is accounting for a greater amount of variance than had been contributed by one variable and as such is retained. However on applying the scree plot test (Fig.4.8) and looking for a break between the components with relatively large eigenvalues and those with small eigenvalues, only 4 components could have been retained.

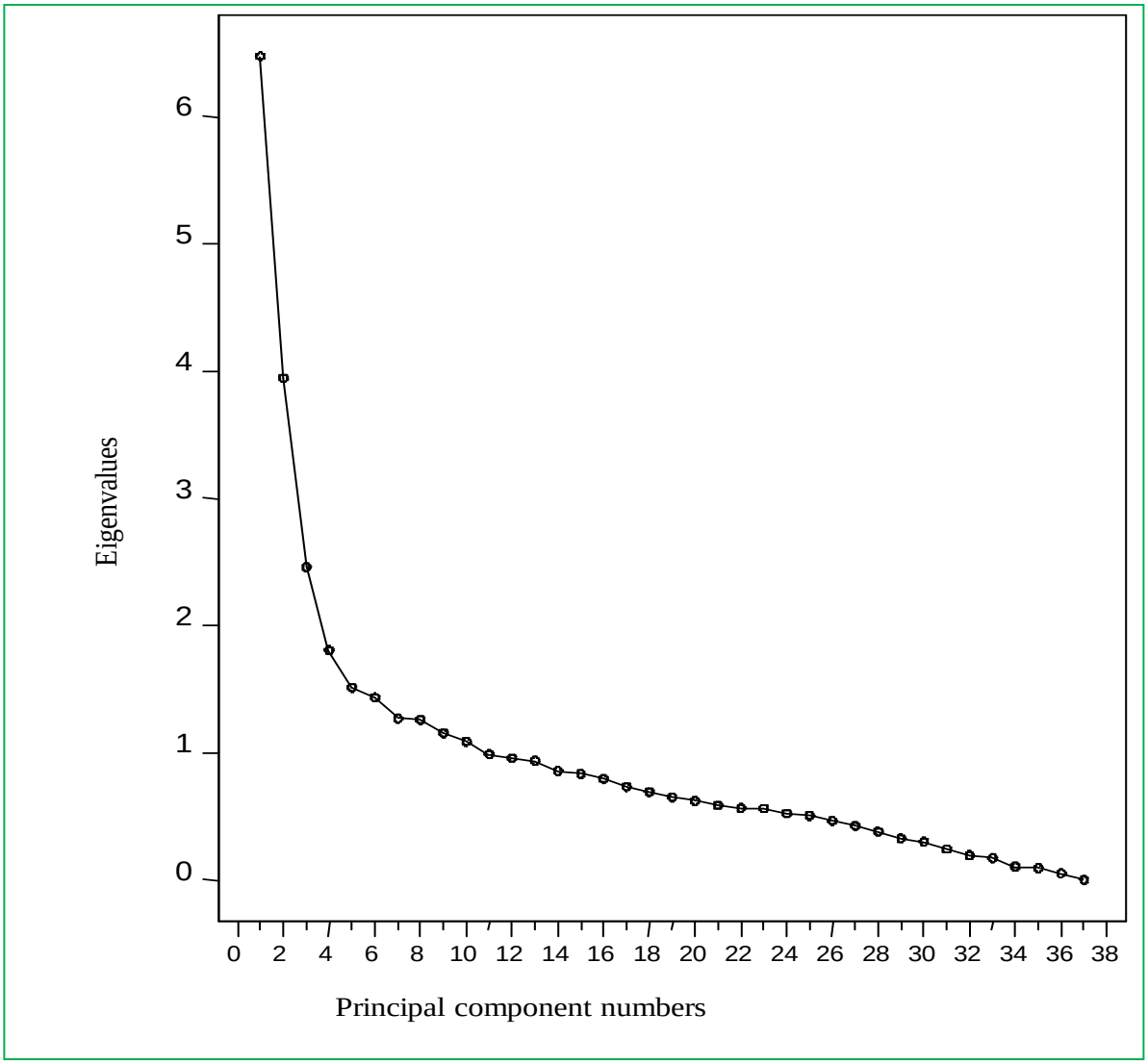

Figure 4.8: Scree plot for all extracted principal components

In the application of the interpretability criterion, 8 principal components were retained for further analysis. This was after verifying that the various variables that significantly loaded on a given component made sense in terms of what was known about the construct under investigation(Table 4.2) and the cumulative percentage of variance accounted for by the retained components was substantive.

Table 4.2: The 8 retained principal components and variables with significant factor loadings

Variables	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Age of the respondent	-0.0977	0.0967	**0.4987**	-0.0905	0.0105	-0.095	0.1705	0.0701
Crevices in housing structure	**0.4462**	-0.0451	-0.0112	-0.0128	-0.0485	0.0539	0.0234	0.0225
District peanut sampled	-0.0524	**0.4639**	0.0057	0.0033	-0.0487	-0.086	-0.0111	-0.0998
Education level of respondent	-0.0391	-0.0202	**-0.4793**	-0.0014	-0.0516	0.1183	-0.0134	0.0239
Enough light in house	-0.1491	0.0101	0.0769	0.1477	**0.312**	-0.0163	0.0494	-0.0505
Floor cracked	0.0291	**-0.3254**	-0.0919	-0.0502	-0.1751	-0.0434	0.1422	-0.1382
Insects present on the floor	-0.0142	-0.0106	-0.0684	0.0302	**-0.4823**	-0.0702	0.0724	-0.0686
Floor type	**-0.3974**	0.0568	0.0059	0.0341	-0.0464	0.0392	-0.0039	0.0125
Floor clean or not	0.2452	0.0378	-0.0416	0.0538	**0.3996**	-0.1702	0.0735	-0.0769
Gender of respondent	0.133	-0.026	**0.4653**	-0.0243	0.1099	0.1666	-0.0415	0.0048
Month for harvest	-0.0503	**-0.4912**	0.0761	0.0294	-0.0005	-0.1075	-0.0807	-0.0346
Year when peanuts harvested	0.0057	**0.492**	0.0231	0.0188	-0.0224	0.0014	-0.0213	-0.0733
House used fulltime/part time	0.1352	0.0519	-0.031	-0.0198	-0.0278	**-0.2491**	-0.1133	-0.0642
Mode of transport for peanuts	0.0984	0.0856	-0.048	0.0382	0.0738	0.0133	**-0.3782**	-0.0787
Musty smell in house	0.0706	0.0269	-0.0316	-0.0304	**-0.467**	-0.0859	-0.0052	0.0039
Package type for peanuts	-0.0344	0.057	0.0184	-0.0023	-0.0679	-0.1126	**-0.4172**	0.1304
Pallets whether used or not	-0.0128	-0.033	0.045	0.0109	-0.0715	**-0.5581**	0.0065	0.0312
Province	-0.0428	-0.2276	**0.3268**	0.175	-0.2413	0.0194	-0.0174	-0.094
Roof leaking	**0.3243**	0.0761	0.0255	0.0921	-0.136	-0.071	0.0063	0.0337
Materials used for the roof	**-0.3066**	-0.036	-0.0155	-0.0377	0.033	0.1343	0.0392	0.0082
Peanut sample type	-0.0574	-0.0873	0.131	0.0435	0.013	-0.0076	-0.3107	**0.3178**
Source of the peanut samples	0.0694	-0.052	-0.0082	0.0415	0.0055	0.1128	0.1841	**0.4466**
Transaction mode	0.0008	0.0465	-0.1752	-0.0315	0.0747	-0.1128	0.0894	**0.498**
Peanut variety	0.0239	0.0483	0.1024	0.0612	**-0.1458**	0.0881	0.0881	0.0057
Vendor type	-0.027	**0.1173**	0.0977	-0.0004	-0.0195	-0.0068	-0.0451	0.0714
Material types for wall structure	**-0.3601**	-0.0742	-0.0169	0.0937	0.0289	-0.0729	-0.0547	-0.0472
Presence of windows in the house	**0.3772**	-0.1446	0.0188	0.0338	0.1094	0.0812	-0.0734	-0.0452
Drying as protection method	0.0295	0.0861	0.0072	0.0151	-0.0187	-0.0787	0.1185	**-0.4032**
Moisture content	-0.0188	-0.0245	0.2082	-0.0823	0.053	0.0911	**-0.2808**	-0.1589
Non-use of protection methods	-0.0335	-0.0859	-0.0679	**0.5728**	0.1087	-0.0139	-0.029	0.0988
Pallets used are clean	-0.0304	-0.0437	0.0312	0.0307	0.056	**-0.5986**	0.0283	0.0081
Sieving as protection method	-0.0593	-0.132	-0.1084	**-0.3927**	0.2257	-0.0581	0.183	-0.1357
Sorting as protection method	-0.0039	-0.0398	0.0522	**-0.5968**	-0.081	0.0489	-0.1148	0.1437
Length of storage of peanuts before sell	-0.0359	-0.1008	-0.1114	-0.0034	0.0947	0.061	-0.0267	**-0.3194**
Duration taken before storage of peanuts	0.0286	0.0512	0.1052	0.1863	-0.099	0.1588	**0.4182**	0.0063
Tumbling as protection method	0.0053	0.0124	0.0887	-0.133	0.1049	-0.1403	**0.3504**	0.1004

