

Evaluation of Data Capture Methods used in Plant Health Clinics in Kenya

Linda Amondi Ohulo

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

2013

DECLARATION

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

Signature:

Date:

Linda Amondi Ohulo

Declaration by supervisors

This dissertation has been submitted for examination with our approval as supervisors.

Signature:

Date:

Dr. J.M. Kihoro, JKUAT, Kenya.

Signature:

Date:

Dr. Roger Day, CABI Africa.

DEDICATION

I dedicate this dissertation to my fiancée, Akili Lung'anzi for his support, patience and encouragements, my parent and siblings for their moral support and for believing in me.

ACKNOWLEDGEMENT

I wish to acknowledge the following institutions for making this research a reality. CABI Africa for giving me support, time and opportunity to carry out this research as a student intern. Regional Universities Forum for Capacity Building in Agriculture (RUFORUM) for provision of scholarship to fund my studies and Jomo Kenyatta University of Agriculture and Technology for giving me the opportunity to advance my education. First I wish to thank my supervisors; Dr. J.M. Kihoro and Dr. Roger Day for their immeasurable contribution. Dr. Charles Agwanda for his guidance and contribution. My fiancée, Akili Lung'anzi for his support, patience and encouragements. My parent and siblings for their moral support and for believing in me. Lastly, I wish to give thanks to the almighty God for seeing me through my research work.

Finally, I would like to thank the Knowledge bank and Plantwise Programme team (Dr. Rob Reeder, Dr. Claire Beverley, Dr. Marylucy Oronje, Ms. Florence Chege and Mr. Peter Karanja) for introducing and implementing the ABBYY software in plant clinic operations in Kenya.

Table of contents

Declaration	II
Dedication	III
Acknowledgement	IV
List of tables.....	VIII
List of figures	IX
List of abbreviations and acronyms	X
Abstract	XI
CHAPTER ONE	1
1.0.INTRODUCTION	1
1.1. BACKGROUND.....	1
1.2. PROBLEM STATEMENT.....	4
1.3. RESEARCH OBJECTIVES	5
1.3.1. Main Objective.....	5
1.3.2. Specific Objective	5
1.4. RESEARCH HYPOTHESIS	5
1.5. JUSTIFICATION OF THE STUDY	6
1.6. SCOPE AND LIMITATION OF THE STUDY	7
CHAPTER TWO	8
2.0.LITERATURE REVIEW	8
2.1. LOSSES OF MAJOR CROPS TO PESTS AND THE ROLE OF CROP PROTECTION	8
2.2. AGRICULTURAL EXTENSION APPROACH IN PLANT PEST MANAGEMENT.....	9
2.3. PLANT HEALTH CLINICS APPROACH	11
2.3.1. Definition of plant health clinic.....	11
2.3.2. Plant health clinics evolution.....	12
2.3.3. Plant health clinics in Kenya.....	13

2.4.	DATA MANAGEMENT.....	15
2.4.1.	Concept of data management.....	15
2.4.2.	Manual data capture and how to avoid error transfer.....	16
2.4.3.	Manual data cleaning and validation process.....	18
2.4.4.	Machine data capture technologies.....	20
CHAPTER THREE		31
3.0. RESEARCH METHODOLOGY		31
3.1.	DATA COLLECTION.....	31
3.1.1.	Data sources, collection method and sample design.....	31
3.2.	DATA CAPTURE/ENTRY, CLEANING AND VALIDATION	32
3.2.1.	Data entry/capture process.....	32
3.2.2.	Data cleaning and validation procedures for objectives 1 and 2.....	33
3.2.3.	Determining the differences between machine capture and manual capture methods (Objective 3).....	35
3.3.	Data analysis.....	36
3.3.1.	Data analysis for objectives 1 and 2.....	36
3.3.2.	Data analysis for objective 3.....	38
CHAPTER FOUR.....		40
4.0. RESULTS AND DISCUSSION		40
4.1.	NUMBER OF ERRORS MADE BY PLANT DOCTORS THROUGH USE OF THE TWO HAND-FILLED REGISTER FORMS	40
4.2.	DETERMING THE FACTORS AFFECTING THE NUMBER OF ERRORS MADE BY PLANT DOCTORS IN GIVING CROP DIAGNOSTICS TO FARMERS	45
4.3.	DETERMINING THE DIFFERENCES BETWEEN MACHINE CAPTURE AND MANUAL CAPTURE METHODS	54
4.4.	DISCUSSION	57
CHAPTER FIVE		62
5.0. GENERAL DISCUSSION, CONCLUSION AND RECOMMENDATIONS		62

5.1.	GENERAL DISCUSSION.....	62
5.2.	CONCLUSION.....	63
5.3.	RECOMMENDATIONS.....	66
	REFERENCES.....	68
	APPENDICES	74
	APPENDIX 1: PLANT CLINIC REGISTER FORM 1.....	74
	APPENDIX 2: PLANT CLINIC REGISTER FORM 2.....	75
	APPENDIX 3: OVERVIEW OF PROCESSES FOR PLANT CLINICS DATA COLLECTION AND ANALYSIS.....	76
	APPENDIX 4: AUTOMATIC DATA VERIFICATION METHOD	77
	APPENDIX 5: SYMPTOM AND DIAGNOSIS GUIDELINES FOR CROP PEST	78
	APPENDIX 6: VALIDATION FOR DIAGNOSIS	81
	APPENDIX 7: VALIDATION FOR RECOMMENDATION	82

LIST OF TABLES

Table 1: Paired T-test of Form 1 and Form 2 based on correctness mean number in data capture for all crops	41
Table 2: Paired T-test of Form 1 and Form 2 for individual crops.....	42
Table 3: ANOVA table showing plant doctors diagnostics using Form 1 and Form 2	45
Table 4: Paired sample T-test for plant doctors in giving crop diagnostics using two hand-filled register forms.....	47
Table 5: Group Statistics for gender of plant doctor for GAPD and GAPE.....	48
Table 6: Group Statistics for method of choice for GAPD and GAPE.....	50
Table 7: Group statistics for experience in years of plant doctors in data capture for GAPD and GAPE	53
Table 8: ANOVA table showing error rates of data capture methods	56

LIST OF FIGURES

Figure 1: Map showing plant health clinics in Kenya	15
Figure 2: Error bar chart showing the frequency of incorrectly captured field data using the hand-filled register forms.....	40
Figure 3: Aggregated correct mean assessment of crop diagnostics by plant doctors based on form used in capturing variable fields	46
Figure 4: Influence of plant doctors' gender on crops data capture.....	48
Figure 5: Influence of plant doctors' gender on giving crop diagnostics	49
Figure 6: Influence of plant doctors' method choice on crops data capture	51
Figure 7: Influence of plant doctors' method choice on giving crop diagnostics.....	52
Figure 8: Influence of plant doctors' years of experience on crops data capture	53
Figure 9: Influence of plant doctors' years of experience on giving crop diagnostics	54
Figure 10: Percentage mean error by method of data capture and error types	55

LIST OF ABBREVIATIONS AND ACRONYMS

AIDC - Automatic Identification and Data Capture

ASCII - American Standard Code for Information Interchange

CBO – Community Based Organization

CMS – Content Management Systems

DQM – Data Quality Management

ECM – Enterprise Content Management

FAO – Food and Agriculture Organization

GDP – Gross Domestic Product

ICR - Intelligent Character Recognition

ID - Identification

IPM - Integrated Pest Management

KARI – Kenya Agricultural Research Institute

KEPHIS – Kenya Plant Health Inspectorate Services

KFI – Key from Image

MoA – Ministry of Agriculture

OCR - Optical Character Recognition

OMR - Optical Mark Reading

PDA - Personal Digital Assistant

RFID - Radio Frequency Identification

SSA – Sub Saharan Africa

UoN – University of Nairobi

ABSTRACT

Plant clinics, which are context and situation specific compared to the existing extension approach in the Ministry of Agriculture (MoA), can deliver advice to farmers at the point of demand to reduce crop losses to pests. They can also address the existing disadvantages associated with manual data capture methods in the extension service. The study had three objectives: comparing the average number of errors made by plant doctors using the two hand-filled clinic register forms, to determine the factors affecting the number of errors made by plant doctors in giving crop diagnostics and to determine the differences between machine capture (OCR/ICR) and manual capture methods in terms of error rate and error type made. The primary data employed was from register forms filled in through purposive sampled plant doctors who met the set criteria and are based in central and eastern provinces. The data was subjected to a paired t-test, ANOVA and independent t-test through gap computation. The study showed significant differences between plant doctors in their diagnosis of the plant health problems ($p=0.002$) and recommendations to farmers ($p=0.040$). The OCR/ICR method had a higher mean score in giving crop diagnostics (11.25) and for correctly capturing data (21.67) than manual method (3.57 and 12.38). Significant variation ($p<0.001$, $F=39.74$) was observed among the two methods of data capture. The study has shown that OCR/ICR method has increased the speed and reliability of producing plant pest data in an accurate, timely and cost effective manner. The study recommends advanced training for plant doctors to improve diagnosis and recommendation, training on data capture methods to accurately capture data.

CHAPTER ONE

1.0. INTRODUCTION

1.1. Background

Currently the world population is estimated to be over 7 billion by the United States Census Bureau (USCB) with Sub-Saharan Africa (SSA) being the second-most populated continent with around 1 billion people (USCB, 2012). In SSA about 239 out of 925 million people in the world go hungry every day and live below one dollar per day (FAO, 2010). This is due to the fact that governments and international agencies have neglected agriculture as the primary focus of development for very poor farmers (FAO, 2010). Farmers incur about 40% global potential yield loss to crop pests; the situation is worse in developing countries with an estimate of 70% of the yield loss in some cases due to highly favorable climate (Ellis and Boehm, 2008). Unfortunately, the combating process is slow because the extent of the crop loss based on plant health problems has not been fully exhausted. This is due to limited data derived from estimates which has not been systematically collected (Strange and Scott, 2005; Flood, 2010).

In Kenya, extension services division under the Ministry of Agriculture is a major way of disseminating agricultural information and appropriate technology to different farmers (Taylor, 1998). However, due to many challenges experienced in the extension services of Kenya, the outcomes are poor collection, storage, packaging and dissemination of agricultural information and technology (MoA,

2012). This has lowered their value of services to the intended users besides not having updated crop protection data. Currently, the limited national pest lists and data (because most are out of date, incomplete or contain errors) has reduced the efforts of creating new markets and business opportunities for agricultural products (MoA, 2012).

This demanded for sustainable production and crop protection systems (Cooke and Savary, 2006) like plant clinics to be put into place so that the available scarce resources can be exploited sparingly or the resource base enhanced. The Plantwise initiative based on plant clinics, was started in Kenya in the year 2010 to complement extension services by providing farmers with advanced advisory services on plant health management (Flood, 2010). Plant clinics are operated by trained extension agents in locations readily accessible to farmers. They are open to all farmers who have any plant health problem. Once a farmer reports a disease or pest problem, the plant doctor managing the plant clinic session gives diagnostics achieved on the observable symptoms on the plant/specimen brought or described by the farmer. The plant doctor then uses designed plant clinic register form to record the details of the farmer and pest affecting the specified crop.

The introduction of an improved data capture technology like Optical Character Recognition/Intelligent Character Recognition (OCR/ICR) by Plantwise initiative can be used to address the existing disadvantages associated with manual data capture methods. Studies have shown that OCR/ICR have proven to foster accuracy, completeness and reliability of collection, storage and reuse of field data (Kiziltas *et al.*, 2008). This is in addition to improved operational efficiencies, time savings,

reduced manual entry costs, increased data integrity and security (Parascript, 2011). Which has made remarkable improvement in data quality in many disciplines besides agriculture sector data processes in Kenya; by overcoming the limitations associated with manual data collection approaches (Kiziltas *et al.*, 2008).

As a national initiative in Kenya, plant clinic record can be used to generate useful information on farmers' priority problems and changing status of pests of various crops (Negussie *et al.*, 2011), hence facilitate timely and specific deployment of prevention and protection measures. The records can also help in deriving the extent of crop loss to plant health problems based on the farmers influx and reported cases in the plant clinics. Nonetheless, little can be achieved from plant clinics records because the data management process has not been fully developed to ensure good quality and valid data, which can be analyzed to give useful information. These records, if well managed using an effective and efficient data capture process and mode would be useful to different possible end-users like extension agents in MoA, agricultural research institutions like Kenya Agricultural Research Institute (KARI) and Kenya Plant Health Inspectorate Services (KEPHIS), public educational institutions, agrochemical dealers and government policy makers.

The present study therefore attempted to evaluate and examine the use of new methodologies like OCR/ICR and its tools for data capture over manual method. How OCR/ICR technology can contribute to better data quality in plant clinics. This is by first identifying the basic causes of errors in the plant clinic data, then using the information to improve the data capture process so as to prevent the errors from re-occurring. The analysed data will be used to suggest the best data capture system and

mode to be used in the country by plant doctors/extension agents. This will ease data collection process plus help plant doctors get accurate crop pest data based on farmers queries brought to plant clinics.

1.2. Problem statement

The declining effectiveness and inappropriateness in agriculture extension service has been identified as a major cause hampering growth of agriculture sector in Kenya (Muyanga and Jayne, 2006; Davis, 2006a). The experienced challenges existing through manual data capture method have resulted to poor collection, storage, packaging and dissemination of agricultural information and technology (MoA, 2012). This has resulted to crop protection data which has not been updated for a while and would rather be used in giving feasible, effective and safe recommendation measures.

This necessitated the search for complementary, guided and innovative data capture system (OCR/ICR) through use of plant clinics to address the disadvantages associated with manual data capture method. Despite the fact that benefit to solve data system challenges can be derived from use of the plant clinic records. The reward is lost due to errors introduced by poor data capture method and mode, plus the undeveloped data management processes in plant clinics. This study was therefore carried out to evaluate the data capture methods of plant health clinics in Kenya.

1.3. Research Objectives

1.3.1. Main Objective

The overall objective of the study was to evaluate the data capture methods and mode of plant health clinics dataset in Kenya.

1.3.2. Specific Objectives

- 1) To compare the average number of errors made by plant doctors through use of the two hand-filled register forms.
- 2) To determine the factors affecting the number of errors made by plant doctors in giving crop diagnostics.
- 3) To determine the differences between machine capture and manual capture methods in terms of error rate and error type made.

1.4. Research Hypothesis

- 1) There is no association between the average number of errors made by plant doctors in the data fields and form used.
- 2) There is no significance difference between plant doctors in the number of errors made when giving crop diagnostics.
- 3) There is no significant difference between machine capture technology method and manual capture method in the error rate and type when capturing data.

1.5. Justification of the study

The distribution and severity of plant pests is essential in making strategic decisions and this can be achieved through accurate and regularly updated records from plant clinics. It can lead to development of more effective pest management measures, more judicious use of the tools available to farmers today (Cooke and Savary, 2006) and knowing the extent of crop losses due to plant health problems in Kenya. However, this is not achievable due to poor quality data with errors that have been neglected within a database, modeling systems or decision support systems out of poor manual data capture method in plant clinics. Chapman (2005) study show that uncritical use of any dataset without consideration of errors contained within can lead to erroneous results, misleading information, improper environmental decisions and increased costs. He therefore advises that one should keenly check on the quality and validity of every data point which are obtained from field studies. This can only be achieved if proper data capture method and data management process are applied.

Nevertheless, it is widely known that improved data capture technologies like OCR/ICR can help in data capture/collection, verification and validation, storage and reuse of field data accurately, completely and reliably (Kiziltas *et al.*, 2008). The up-to-date systems of ABBYY FlexiCapture10 software with Canon P-150 scanner have all the components interrelated smoothly to identify and resolve existing data problems causing errors in any dataset. With quick and easy navigation to any screen which help data analysts to get any auxiliary data, inflexibilities in the system out of errors contained due to manual data capture are also reduced. Due to the reasons stated earlier and the challenges faced out of manual data capture method in clinics,

this study aimed at evaluating the data capture methods and mode of plant health clinics dataset in Kenya. This is to give results of effective and efficient data capture processes and modes which could increase the speed and reliability of producing crop pest data, in an accurate, timely and cost effective manner based on farmers queries brought to plant clinics.

