2020-04

Use of agent – based models in characterizing farm types and evolvement in smallholder dairy systems

Nyambo, Devotha

NM-AIST

http://doi.org/10.58694/20.500.12479/895

USE OF AGENT – BASED MODELS IN CHARACTERIZING FARM TYPES AND EVOLVEMENT IN SMALLHOLDER DAIRY SYSTEMS

Devotha Godfrey Nyambo

A Thesis Submitted in Fulfilment of the Requirements for the Degree of Doctor of Philosophy in Information and Communication Science and Engineering of the Nelson Mandela African Institution of Science and Technology

Arusha, Tanzania

April, 2020

ABSTRACT

The ever-increasing demand for milk and dairy products has attracted research interventions on how milk yield can be increased for the context of smallholder farmers. While bearing significant contribution on milk production and fulfilment of the market demand, the smallholder dairy farmers are faced with challenges that hinder productivity. Among the challenges is the inadequate characterization of the dairy production systems and lack of knowledge on factors attributing to their growth. This has resulted in aggregation of the smallholder dairy farmers and lack of interventions tailored to suit particular farm types. By using Tanzania and Ethiopia as case studies, this research identified the main determinants for evolvement of smallholder dairy farmers. Evolvement in this research refers to, gradual increase in milk yield. The factors that determine evolvement for individual farm typologies were identified by using cluster and frequent pattern analysis. The differential influence of the identified determinants towards increase in milk yield was studied by using Agent-based modelling and simulation where each factor was observed.

Six farm types were identified for Tanzania and four for Ethiopia. The characteristics of the farm types were enriched by frequent pattern analysis with confidence level 60% - 97%. Agent-based modelling revealed that, income and farm-based determinants influenced an increase of up to 7.58 litres above the average (13.62 ± 4.47) for Ethiopia. For Tanzania, farm and farmer-based determinants influenced an increase of up to 7.72 litres of milk above the average (12.7 ± 4.89). The identified determinants could predict up to 96% and 93% of the variances in milk yield for Tanzania and Ethiopia, respectively. There was an increase in milk yield based on the identified evolvement determinants; from baseline data average milk yield of 12.7 ± 4.89 and 13.62 ± 4.47 to simulated milk yield average of 17.57 ± 0.72 and 20.34 ± 1.16 for Tanzania and Ethiopia, respectively. Dairy development agencies should consider the disaggregation of dairy farmers and prioritization of the determinants identified in this research for evolvement of dairy farms. In future, it is important to develop a web or mobile application that can inform smallholder dairy farmers about the identified evolvement determinants to aid on-farm decision making.

DECLARATION

I, Devotha Godfrey Nyambo do hereby declare to the Senate of Nelson Mandela African Institution of Science and Technology that this thesis is my own original work and that it has neither been submitted nor being concurrently submitted for degree award in any other institution.

15/04/2020

Devotha Godfrey Nyambo

Name and Signature of Candidate Date

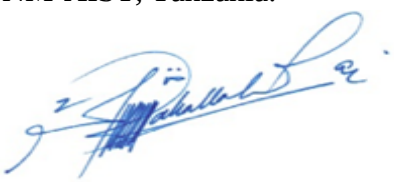
The above declaration is confirmed

15/04/2020

Dr. Edith T. Luhanga Date

Supervisor, School of Computation and Communication Science and Engineering,

NM-AIST, Tanzania.



15/04/2020

Eng. Dr. Zaipuna O. Yonah Date

Supervisor, Applied Engineering & ByteWorks (T) Limited, Tanzania.

15/04/2020

Dr. Fidalis D. N. Mujibi Date

Supervisor, USOMI Limited, Kenya.

COPYRIGHT

CERTIFICATION

The undersigned certify that they have read the thesis titled **"Use of agent – based models in characterizing farm types and evolvement in smallholder dairy systems"** and approve for submission to the Nelson Mandela African Institution of Science and Technology senate for award consideration of the PhD degree in Information and Communication Science and Engineering of the Nelson Mandela African Institution of Science and Technology.

Dr. Edith T. Luhanga 15/04/2020

Name and signature of supervisor Date

Eng. Dr. Zaipuna O. Yonah 15/04/2020

Name and signature of supervisor Date

Dr. Fidalis D. N. Mujibi 15/04/2020

Name and signature of supervisor Date

ACKNOWLEDGEMENTS

I thank God Almighty for His grace to undertake this research to its accomplishment.

My sincere gratitude goes to my lovely husband Mr. Lawrence Nehemiah, my children Ethan and Elsie, my young brother Nickson Nyambo and my family in its entirety. Your love and support during the research are unmeasurable and will forever be cherished. Specifically, I thank you for your patience and understanding for the times I had to travel and leave you for extended number of days and months. The daily motivation and encouragement on the many times of my breakdown is evident as I finalize and mark this milestone.