The figures in bold are for variables with significant factor loadings for each retained principal component. PC1-Openings in the housing structure; PC2-Conditions of the harvest year; PC3-Respondent attributes; PC4-Nonuse of peanut protection methods; PC5-Cleaning of peanut housing structure; PC7-Crop protection methods; PC8-Mode of transacting peanuts.

The 8 retained principal components all qualified according to the interpretability and Kaiser's eigenvalue criterion and cumulatively accounted for 56.02% of the total variance in the data set (Table 4.3).

Table 4.3: Latent roots and percentage variance for the 8 retained principal components (PC)

Parameter of study for PCA	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Latent roots(eigenvalues) per component	6.48	3.95	2.46	1.81	1.51	1.44	1.27	1.26
Percentage variance for retained components	17.99	10.96	6.83	5.02	4.20	3.99	3.53	3.50

The variables that significantly loaded on each principal component were further subjected to ANOVA as predictors to determine whether they were statistically significant (at 0.05α) in explaining aflatoxin contamination in the peanuts (Table 4.4).

From Table 4.4, in the last column (surviving variables), the determination of the presence of windows, enough light(ventilation), duration taken before storage of the peanuts, application of peanut protection methods in addition to the province and district where the peanuts were sampled were all statistically significant (at 5% α) in explaining aflatoxin contamination in the peanuts. Hence, the utilization of PCA facilitated the statistical reduction of the variables that were under investigation from 37 to just six that could then be subjected to other types of statistical analysis where applicable.

Table 4.4: Determination of significant variables from those that loaded significantly on each principal component

PC nos. and Names	PCA variables	PC loadings	F probability value[a]	F probability value for PCA[b]	Surviving variables
1. Openings in housing structure	Crevices in housing structure	0.4462	0.2945	0.0046	
	Roof leaking	0.3243	0.1523		
	Presence of windows in the house	0.3772	0.0010		*
2. Conditions of the harvest year	District peanut sampled	0.4639	0.0005	0.0022	*
	Year when peanuts harvested	0.4920	0.9846		
	Vendor type	0.1173	0.3380		
3. Respondent attributes	Age of the respondent	0.4987	0.1042	0.0000	
	Gender of respondent	0.4653	0.6761		
	Province	0.3268	0.0000		*
4.Non use of peanut protection methods	Non-use of protection methods	0.5728	0.013	0.0000	*
5. Cleaning of housing structure	Enough light in house	0.3120	0.0068	0.0222	*
	Floor clean or not	0.3996	0.2356		
7. Crop protection methods	Duration taken before storage of peanuts	0.4182	0.0014	0.0135	*
	Tumbling as protection method	0.3504	0.3018		
8. Mode of transacting different sample types	Peanut sample type	0.3178	0.0739	0.1960	
	Source of the peanut samples	0.4466	0.5923		
	Transaction mode	0.4980	0.2886		

a - ANOVA for determining significant variables as indicated by the p-values ($p \leq 0.05$).
b - Determination of whether ANOVA model was significant ($p \leq 0.05$) for each principal component.
* - Surviving variables from the various principal components after ANOVA.

4.2.1 Summary on applicability of Multiple correspondence (MCA) and Principal component analyses (PCA) in interpretation of aflatoxin contamination of peanuts

In trying to establish if there could be any association for more than two categorical variables, multiple correspondence analysis was applicable. To facilitate the interpretation of the relationships, MCA searched for groups of objects that were homogeneous in their responses to the variables. It produced plots in which both objects and categories were represented as points in a low-dimensional space. Similar objects were represented close to each other in the object space and objects that had different scores on the variables were represented far apart. Also, categories which were at close distance revealed that they had a particular object in common (Groenen, Commandeur and Meulman, 1998). The advantage

of MCA was then to obtain a simplified representation of multiple associations characterizing attributes as well as to remove noise and redundancies in the data.

In agreement with Glynn (2012) and Akturk et al. (2007), Multiple correspondence analysis (MCA) is an exploratory statistical method and an excellent heuristic for getting into complex multi-factorial data and in identifying patterns but not for establishing their significance. In MCA biplots some data points could overlap causing problems for interpretation. This was a natural result of visualizing association through the proximity of data points and the biplots needed to be enlarged in order to discern what data points were overlapping. A different approach according to Glynn (2012) and was to analyze the data using *R* statistical software with *FactorMineR* package which has an option for dynamic graphing which allows movement of the labels as opposed to small data points interactively so that they don't overlap.

Again, for MCA to be more useful the data had to be stratified into subgroups due to the limitation of Singular Value Decomposition (D'Enza and Greenace, 2012). In developing the subgroups, MCA needed to be used in reasoned fashion. When too many factors were examined simultaneously, the results could not be interpretable since visualization of so many factors became impossible to decipher. Moreover according to Greenacre (2006; 2010) the chance of false associations increases dramatically when more than 15 variables that are considered simultaneously.