1.6. Scope and limitation of the study

The study was carried out with the following limitations. The dataset was limited in size due to irregularities in data capture by plant doctors. This is because previously (2010 - 2011) the Plantwise programme had not been officially launched in Kenya as a form of the Ministry of Agriculture activity. Most of the OCR/ICR dataset which compose the first data source collected in 2011 came from Central province. This is where the OCR/ICR data method had been introduced in the year 2011 before scaling to the other provinces, where plant clinics are based. Plant doctors also differed in experience on data capture methods due to differences in exposure in data capturing. The automated OCR/ICR technology with scanners and the manual data method were assumed to be prone to some errors which may compromise data quality of plant clinic data. Finally, OCR/ICR data capture approach has not been researched/tried elsewhere in plant pest data in Kenya. So limitations included lack of similar undertakings with which one can draw lessons and compare results. Therefore this study was undertaken to meet the stated objectives within the limitations mentioned.

CHAPTER TWO

2.0. LITERATURE REVIEW

2.1. Losses of major crops to pests and the role of crop protection

Most developing countries are producing low crop yield due to losses from insect pests, pathogens and weeds. The major production constraints and associated damage is said to vary by region, country and crop (Nwilene *et al.*, 2008). While this is true, according to United Nations (2002) African human population has been projected to increase yearly. This pattern of population increase demands a matching increase in food production.

Crop protection can play a key role in this by reducing crop losses thereby facilitating provision of food and feed (for animals) in sufficient quantity and quality. This would subsequently lead to increased food security and reduced rural poverty. The potential and actual losses of pest groups, despite the present crop protection practices, among crops worldwide have been estimated to range from 25% to over 50% depending on the crop being analyzed (Oerke, 2000). According to the study carried out by Oerke and Dehne (2004) analyzing production data from 1996 to 1998 for eight major crops, the estimated pre-harvest losses due to pests was (42%) of the potential value of the overall output of which insects accounted for 15% while weeds and pathogens accounted for 13% loss each. A further 10% loss was estimated at postharvest. Weeds had the highest loss potential (32%) but can be mechanically or chemically controlled, while animal pests and pathogens were less important with

18% and 15% losses, respectively. Viruses which can cause serious problems in some areas had a worldwide loss averaged 1-3% on crops.

Another study carried out by DeGroote (2002) showed pre-harvest crop losses due to stem borers account for over 20–40% in maize and sorghum production in Africa. Some of the major pests causing serious yield loss in Africa are stem-borers and grain weevils in stored cereals (Gressel *et al.*, 2004). The extent of the crop losses based on plant health problems has not been fully exhausted. There is little available information found apart from an estimate of 30-40% loss annually from “field to fork” which is common (Strange and Scott, 2005; Flood, 2010), estimates that ca.10% of the world’s harvest is lost to disease alone. Studies indicate that the world could be losing about third of the potential harvest (Bentley *et al.*, 2009). Much of the information on losses is derived from the estimates or projections on limited data and is not systematically collected. Scientists are thereby advising that future solutions regarding improved global food security should address crop losses by improving plant health. Oerke and Dehne (2004) recommend intensification of crop production through enhanced productivity per unit area without involving intense pest and disease control.

2.2. Agricultural extension approach in plant pest management

Approximately 75-80% of the world’s farmers are smallscale, they provide more than half of the world’s food supply (Altieri, 2009) and contribute over 90% production for domestic consumption in Latin America (Nagayets, 2005; Altieri, 2009). Majority of the food supply include maize, beans and potatoes. Nevertheless, these farmers have not been the primary focus of agricultural development (FAO,

2010). The smallscale, resource poor farmers are those farmers with limited resources that have difficulty obtaining sufficient inputs, to allow them to adequately use the available technology (Shanner *et al.*, 1982). Moreover they have less than one hectare of land for agricultural production (FAO, 1996) and are located worldwide in Asia, Africa, Latin America and the Caribbean.

The use of an extension service is one approach through which many smallscale farmers can be reached and provided with innovative technology transfer; using knowledge and skills on sustainable agriculture to increase agricultural production. Though research by Taylor (1998) shows that their services are provided to the innovative, progressive, wealthy and educated commercial farmers. A report from FAO (1996) indicated many smallscale farmers are not being reached despite constituting a bigger percentage in farming and experiencing more limitation in production. The importance of agricultural extension in the fight against poverty has been under-scored while still being considered a crucial area that requires immediate intervention. The declining effectiveness and inappropriateness has been identified as a major cause hampering growth to agriculture sector in Kenya (Muyanga and Jayne, 2006; Davis, 2006a). These gaps in extension innovations to smallscale farmers are said to have been contributed by the poor government services in Kenya (Davis, 2006a). They include changing policies, declining public funding, decline in new thinking and approaches in agriculture and rural development. These are apart from other factors like decentralization, privatization, market orientation, pluralism, client participation and demand responsiveness, climate change and other environmental factors (Negussie *et al.*, 2011). With these in consideration, calls for demand driven

agricultural extension has existed for several decades now. Innovative modes of reaching out to farmers through agricultural extension approaches that are context and situation-specific are presumed to have impact in Kenya, as they might better reflect the local information needs of farmers. One such innovation is plant clinics, for more efficiently delivering information for diverse smallholder farmers. Plant clinics aim to provide regular, relevant and practical advisory services on plant health management to curb plant pest problems. With sustainable production and crop protection systems in place through advice from plant clinics, the available scarce resources can be exploited sparingly or the resource base enhanced (Cooke and Savary, 2006).

Plant clinics have introduced use of the relatively new technology of OCR/ICR data capture technology. The ABBYY FlexiCapture10 software with Canon P-150 scanners has been tested for plant clinic data capture to reduce challenges associated with manual data capture method. In future, this technology can be used to improve data management system existing in extension service under MoA.

2.3. Plant Health Clinics Approach

2.3.1. Definition of plant health clinic

A plant health clinic is a form of primary healthcare which gives advisory services/recommendations to farmers. This is through discussions based on the diagnosis from a crop sample carrying the pest. Like human health clinics operate, plant health clinics provide demand-driven advice to farmers, rather than promoting pre-packaged and technology centered solutions. This is a bottom-up approach which

differs from other project-led interventions; plant clinics offer regular service where demand is defined by the queries farmers present and not by extension workers or researchers (Negussie *et al.*, 2011).

2.3.2. Plant health clinics evolution

The plant health clinic approach was first started in Bolivia in 2003 then expanded to Bangladesh, Uganda and Nicaragua as pilot projects (Danielsen and Ferna'ndez, 2008; Bentley *et al.*, 2009; Danielsen and Mutebi, 2010). Clinics have now been piloted in several developing countries like Kenya, Rwanda, Tanzania, Uganda, DRC, Ghana and Sierra Leone as a new way of providing plant health services to farmers. Its development has undergone several stages conforming to the cycle of experiential learning involving experimentation, gaining of experience, reflective observation and conceptualization (Negussie *et al.*, 2011).

Plant clinics initially started as an experimental initiative, with a few operational guidelines so as to reach more smallscale farmers with plant health services. The development procedures for clinic operations were taken through experimentation and iterative learning with demands of local adaptation. Nowadays the scaling up of plant clinics in some countries is due to the success realized during experimentation with the clinics and this has stimulated the development of a stronger plant health care system in different supportive organizations. For instance, with the May 2012 launch of Plantwise initiative in Kenya , a grassroots plant clinic experiment has become a nationwide initiative involving local service providers leading to the creation of a national system offering regular advice to farmers. They include

Ministry of Agriculture, KEPHIS, KARI, Public universities (JKUAT, UoN and Egerton) and Katumani CBO. Although different countries proceed at different pace and ways, in each case experience is gained from trying out the approach, reflecting on the results and designing the next iteration. The nature of local institutions and working cultures in a particular country can affect how the approach develops.

2.3.3. Plant health clinics in Kenya

Since mid-2010, 35 operating plant clinics have been established in Kenya, with the number still growing. Plant clinics were started after the realization of inadequate conventional extension approaches which had failed to reach and address the varying needs of the resource poor farmers. Secondly, most of the smallholder farmers in Kenya are located in rural areas and they experience diverse agro-ecological and socio-economic circumstances, yet they do not access effective, reliable, practical and timely advice. In addition, farmers have inadequate support services on a regular basis including those that can enable them to address threats of crop pests.

In Kenya, plant clinics run regularly in public places like market centers and are operated by trained staffs that are mainly extension workers from MoA. Other local organizations and institutions offering plant clinic services include KARI, KEPHIS, public universities and Katumani CBO. Majority of plant clinics have been operating weekly or fortnightly on fixed days. Few clinic sessions are held during events like farmers' field days, agricultural exhibitions or agricultural shows. Farmers who visit the plant clinics come from the locality of surrounding villages, markets and nearby centres where clinics are based. Farmers come with crop health problems which

normally range in different clinics based on the agro-ecological zones, crop production (land uses), precipitation, altitude, among other factors.

Different potential organizations to run the plant clinics are identified through interest expressed to implement plant clinic, or through approach of relevant and suitable local organizations which are already involved in supporting farmers in different ways. Training and backstopping of plant doctors is then carried out in different phases/modules. These include field diagnosis to identify common pest problems of both biotic and abiotic cause.

Simple approaches based on observations of plant samples, interview with the farmer, photo-sheets and reference materials are used by plant doctors to give suitable recommendation. This is highlighted according to availability of inputs from the farmer. Plant doctors are linked to reliable sources of technical expertise, agrochemical dealers and diagnostic laboratories, which assist them in identifying complex crop pest problems to find the best solutions.

Publicity for plant clinics is important prior to the set up and during the implementation of the clinics. This is done to sensitize farmers about the clinics and the common methods used include posters, banners, radio announcements, meetings, agricultural shows, writing letters to churches and schools, field days, farmers groups, and megaphone announcements in the market.

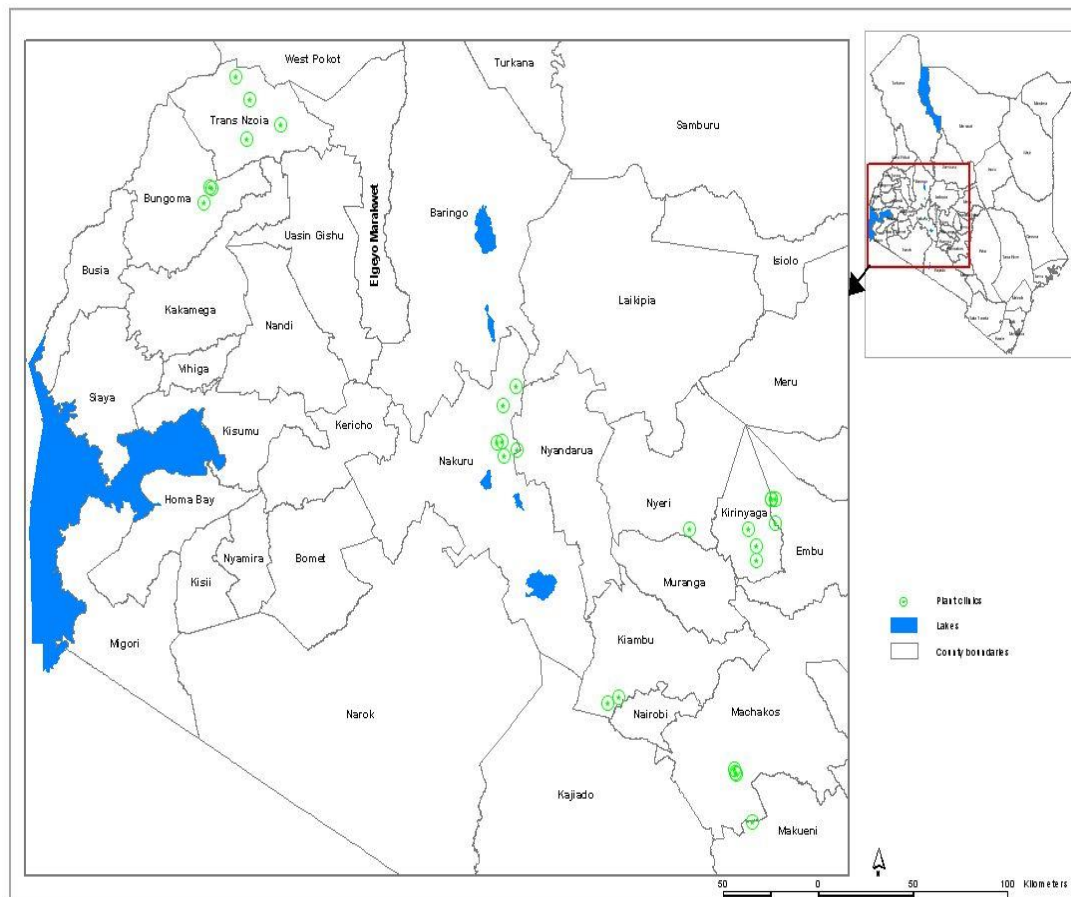


Figure 1: Map showing plant health clinics in Kenya

2.4. Data Management

Quote “*Experience has shown that treating data as a long-term asset and managing it within a coordinated framework produces considerable savings and ongoing value*”. (NLWRA, 2003).

2.4.1. Concept of data management

Data management is concerned with handling and processing data in respect to efficiency, effectiveness, quality and availability for proposed use. It is a key area in any research and if not done well, it can limit the usefulness of the data to fulfill the

purpose of the research being undertaken. The principles of data quality need to be applied at all stages of the data management process like in capture, digitization, storage, analysis, presentation, retrieval and use. This is because factors which affect data quality occur in different ways with the common ones being: poor data handling procedures and processes, inability to follow data entry procedures, errors in the migration process from one system to another and unstructured data that may not look like the required data (Manjunath *et al.*, 2011). The key factors in the improvement of data quality include prevention and correction, where error prevention is closely related to both the collection of the data and the entry of the data into a database, while data validation and correction cannot be ignored and comes afterwards. Dalcin (2004) reported that error prevention is considered to be much better than error detection since detection is often costly and can never guarantee to be 100% successful. Good data management, resulting in high quality data which cuts cost and improves organization processes, is a key principle of data management policy (Arbor, 2009).

2.4.2. Manual data capture and how to avoid error transfer

Data entry is a powerful tool and stage in any organization to manage information, increase the effectiveness of employees and to cut about 60% costs through sharing information. Costs are reduced when multiple people can have access to the same file stored in digital format or use emails instead of stationery (Redman, 1998). Poor data collection and entry can lower information availability which is the key to economic growth. Data entry normally starts immediately after collection/capture in the field because the data is still fresh in the data collector's mind. This may be done through

use of skilled or unskilled data entry operators or through electronic scanning of information (Maletic and Marcus, 2000).

Some initial data quality checks should be done to avoid obvious errors during data entry. Missing values should be carefully treated such as ensuring non-available data appears as blanks in the worksheet and not as zero since zeros are included in statistical calculation. The development of a specific data entry user interface can be incorporated to quickly search authority fields, existing entries in the database and other related databases. Use of search engines such as Google can help increase the accuracy in data entry. It can help an operator decide on the correct spelling or terminology where they may have difficulty reading a label or determining what should and shouldn't go into particular fields (Arbor, 2009). More error levels due to data entry can often be reduced through double-keying, using learning and training software associated with scanning and through using experts or supervisors to carry out testing of entry on a sample basis. Spreadsheet packages can be programmed to perform integrity checks as data is being entered (Arbor, 2009). In addition to spreadsheet packages, a variety of database packages such as Microsoft Access 2010, Microsoft SQL server 2012 and Oracle database 11g have been used in the past for both data entry and documentation. Word processing packages like Microsoft Word 2010 and WordPerfect Office X5 have been used for data entry and maintenance of dataset documentation but are not considered suitable tools for data integrity checks (Arbor, 2009). Apart from the above, there are a number of techniques to aid data entry and avoid transfer errors of which some are discussed later in automated data capture. Data collection and entry process is followed by data organization, checking

and acquisition where the person entering the data can ask some queries concerning the data/data clarification to be answered by the data collectors.

2.4.3. Manual data cleaning and validation process

As stated by Chrisman (1991) and Strong *et. al.* (1997), data quality normally relates to use and therefore cannot be assessed independently of the user. Information quality relates to its ability to satisfy its end-user and to meet their needs (English, 1999). Therefore, for data to be fit for use they must be accessible, accurate, timely, complete, reliable, precise, consistent with other sources, relevant, comprehensive, provide a proper level of detail, be easy to read and easy to interpret (Chapman, 2005).

The general framework for data cleaning (after Maletic and Marcus, 2000) is to define and determine error types, search and identify error instances, correct the errors, document error instances and error types then modify data entry procedures to reduce future errors. The term data cleaning will encompass three sub-processes; data checking and error detection, data validation and error correction. It can also guide the improvement of the error prevention processes.