The genesis of this research idea started during analysis of the PEARL project's data under the leadership of Prof. Morris K. Agaba and Dr. Fidalis (Denis) N. Mujibi. I sincerely thank you for the guidance and supervision. Dr. Mujibi, your time and supervision towards my understanding of smallholder dairy production systems cannot be underestimated and I am very grateful. Eng. Dr. Zaipuna O. Yonah, your name and expertise has been traveling with my career. Thank you for being such a great supervisor and mentor since my MSc. degree. During the research journey I found an inspirational young faculty who has played a very significant role in the modelling and simulation not to mention the times she lifted me up with encouragement and trust, Dr. Edith T. Luhanga. I am very grateful for your support, inputs during papers publications and the agent-based models' development and simulations.

I sincerely thank the management of the Nelson Mandela African Institution of Science and Technology (NM-AIST) for granting me the admission, study leave and a scholarship to undertake this research. I equally thank the management of the African Development Bank project at NM-AIST for funding the research and their support.

The agent-based models and simulations done were not possible if not by the support from the MARS laboratory in Hamburg, German. I highly appreciate the support given by the laboratory under the leadership of Prof. Thomas Clemen. I sincerely acknowledge the support of Mr. Julius Weyl, Mr. Daniel Glake and Ms. Ulfia Lenfers during the agent-based models' conceptualization and implementation. Last but not least, I am very grateful to all my friends and classmates whom we shared several challenges and achievements. Of great importance, those who kept me in their prayers and encouraged me from time to time.

May God bless you all.

DEDICATION

To my parents, Eng. Godfrey Basil Nyambo and Esther Sanga (R.I.P).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS AND SYMBOLS

ABMS	Agent-Based Modelling and Simulation
AI	Artificial Insemination
ASC	Action Script Communication
CSV	Comma Separated Values
DSL	Design Specific language
FAO	Food and Agriculture Organization
IDE	Integrated Development Environment
KM	Kilometre
KQML	Knowledge Query and Manipulation Language
LHS	Left Hand Side
MARS	Multi Agent Research and Simulations
MLFD	Ministry of Livestock and Fisheries Development
NAIC	National Artificial Insemination Centre
NM-AIST	Nelson Mandela African Institution of Science and Technology
PEARL	Program for Emerging Agricultural Research Leaders
RHS	Right Hand Side
SAS	Statistical Analysis Software
SDF	Small-holder Dairy Farmer
SOM	Self-Organizing Maps

CHAPTER ONE

INTRODUCTION

1.1 Background of the problem

In Africa, smallholder farmers accounts for 75% of the agriculture production and contribute about 50% of the total livestock output (Salami *et al*., 2010). Livestock keeping continues to be a key component to the livelihoods of many rural and peri-urban populations in Sub-Saharan African countries (Otte *et al*., 2012; Smith *et al*., 2014). For example, in Tanzania, 50% of the total population keep livestock to support livelihoods (Tanzania, 2016); and the livestock sector contributes about one-third of the total agriculture value (FAO, 2014). However, despite the expected increase in future demand for milk and other livestock products in developing countries, the sector still faces major challenges including technology generation and dissemination as well as infrastructure. Observed globally is that dairy farming is practiced by crop-livestock farmers mainly at small scale, and some medium to large scale farms; and it has been acknowledged that, the small-scale producers dominate the quantity of produce (Cortez *et al*., 2014).

According to the Food and Agriculture Organization (FAO), smallholder dairy farm is comprised of 2-5 milking cows (FAO, 2010). Researchers have revealed similar structures across the Eastern Africa region; characterized by small herds with 1-5 cows, of which 3 are pedigree or crosses of exotic breeds (Swai, Mollel & Malima, 2014). Majority are crop-livestock farmers practicing mixed cropping (staple food and cash crops) (Kurwijila, 2001; Thorpe & Muriuki, 2001; Herrero *et al*., 2014). Key differences among farmers in the region are centred around management practices, market structures and household demographics.

In Africa Ethiopia is the highest ranked in terms of dairy output, followed by Kenya and Sudan (Karmella & Dolecheck, 2015), while Tanzania is the 10^{th} largest milk producer (Allan, 2019). Dairy production in Tanzania, Ethiopia and Kenya is dominated by smallholder dairy farmers operating in very diverse production systems. It is therefore imperative to study these systems to identify new opportunities to improve the sector in these countries. Generally, the cattle population for Ethiopia and Tanzania is comprised of indigenous and cross breed animals (Gondro *et al*., 2018). However, the production metrics signifies there is so much to be done for these countries to meet the growing demand.