Principal component analysis was applicable in the reduction of the large data set into a lower dimension of few but significant variables that could then be subjected to other types of statistical analysis where applicable. This could simplify the statistical analysis of the

data by focusing on important variables as the most contributing factors in aflatoxin contamination of the peanuts. Principal component analysis was more useful when the variables were correlated otherwise in agreement with Shlens (2005), it could produce small factor loadings that could make it difficult in deciding variables to be retained for each principal component. In producing valid results, both MCA and PCA needed to be complemented with other statistical methods such as multiple linear regression and ANOVA in increasing the validity of findings.

4.3 Evaluation of the applicability of Multiple correspondence analysis in interpretation of identified significant variables compared to contingency table analysis (Pearson's Chi-square).

In establishing an association between any two categorical variables, contingency tables (Pearson's chi-square) could be used. For instance, in the study about the impact of education level of the vendors in relation to aflatoxin contamination level, the following results were obtained. There was a significant association between education level and the aflatoxin category (χ^2= 88.05; p< 0.001). Respondents with tertiary level of education were associated with the highest percentage (68.8%) of peanut product samples which did not meet the KEBS standards regarding total aflatoxin levels (>10µg/kg).The proportion of other vendors with different education levels were in the following decreasing order (Table 4.5): secondary (46%), primary (32.6%) and no formal education (17.6%). This could imply that there was a given peanut vendor attribute(s) that needed further investigation especially the kind of peanut packaging materials that were preferred at different education levels.

Table 4.5: Association between education level of vendors and aflatoxin contamination category

			Aflatoxin contamination category				Total
				<4µg/kg	>10µg/kg	4-10µg/kg	
Education level		Count	104a	0b	0b	0b	104
		% within education level	100.00%	0.00%	0.00%	0.00%	100.00%
	No formal education	Count	0a	106b	24a, c	6b, c	136
		% within education level	0.00%	77.90%	**17.60%**	4.40%	100.00%
	Primary	Count	0a	350b	176c	14b, c	540
		% within education level	0.00%	64.80%	**32.60%**	2.60%	100.00%
	Secondary	Count	0a	188b	178c	21c	387
		% within education level	0.00%	48.60%	**46.00%**	5.40%	100.00%
	Tertiary	Count	0a	28a	66b	2a, b	96
		% within education level	0.00%	29.20%	**68.80%**	2.10%	100.00%
Total		Count	104	672	444	43	1263
		% within education level	8.20%	53.20%	35.20%	3.40%	100.00%

NB.Each subscript letter denotes a subset of aflatoxin categories whose column proportions do not differ significantly from each other at the .05 level.

The results for the relationship between the type of packaging for peanuts and aflatoxin contamination category are shown in Table 4.6. There was significant association between the type of packaging for peanuts and aflatoxin category (χ^2=85.96; $p < 0.001$). The commonly used packaging materials for the peanuts were propylene bags which constituted 35.3% of the total packaging materials used. Peanut samples from propylene bags contained aflatoxin levels mostly in aflatoxin contamination category ≤4µg/kg and category

>10 µg/kg. Another commonly used packaging material was the PVC bag that constituted 33.9% of the total packaging materials. Compared with the rest of other packaging materials, plastic jars and metal tins had the highest percentage for aflatoxin contamination in the category >10 µg/kg. If these packaging materials were the most favorite by a section of peanut vendors (Fig 4.6), then this might explain why aflatoxin contamination level increases with higher education level of peanut vendors. Hence further research was needed on the type of peanut packaging materials which are susceptible to aflatoxin contamination but which are the most favorite in peanut markets to confirm the findings as to whether this could be a predisposing factor in aflatoxin contamination of peanuts.

Table 4.6: Association between peanut package materials and aflatoxin contamination level category

			Aflatoxin contamination category				
				< 4µg/kg	>10µg/kg	4-10µg/kg	Total
Package		Count	104_a	0_b	0_b	0_b	104
		% within Package	100.0%	0.0%	0.0%	0.0%	100.0%
	jute bags	Count	0_a	6_a	0_a	0_a	6
		% within Package	0.0%	100.0%	0.0%	0.0%	100.0%
	metal tins	Count	0_a	10_a	8_a	0_a	18
		% within Package	0.0%	55.6%	**44.4%**	0.0%	100.0%
	Others	Count	$0_{a, b}$	$1_{a, b}$	0_b	1_a	2
		% within Package	0.0%	50.0%	0.0%	50.0%	100.0%
	Paper	Count	0_a	18_a	17_a	6_b	41
		% within Package	0.0%	43.9%	41.5%	14.6%	100.0%
	plastic basin	Count	0_a	66_b	$26_{a, b}$	3_b	95
		% within Package	0.0%	69.5%	27.4%	3.2%	100.0%
	plastic jars	Count	0_a	27_a	62_b	$2_{a, b}$	91
		% within Package	0.0%	29.7%	**68.1%**	2.2%	100.0%
	Propylene	Count	0_a	249_b	148_b	12_b	409
		% within Package	0.0%	60.9%	36.2%	2.9%	100.0%
	PVC bags	Count	0_a	219_b	160_b	14_b	393
		% within Package	0.0%	55.7%	40.7%	3.6%	100.0%
	Reeded basket	Count	0_a	76_b	$23_{a, c}$	$5_{b, c}$	104
		% within Package	0.0%	73.1%	22.1%	4.8%	100.0%
Total		Count	104	672	444	43	1263
		% within Package	8.2%	53.2%	35.2%	3.4%	100.0%

Each subscript letter denotes a subset of aflatoxin contamination categories whose column proportions do not differ significantly from each other at the .05 level.

When the relationship between educational level of peanut vendors and aflatoxin contamination category was investigated by use of multiple correspondence analysis biplot (Fig 4.9), similar results like those obtained from contingency tables (Table 4.5) were

obtained. Respondents with tertiary level of education were associated with peanut product samples that did not meet the KEBS standards regarding total aflatoxin levels (>10µg/kg).Those with secondary, primary and without formal education were associated with decreasing level of association with aflatoxin contamination categories.