Data cleaning which is also known as data cleansing or scrubbing improves the quality of data by dealing with detection and removal of errors or omissions and any inconsistencies from data. The process as described by Chapman (2005) may include format checks, completeness checks, reasonableness checks, limit checks, review of the data to identify outliers (geographic, statistical or environmental) or other errors, and assessment of data by subject area experts (e.g. through use of a pathologist).

Since data quality problems are present in any single data collection such as files and databases, example; due to misspellings during data entry, missing information or other invalid data. Multiple data sources need to be integrated and the need for data cleaning will increase significantly. This is because the sources often contain redundant data in different representations (Rahm, 2000).

To obtain clean and reliable data it is imperative to focus on data quality so as to ensure accuracy of the complete database. It is also important during the data cleaning process that data is not inadvertently lost and changes to existing information are carried out very carefully. This can be done by retaining both the old (original data) and the new (corrected data) in the database incase mistakes are made during the cleaning process, the original information can be recovered (Chapman, 2005).

2.4.3.1. Manual data validation

Validation is a process used to determine if data are inaccurate, incomplete or unreasonable. Validation checks may also involve checking for compliance against applicable standards, rules and conventions which are mainly done for consistency accuracy of the data. There are essentially two parts to data validation, data screening and data verification. A key stage in the two processes (cleaning and validation) is to identify the root causes of the errors detected and to focus on preventing those errors from re-occurring (Redman, 2001).

Data validation tools are mainly used when designing the data entry program for a project. It encompasses extensive use of field validations including table lookups,

formulas and data-type restrictions. The various techniques for data checking after entry into a spreadsheet may include: the use of pivot tables (for data consistency), graphs (like scatter plot) and validation checks. Scatter plots are useful tools for spotting outliers. Such errors could be trapped at data-entry if validation rules are set up.

Re-key verification of data is used when the very highest level of accuracy is critical in forms processing, selected fields or entire documents. This can be done through re-keying the data and this applies to recognized text from OCR/ICR/OMR engines. After data has been keyed once (or recognized), it is keyed again by a second independent operator. As each keystroke is made, the system flags any differences. These are immediately verified and corrected.

2.4.4. Machine data capture technologies

Data capture is the activity that combines collection, checking, entry and saving data in some permanent electronic medium. While data capture technologies are methods which automatically identify an object, image or sound, collect and enter the data directly into computer systems without manual data entry (Waldner, 2008). According to Waldner (2008), few categories still exist for data identification and validation at the source tracking and interfacing to the other systems. These Automatic Identification and Data Capture (AIDC) technologies include bar codes, Radio Frequency Identification (RFID), biometrics, magnetic stripes, smart cards and OCR.

2.4.4.1. Optical character recognition (OCR)

OCR is a form of computer system used mechanically or electronically to extract alphanumeric characters from a digital image and recognize. It then converts the scanned and/or handwritten text characters into a machine-encoded text format, such as ASCII. This technology makes digital images searchable for text and thereby is used as a form of data entry from some sort of original paper data source. OCR software works by analyzing a document and comparing it with fonts stored in its database and/or by noting features typical to characters. Some OCR software puts a document through a spell checker to “guess” unrecognized words and the software accuracy is about 100%.

In form processing, information entered into data fields is converted into electronic form either through the entered data being captured in their respective fields or by forms being digitized and saved as images. Form processing is considered complete when the data from all the forms have been captured, verified and saved into a database. Two main approaches to data collection from the paper forms include manual data keying and automated forms processing/input system. One application of data capture is the collection of information from forms and saving it into the databases like CMS, ECM and other systems, for further processing.

There are several types of basic technologies used for data capture according to the data type (Waldner and Baptiste, 2008): OCR – for printed text recognition, ICR – for hand print text recognition, OMR – for marks recognition, OBR – for barcodes recognition and BCR – for business cards recognition. At the same time, documents

for data capture can be divided into 3 groups: Structured documents like questionnaires, tests, insurance forms, tax returns and ballots which have completely the same structure and appearance. They are considered the easiest type for data capture because every data field is located at the same place for all documents. Secondly, semi-structured documents like invoices, purchase orders, waybills and business documents that have the same structure with appearance that is dependent on number of items and other parameters. Capturing data from the semi-structured documents is a complex but solvable task. Finally, unstructured documents like letters, contracts and articles that are flexible with structure and appearance. The different developers who come up with some of these technologies include: Accusoft Pegasus, ABBYY, anyDoc Software, Artsyl, Cvision Technologies, Expervision, IRIS group, Nuance communications and PSIGEN software (Waldner and Baptiste, 2008).

2.4.4.2. Application of OCR in data collection, entry and validation

Although the market for OCR is still not large, many developers are taking advantage of the basic principles. Currently, the largest use of OCR among consumers is from Adobe Acrobat and Microsoft OneNote. This allows for easy character recognition and conversion to digital form (Condrea *et al.*, 2011). But the techniques can be virtually used in any business, especially those in information-intensive industries, including: Banking – loan, mortgage and account applications, Education– student applications, exams and transcripts, Government – tax forms, voter registrations and more, Healthcare– patient records, and others, Hospitality – guest registrations, Insurance – claims forms and Retail – timecards and safety

observation cards. Advancements in field data capture technologies (such as smart tags, laser scanners, and embedded sensors) to enable collection, storing and reuse of field data accurately, completely and reliably has made a remarkable improvement in data quality. The limitation due to current manual data collection approaches in terms of speed, completeness and accuracy, and the implications can now be avoided (Kiziltas *et al.*, 2008). These limitations are besides the non-avoidable ones in manual data entry like delays in data capture, great amount of misprints, high labor costs, equipment spending and rental charges, to name but a few. Most of these costs can be avoidable with the use of data capture technologies.

Some mobile devices have been used in the past in data collection applications. The few which have shown some remarkable promise in the medical branch and mainly used to collect clinic information include: personal digital assistant (PDA), tablet personnel computer (tablet-PC) and digital pen (d-pen). The mobile phone also has potential of running many of the software functions of a PDA and having the advantage of connectivity to the internet via mobile networks.

Digital pens

This is a writing instrument which uses either battery or Bluetooth to digitally capture handwritten note or drawing into typed text. The captured data can be interpreted into characters for database input by software like OCR and used in different applications or as graphics (Shelly and Misty, 2009). The pen uses a pre-printed paper with microdots and camera that read the dots for data to be captured.

Some of the advantages the digital pen possesses are that it is portable, easier to learn and use, it saves time through capturing data electronically. Pen strokes are used to enter data and easily create structured documents which can be sent directly the user's PC or mobile phone. Printed record can be given out for use by client without further printing or scanning. Carbonated pads which are used serves as a back-up in case electronic data is lost or corrupted (Reeder, 2012). Nevertheless pen can easily be stolen due to the small size, has no other functionality like adding text or image and it's impossible to update the captured data. The pen depends on OCR software and it expensive because of other unforeseen costs in terms of license agreement for the paper to be used and pen technology (Reeder, 2012).

Personal digital assistant (PDA)/ personal data assistant

A PDA is a mobile device that is light and functions as personal information manager. It acts as a handheld computer which has a sensitive touch screen (Viken, 2009). PDA has electronic visual display and can run many types of software. This gives it ability to display high resolution images and video. Software in PDAs includes voice recognition and OCR while others can access internet, intranet or extranets by having an internal mobile phone and modem/Wi-Fi which connect other computers or networks.

Apart from advantages found with d-pen, PDA can upload instant information to a database based on its ability of wireless internet usage. Inbuilt OCR software can recognize and convert hand writing immediately although it will not store original data source. With a virtual keyboard, text or handwriting can be inputted directly by

tapping the on-screen with a finger or stylus (Viken, 2009). Accuracy of data and battery shelf life can be increased through inbuilt validation and logic checks (Reeder, 2012). Although with the small screen, PDA is difficult to use for reading in bright light plus it is not designed for lengthy data. It can be stolen or lost easily.

PDA requires a printer due to lack of information copy or data back-up, in case a device is corrupted before data storage into a database. For efficient use of PDA some training should be carried out and new working practices (Reeder, 2012).

Tablet Personal computer

This is a small and slate type, portable personal computer with a touch screen interface. Examples include Apple's iPad and Microsoft's Surface. The device has a wireless function and allow user to work on the touch screen using a stylus or a fingertip instead of a keyboard or mouse (Rouse, 2010). A tablet has a large screen with a display resolution which helps in entering data easily including long data lists. Data accuracy can be increased through inbuilt validation and logic checks. This is after direct input of data into the device as text characters or handwriting using the stylus. It saves time by capturing data electronically and can easily create structured documents in a non-native language (Rouse, 2010). With increased functionality, tablet can be used to display ancillary information like data lists, images and video. The use of wireless internet enables it to instantly upload information into a database while the inbuilt OCR software help in immediate conversion of information keeping the original source text. It has a larger memory than PDAs and battery shelf life is advance (Reeder, 2012).

Nevertheless, it is expensive than PDA and digital pen, and may be relatively bulky and require IT support in case of machine problems. There is no data back-up and data may be lost before saving into a database, because it cannot produce a direct copy of information. Proficient training is a required for efficient use (Reeder, 2012).

2.4.4.3. Machine readable form and ICR software

A form is a document with blank spaces which are to be filled in with particulars using an ordinary ballpoint pen, before it is executed. A machine readable form is a

type which data can be extracted, or captured automatically by computer (Reeder, 2012). The form has entry/data field consisting of text fields, check boxes and group of check boxes which the user fill to capture information. Text field contain certain number of character spaces which stand apart for each to be completed by writing a single letter or digit. Check boxes consist of fields in various geometric shapes (mainly square) for capturing particular options through use of a mark such as a check, tick or a cross. Group of check boxes are used to capture multiple choices data. The use of check boxes has been widely adopted on all sorts of form, because they make their completion easier and can increase the reliability of data output up to 99.9% (Gopisetty *et al.*, 1996).

The filled form is scanned using a high-speed scanner to extract data, then the data is converted into a PDF document. Inbuilt ICR software rewrite/transcribe the hand written characters (automatically) into ASCII character, by recognizing data fields on the form and placing into a correct field in the database. Data recognition accuracy is improved through use of “drop out” colours and through relationship of fields to fixed labels on the form which recognize skewed data and check-box overlapping characters (Reeder, 2012). Few characters which the programme is uncertain are passed on to a human operator for manual verification before data is saved and exported into a database.

2.4.4.4. An overview of data collection and analysis in plant health clinics in Kenya

A personal observation taken from attendances of plant clinics sessions indicates the following. Once a farmer reports a disease or pest problem, the plant doctor managing the plant clinic session gives diagnostics achieved on the observable symptoms on the plant/specimen brought or described by the farmer. The plant doctor then uses designed plant clinic register form to record the details of the farmer and pest affecting the specified crop. The field data collected include: farmer's name, gender, location, first time visitor and frequency of visits, crop type, crop age and observable symptoms on the crop/specimen. Crop diagnostic; recommendation and comments given by plant doctors are also recorded. Other data fields collected include: dates of plant clinic sessions, name of plant clinic, the plant doctor's name, and the code of the register form used (Appendices 1 and 2). The data from the register form is either entered into Microsoft Excel spreadsheet by data personnel/plant doctor or scanned for optical data capture process in some clinics. Appendix 3 shows an overview of the processes for plant clinics data collection and analysis. Data verification and validation is done as described in the research methodology, through a consensus with the plantwise team members.

The following process is used for machine data capture using automated form through ABBYY FlexiCapture10 software with Canon P-150 scanner. First, a batch of completed forms is scanned using a high-speed scanner (usually scanners that scan at least 10 pages per minute are used). Secondly, data are recognized automatically and a few characters about which the program is uncertain are passed on to a data

personnel operator who verifies and corrects the data manually with the image loaded in the software. Finally the group verified data are saved into Knowledge bank database and the data can be exported in Microsoft excel spreadsheet (XLS) or as comma delimited (CSV) format for further analysis. Data from these formats can be easily imported into SPSS and other statistical software packages for further analysis. Normally, the initial or “raw” recognition results in step two are validated using edit procedures such as grammatical rules, spell-checkers, dictionaries, checksum routines and look-up tables. ABBYY software interprets machine-print (OCR), isolated handprint (ICR) including alpha and numeric, checkboxes and checkmarks (OMR) and barcodes from paper forms gathered from a scanner or a fax machine.

2.4.4.5. Sources of errors in plant clinics data flow process in Kenya

The processes involving data collection and entry in the plant clinics are prone to errors. This is through a personal observation taken from attendances of plant clinics sessions and data entry, verification and validation process. In the planning stage of plant clinics data collection, schedules of plant clinic sessions in some areas are done once or twice a month.

The number of plant doctors in a given plant clinic is limited compared to the number of clients seeking advice. The number of farmers coming for advisory services is not taken into account and in case of influx in farmers number, plant doctors become overwhelmed and this interferes with the quality of advisory services provided to farmers. The standard format in the two hand-filled plant clinic register forms also interferes with the manner plant doctors give diagnostics and collect data. While the crop seasonality using agriculture calendar and plant health problems

arising from them, are not well known to all plant doctors. This contributes to less deliberation in giving crop diagnostics to farmers. Some farmers also do not carry crop samples/pest specimen which have plant health problems when seeking advice to the plant clinic. This interferes with the manner crop diagnostics are administered.

According to personal observations taken from data entry, verification and validation process, at times data is captured wrongly during data entry process by different data personnel stationed in the plant clinics. This is by giving own diagnosis and not what is written in the hand-filled register form by plant doctors. In addition there are a lot of misspelling, blanks/missing information (from plant doctors), use of invalid data like local languages to name crops, double entries, other simply identified format errors like undefined key characteristics (example, typographic errors) and not-allowed characters (example, child in gender column). Most of these error sources are besides the aging manual data management system which still exists for plant doctors sources, from the local extension services division under the MoA. This study was therefore done to reduce or eliminate the approaches which cause errors in plant clinics dataset, use the information to improve the data capture process and some problematic areas in the available plant clinic database so as to prevent the errors from re-occurring. This is through evaluating the efficiency of the automated form processing (OCR/ICR data capture) and its applicability compared to manual data capture in use.

CHAPTER THREE

3.0. RESEARCH METHODOLOGY

3.1. Data collection

3.1.1. Data sources, collection method and sample design

This study obtained primary data from two main sources. The first data source consisted of available and unprocessed 2011 dataset from the month of September to December. The second data source was collected during plant clinic sessions in the year 2012 from the month of February to May. Purposive sampling was undertaken to target nine specific plant doctors who had enough experience in carrying out plant clinic sessions and knowhow of both OCR/ICR and manual data capture methods. These criteria were met with the help of plantwise programme team representatives who were more familiar with plant doctors in the study site of eastern and central provinces. A dataset of 1620 completed forms of type 1 (manual form) and 1620 completed forms of type 2 (automated form) were collected from hand-filled plant clinic register (Appendices 1 and 2) through systematic random sampling. This raw data was collected for 10 crops (bean, maize, tomato, potato, kale, cabbage, pigeon pea, coffee, mango and banana) whose plant health problems and crop protection control measures were widely known. All data capture and entry of data was done as described below.

In addition, plant doctors' available Biodata information which was in the CABI records were considered for analysis to give more clarification on the contribution to

errors for the two data capture methods evaluation. The described data capture methods were considered for plant clinics data collection and entry. They were used for research data analysis.

3.2. Data capture/entry, cleaning and validation

3.2.1. Data entry/capture process

The data capture methods are as follows. All of the main methods have been used in one form or another; in some cases multiple methods have been implemented to complement each other. These methods include:

1. Manual entry from hand-filled register Form 1 – Data personnel in different plant clinic type in responses from farmers they see on the physical plant clinic hand-filled register Form 1 into the computer system using the Microsoft excel spreadsheets.
2. Key From Image (KFI)/ Manual data capture method – This is where data from the scanned register Form 2 is entered into a Microsoft excel spreadsheet. This is done by a different operator before verification and validation is carried out.
3. Optical Mark Reading (OMR)/Machine data capture method – Use of special scanners (Canon P-150) where data is automatically extracted from the register Form 2 at the point of scanning by the recognition of marks (such as tick boxes or multiple choice lozenges) in specific locations on the form.
4. Intelligent Character Recognition (ICR) – ABBYY FlexiCapture software is used to recognize handwritten text on each register forms' scanned image from Form2.

The latter two methods of data capture were combined and referred to as OCR/ICR data capture method in this research.

3.2.2. Data cleaning and validation procedures for objectives 1 and 2

To determine the factors affecting the number of errors made plant doctors in giving crop diagnostics (objective 2). The available raw data in the Microsoft Excel spreadsheet from plant clinics were first edited to avoid duplication and to separate them into logical portions for cleaning, validation and analysis process to get the major affected crops. Cleaning of the data for major affected crops considered for this study was done under the general framework as described by Maletic and Marcus (2000).

The assumptions applied before data examination, cleaning and validation included: the presence of a sample at the plant clinic before diagnostics and recommendations are given to a farmer, plant doctors are all equally well qualified from the module training received and are capable of discerning plant pest and diseases using field symptoms and other available resources in the plant clinics, and lastly, all the required materials for further distinction of plant pests are equally available to all plant doctors.

Data screening and verification was done to cross check and confirm the data for its accuracy through general system checks. The two hand-filled register forms data were verified through visual checks if all the fields are available or automatically through computer based portals after Form 2 scan (Appendix 5). This was to give the association in the average number of errors made by plant doctors in data capture

between Form 1 and Form 2 based on the captured fields. This was done in achievement of objective one and two.

Validation checks were done through consensus with the Plantwise team members at CABI Africa based on the module trainings which had been administered to plant doctors' earlier (Appendices 6 and 7). This was to assess if the data collected by plant doctors are sensible; if diagnosis matches the symptoms (problem description) and recommendation measures were effective, practical and safe. These were done for three fields; description of the crop problem field, diagnosis of the problem field and recommendation for the problem field. The criteria used in data validation are as follows: Diagnosis of the plant pests to group levels (rodents, bacteria, fungi, arthropods) were not accepted, common names of the pests and diseases were only accepted where they are specific to identify the problem, lastly, common insect pests, diseases or disorders that have characteristic symptoms were validated despite the incomplete descriptions (See appendix 5 for symptoms and diagnosis guidelines of crop pest). Validation of plant doctors' diagnosis based on the module trainings administered earlier, were done by checking if the symptoms are typical of the pest or has been confused with another problem, if the right plant parts had been examined and the descriptions were precise/specific enough (See appendix 6). The advice recommended by plant doctors were also validated to check if the recommended measures were effective and safe, feasible for the farmer to carry out, if the instructions given were precise and there are no obvious control measures missing (See appendix 7).

Options range like valid (retain record/good recommendation), invalid (reject record/unsafe) and/or redundant (further investigation), were used then coded into binary data (0, 1) for objective one and two to be realized. This was done for the 10 crops (bean, maize, tomato, potato, kale, cabbage, pigeon pea, coffee, mango and banana) whose pest problems and control measures are widely known. Other fields considered and perceived to increase precision when modeling crop losses data in future included: location of the farmer from plant clinic (Counties in Kenya) and frequency of visit to the plant clinic.

3.2.3. Determining the differences between machine capture and manual capture methods (Objective 3)

To determine the differences between machine capture technology method and manual capture method in objective three. Hand filled register Form 2 data had to undergo a two stage filtering process so as to assess the rate of errors associated with Manual data capture (Key From Form) and Machine data capture (OCR/ICR method). 100 Form 2 sampled (in a batch of 10 register forms each) and scanned copies were entered (Key From Image) into Microsoft Excel template, the same copies were run through the ABBYY software data capture system (Machine capture). The content in terms of correctly captured character, check boxes and mismatch characters, from KFI/Manual capture and from the automated technology ABBYY software/Machine capture was compared. This was to determine the rate of errors omitted/made during data capture processes. In addition, comparisons between the automated technology/Machine capture and Key From Image/Manual capture were done to determine the time taken to verify captured characters.

Both automated technology of ABBYY software (Machine capture) and Key From Image (Manual capture) method were assumed to be prone to any one of the following errors:

1. Incorrectly capturing a character on the form. This is where automated technology or KFI miss characters or add some characters not in the scanned Form 2.
2. Capturing a mark on the check box and trying to interpret as stray mark. The automated technology or KFI could have reported a box apart from the one chosen by the plant doctor on Form 2.
3. Incorrectly capturing the content on Form 2 (mismatch). This is where a word from captured characters is interpreted wrongly by the operator or ABBYY software to mean another thing.

Hard match are omitted by picking the wrong check-box on the form. They were determined through placing the content read by both methods against the true response on the register Form 2. The two check-boxes are then compared to see if they match before being considered as a match. Soft match errors are omitted by either adding or dropping characters which are not in the scanned Form 2. Soft match was ascertained either through comparison of character by character or comparison of content recognized by automated technology or operator from KFI against the true response. The result number was then taken for further analysis.

3.3. Data analysis

3.3.1. Data analysis for objectives 1 and 2

After data cleaning and validation processes, the content in different datasets from the two systems were used to extract data for analysis for achievement of objectives 1 and 2. The six fields on the two different forms were classified into separate categories and the frequency of the field based on data capture tabulated. Rate of accuracy and errors made on data capture for specific fields by different plant doctors were identified before analysis.

Data analysis and exploration was done using PSPP, SPSS version 16 and R 2.13.2 statistical packages. First the data file was split by crop to ensure each doctor handled a crop only once so as to achieve independence of the observation. Summation/computation of error/accuracy rate for all the fields handled by the nine plant doctors was then done using PSPP. This was to give the frequency of correct and incorrect capture in all the fields before further analysis could be done. The six fields include crop field, description of the problem field, diagnosis of the problem field, recommendation advice for the problem field, location of the farmer field and frequency of farmers' visit to plant clinic field. The dataset comprising the rate for correct data capture was imported to SPSS version 16. Three extra fields computed are Errors, Diagnostics and Gap. Errors Form 1 and Errors Form 2 field consisted summation of correct rate for all the six variable fields divided by 24 times 100. Diagnostics Form 1 and Form 2 field consisted of three variable fields (description of the problem, diagnosis of the problem and recommendation advice given for the problem) divided by 12 times 100, while GapD for diagnostics and GapE for data accuracy were computed by getting the difference of the two previous fields

simultaneously, that is $GAPD = \text{DIAGNOSTICS}_{\text{Form 2}} - \text{DIAGNOSTICS}_{\text{Form 1}}$ while $GAPE = \text{ERRORS}_{\text{Form 2}} - \text{ERRORS}_{\text{Form 1}}$.

The association between the average number of errors made by plant doctors in data capture fields between Form 1 and Form 2 was analyzed using paired sample T-test on each crop. The significance difference between doctors in the number of errors made when giving crop diagnostics and module trainings undertaken was done through plant description, diagnosis and recommendation fields using paired sample T-test and other descriptive in R software. Further analysis to achieve objectives 1 and 2 were done using the independent sample T-test through gap computation. This was done through incorporation of additional plant doctors' Biodata information extracted from the CABI records. It was to give more clarification on the accuracy of data capture by each plant doctor based on crops. The Biodata included: gender, experience level earned in plant clinic data capture and preferred method of choice for each plant doctor.

3.3.2. Data analysis for objective 3

For realization of objective 3, the following analyses were done to determine the differences between machine capture and manual capture methods in term of error rate and type made when capturing data. Analysis were done to assort if a check box field is a hard match error, assort if a write-in field (mismatch and erroneous character) is a soft match error, calculate the overall hard match and soft match error rates by form and field, then test for statistically significant relationships between

error rates and factors such as form and field category. This was to determine if both methods accurately record the contents of a field.

Operator/Manual capture for content was set at 0-2% error rate. Since the accuracy to capture content of a field by OCR/ICR ABBYYFlexi Capture10 software depended on complexity of the plant clinic form layout and the inbuilt ICR software ability to transcribe the hand written characters (automatically) into ASCII character, through recognizing data fields on the form and placing into a correct field in the database. The assumptions included were if OCR/ICR technology ascertains the content of a write-in field is clear, it processes it with a typical error rate of 0.5 percent. If OCR/ICR technology ascertains the content of a check box field is clear, it processes it with a typical error rate of 0.1 percent (Gopisetty *et al.*, 1996). T-test analysis of the OCR/ICR process was then conducted at 95% confidence level after getting summaries of error rates average over all fields. The procedure done was as described by Conklin (2010).

CHAPTER FOUR

4.0. RESULTS AND DISCUSSION

4.1. Number of errors made by plant doctors through use of the two hand-filled register forms

There were significant differences between Form 1 and Form 2 (Figure 2). These results confirm that there was association between the average number of errors made by plant doctors in data fields and form used.

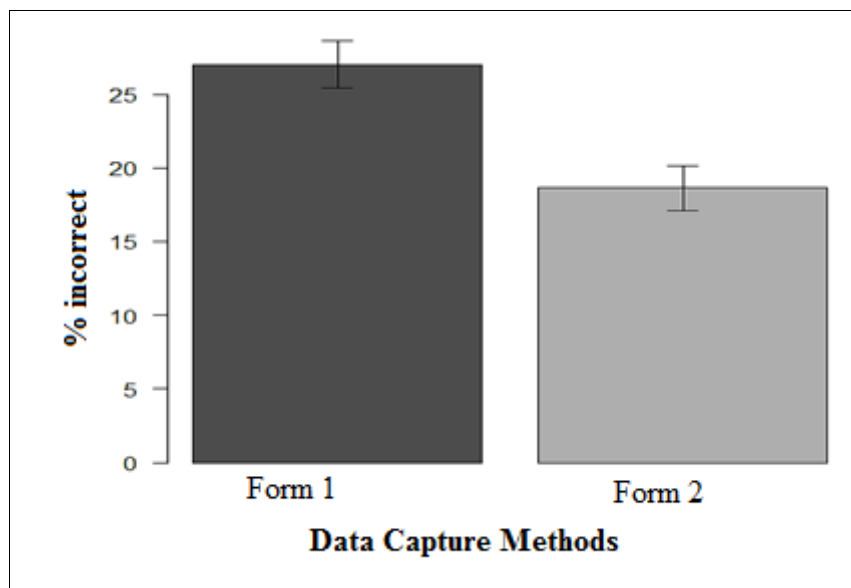


Figure 2: Error bar chart showing the frequency of incorrectly captured field data using the hand-filled register forms

Mean score in number of correctly captured data using Form 2 is significantly higher ($P < 0.05$) than the mean score for correctly capturing data using Form 1, for all the crops (Table 1).

Table 1: Paired T-test of Form 1 and Form 2 based on correctness mean number in data capture for all crops

Crop Name	Form	Sample No.	Correctness mean	Correctness Std.Dev.	Correctness Std. Error Mean	p-value
Banana	Form 1	9	76.39	10.83	3.61	p<0.00
	Form 2	9	91.67	12.33	4.11	1
Bean	Form 1	9	70.37	7.35	2.45	p<0.00
	Form 2	9	93.06	7.22	2.41	1
Cabbage	Form 1	9	74.07	7.73	2.58	p<0.00
	Form 2	9	93.98	6.29	2.10	1
Coffee	Form 1	9	73.15	15.88	2.10	p=0.02
	Form 2	9	89.35	10.23	3.41	6
Kale	Form 1	9	79.63	12.05	4.02	p=0.00
	Form 2	9	93.06	5.89	1.96	7
Maize	Form 1	9	78.24	11.74	3.91	p=0.00
	Form 2	9	91.67	10.21	3.40	2
Mango	Form 1	9	78.70	9.42	3.14	p=0.00
	Form 2	9	91.20	7.92	2.64	9
Pigeon pea	Form 1	9	78.24	12.46	4.15	p=0.03
	Form 2	9	89.35	8.10	2.70	9
Potato	Form 1	9	76.39	11.22	3.74	p=0.01
	Form 2	9	88.43	9.94	3.31	5
Tomato	Form 1	9	81.94	13.34	4.45	p=0.02
	Form 2	9	89.81	9.34	3.11	7

The main contributors in average number of errors made by plant doctors using the two hand-filled register forms were frequency of a farmer to the plant clinic, location of the farmer from where a plant clinic is situated, description of the plant health problem and crop field (Table 2). The frequency of farmers to the plant clinic field arises through the use of Form 1 to capture data. This was observed by lower mean scores and standard deviation for Form 1 when compared to mean score in Form 2.

Table 2: Paired T-test of Form 1 and Form 2 for individual crops

BANANA		Mean	SD	Std. Error	Sig	BEAN		Mean	SD	Std. Error	Sig
FIELDS						FIELDS					
Crop	F1	3	0.866	0.289	p=0.282	Crop	F1	2.78	0.833	0.278	p=0.051
	F2	3.33	0.866	0.289			F2	3.33	0.707	0.236	
Descr	F1	3.11	0.928	0.309	p=0.050	Descr	F1	3.67	0.5	0.167	p=1.000
	F2	3.78	0.441	0.147			F2	3.67	0.5	0.167	
Diag	F1	3.33	1	0.333	p=0.782	Diag	F1	2.44	0.726	0.242	p=0.009
	F2	3.44	1.014	0.338			F2	3.44	0.527	0.176	
REC	F1	3.56	0.527	0.176	p=1.000	REC	F1	3.44	0.726	0.242	p=0.169
	F2	3.56	1.014	0.338			F2	3.89	0.333	0.111	
LOC	F1	3.44	0.527	0.176	p=0.104	LOC	F1	3.22	0.833	0.278	p=0.023
	F2	3.89	0.333	0.111			F2	4	0	0	
FREQ	F1	1.89	1.453	0.484	p=0.002	FREQ	F1	1.33	1	0.333	p<0.001
	F2	4	0	0			F2	4	0	0	

Cabbage		Mean	SD	Std. Error	Sig	Coffee		Mean	SD	Std. Error	Sig
Fields						Fields					
Crop	F1	2.67	1.118	0.373	p=0.010	Crop	F1	2.78	0.972	0.324	p=0.377
	F2	3.89	0.333	0.111			F2	3.22	0.972	0.324	
Descr	F1	3.11	1.054	0.351	p=0.022	Descr	F1	3.67	0.5	0.167	p=0.081
	F2	3.78	0.441	0.147			F2	3	0.707	0.236	
Diag	F1	2.78	0.833	0.278	p=0.276	Diag	F1	2.33	1.658	0.553	p=0.084
	F2	3.33	1	0.333			F2	3.56	0.726	0.242	
REC	F1	3.89	0.333	0.111	p=0.347	REC	F1	2.89	1.054	0.351	p=0.028
	F2	3.67	0.5	0.167			F2	3.89	0.333	0.111	

LOC	F1	3.56	0.726	0.242	p=0.104	LOC	F1	3.56	0.726	0.242	p=0.282
	F2	4	0	0			F2	3.89	0.333	0.111	
FREQ	F1	1.78	1.394	0.465	p=0.002	FREQ	F1	2.33	1	0.333	p=0.001
	F2	3.89	0.333	0.111			F2	3.89	0.333	0.111	

Maize		Mean	SD	Std. Error	Sig	Kale		Mean	SD	Std. Error	Sig
Fields						Fields					
Crop	F1	2.44	1.13	0.377	p=0.040	Crop	F1	3.33	0.866	0.289	p=0.594
	F2	3.44	1.333	0.444			F2	3.44	0.726	0.242	
Descr	F1	3.44	0.882	0.294	p=0.813	Descr	F1	3.33	0.866	0.289	p=0.438
	F2	3.33	0.866	0.289			F2	3.67	0.5	0.167	
Diag	F1	3.11	1.269	0.423	p=0.622	Diag	F1	3.11	0.782	0.261	p=0.195
	F2	3.33	0.707	0.236			F2	3.44	1.014	0.338	
REC	F1	3.56	0.527	0.176	p=0.081	REC	F1	3.67	0.707	0.236	p=0.729
	F2	3.89	0.333	0.111			F2	3.78	0.441	0.147	
LOC	F1	4.00 ^b	0	0	ND	LOC	F1	3.44	0.882	0.294	p=0.095
	F2	4.00 ^b	0	0			F2	4	0	0	
FREQ	F1	2.22	1.202	0.401	p=0.002	FREQ	F1	2.22	0.972	0.324	p=0.001
	F2	4	0	0			F2	4	0	0	
ND-Not determined											

Average number of errors for crop field and description of the problem field arises from use of both Form 1 and Form 2. Example: cabbage and banana description of the plant health problem field's means are 3.11 (Form 1) and 3.78 (Form 2) respectively, with a standard deviation which is approximately the same (1.05). This was contrary to the bean crop field in Form 2 which had a slightly higher mean score (3.33) and lower standard deviation (0.707) compared to its mean score (2.78) and standard deviation (0.833) in Form 1.