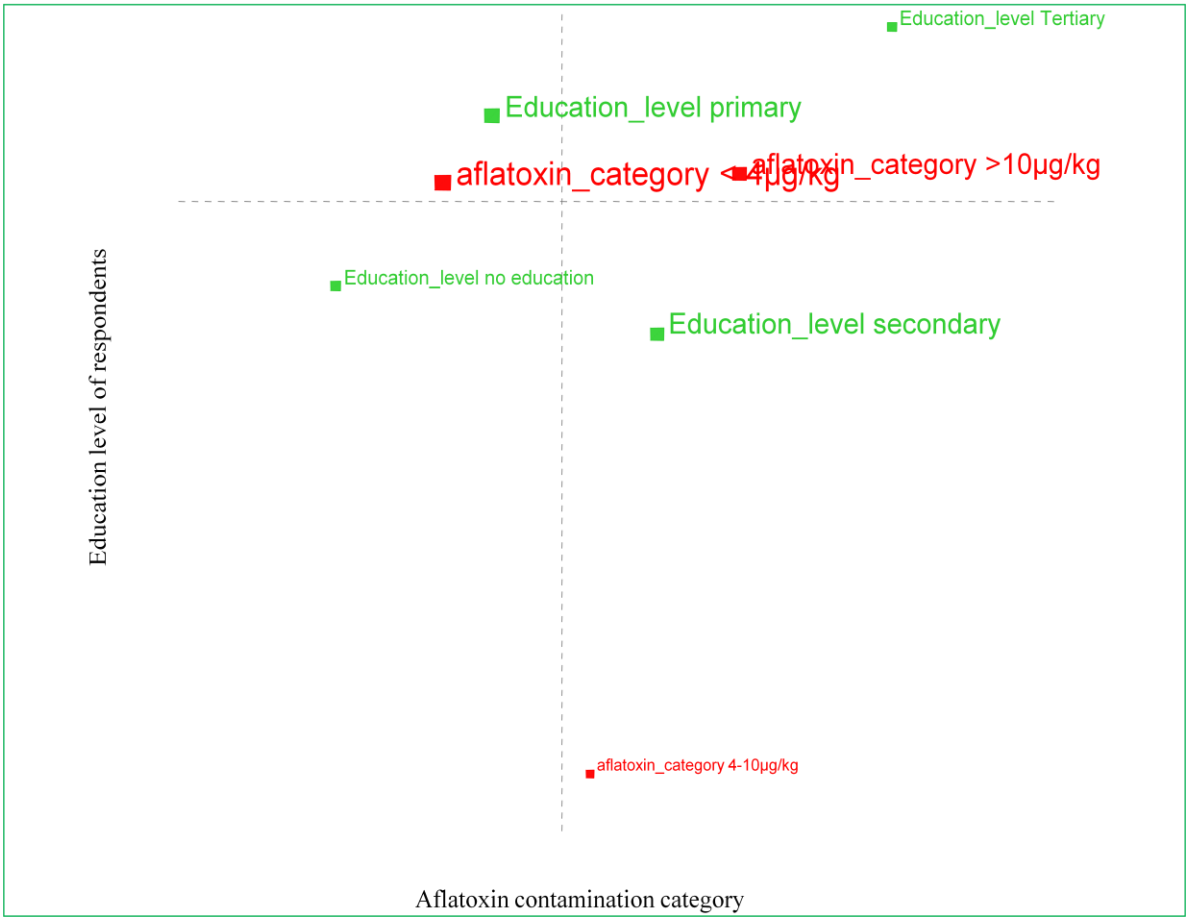

Figure 4.9: Multiple correspondence analysis plot on the association between education level of vendors and aflatoxin contamination category

Unlike contingency tables, Multiple correspondence analysis could allow the study of multiple associations characterizing more than 2 categorical variables. For instance when 4 categorical variables namely aflatoxin contamination category, province where peanuts were sampled, non-use of peanut protection and transaction mode were investigated together, similar results were obtained in the MCA biplot (Fig 4.10) as those previously

obtained (Table 4.1). Western and Nyanza provinces were associated with aflatoxin contamination category >10 µg/kg while Nairobi was associated with aflatoxin contamination category ≤4µg/kg. Use of peanut protection methods was associated with aflatoxin contamination category ≤4µg/kg while non-use of those methods was associated with aflatoxin contamination category >10 µg/kg. Purchasing peanuts directly from peanut farmers was associated with aflatoxin contamination category ≤4µg/kg while transacting the peanuts through middlemen was associated with aflatoxin contamination category ≤4µg/kg.

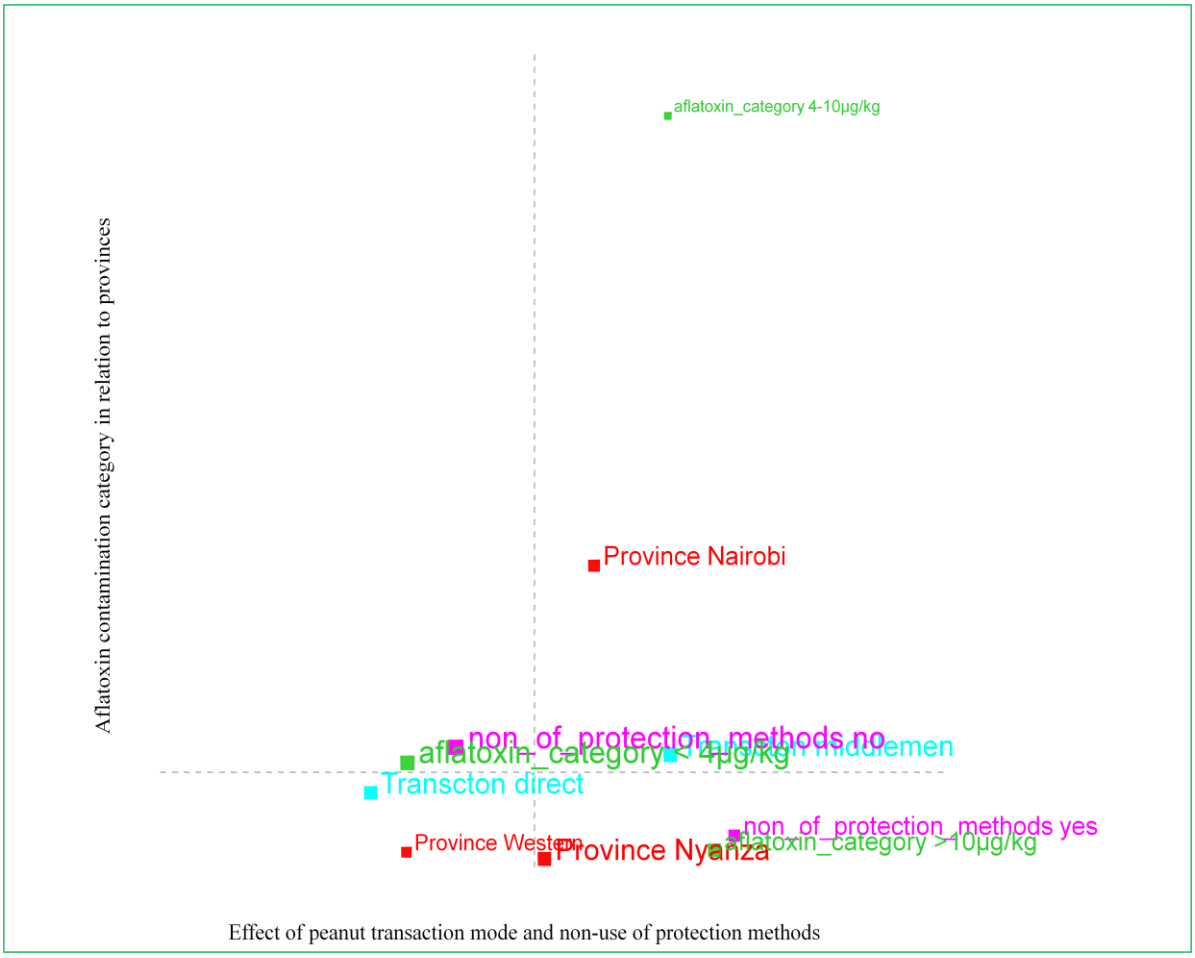

Figure 4.10: Multiple correspondence analysis plot on effect of peanut transaction mode, non-use of peanut protection methods in the 3 provinces of Kenya and aflatoxin category

Since contingency tables could be used to establish an association between any 2 categorical variables only, multiple correspondence analysis was more appropriate in the current study in establishing multiple associations since the data was large and categorical in nature. It could help simplify the analyses by establishing patterns in the data before confirmatory test are be done. In interpretation of the identified significant variables from other statistical methods like multiple regression and ANOVA, MCA is an excellent heuristic for getting into complex multi-factorial data than contingency tables.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

The factors that played a significant role in aflatoxin contamination of the peanuts in Kenya included the province where peanuts were sampled from with Western and Nyanza provinces having the highest contamination and Nairobi the lowest level of aflatoxin. The education level of peanut vendors influenced aflatoxin contamination by through vendors preferring some packaging materials that were more prone to aflatoxin contamination than others. The type of peanut variety influenced the level of aflatoxin contamination with the varieties Red mixed, Red small, Uganda red and Tatu tatu being susceptible while Brown medium, Homabay local and Brown kubwa being resistant to aflatoxin contamination. The application of peanut protection methods significantly reduced aflatoxin contamination of peanuts and when the peanut housing structure was used part time as opposed to full time, it led to less aflatoxin contamination of the peanuts due to reduced sources for contamination.

Materials used for the peanut housing structure influenced aflatoxin contamination whereby the use of bricks and blocks as wall materials, cemented floors as well as leaking of the roof enhanced aflatoxin contamination of peanuts. The season when peanuts harvested influenced the level of aflatoxin contamination whereby peanuts harvested in the month of October were more likely to be contaminated than the other months of the year. The type of peanut product also determined the level of aflatoxin contamination with spoilt peanuts having the highest level of aflatoxin contamination. The storage of peanuts for a period of 6

months before selling them as well as purchasing of peanuts through middlemen significantly enhanced the chances for aflatoxin contamination.

Principal component analysis was applicable in interpretation of aflatoxin contamination of peanuts because some of the variables were correlated and measuring the same construct. It was therefore applicable in the reduction of the large data set into a lower dimension of few but significant variables that could then be subjected to other types of statistical analysis where applicable. This could simplify the statistical analysis of the data by focusing on important variables as the most contributing factors in aflatoxin contamination of the peanuts. Principal component analysis could also be used in constructing data composites for multiple correspondence analysis.

Multiple correspondence analysis was applicable in the interpretation of aflatoxin contamination of peanuts by establishing associations for more than two categorical variables in a low-Euclidean dimensional space. It could produce a simplified representation of multiple associations characterizing attributes as well as to removing noise and redundancies in the data. It could help simplify the analyses by establishing patterns in the data before confirmatory test are be done.

In interpretation of the identified significant variables from other statistical methods like multiple regression and ANOVA, Multiple correspondence analysis was an excellent heuristic for getting into complex multi-factorial data than contingency tables. It was more appropriate in the current study in establishing multiple associations since the data was large and categorical in nature and therefore could make data analysis and interpretation easier and quicker than contingency table analysis.

There was need for further studies on those variables that were identified as having played a significant role in the aflatoxin contamination of the peanuts especially those to do with the peanut storage and housing conditions, the nature of association between education level of peanut vendors and peanut packaging materials and aflatoxin contamination level in order to qualify the findings.

Further studies were needed to identify the type of multiple correspondence analyses that could be most applicable in terms of accuracy and interpretation in the study of aflatoxin contamination in peanuts since there were at least three types: Indicator matrix MCA, Burt matrix MCA and Joint MCA.

REFERENCES

Abbas**,** H.K., Wilkinson, J.R., Zablotowicz, R.M., Accinellic, C., Abel, C.A., Bruns, H.A.,Weaver, M.A. (2009). Ecology of Aspergillus flavus, regulation of aflatoxin production and management strategies to reduce aflatoxin contamination of corn. *Informa Healthcare* 28, 142-153.

Abeyasekera, S. (2003). Multivariate methods for index construction. *The University of Reading*, 1-21

Alonso, V.A., Gonzalez, P.M.L., Armando, M.R., Dogi, C.A., Dalcero, A.M., Rosa, C.A.R., Chiacchiera, S.M. and Cavaglieri, L.R. (2011). Silage contribution to aflatoxin B1 contamination of dairy cattle feed. In: Pacheco, I.T.(Ed). *Aflatoxins-Detection, Measurement and Control*. New York: InTech.

Akturk, D., Sema, G and Tanner, K. (2007). Multiple correspondence analysis technique used in analyzing the categorical data in social sciences. *Journal of Applied Sciences* 7, 585-588.

Anon, (n.d**)**.About multivariate methods [pdf] Available at http://www.oxfordjournals. org/ our_journals /tropej/online/ma_chap1.pdf [Accessed 4th March 2012]

Anon, (1996).Principal component manual [pdf] Available at http://www.itl.nist. gov/div898 /software/dataplot/refman2/ch4/prin_com.pdf [Accessed 1st May 2012]

Awuah, R.T., Fialor, S.C., Binns, A.D., Kagochi, J. and Jolly, C.M. (2006). Factors influencing market participants decision to sort groundnuts along the marketing chain in Ghana. *Peanut Science* 36, 68–76.

Ayanbimpe, G.M., Danjuma, W.S., and Okolo, M.O.(2012). Relationship Between Fungal Contamination of Indoor Air and Health Problems of Some Residents in Jos. In: Kumar, S. and Kumar, R.(Eds). *Air Quality–Monitoring and Modeling* (PP.1-19). Rijeka: InTech

Barrett, J.R.(2005). Liver cancer and aflatoxin: New information from the Kenyan outbreak. *Environmental Health Perspectives* 113, A837.

Bower, K.M. (2003).When to use Fisher's Exact Test. *Six Sigma Forum Magazine* 2, 35-37

Choudhury, A. (2009). Multiple regression analysis. *Experiment Resources.*[online]Available at http://www.experiment-resources.com/multiple-regression-analysis.html [Accessed 26 August 2012]

CDC-Center for Disease Control and Prevention. (2004). Outbreak of aflatoxin poisoning-eastern and central provinces, Kenya. *MMWR Morb. Mortal. Weekly Rep.* 53, 790-792.

CDC-Center for Disease Control and Prevention. Available at http://www.cdc.gov/nceh/hsb/chemicals/aflatoxin.htm [Accessed 20th July 2012].

Clark, G.A.(1976).More on contingency table analysis, decision making criteria and the use of log linear models, *American Antiquity* 41, 3

Cohen, J., Cohen, P., West, S.G. and Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd Ed). Mahwah, NJ: Lawrence Erlbaum Associates

Cornell University (2008).Aflatoxins: Occurrence and health risks. Available at http://www.ansci.cornell.edu/plants/toxicagents/aflatoxin/aflatoxin.html [Accessed 16th July 2012]

Coulibaly, O., Hell, K., Bandyopadhyay, R., Hounkponou, S. and Leslie, J.F.(2008). Economic impact of aflatoxin contamination in Sub-Sarahan Africa. In: Leslie, J.F., Bandyopadhyay, R. and Visconti, A.(Eds). *Mycotoxins: Detection Methods, Management, Public Health and Agricultural Trade* (pp. 67-75). Trowbridge: Cromwell press

Cramer, R.D. (1993). Partial least squares (PLS): its strengths and limitations. *Perspectives in drug discovery and design* 1, 269-278.