The results indicate that other factors may have influenced the differences in the form used for data capture. One of the reasons may be differences in plant doctors when capturing data as opposed to form use, especially, when capturing the crop name variable and giving advisory services. Secondly, sampled registered forms used to extract data for the analysis showed that most plant doctor leave crop field variable either blank/missing data, misspelled, use invalid data – local names or pluralize crop name (tomatoes instead of tomato, mangoes instead of mango, and so forth). Thirdly, discrepancies in plant doctors when describing plant health problems may be due to factors like gender, education level, module training and experience in data capturing. Poor data capture in location of the farmer and frequency of farmer to the plant clinic may be due to the form layout structure. Personal observations from engagements in plant clinics data management processes shows that it is easier for most plant doctors to use the Form 2 register form because of the design elements incorporated. These include check boxes, text fields and identifiers fields which are not availed in the Form 1 register form layout.

There are unforeseen implications resulting from poor data capturing of the two variable fields: location of farmer from plant clinic and frequency of farmer visiting plant clinic. These may include the difficulty to report how many times a problem has emerged from the locality of the plant clinic, difficulty to detect the severity or prevalence of a plant health problem or if the problem reported is the same and is out of negligence to follow recommendations given by plant doctors.

4.2. Determining the factors affecting the number of errors made by plant doctors in giving crop diagnostics

There was significant variation ($p < 0.001$) between plant doctors in giving crop diagnostics using the two forms (Table 3).

Table 3: ANOVA table showing plant doctors diagnostics using Form 1 and Form 2

		Sum of Square	df	Mean Square	F	p-value
List of plant doctors (Form 1)	Between Groups	91.26	7	13.04	2.10	0.05
	Within Groups	508.74	82	6.20		
	Total	600.00	89			
Lists of plant doctors (Form 2)	Between Groups	145.22	5	29.04	5.31	0.00
	Within Groups	454.78	84	5.41		
	Total	600.00	89			

There was a significant variation in problem description in plant doctor 1 (Figure 3). The rest of the plant doctors did not show a significant difference as indicated by the corrected mean score, standard deviation and probability value (3.4 ± 0.8 , $p = 1.00$).

On the contrary, significant differences were observed in plant doctors when giving crop diagnostics. These were observed in diagnosis for plant health problem field (2.9 ± 1 for Form 1, 3.4 ± 0.8 for Form 2, $p=0.002$) and recommendation for plant health problem field (3.6 ± 0.7 for Form 1, 3.8 ± 0.5 for Form 2 $p=0.040$).

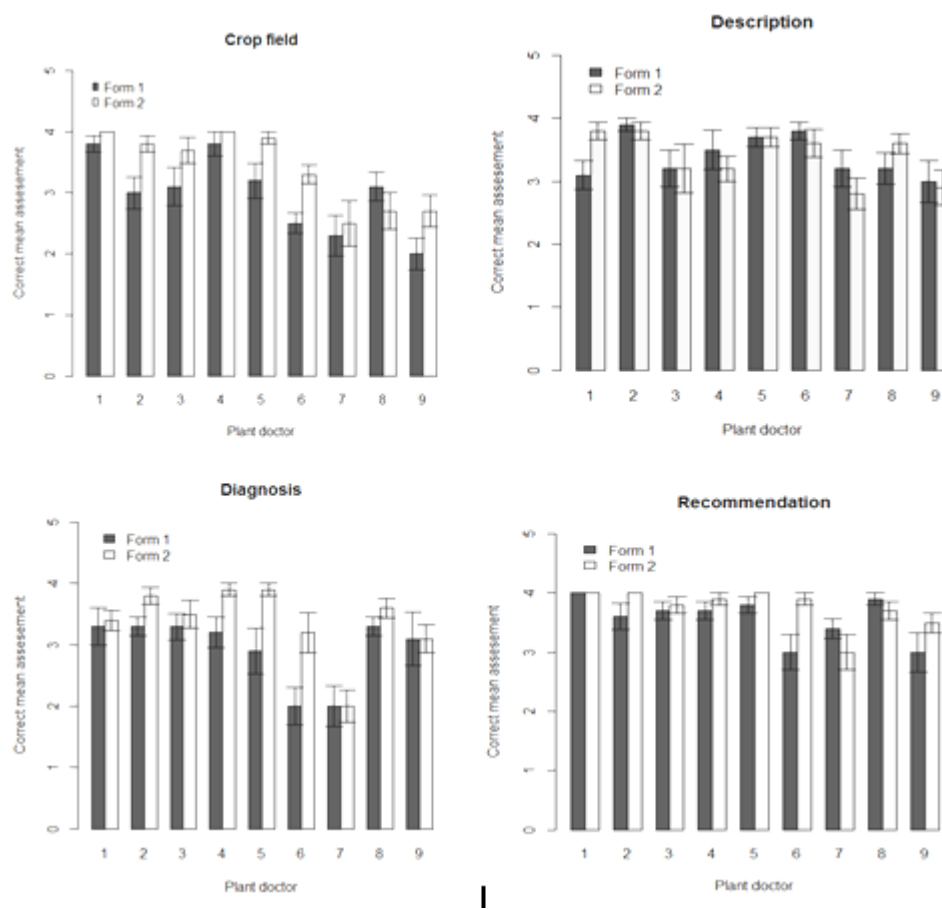


Figure 3: Aggregated correct mean assessment of crop diagnostics by plant doctors based on form used in capturing variable fields

Crop diagnostics assessment result shows a significant difference ($t=2.490$, $p=0.038$) between bean crop in plant doctors using Form 1 and Form 2 (Table 4). The probability values for the other crops are not significant ($p \geq 0.05$). These results are

based on correct mean number of errors made by plant doctors when giving crop diagnostics.

Table 4: Paired sample T-test for plant doctors in giving crop diagnostics using two hand-filled register forms.

Crop name	$\bar{X}$ Form1	$\bar{X}$ Form2	$\bar{X}$ Form 2- $\bar{X}$ Form 1	$\bar{X}$ Form 2- $\bar{X}$ Form 1 Std. Dev	t-value	p-value (Sig.2- tailed)
Banana	83.333	89.815	6.481	13.679	1.421	0.193
Bean	79.630	91.667	12.037	14.500	2.490	0.038
Cabbage	81.482	89.815	8.333	14.434	1.732	0.122
Coffee	74.074	87.037	12.963	21.291	1.827	0.105
Kale	84.259	90.741	13.426	11.176	1.257	0.244
Maize	84.259	87.963	3.704	13.889	0.800	0.447
Mango	81.482	87.963	6.481	17.568	1.107	0.301
Pigeon pea	87.037	83.333	-3.704	12.577	-0.883	0.403
Potato	83.333	83.333	0.000	19.543	0.000	1.000
Tomato	86.111	86.111	0.000	10.206	0.000	1.000

$\bar{X}$ indicate Mean

The difference in number of errors made while giving crop diagnostics and in capturing data for both male ($t = 0.374$, $p = 0.709$) and female ($t = -1.76$, $p = 0.860$) was not significant (Table 5).

Table 5: Group Statistics for gender of plant doctor for GAPD and GAPE

Field	Gender of plant doctor	N	Mean	Std. Deviation	Std. Error Mean
GAPD	Male	50	5.8333	14.98771	2.11958
	Female	40	4.5833	16.66132	2.63439
GAPE	Male	50	14.2500	12.08465	1.70903
	Female	40	14.6875	11.16390	1.76517

GAPD is Gap Diagnostics Computed Data while GAPE is Gap Error Computed Data

There was no significant variation in number of errors made by plant doctors in both genders when capturing data accurately (Figure 4).

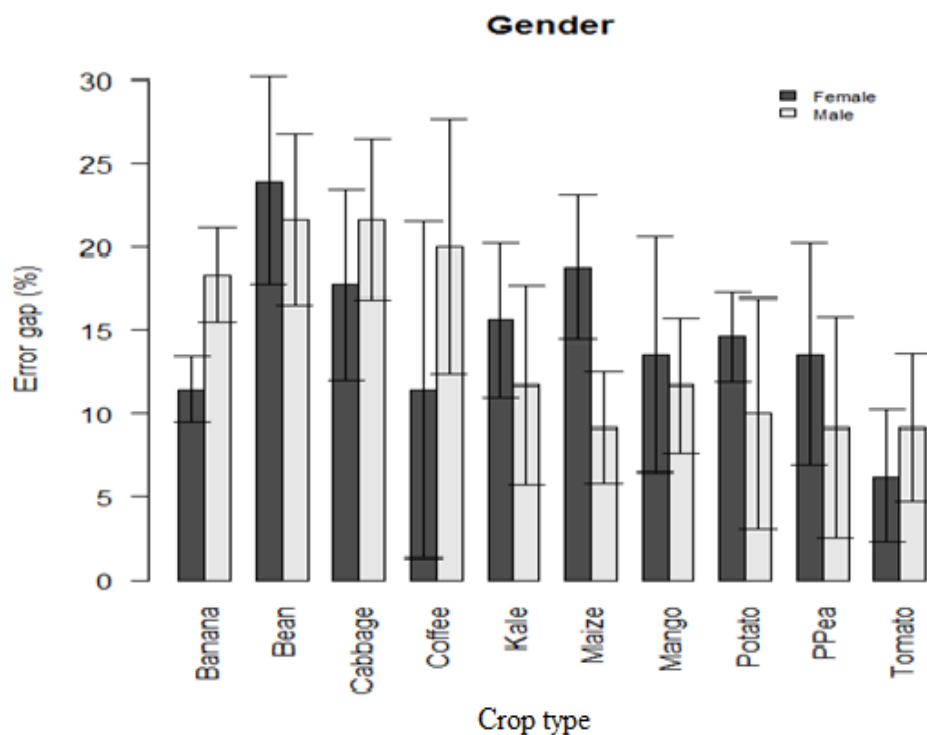


Figure 4: Influence of plant doctors' gender on crops data capture

Variation in number of errors made by plant doctors in both genders when giving crop diagnostics was not significant (Figure 5).

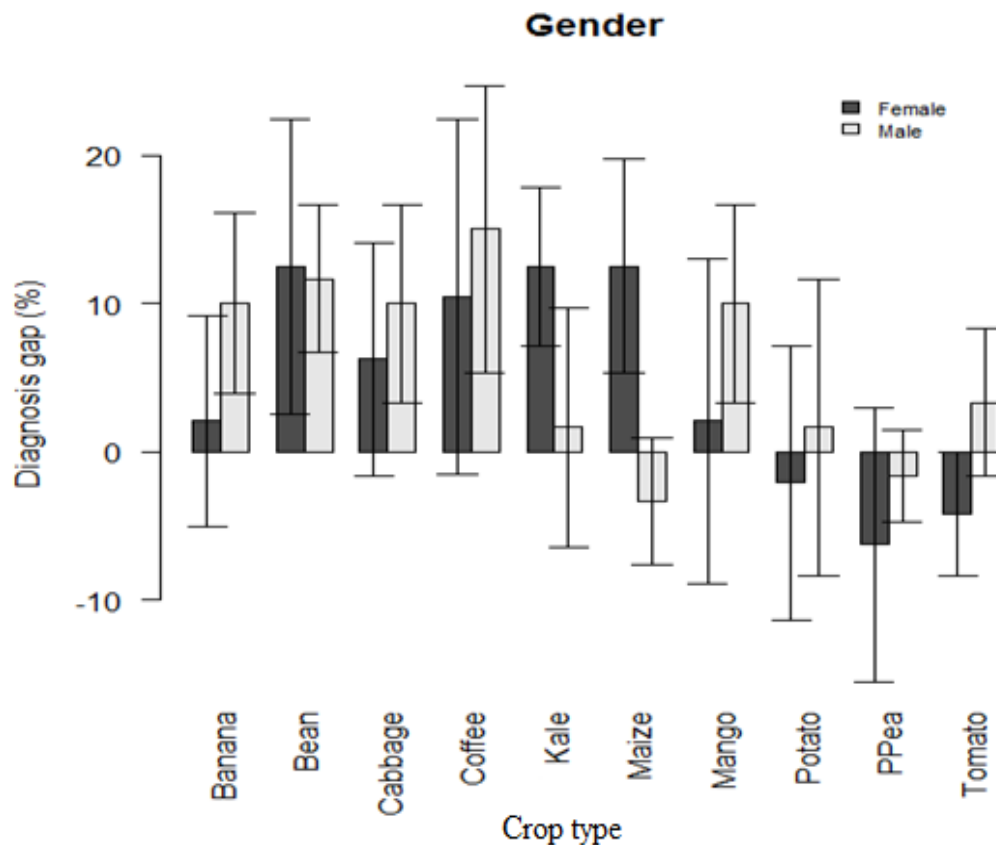


Figure 5: Influence of plant doctors' gender on giving crop diagnostics

The standard deviations for the two groups of gap diagnostics (15.64 and 14.63) and gap error (10.80 and 11.75) in plant doctors' method of choice for data capture was almost equal therefore "equal variances assumed" test was used. There was a significant difference between the mean frequency of captured error for giving diagnostics ($t=-1.963$, $p=0.053$) and the accuracy in data capture ($t=-3.324$, $p=0.001$) (Table 6). OCR/ICR method had a significantly higher mean score when used for giving diagnostics and correctly capturing data (11.25 and 21.67) than Manual method (3.57 and 12.38).

Table 6: Group Statistics for method of choice for GAPD and GAPE

Field	Preferred method	N	Mean	Std. Deviation	Std. Error Mean
GAPD	OCR/ICR	70	3.5714	15.64318	1.86972
	MANUAL	20	11.2500	14.62869	3.27107
GAPE	OCR/ICR	70	12.3810	10.80430	1.29136
	MANUAL	20	21.6667	11.75406	2.62829

GAPD is Gap Diagnostics Computed Data while GAPE is Gap Error Computed Data

There was a significant difference between plant doctors in the frequency of errors made in data capture accuracy (Figure 6) for crops. This implies that method of choice by a plant doctor contributes to the average number of errors made when capturing field data based on crops.

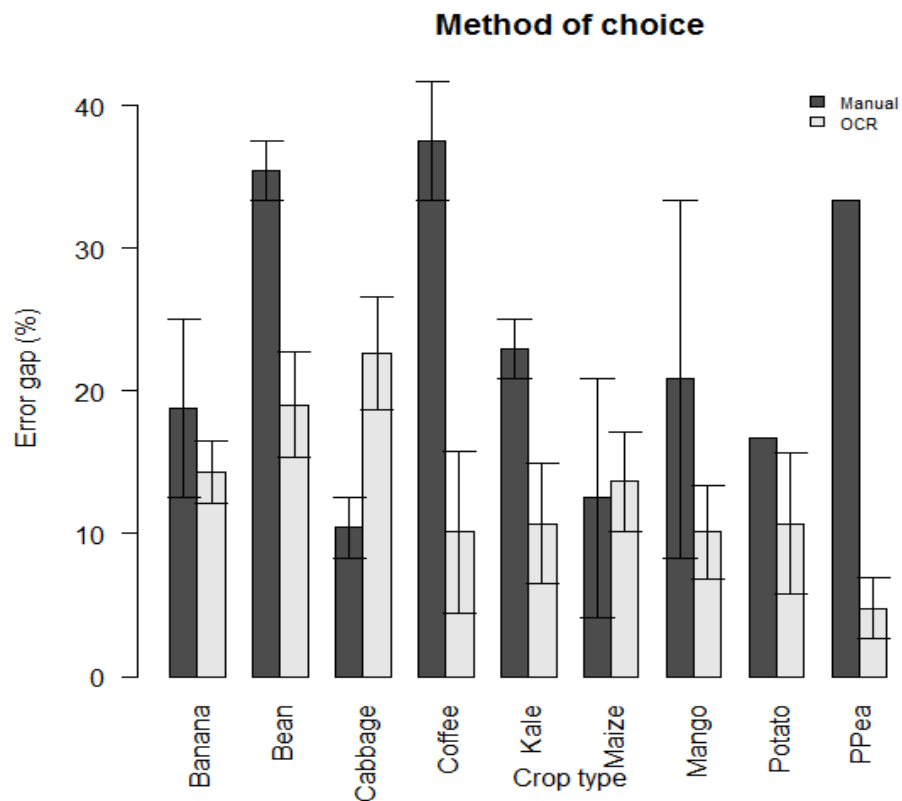


Figure 6: Influence of plant doctors' method choice on crops data capture

Difference between plant doctors in the average number of errors made when giving diagnosis for most of the crops was significant (Figure 7). This indicates that method of choice determine the number of errors made by a plant doctor when giving diagnosis.