Dallal, G. E. (2000). Contingency tables. *The little handbook of statistical practice*.

Debrah, S.K. and Waliyar, F. (1996). Groundnut Production and Utilization in Africa; Past Trends; Projections and Opportunities for Increased Production. Paper Delivered at the 5th Regional Groundnut Workshop for West Africa. No. 18-21, 1996. Accra, Ghana.

D'Enza, A.I and Greenace, M. (2012). Multiple correspondence analysis for the quantification and visualization of large categorical data sets. In: Ciaccio, A.D, Coli, M. and Ibanez, J.M.A. (Eds), *Advanced Statistical Methods for the Analysis of Large Data-Sets* (p.453).Berlin: Springer

Diener, U.L., Cole, R.J., Sanders, T.H., Payne, G.A., Lee, L.S. and Klich, M.A.(1987). Epidemiology of Aflatoxin formation by Aspergillus Flavus. *Annual Review of Phytopathology 25*, 249-270.

EC-European Commission (2006). The commission decision 2006 / 504 / EC. *Official Journal of the European Union* L199, 21 – 32.

Everitt, B.S and Graham, D. (2001). *Applied multivariate data analysis* (2nd Ed). London: Arnold

Fernandez, G.C. (2002). Discriminant analysis, a powerful classification technique in data mining. In: SAS Institute Inc., *Proceedings of the SAS Users International Conference SUGI 27.*(pp. 247-256). Orlando USA, 14–17 April, Florida: Cary, NC

Fieller, N.(2010). Multivariate data analysis [pdf] Available atwww.nickfieller. staff. shef. ac. uk/sheff-only/mvaslideschs0-1.pdf [Accessed 5th March 2012]

Fu, Y.X and Arnold, J. (1992). A table of exact sample size for use with Fisher's exact test for 2×2 tables, *Biometrics* 48, 1103-1112.

Glynn, D. (2012).Correspondence analysis, identifying for patterns of correlation. In: Glynn. D. and Robinson.(Eds). *Polysemy and synonymy: Corpus methods and applications in cognitive linguistics*. Amsterdam: John Benjamins

Gong, Y.Y., Cardwell, K.F., Hounsa, A., Egal, S., Turner, P.C., Hall, A.J.and Wild, C.P. (2002). Dietary aflatoxin exposure and impaired growth in young children from Benin and Togo: cross sectional study. *British Medical Journal* 325, 20-21.

Gong,Y.Y., Egal, S., Hounsa, A., Turner, P.C., Hall, A.J., Cardwell, K.F. and Wild, C.P. (2003). Determinants of aflatoxin exposure in young children from Benin and Togo, west Africa: The critical role of weaning. *International Journal of Epidemiology* 32, 556-562.

Greenacre, M.(1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Greenacre, M. (2006). From Simple to Multiple Correspondence Analysis. In: M. Greenacre & J.Blasius (Eds), *Multiple Correspondence Analysis and Related Methods* (pp.41-76). London: Chapman & Hall

Greenacre, M. (2007). *Correspondence Analysis in Practice.* London: Chapman & Hall.

Greenacre, M. (2010). *Biplots in Practice*. Bilbao: Fundación BBVA.

Greenacre, M. and Blasius, J.(eds)., (2006). *Multiple Correspondence Analysis and Related Methods*. London: Chapman & Hall.

Greenacre, M.and Nenadić, O.(2010). Ca: Simple, Multiple and Joint Correspondence Analysis, R package version 0.33. Available at http://CRAN.Rproject.org/package=ca[Accessed 2nd November 2012]

Groenen, P.J.F., Commandeur, J.J.F and Meulman, J.J.(1998). Distance analysis of large data sets of categorical variables using object weights. *British Journal of Mathematical and Statistical Psychology* 51, 217-232.

Groopman, J.D. and Wild, K**.** (1996). Temporal patterns of aflatoxin-albumin adducts in hepatitis B surface antigen-positive and antigen-negative residents of Daxin, Qidong County, People's Republic of China. *Cancer Epidemiology Biomarkers and Prevention* 5, 253-261.

Guo, B., Yu, J., Holbrook, C.C., Cleveland, T.E., Nierman, W.C. and Scully, B.T. (2009). Strategies in Prevention of Preharvest Aflatoxin Contamination in Peanuts: Aflatoxin Biosynthesis, Genetics and Genomics. *Peanut Science* 36, 11-20.

Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. Cary, NC: Sas Institute Inc.

Hell, K., Fandohan, P., Bandyopadhyay, R., Cardwell, K., Kiewnick, S., Sikora, R. and Cotty, P. (2008). Pre- and post-harvest management of aflatoxin in maize (pp.413–422). In: Leslie, J.F., Bandyopadhyay, R. and Visconti, A.(Eds). *Mycotoxins: Detection Methods, Management, Public Health and Agricultural Trade*. Wallingford: CABI Publishing.

Hendrickse, R.G. (1984). The influence of aflatoxins on child health in the tropics with particular reference to Kwashiorkor. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 78, 427 – 435.

Horn, B.W.(2003).Ecology and population biology of aflatoxigenic fungi in soil. *Informa Healthcare* 22, 351-379.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *The Journal of educational psychology*, 498-520.

Howell, D.C (n.d).Chi-square test- analysis of contingency tables. *Women 35*, 28-83.

ICRISAT, (2000). Properties of aflatoxin and it producing fungi**,** aspergillus and aflatoxin in groundnut [online] Available at: http://www.icrisat.org/aflatoxin/aflatoxin.asp [Accessed 9th Feb 2012]

International Agency for Research on Cancer-IARC.(2002). Summaries and evaluations 82, 171

Jason, W. O and Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation* 8, 1-9.

Jay, J.M., Loessner, M.J. and Golden, D.A. (2005). *Morden food microbiology*(7th Ed), New York: Spring science business media.

KEBS-Kenya Bureau of Standards (2007). *Kenya Standard KS 694-1:2007. Shelled Groundnut (Arachis hypogaea Linn.) – Specification Part 1: Raw Groundnut for Table Use.* Kenya Bureau of Standards Documentation Centre, Nairobi: KEBS

Kessler, W. (2007). *Multivariate Datenanalyse: fr die Pharma-, Bio-und Prozessanalytik.* Weinheim: Wiley-VCH.

Key, J.P (1997). Module s7 - chi square, research design in occupational education. *Oklahoma State University*

Khatoon**,** S., Nafeesa, Q.H., Iffat, T., Nighat, S., Kishwar, S.and Najma, A. (2012). Natural occurence of aflatoxins, zearalenone and trichothecenes in maize grown in Pakistan. *Pakistan Journal of Botany* 44, 231-236.

Klich, M.A. (2007). Aspergillus flavus: The major producer of aflatoxin. *Molecular Plant Pathology* 8, 713-722.

Lewis, L., Onsongo, M., Njapau, H., Schurz-Rogers, H., Luber, G., Kieszak, S., Nyamongo, J., Backer, L., Dahiye, A.M., Misore, A., DeCock, K. and Rubin, C.(2005). Aflatoxin contamination of commercial maize products during an outbreak of acute aflatoxicosis in eastern and central Kenya. *Environ. Health Perspect* 113, 1763-1767.

Liang, X.Q., Luo**,** M. and Guo**,** B.Z. **(**2006**)**. Resistance mechanisms to aspergillus flavus infection and aflatoxin contamination in peanut (*Arachis hypogaea*). *Plant Pathology Journal* 5, 115-124.

Liu, Y. and Wu, F. (2010). Global burden of aflatoxin-induced hepatocellular carcinoma: A risk assessment. *Environmental Health Perspectives* 118, 818-824.

Lowry, R.(1999). *Concepts and applications of inferential statistics*. Poughkeepsie. NY: Vassaar College

Manly**,** B.F.J.(1986). *Multivariate statistical methods*. London: Chapman& Hall

Maxwell, S.M., Apeagyei, F., De Vries, H. R., Mwanmut, D. D. and Hendrickse, R. G. (1998). Aflatoxins in breast milk, neonatal cord blood and sera of pregnant women. *Toxin Reviews 8*, 19-29.

Mead**,** R., Robert, N. and Anne, M.H.(2003). *Statistical methods in agriculture and experimental biology (*3rd Ed).Washington: Chapman&Hall/CRC

Mejia, D.and Lewis, B.(2002). Groundnut: post-harvest operations [pdf] Available at http://www.fao.org/fileadmin/user_upload/inpho/docs/post_harvest_compendium_-_groundnut.pdf [Accessed 17th Feb 2012]

Meng, X.L., Rosenthal, R. and Rubin, D.B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin* 111, 172.

Michael, R.S (2001). Crosstabuslation & chi square. *Indiana University*.

Morrison, D.F(1990). *Multivariate statistical methods* (3rd Ed). New York: McGraw-Hill Publishing Company

Mutegi, C. K., Ngugi, H. K., Hendriks, S. L. and Jones, R. B. (2009). Prevalence and factors associated with aflatoxin contamination of peanuts from Western Kenya. *International Journal of Food Microbiology* 130, 27-34.

Mutegi, C.K., Ngugi, H.K., Hendriks, S.L. and Jones, R.B. (2012). Factors associated with the incidence of Aspergillus section Flavi and aflatoxin contamination of peanuts in the Busia and Homa bay districts of western Kenya. *Plant Pathology* 61, 1143-1153.

Mutegi, C.K., Wagacha, M., Kimani, J., Otieno, G., Wanyama, R., Hell, K. and Christie, M.E. (2013). Incidence of aflatoxin in peanuts (Arachis hypogaea Linnaeus) from markets in

Western, Nyanza and Nairobi Provinces of Kenya and related market traits. *Journal of Stored Products Research* 52, 118 – 127.

Namuth-Covert, D. M., Merk, H. L. and Haines, C. (2012). Chi-Square Test for Goodness of Fit in a Plant Breeding Example. *Journal of Natural Resources & Life Sciences Education* 41, 22-22.

National Institute of Environmental Health Sciences-NIEHS.(2007). Aflatoxin & liver cancer. Available at http://www.niehs.nih.gov/health/impacts/aflatoxin/index.cfm [Accessed 16th July 2012]

Nautiyal, P. C. (2003). Groundnut: post-harvest operations. *Chapter XXI*.

N'dede, C.B., Jolly, C.M., Vodouhe, S.D. and Jolly, P.E.(2012). Economic Risks of Aflatoxin Contamination in Marketing of Peanut in Benin. *Economics Research International 2012*, 12

Nigam, S. N., Waliyar, F., Aruna , R., Reddy, S. V., Kumar, P. L., Craufurd, P. Q., Diallo, A. T., Ntare, B. R. and Upadhyaya , H. D. (*2009*). Breeding peanut for resistance to aflatoxin contamination at ICRISAT. *Peanut Science* 36, 42-49.

Obuchowski, N.A. (2005).Multivariate statistical methods. *American Journal of Roentgenology* 185, 299-309.

Okoko, N., Kidula, N., Wasilwa, L., Makini, F., Murithi, F. and Graham, K. (2009). Participatory evaluation and dissemination of improved groundnut varieties and technologies for processing and utilization [pdf].Available at .kari.org/biennial conference/conference12 /docs/participatoryevaluationanddisseminationofimprovedgroundnut.pdf [accessed 17th feb 2012]

Otsuki, T., Wilson, J.S and Sewadeh, M. (2001). Saving two in a billion: Quantifying the trade effect of European food safety standards on African exports. *Food Policy* 26, 495-514.

Pearson, K. (1901). Principal components analysis. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science 6*, 559.

Rachier, G.O., Orondo, K.0. and Nyakundi, B.N.S. (2010). Identification and on-farm evaluation of groundnut lines tolerant to Rosette virus and Leaf spot diseases in west Kenya [pdf] Available at http://publications/10thProceedings /Volone/IdentficnOnfarmEvaln.pdf [Accessed 17th Feb 2012]

Reddy, T.Y., Reddy, V.R and Anbumozhi, V. (2003). Physiological response of groundnut (*Arachis hypogeal* L) to drought stress and its amelioration: a critical review. *Plant Growth Regulation* 41, 75-88.

Rensburg, V.S.J., Cook, P., Van, S. D.J., Van,W. J.J. and Vincent, T.J.(1985). Hepatocellular carcinoma and dietary aflatoxin in Mozambique and Transkei, *Br.J.Cancer* 51, 713-726.

Routledge, R. (2005). Fisher's exact test. In: Armitage. P. and Colton.T. (Eds). *Encyclopedia of biostatistics* 2 (pp.1519-23). New York: John Wiley.

Rustom, I.Y.S.(1997). Aflatoxin in food and feed: occurrence, legislation and inactivation by physical methods.*Food Chemistry* 59, 57-67.

Saad, N. (2004). Aflatoxins: Occurrence and health risks. *Retrieved November 19*, 2004.

Schriever, B.F. (1986). Multiple correspondence analysis and ordered latent structure models. *Kwantitatieve Methoden* 21, 117-131.

Shepard, J.(2008). Fisher's exact test [PowerPoint] Available at http://www.slideserve.com/jenna/fisher-s-exact-test [Accessed 2nd August 2012]

Shiker, M.A.K. (2012). Multivariate statistical analysis. *British Journal of Science* 6, 55-56.

Shimeles, A. and Thoenen, R. (2005). Poverty profiles: A methodological note on measuring poverty. *Poverty and Social Policy Team, Economic and Social Policy Division, United Nations Economic Commission for Africa, ESPD/NRP/04/05*.

Shlens, J. (2005). A Tutorial on Principal component analysis. *Systems Neurobiology Laboratory, University of California at San Diego*.

Smith, L.I (2002). A tutorial on principal component analysis. *Cornell University USA* 51, 52.

Smith, w.and Solow, A.R.(1996). An exact McNemar test for paired binary Markov chains. *Biometrics* 52, 1063-1070.