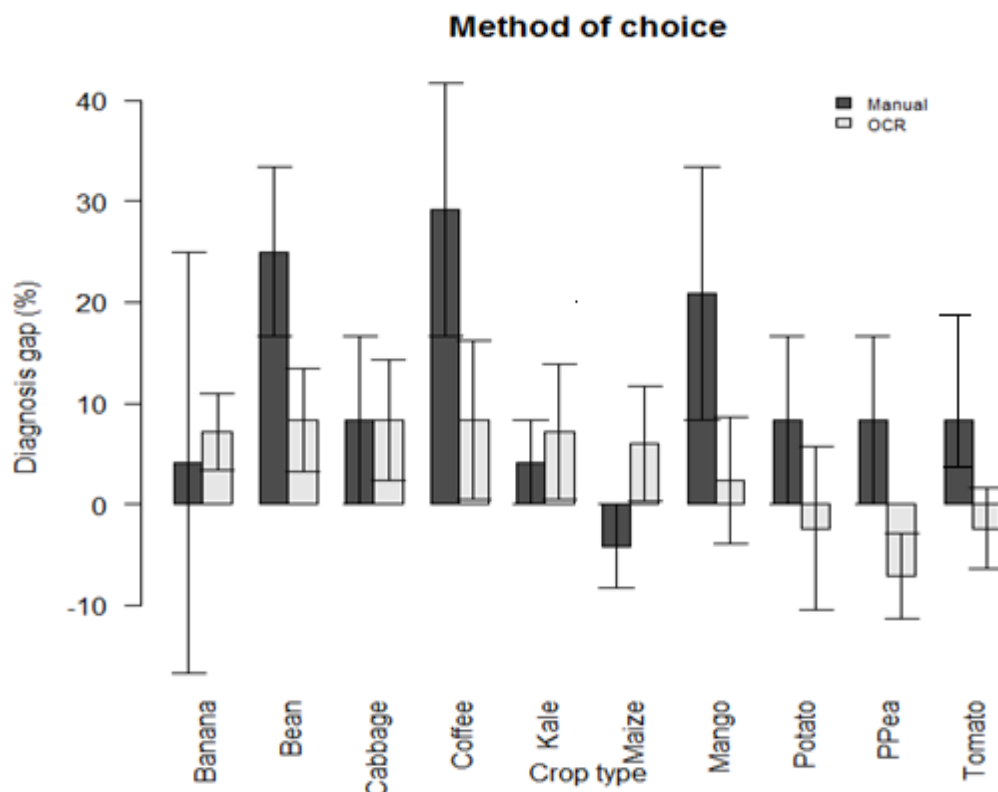


Figure 7: Influence of plant doctors' method choice on giving crop diagnostics

There was no significant difference in plant doctors data capture ($p=0.556$) and for giving crop diagnostics ($p=0.411$) based in the years of experience earned (Table 7). An estimated change based on the difference in standard deviation for use of OCR method (18.09 and 11.14) and manual method (13.49 and 12.06) was present. However, there was insufficient evidence ($t=-0.826$, $p=0.411$ and $t=0.556$, $p=0.580$) to suggest that diagnostics given and accuracy in data capture depended on the number of years of experience earned by the plant doctor.

Table 7: Group statistics for experience in years of plant doctors in data capture for GAPD and GAPE

Field	Experience in years of plant doctor in data capture	N	Mean	Std. Deviation	Std. Error Mean
GAPD	1.5	40	3.7500	18.09578	2.86119
	2	50	6.5000	13.49918	1.90907
GAPE	1.5	40	15.2083	11.14644	1.76241
	2	50	13.8333	12.06558	1.70633

GAPD is Gap Diagnostics Computed Data while GAPE is Gap Error Computed Data

There was no significant difference between plant doctors based on frequency in numbers of errors made when capturing data accurately (Figure 8).

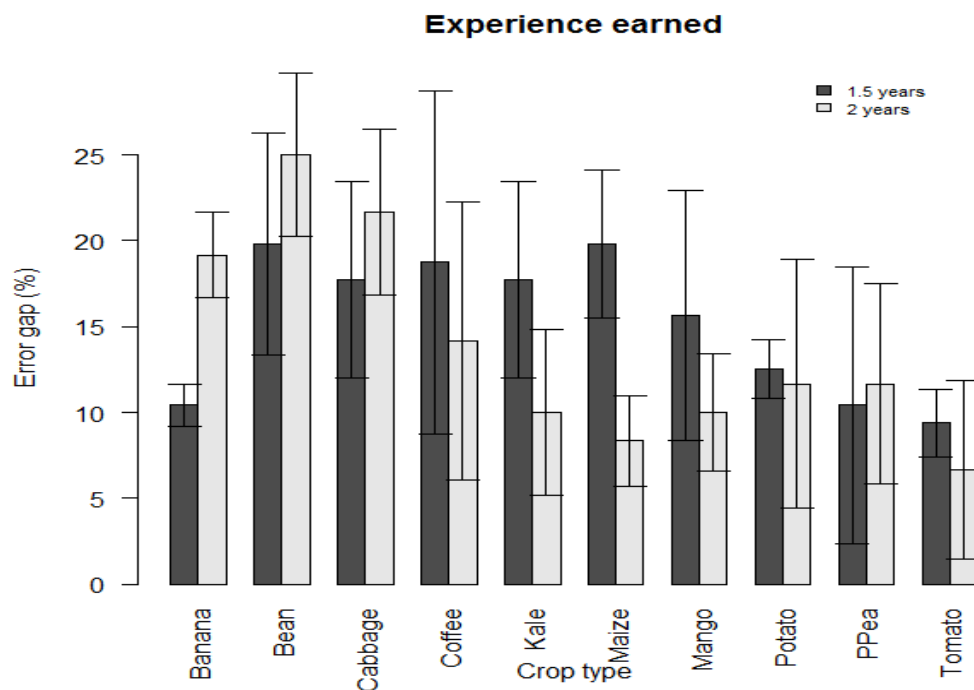


Figure 8: Influence of plant doctors' years of experience on crops data capture

Difference in giving crop diagnostics was not significant (Figure 9). This implies that years of experience earned by plant doctor do not affect the frequency in number of errors made for data capture and in giving crop diagnostics.

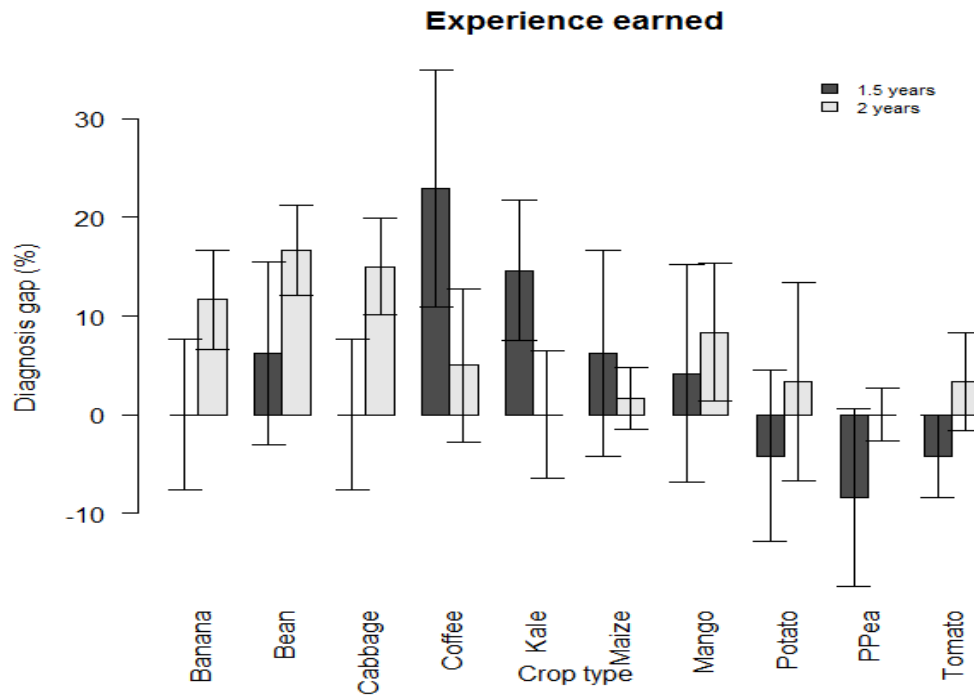


Figure 9: Influence of plant doctors' years of experience on giving crop diagnostics

4.3. Determining the differences between machine capture and manual capture methods

Substitution error rates for manual data capture from image were relatively higher when compared to substitution error rate for machine data capture (Figure 10). There was significant difference among the methods of data capture and error types. The mean difference for data capture and error types is 1.0 for hard match error and 2.59 for soft match error. Among the data capture methods, machine method had an

overall mean difference of 1.11 with 0.37 and 1.84 for hard match and soft match, respectively. Manual method had an overall of 2.48 with 1.63 for hard match errors and 3.33 for soft match errors.

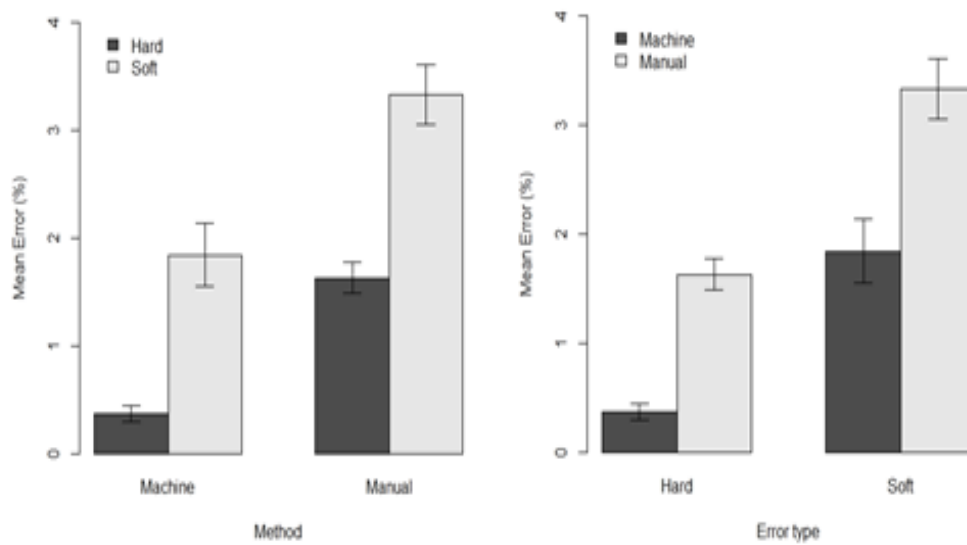


Figure 10: Percentage mean error by method of data capture and error types

Significant variation was observed among methods of data capture ($p < 0.001$, $F=39.74$) and error types/factors ($p < 0.001$, $F=53.01$) (Table 8). Variation among interaction between methods of data capture and error types was not significant ($p=0.601$).

Table 8: ANOVA table showing error rates of data capture methods

Source of variation	D.F.	S.S.	M.S.	V.R.	F pr.
Method of capture	1	9.4311	9.4311	39.74	<0.001
Factor	1	12.5803	12.5803	53.01	<0.001
Method of capture*Factor	1	0.0676	0.0676	0.28	0.601
Residual	16	3.7973	0.2373		
Total	19	25.8762			

4.4. Discussion

Even though the OCR/ICR market is still not large, the basic techniques of automated forms processing can be virtually used in any business that have a significant data processing requirements including: governments, insurance companies, census bureaus and tax offices. This approach has not been tried elsewhere especially in the field of agriculture for plant pests data capture in Kenya. Similar experiments in extension service approach data capture requirements which one can draw lessons and compare results are limited apart from the research carried out in census 2000 evaluation by Conklin (2010) and Canadian census of Agriculture by Mudryk and Xie (2001).

In this current study the association between the number of errors made by plant doctors in the data fields and form used were determined. Analysis done for the two hand-filled plant clinic register forms using paired sample T-test shows that the mean score in the average number of correctly captured data using Form 2 is significantly higher compared to the mean score in the average number for correctly captured data using Form 1, for all the crops ($P < 0.05$). This could be due to ease in use of the automated processing form as opposed to manual data entry forms which exist in plant clinics. This is based on the fact that Form 2 structure is composed of different design elements like entry fields, service fields, and identifiers fields among others which some are not found in Form 1.

Results of paired T-test ($P < 0.05$) of Form 1 and Form 2 for individual crops shows that the main contributors in the average number of errors made by plant doctors

using the two hand-filled register forms are: frequency of a farmer to the plant clinic, location of the farmer from where a plant clinic is situated, description of the plant health problem and crop field. Poor capturing of the variable fields like frequency of a farmer to the plant clinic and location of the farmer from where a plant clinic is stationed, may contribute to unforeseen implications. These may include the difficulty to report how many times the problem have emerged from the locality of the plant clinic, difficulty to detect the severity or prevalence of a plant health problem or if the problems reported are the same and are out of negligence to follow recommendation given by plant doctors.

The study has also shown that the significance differences between plant doctors when giving crop diagnostics to farmers resulted from diagnosis of the plant health problems (2.9 ± 1 for Form 1, mean 3.4 ± 0.8 for Form 2 and $p=0.002$) and recommendation given for plant health problems (3.6 ± 0.7 for Form 1, mean 3.8 ± 0.5 for Form 2 and $p=0.040$) as indicated by correct mean score, standard deviation and probability value. Further analysis done using Paired sample T-test analysis showed that plant doctors make wrong diagnostics on bean crop as opposed to the other crops. This difference in plant doctors when giving crop diagnostics may be due to many reasons. These reasons could include: inadequate module trainings administered to a plant doctor when starting a plant clinic, lack of a plant sample from farmers when seeking advice from plant doctors, inadequate knowhow of the most recent plant health problems due to inaccessibility to the information, insufficient knowhow of the crop seasonality, influx in clients/farmers number during plant clinic days and use of a standard format to give diagnostics.

Nevertheless, description of the plant health problem given by plant doctors using the two hand-filled forms was not significantly different. This is shown by correct mean score, standard deviation and probability value (3.4 ± 0.8 , $p=1.00$). The main reason could arise from independence association which exists in plant doctors when giving description of plant health problem using the two forms.

Statistical difference was done using available Plant doctors Biodata from the CABI records. This was carried out using independent sample T-test analysis based on crop diagnostics given to famers and accuracy on data capture for all crops. Results indicated a significant difference between plant doctors in the manner of giving crop diagnosis and data capturing using the two hand-filled register forms. This was based on preferred method of data capture ($t=-1.963$, $p=0.053$ and $t=-3.324$, $p<0.001$). Significant difference was not observed between plant doctors based on gender of plant doctors ($t=0.374$, $p=0.709$ and $t=-1.76$, $p=0.860$) and the years of experience ($t=-1.963$, $p=0.411$ and $t=0.556$, $p=0.556$).

These findings from objective three contradicts the ABBYY software substitution error rates measured during data verification, which were maintained near the target levels established for the data fields. These were as follows: Hard match error rates (0.37% not relative to the target of 0.1%); Soft match error rates (1.84% not relative to the target of 0.5%). For manual data capture method, substitution error rates were maintained near the target level of 0 – 2%. These could result from lower data quality in terms of completeness and accuracy of captured information. The higher the correspondence between the data exported into the database and the data entered into the fields of the register forms, the higher the quality of data. There are a number

of factors which may contribute to the substitution error rate. Sloppy writing when a plant doctor writes carelessly, makes corrections or merges some letters, which affect the chances of recognition errors. Typos committed when manually entering data and rekeying, and recognition errors/hidden errors committed by a programme when reading content from the field.

The software substitution error rate finding is different from the study done by Mudryk and Xie (2001). This could be because of lower target levels set for ABBYY FlexiCapture10 in this study as compared to the ones specified in the previous study. The other reason may be because of different ICR software used in the two studies. From the study carried out by Mudryk and Xie (2001), the target levels of ICR software depend on the equipment settings of the scanning system and Statistical Process Control (SPC) operation control.

Nevertheless, results from scanned plant clinic register form shows that extracted data using ABBYY FlexiCapture10 software matches the one from manual system but in a faster way. One operator could process from 1000 to 3000 forms in a day. This may depend on the complexity of the form layout, for other operators not using plant clinic form. These indicates captured content is more accurate and reliably correct (higher quality) due to reduced/minimized role of human factor, which is a critical phenomenon for any organization. FormReader technology in ABBYY software used for plant clinic data has been specifically designed, to have validation rules which ensure higher data integrity and reliability therefore reducing errors associated to typos. Throughout this evaluation we present results and comments based on the analysis of data capture error rates.

Previous research observation has shown that there is inherited aging system for data management in the extension service under MoA, which has not been modified for a longer period of time. This system may result from incomplete and errors contained from manual data capture which involves using navigating combination of function and control keys. Furthermore with many challenges, the value of MoA services to the intended users has lowered rapidly. This is besides not having updated crop protection data to help in giving feasible, effective and safe recommendation measures (MoA, 2012). The website is neither interactive nor user-friendly enough to access important information like crop economic analysis, yearly crop production and national pest lists/crop protection information.

Research findings in other discipline apart from agriculture have shown that automated form processing, has a lot of benefits when compared to manual method for data capture (Consulting and Gingrande, 2005). These benefits include improved productivity with minimal human operation since most stages in data capture are automated apart from verification (Gopisetty *et al.*, 1996). This has helped in elimination of manual data drawbacks like time consumption in manual data extraction and entry, elimination of operator's misprints due to the fact that most of the job is done by a computer. Reduction in operation costs through faster processing, rental charges, less spending in equipment and redeployment of staff to other tasks. With automated form processing technology in place, more FormReader modules can be trained for less duration (Gopisetty *et al.*, 1996). The findings comply with this study findings carried out using ABBYY FlexiCapture10 software technology.

CHAPTER FIVE

5.0. GENERAL DISCUSSION, CONCLUSION AND RECOMMENDATIONS

5.1. General discussion

The study was carried out to evaluate the data capture methods of plant clinic in Kenya in order to give results of effective and efficient data capture processes and modes. The objectives were to compare the average number of errors made by plant doctors through use of the two hand-filled register forms, to determine the factors affecting the number of errors made by plant doctors in giving crop diagnostics to farmer and to determine the differences between machine capture (OCR/ICR) and manual capture methods in terms of error rate and error type made.

The extension services division in the Ministry of Agriculture in Kenya is one major means of disseminating agricultural information and appropriate technology to different farmers. However, due to many challenges experienced in the extension services of Kenya, the outcomes are poor collection, storage, packaging and dissemination of agricultural information and technology, among others. This has lowered the value of their services to the intended users besides not having updated crop protection data to help in giving feasible, effective and safe recommendation measures. Their website is neither interactive nor user-friendly enough to access important information like crop economic analysis, yearly crop production and national pest lists/crop protection information. This necessitated the search for complementary, guided and innovative data capture system through use of plant

clinics to address the disadvantages associated with manual data capture methods. This study employed primary data from two main sources collected using register forms through purposive sampling of specific plant doctors who met the set criteria and are based in central and eastern provinces. Additional plant doctors' Biodata information in the CABI records was considered for analysis. The data was subjected to a detailed statistical analysis using paired sample T-test, ANOVA and independent sample T-test through gap computation.

The research showed that significance differences between plant doctors when giving crop diagnostics to farmers resulted from diagnosis of the plant health problems (2.9 ± 1 for Form 1, mean 3.4 ± 0.8 for Form 2 and $p=0.002$) and recommendation given for plant health problems (3.6 ± 0.7 for Form 1, mean 3.8 ± 0.5 for Form 2 and $p=0.040$) as indicated by correct mean score, standard deviation and probability value. OCR/ICR method has a higher mean score when used for giving crop diagnostics and correctly capturing data (11.25 and 21.67) than Manual method (3.57 and 12.38). Significant variation was observed among methods of data capture ($p<0.001$, $F=39.74$) and error types/factors ($p<0.001$, $F=53.01$). Variation among interaction between methods of data capture and error types was not significant ($p=0.601$)

5.2. Conclusion

The findings of objective one showed that automated/OCR/ICR register Form 2 made fewer errors when used to capture data compared to the manual register Form 1. The filling of the automated form which has novel reporting and data visualization techniques, can be used to make farmers understand more on plant health problems

which are affecting them, plus know how their crops are performing. This was because farmers were given a copy of the automated register sheet after visiting plant clinics to refer when they visit agrochemical retail shops.

In objective two, significant differences between plant doctors were found in their diagnosis of the plant health problems, and in the recommendations given to farmers. The records from plant health clinics are useful and are based on the validity of the information collected by plant doctors using the register forms. The study therefore gave various recommendations on how to improve the accuracy of data capture. This was due to the fact that these records can be used to generate useful information on farmers' priority problems and changing status of plant pests, to identify new and emerging crop pests, to help scientists have an idea of farmers' knowledge and perceptions concerning crop pest management and in identification of new research niche' plus documentation of the plant health clinics work. The information derived from plant clinic records would be useful to different possible users such as agricultural extension agents under the Ministry of Agriculture and other parastatals, public universities and agricultural research bodies (KEPHIS, KARI and Coffee Research Foundation), agrochemical dealers and non-governmental institutions like CABI - for further Research developments.

The automated form (OCR/ICR) was better than manual form when used for giving crop diagnostics and in capturing data. The study has thus shown that the technology was very efficient and can be used to make an impact in improving data quality in plant health clinics in Kenya; other fields in the agriculture department that handle intensive data from surveys. This would reduce the limitation associated with manual

data collection approaches in terms of speed, completeness and accuracy, which have been a drawback to researchers. Speed of OCR/ICR software allowed an operator to process about 1000 to 3000 forms in a day and this was beside reduction in misprints plus manual data extraction and entry due to less human involvement. Even though ABBYY software required a license for use, the study showed that the technology was cost effective. The cost incurred for data extraction, entry or verification had been estimated to be three times less compared to manual method, for the same roles in plant health clinic dataset.

There were multitude of reasons which would contribute to relatively slow adoption of OCR/ICR technology in Kenya plant health clinics and extension service under MoA. ABBYY FlexiCapture10 software with Canon P-150 scanner is a hand print recognition technology, the accuracy and productivity of the forms processing is highly dependent on the form design/layout. With many factors to consider when designing a form to collect the data, the form needed to be understood by plant doctors. This was because a problem may arise through plant doctors who had sloppy writing, make a lot of typos or use invalid data. Meanwhile CABI data content group were making progress in testing appropriate form use in the plant clinics; especially to plant clinics which entirely rely on manual method.

The automated forms processing using ABBYYFlexiCapture10 software was not totally an error-free process. A small potential for incorrect data extraction still existed, particularly due to erroneous optical character recognition (hidden errors), typos and sloppy writing. Example; research findings has shown that substitution error rate is higher than the target level of 0.5% and 0.1% for soft match and hard

match error rates, respectively. Therefore, this demanded a simple and effective way for end users to review the processed data beside internal validation checks, to ensure its accuracy and correct any mistakes made. Because of this reason, envision role for paper user interface still existed. Other limiting factors included power outages in rural areas where some plant clinics were stationed, risks associated with theft of scanners, capacity of plant doctors to use the technology given that some were computer illiterate and malfunction of scanners or computer which arouse due to technical problems (repair issues, software failure, among others).

5.3. Recommendations

For the average number of errors made by plant doctors to be reduced, plant doctors were recommended to accurately capture data for all fields and especially, on location of farmers and farmers' frequency of the plant clinics. The information was perceived to be useful to various stakeholders besides agricultural extension agents/plant doctors. This was also done for provision of timely and risk assessments in different areas, so as to localize strategies to control losses based on plant health problems. In addition, plant clinics records could help in knowing if the problem was from specific farmers' negligence or a recurring problem due to insect pest or disease resistance. It may advocate for research of new approaches to be tried out in persistent plant health problems and the records used to predict pests alerts.

Diagnostics for plant health problems and recommendation given to farmers were the main contributors of errors made by plant doctors when giving crop diagnostics despite having different module trainings. For improved and accurate crop diagnostics to be provided by plant doctors, farmers were encouraged to always carry

a sample of affected crop part when visiting a plant clinic session. In addition, frequent and effective modules trainings should be administered to all plant doctors. The study observed that the training should be narrowed down to local and seasonal plant health problems based on yearly crop calendars of Kenya. This study also recommended provision of solar mini-laptops to plant doctors so as to enhance their chances of giving accurate, timely and advanced diagnostics to solve plant health problems. This was through quick access of the websites which had outlines for diagnostics like: Plantwise website, Knowledgebank website and Biovision website.

Finally, the study recommended advanced training to plant doctors on an overview of how well clinics forms was to be filled. For this would reduce the average number of errors made plus reduce the impact incurred in making errors through improving plant clinic planning approaches. This was based on the fact that no method of data capture could be recommended for use in association with a particular error type. The study showed that introduction of tablet computers for use of capture data was appropriate in plant clinics. For it would solve the higher substitution error rate in machine capture and eliminate the paper user interface.

With the following recommendations checked, Plantwise programme would efficiently manage to accomplish their aim of increasing food security through reduction of crop loss, by providing knowledge to farmers. This would be achieved through expansion of the established plant clinic network in CABI and support creation from global knowledge bank of information on plant health.

6.0 REFERENCES

- Altieri, M. (2009). Small farms as a planetary ecological asset: Five key reasons why we should support the revitalisation of small farms in the global south, Penang: Third World Network.
- Arbor, A. (2009). Guide to Social Science Data Preparation and Archiving. Practices pp 47 in Inter-university Consortium for Political and Social Research (ICPSR).
- Bentley, J. W. (2009). Impact of IPM extension for smallholder farmers in the tropics in: R. Peshin, A. K. Dhawan (eds), Integrated Pest Management: Dissemination and Impact, Springer Science + Business Media B.V, The Netherlands.
- Bentley, J.W., Boa, E., Danielsen, S., Franco P., Antezana, O., Villarroel, B., Rodríguez, H. (2009). Plant health clinics in Bolivia 2000-2009: Operations and preliminary results. *Food Security*, 1(3), 371-386.
- Chapman, A. D. (2005). Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.
- Chrisman, N.R. (1991). The Error Component in Spatial Data. pp. 165-174 in: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical
- Conklin, J. (2010). Evaluation of the Quality of the Data Capture System and the Impact of the Data Capture Mode on the Data Quality.
- Consulting, I. and Gingrande, A. (2005). Automated Forms Processing. <http://www.imergeconsult.com/img/101ag.pdf>

- Cooke, B.M. and Savary, S. (2006). Plant disease epidemiology: Facing challenges of the 21st Century Under the aegis of an International Plant Disease Epidemiology Workshop held at Landernau, France, 10–15th April, 2005. Reprinted from European Journal of Plant Pathology, Volume 115 Issue 1, 2006.
- Danielsen, S. and Mutebi, E. (2010). The introduction of mobile plant clinics to Uganda: First results and lessons learned 2005-2010. Work Paper 1. Faculty of Life Sciences, University of Copenhagen, Denmark. 33 pp.
- Danielsen, S. and Fernández, M. (eds) (2008). Public Plant Health Services for All, Plant Healthcare and Diagnostic Network, Managua, Nicaragua.
- Dalcin, E.C. (2004). Data Quality Concepts and Techniques Applied to Taxonomic Databases. Thesis for the degree of Doctor of Philosophy, School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southampton. November 2004. 266 pp. http://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf. Accessed January, 2012.
- Davis, K. (2006a). Institutional arrangements for increasing the dissemination role of farmer groups in agricultural innovation system. A paper presented on the 22nd AIAEE Annual Conference. 14 – 19 May, 2006. Clearwater Beach, Florida.
- DeGroote, H. (2002). Maize yield losses from stem borers in Kenya. *Ins. Sci. Appl.* 22, 89–96.
- Ellis, S. D., and Boehm, M. J. (2008). Factsheet. Introduction to Plant Pathology, 1-5. PP401.01: Plants Get Sick Too! An Introduction to Plant Diseases

- English, L.P. (1999). Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. New York: John Wiley & Sons, Inc. 518pp.
- Flood, J. (2010). The importance of plant health to food security. Food Security, 2(3), 215-231.
- Food and Agricultural Organization of the United Nations, (1996). Mahasahana: A Participatory Extension Network in Sri Lanka. Rome, Italy: FAO.
- Food and Agricultural Organization of the United Nations, (2010). <http://www.worldhunger.org>. Accessed April, 2013.
- Gressel, J., Hanafi, A., Head, G., Marasas, W., Obilana, A.B., Ochanda, J., Souissi, T. (2004). Major heretofore intractable biotic constraints to African food security that may be amenable to novel biotechnological solutions. Crop Protection, 23(8), 661-689.
- Gopisetty, S., Lorie, R., Mao, J., Mohiuddin, M., Sorin, A. and Yair, E. (1996) Automated forms-processing software and services. IBM Journal of Research and Development, 40(2), 1996.
- Kiziltas, S., Akinci, B., Ergen E., Tang P., Gordon, C. (2008). Technological assessment and process implications of field data capture technologies for construction and facility/infrastructure management, ITCon Vol.13 (2008), pg. 134
- Maletic, J.I. and Marcus, A. (2000). Data Cleansing: Beyond Integrity Analysis pp. 200-209 in Proceedings of the Conference on Information Quality (IQ2000). Boston: Massachusetts Institute of Technology. <http://www.cs.wayne.edu>. Retrieved March 2012.

- Manjunath, T.N., Ravindra, S. H., Ravikumar, G.K. (2011). Analysis of Data Quality Aspects in DataWarehouse Systems. International Journal of Computer Science and Information Technologies, Vol. 2 (1), 2011, 477-485. <http://www.academia.edu>. Retrieved March 2012.
- Ministry of Agriculture website. (2012). <http://www.kilimo.go.ke>. Accessed October 2012.
- Mudryk, W. and Xie, H. (2001). Quality Control Application in ICR Data Capture for the 2001 canadian Census of Agriculture.
- Muyanga, M. and Jayne, T.S. (2006). Tegemeo Institute of Agricultural Policy and Development Agricultural Extension in Kenya: Policy, 20498(00200).
- Nations, U. (2002). World Population Prospects. The 2002 Revision Volume III : Analytical Report, III. <http://www.un.org/esa/population/publications>. Accessed June 2012.
- Negussie, E., Karanja, P., Romney, D., Muriithi, C., Kamau, R., Boa, E., Zombe, N., Wanjiku, R., Murage, N., Mulaa, M., Day, R., Mutisya, J. (2011). Role of plant health clinics in enhancing adaptive capacity to climate induced plant health problems: Experiences from Kenya. A paper presented at the Africa Adapt Climate Change Symposium, 09-11 March, 2011, UNECA, Addis Ababa, Ethiopia.
- NLWRA, (2003). Natural Resources Information Management Toolkit. Canberra: National Land and Water Resources Audit. <http://www.nlwra.gov.au/toolkit/contents.html> Accessed February 2012.
- Nwilene, F.E., Nwanze, K.F. & Youdeowei, A. (2008). Impact of integrated pest management on food and horticultural crops in Africa. Entomologia

- Experimentalis et Applicata, 128(3), pp.355-363. Available at: <http://doi.wiley.com/10.1111/j.1570-7458.2008.00744.x> Accessed March, 2012.
- Oerke, E.C. (2000). Infection, damage, loss: An assessment of crop losses to pests and their avoidance. Habilitation thesis, University of Bonn, 313pp (in German).
- Oerke, E.C. and Dehne, H.W. (2004). Safeguarding production: losses in major crops and the role of crop protection
- Parascript L.L.C. (2011). White paper 1 of 2 – Identifying the functionality that will save your company time and money.
- Rahm, E. (2000). Data Cleaning : Problems and Current Approaches. Informatica, 1-11.
- Redman, T.C. (1998). The Impact of Poor Data Quality on the Typical Enterprise. Communications of the ACM 41 (2), 79-82.
- Redman, T.C. (2001). Data Quality. The Field Guide. Digital Press, Boston.
- Reeder, R. (2012). Data management training. Global Plant Clinic - CABI Report
- Rouse M. (2010). <http://searchmobilecomputing.techtarget.com/definition/tablet-PC>. Accessed November 2012.
- Shanner, W.W., Philipp, P. and Schmehl, W. (1982). Farming Systems Research and Development: Guidelines for Developing Countries. Boulder, Colorado: Westview Press.
- Strange, R.N. and Scott, P.R. (2005). Plant disease: A threat to global food security. Annual Review Phytopathol 43:83–116

Taylor, J. E. (1998). An Exploratory Literature Review of Efforts to Help the Small-scale, Resource Poor Farmer in International Agricultural Development.

United States Census Bureau, (2012).
http://en.wikipedia.org/wiki/World_population. Accessed April, 2013.

Viken, A. (2009). "The History of Personal Digital Assistants 1980 – 2000". Agile mobility. Retrieved February, 2012.