Stolzenberg, R.M. (2004). Multiple regression analysis. In: Hardy.M. and Bryman.A.(Eds). *Handbook of data analysis* (pp.165–207). London: Sage Publications.

Sweets, L.E and Wrather, J.A. (2009). Aflatoxin in corn,*University of Missouri*, [online] Available at http://aes.missouri.edu/delta/croppest/aflacorn.stm [Accessed 3rd October 2012]

Takane,Y and Heungsun, H. (2006). Regularized multiple correspondence analysis. In: M.J.Greenace and J.Blasius (Eds), *Multiple correspondence analysis and related methods* (pp.259-279). London: Chapman and Hall

Varga, J., Frisvad, J.C. and Samson, R.A. (2009). A reappraisal of fungi producing aflatoxins.*World Mycotoxin Journal* 2, 263-277.

Wagacha, J.M and Muthomi, J.W. (2008). Mycotoxin problem in Africa: Current status, implications to food safety and wealth and possible management strategies. *International Journal of Food Microbiology* 124, 1-12.

Wang**,** T**.,** Zhang**,** E**.,** Chen**,** X**.,** Li**,** L**.** and Liang**,** X.(2010). Identification of seed proteins associated with resistance to pre-harvested aflatoxin contamination in peanut (*Arachis hypogaea L*). *BMC Plant Biology* 10, 267.

Widstrom, N.W.(1979). The role of insects and other plant pests in aflatoxin contamination of corn, cotton and peanuts-a review. *Journal of Environmental Quality* 8, 5-11.

Wild, C.P and Gong,Y.Y.(2010). Mycotoxins and human disease: A largely ignored global health issue. *Carcinogenesis* 31, 71-82

Wild, C.P and Turner, P.C. (2002).The toxicology of aflatoxins as a basis for public health decisions. *Mutagenesis* 17, 471-481.

Williams, J. H., Phillips, T.D., Jolly, P. E., Stiles, J. K., Jolly, C. M and Aggarwal, D. (2004). Human aflatoxicosis in developing countries: A review of toxicology, exposure, potential health consequences and interventions. *American Society for Clinical Nutrition* 80, 1106-1122.

Wu, F. and Khlangwiset, P. (2010). Health economic impacts and cost-effectiveness of aflatoxin reduction strategies in Africa: Case studies in biocontrol and postharvest interventions. *Food Additives and Contaminants* 27, 496-509.

Yu, J. (2012). Current understanding on aflatoxin biosynthesis and future perspective in reducing aflatoxin contamination. *Toxins 4*, 1024-1057.

Yu, J., Deepak, B. and Kenneth, C. E. (2002). Aflatoxin biosynthesis. *Revista Iberoamericana de Micología* 19, 191-200.

Yu, J., Chang, P.K., Ehrlich, K.C., Cary, J.W., Bhatnager, D., Cleverland, T.E., Payne, G.A., Linz, J.E.,Woloshuk, C.P. and Benneth, J.W.(2004).Clustered pathway genes in aflatoxin biosynthesis, *Applied and Environmental Microbiology* 70, 1253-1262.

APPENDIXES

Appendix 1: Wald tests for dropping terms from regression analysis

Term	Wald statistic	d.f.	F statistic	F pr.
Province peanut sampled	0.32	2	0.16	0.853
District peanut sampled	3.05	8	0.38	0.931
Age of the respondent	3.56	6	0.59	0.735
Crevices in housing structure	0.01	2	0.01	0.994
Education level of respondent	10.54	3	3.51	**0.015**
Enough light in house	0.54	1	0.54	0.462
Floor cracked	1.8	1	1.8	0.18
Insects present on the floor	2.22	1	2.22	0.136
Floor type	5.01	2	2.51	0.082
Floor clean or not	1.35	1	1.35	0.246
Gender of respondent	0.05	1	0.05	0.819
Month for harvest	47.64	12	3.97	**<0.001**
Year when peanuts harvested	0.04	1	0.04	0.851
House used fulltime/part time	4.06	1	4.06	**0.044**
Moisture	1.15	1	1.15	0.285
Mode of transport for peanuts	1.5	6	0.25	0.959
Musty smell in house	0.49	1	0.49	0.483
Package type for peanuts	5.9	8	0.74	0.659
Pallets whether used or not	0.09	1	0.09	0.77
Roof leaking	0.21	2	0.11	0.899
Materials used for the roof	1.11	5	0.22	0.953
Peanut sample type	8.59	8	1.07	0.379
Source of the peanut samples	3.92	8	0.49	0.864
Transaction mode	1.77	1	1.77	0.184
Peanut variety	8.01	17	0.47	0.966
Vendor type	3.8	5	0.76	0.579
Material types for wall structure	323.76	9	35.97	**<0.001**
Presence of windows in the house	0.01	1	0.01	0.905
Drying as protection method	3.69	1	3.69	0.055
Non-use of protection methods	1.07	1	1.07	0.3
Pallets used are clean	0.03	2	0.01	0.987
Sieving as protection method	0.26	1	0.26	0.611
Sorting as protection method	0.08	1	0.08	0.784
Length of storage of peanuts before sell	55.23	8	6.9	**<0.001**
Duration taken before storage of peanuts	4.59	2	2.29	0.102
Tumbling as protection method	0.38	1	0.38	0.54

Note: Figures in bold indicate significant variables which can be retained ($p<0.05$)

Appendix 2: Aflatoxin analysis in the peanut samples

Sample preparation

From each vendor interviewed who had whole nuts that were not commercially packaged, a half kilogram sample was collected for aflatoxin analysis. The product was ground and mixed thoroughly in the laboratory using a dry mill kitchen grinder (Kanchan Multipurpose Kitchen Machine, Kanchan International Limited, Mumbai, India). Products that were sold already packaged were purchased in the quantities they were packaged in. In cases where the package was less than 200 g, more than one packet of the produce was bought in order to make a representative sample size for analysis. Grinding of these products was done in a similar manner as those that were not commercially packaged. In cases where peanut butter paste was sampled, grinding was not necessary.

Analysis of peanut samples for aflatoxin content

A 200 g sub-sample was drawn from each sample after thoroughly mixing. The sample powder was then sub- divided into two equal portions. The powder (or peanut paste) was triturated in 70% methanol (v/v 70 ml absolute methanol in 30 ml distilled water) containing 0.5% w/v potassium chloride (KCl) in a blender, until thoroughly mixed. The extract was transferred to a conical flask and shaken for 30 min at 300 rpm. The extract was then filtered through Whatman No. 41 filter paper and diluted 1:10 in phosphate buffered saline containing 500 µl/Tween-20 (PBS-Tween) and analyzed for aflatoxin with an indirect competitive ELISA (Waliyar *et al*., 2005) by preparing an aflatoxin-bovine serum albumin conjugate in carbonate coating buffer at 100 ng/mlconcentration and dispensing 150 µl in each well of the Nunc-Maxisorp ELISA plates (Thermo Fisher Scientific Inc).

Absorbance was measured at 405 nm in an ELISA plate reader (Multiskan Plus, Labsystems Company, Helsinki, Finland).