APPENDICES

Appendix 1: Plant clinic register Form 1(Manual)

 Global Plant Clinic		REGISTER		Site of plant clinic	Plant Doctor
DATE	CODE	crop (only one); age + variety	description of problem according to client; crop history (management, what was planted previously)		
Name and Telephone no.					
Village- Sub location and Location, Division		diagnosis		action taken	

plantwise

Prescription and record sheet

Clinic Information

Date Code Plant doctor

Farmer Information

Name Sex ☐ Male ☐ Female

District Parish

Village Tel

Sample Information

Crop

Variety Sample brought ☐ Yes ☐ No

Development Stage

☐ Seeding ☐ Intermediate ☐ Flowering
☐ Fruiting ☐ Mature ☐ Post harvest

Part Affected

☐ Seed ☐ Root ☐ Stem base ☐ Stem ☐ Twig/Branch
☐ Leaves ☐ Flower ☐ Fruit/Grain ☐ Whole Plant ☐ Shoot

When first seen and area affected

Year first noticed Area planted Acres ☐ Hectares ☐ Number ☐

% crop affected

☐ 100% ☐ 50%
☐ 75% ☐ 25% ☐ < 25%

Predicted Yield

% yield loss

Major Symptoms

☐ Wilted ☐ Stunted ☐ Streak ☐ Blistered ☐ Leaf blotch / mosaic ☐ Witches broom ☐ Cankers (stem lesions) ☐ Rots
☐ Yellowed ☐ Dieback ☐ Leaf spot ☐ Distorted ☐ Leaf edge scorch ☐ Surface growth ☐ Bore holes (stem / fruit) ☐ Staining
☐ Abnormal colour ☐ Leaf fall ☐ Chewed ☐ Little leaves ☐ Insect / mite seen ☐ Fruit drop ☐ Galls / swellings ☐ Drying

Distribution of Symptoms Within Field

☐ Localised ☐ Scattered ☐ Linear ☐ Field margin ☐ Even ☐ Certain varieties ☐ Individual plants ☐ High areas ☐ Low areas

Describe problem

(Additional information - include key observed symptoms, plant parts affected, etc)

Type of Organism: Biotic

☐ Fungi ☐ Bacteria ☐ Insect/Mite ☐ Nematode ☐ Virus ☐ Phytoplasma ☐ Animal ☐ Unknown

Type of Problem: Abiotic

☐ Nutrient ☐ Environment ☐ Unknown

Diagnosis (start a new sheet for each new problem)

Pest / disease

Current Control

Practices used

Recommendations for management

☐ Monitor problem ☐ Cultural ☐ Biological ☐ Host resistance ☐ Fungicides ☐ Insecticides ☐ Acaricides ☐ Nematicides ☐ Herbicides

Follow up activities:

Sample sent to lab

☐ Yes ☐ No

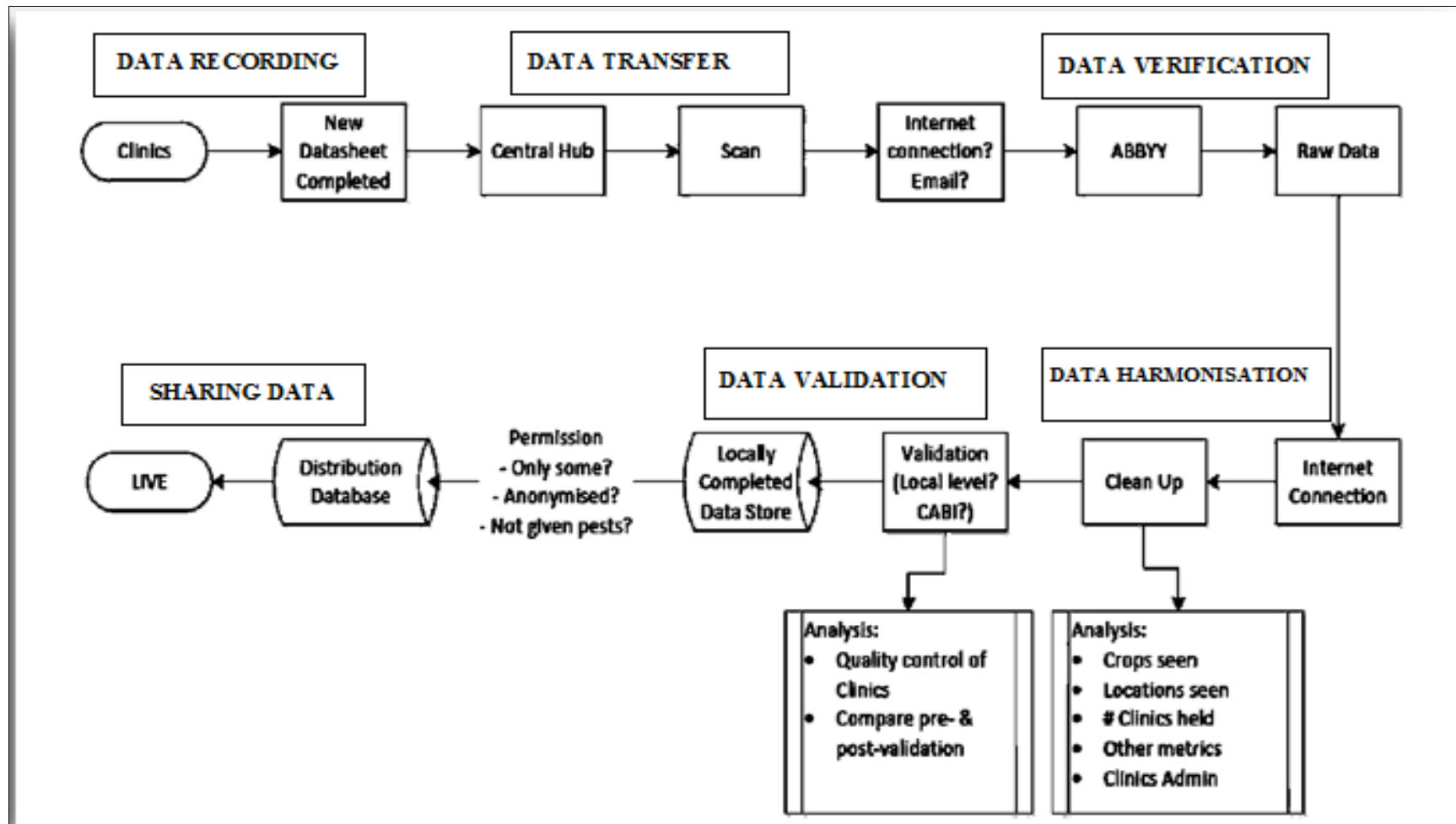
Fact sheet given

☐ Yes ☐ No

Field visit arranged

☐ Yes ☐ No

Appendix 3: Overview of processes for plant clinics data collection and analysis



District - Field Verification

Verification
Edit
View
Tools
Help

Field: "District"

BATISTA BAHAMA

M

B

E

Y

A

M
BEYA

Name	B A T I S T A B A H A M E		SEX	☒ Male	☐ Female
District	M B E Y A	Pansh			
Village	I S E N D E	Tel	0	1	2 3 4 5 6

ISENDE

BANANA

Verified: 0%
ClinicsDemo: Document 2: Page 1
FarmerInformation\District

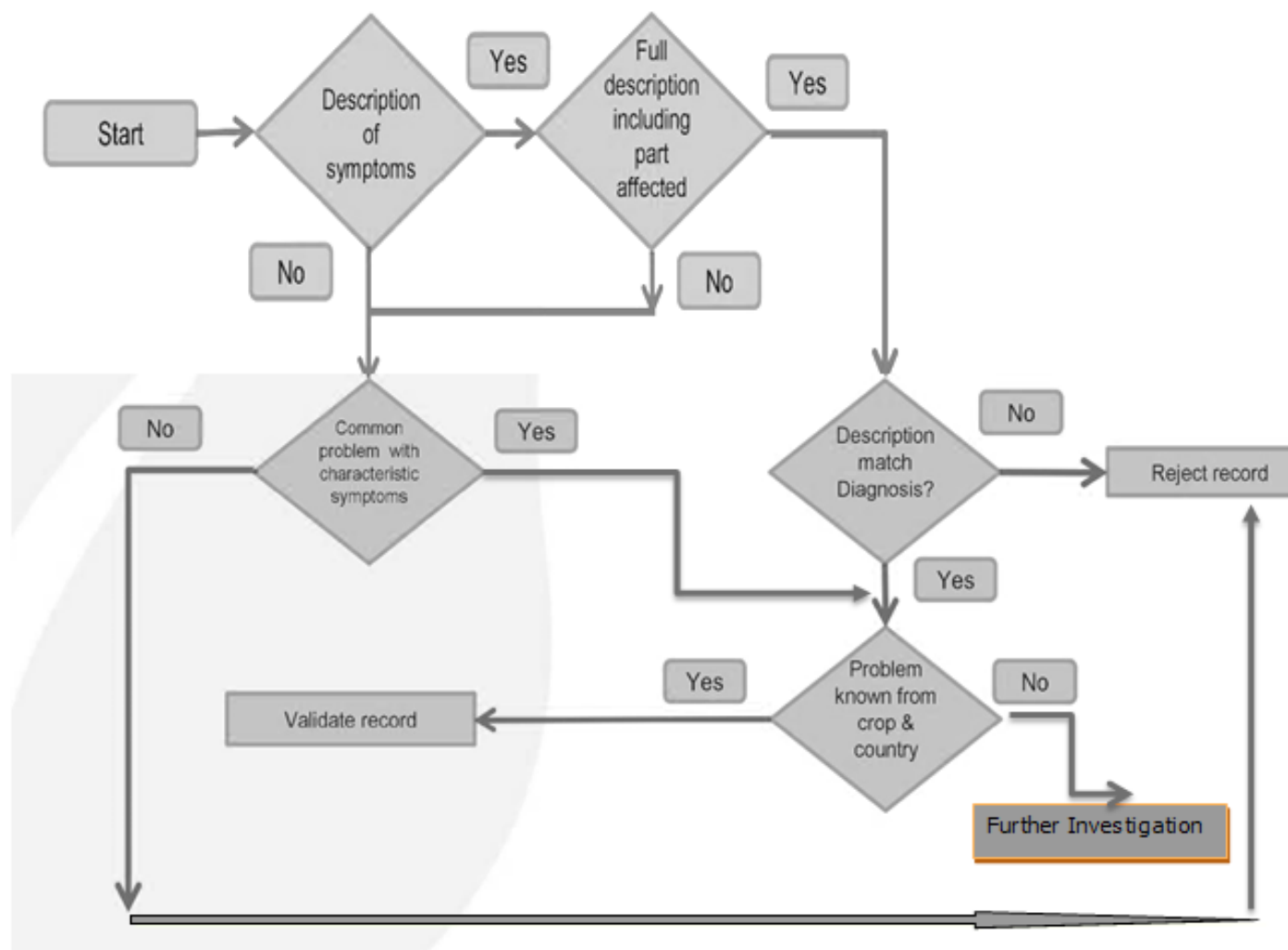
Appendix 5: Symptom and diagnosis guidelines for crop pest

Crop name	Crop pest/Diagnosis	Full description of the symptoms including part of the plant affected	Expected advisory recommendation
BEAN	Bean aphid (<i>Aphis fabae</i>)	Bean aphid is a dark olive green to black in colour. The caudal has hairs and appears bushy. Bean aphid has wings. Infestation of the bean aphid begins on young leaves in the center of the crown before colonizing older leaves. They are mainly found on underside of leaves; rarely on upper surfaces. Infested leaves curl under and inward, become severely distorted, necrotic before dying. Honeydew produced by the aphid on infested leaves have a sticky black mould appearance.	Mention any or a combination of the following methods. Use of recommended insecticide, cultural practices or Integrated Pest Management control measures.
MAIZE	Maize streak virus/MSV	The disease severity depends on the age at which the plant is infected, maize variety and climatic conditions. Initial symptoms emerging after infection consist of scattered, pale, chlorotic, circular spots or flecks on the base of the leaf. Severe cases the spots enlarge and the pale green spot become parallel longer streaks on leaf vein/interveinal necrosis. The disease causes stunted growth, yellow-green appearance on mature plants and deaths. Plants affected early may produce deformed cobs or become barren.	Mention any or a combination of the following methods. Use of recommended insecticide to control the vector leafhopper - <i>Cicadulina mbila</i> causing virus transmission, cultural practices, IPM control measures.
TOMATO	Early blight (<i>Alternaria solani</i>)	It is a common foliar disease caused by <i>Alternaria solani</i> . Foliar damage show one or two spots per leaf. Spots have tan centers with concentric rings in them and yellow halos/chlorotic around the edges. The affected leaves may dry up and drop. On affected fruit there is a dark, sunken irregular lesion which appears near the calyx of fruits while on stem the damage consist of dark, sunken cankers at or above the soil level.	Mention any or a combination of the following methods. Use of recommended fungicides, cultural practices or IPM control measures.
POTATO	Late blight (<i>Phytophthora infestans</i>)	On foliar the lesions/spots start out pale green near the edges of tips of foliage and they are initially small. They later turn brown or dark grey in colour. These lesions enlarge rapidly over entire foliage during moist warm weather. On fruit brown, leathery spots/lesions appear on green fruit on the top and sides of the fruit. In humid conditions, white mold also forms. Stem damage include dark grey and brown spots appearance which may spread to cause death.	Mention any or a combination of the following methods. Use of recommended fungicides, cultural practices or IPM control measures.

KALE	Aphid (<i>Brevicoryne brassicae</i>)	Aphids are soft-bodied insects often seen in colonies on leaves and stem. They are pale grey/green in colour, winged or wingless and vary in size from 1.5 to 3.0 mm. Aphids damage plants by sucking the sap. Heavily infested leaves may curl and wrinkle. The plant may deform, become stunted in growth. Aphids excrete honeydew which can cause blackening areas on leaves as fungal mould grows on it.	Mention any or a combination of the following methods. Use of recommended insecticide, cultural practices or IPM control measures
CABBAGE	Diamondback Moth (<i>Plutella xylostella</i>)	Adult are small thin moths which fold their wings along their backs. Male have yellow diamond-shape joined on a brown background. The female larvae curl and drop when disturbed. Young larvae damage brassica by making small holes underside of the leaf. Older larvae feed on all plant parts and around the growing bud of young plants. They create irregular holes/windowing on leaves.	Mention any or a combination of the following methods. Use of recommended insecticide, cultural practices or IPM control measures
PIGEON PEA	African bollworm (<i>Helicoverpa armigera</i>)	Caterpillars of the pest feed on leaves, buds, growing points, flowers and fruit. They cause small, clean circular holes, with frass waste and decay on pods and fruits. The holes serve as entry points for secondary infection by diseases causing fruit decay. The larvae feed on the tips of flower buds and on young leaves. This can prevent fruit formation plus reduce leaf area which cause slow plant growth. Feeding on flowers and fruit causes the main damage. Often caterpillars feed with the head and forepart of the body inside the fruit/pod and the rest of the body outside. They are found mainly during flowering and fruiting stage.	Mention any or a combination of the following methods. Use of recommended insecticide, cultural practices or IPM control measures. Pesticides should be applied before the caterpillars bore into the fruits/pods.
COFFEE	Coffee berry disease (<i>Colletotrichum kahawae</i>)	Symptoms on green berry include water-soaked sunken dark spots/lesions which enlarge to cause rot of the berry. Pink spores masses are visible on lesion surface. At advanced stage of the disease; the berry turns completely brown to black with the beans inside destroyed. Affected berries shed from the branch at an early stage of the disease. Affected ripe berries have typical dark and sunken anthracnose lesions. The disease infection on flowers can cause brown lesions on petals.	Mention any or combination of the following methods. Use of recommended fungicides, cultural practices or IPM control measures.

MANGO	Anthrachnose (<i>Colletotrichum gloeosporioides</i>)	The disease appears on young leaves, stem, inflorescence and fruits. Leaves show oval or irregular greyish, brown spots which may coalesce to cover larger area of the leaves. Affected leaves dry and shed off. On young stem, grey brown spots develop which enlarge and cause girdling and drying of affected area. Black necrotic areas develop on the twigs from the tip downwards causing a dieback. Tear stain symptom on developed fruit which are linear necrotic may or may not be associated with superficial cracking of the epidermis	Mention any or combination of the following methods. Use of recommended fungicides, cultural practices or IPM control measures.
BANANA	Bacterial wilt (<i>Xanthomonas campestris</i> pv. <i>Musacearum</i>)	The disease causes wilting of banana plant at all ages. Affected plants die prematurely while infection during fruiting stage causes fruit rot. First symptoms are a dull green colour of the lamina which gradually assumes a scalded appearance and wilting back on its midrib. Cross sections of diseased petioles or pseudostems reveal yellowish colouration. There is uneven and premature ripening of the fruit and when fruit are cut, the sections show unique yellowish blotches in the flesh fingers and dark brown placental scars.	Mention any or combination of the following methods. Use of recommended fungicides, cultural practices or IPM control measures.

Appendix 6: Validation for diagnosis



Appendix 7: Validation for recommendation