Ethiopia ranks the first in Africa for the case of cattle population; 54 million heads of cattle forming 3.68% of the total world's cattle population while, Tanzania has 24 million heads of cattle forming 1.67% of the total world's cattle population (Cook, 2015). For Tanzania, regardless of the number of cattle the total dairy yield is far way below current consumer demand (Tanzania, 2016). This situation has been due to factors within the dairy value chain (from farm characteristics, service providers to market systems). Some researchers have identified feeding systems to be among the key hindering factors for the growth in the sector (Maleko *et al*., 2018) and an imbalance of breed types suited in the specific production environments (Mujibi *et al*., 2019). Specifically, this research contributes to the ongoing efforts to improve the smallholder dairy farming systems by understanding the key production characteristics and the factors that promotes low to higher milk yield (evolvement) by using Tanzania and Ethiopia as a case study. Ethiopia has been selected based on the fact that it is leading in Africa's dairy production while Tanzania has been selected based on the fact that it has low dairy production metrics yet its dairy system is similar to other East African countries. The term evolvement as used in this research refer to the gradual process of increasing milk yield from low to mid and high yield.

Researchers generally agree that knowledge on the pathways through which dairy systems may change in the future and how to influence them for maximal productivity is critical for poverty alleviation and food security goals (Herrero *et al*., 2014). Knowledge of farm (and farmer) evolvement is critical specifically for the introduction and adoption of improved farming practices. Additionally, understanding the key factors constraining dairy productivity and how they interact and interplay within different production contexts will sustainably improve milk production (Swai *et al*., 2014). Smallholder dairy farming is constrained by a number of factors including, availability of feeds, no access or incapacity to purchase farm inputs, insufficient land, unreliable markets for milk products, and unreliable or absence of breeding services (Swai, Mollel & Malima, 2014; Wodajo & Ponnusamy, 2015) . However, the level of these constraints varies from farm to farm based on locality, education level of the farm owners, and other household demographics. In view of this variety, there is need to critically characterize the farms and farmers and disaggregate the factors that hinder high milk yield for various types of farms.

Characterization of smallholder farming systems is a study of categories of farms based on their demographics, management practices, trends in total yield and existing production

clusters. Generally, the characterization process involves identification of homogenous groups of farmers. The groups obtained in a characterization process are also known as production clusters. The main goal of characterization is to depict production categories existing in a particular environment or a complex agro-ecological system for appropriate introduction of improved technologies and conversant policy support (Goswami *et al*., 2014). Development of the farm types can allow focus to specific types of farmers and avoid aggregation. Thereafter, the effectiveness of farm management practices designed to improve milk yield for various farm types can be established by subjecting them in real world modelling tools. The progression of varying management practices to improve milk yield can be of trial and error processes, which involve several heuristics, costs time and financial investments (Mbwambo, Nigussie & Stapleton, 2017). Unsuccessful attempts may lead to undermined desire to continue farming. To mitigate that, optimal strategies must be designed and tested in an abstract setting before being applied to the real-world dairy farming.

From modelling point of view, object-oriented real-world modelling is used in development of computer programs that simulate processes in the real world (Isoda, 2001). In such cases, computerized objects are essential in predicting evolvements of such real-world entities when given different inputs. This type of modelling is basically a construction of real-world entities into computer programs often called objects (Larman, 2005). An object refers to any real-world material entity that can be seen and touched. Any specific realization of an object based on different inputs is referred to as an instance. Characterized with encapsulation, inheritance and polymorphism (Armstrong, 2006) object-oriented modelling allows entity data hiding, reuse of predefined characteristics through inheritance and varying evolvement based on different inputs (data sets). In the later character, a function of an object will be able to respond differently when provided with different arguments.

For the case of smallholder dairy farmers, agent-based modelling brings in the capacity to study various characteristics and production trend. Being complicated as it is, a smallholder dairy farm is tied to various components that have differing influence on milk yield and commercialization. Among the various components in a smallholder dairy farm are; the farm itself being composed of dairy cows and sometimes a mixture of small ruminants and poultry, service providers and a farm manager who depends on the farm for daily household needs. Different farms may exhibit unique characteristics or management practices. From the mentioned components, it is important to study the factors that influence milk production (Fig.

1). To better understand the evolvement pattern of a farmer in focus, it then becomes vital to study and disaggregate the farm's internal factors deriving milk yield and the external service providers which have an effect in milk yield. An understanding on these factors would reveal key determinants for a farmer's evolvement from low to mid and high production.

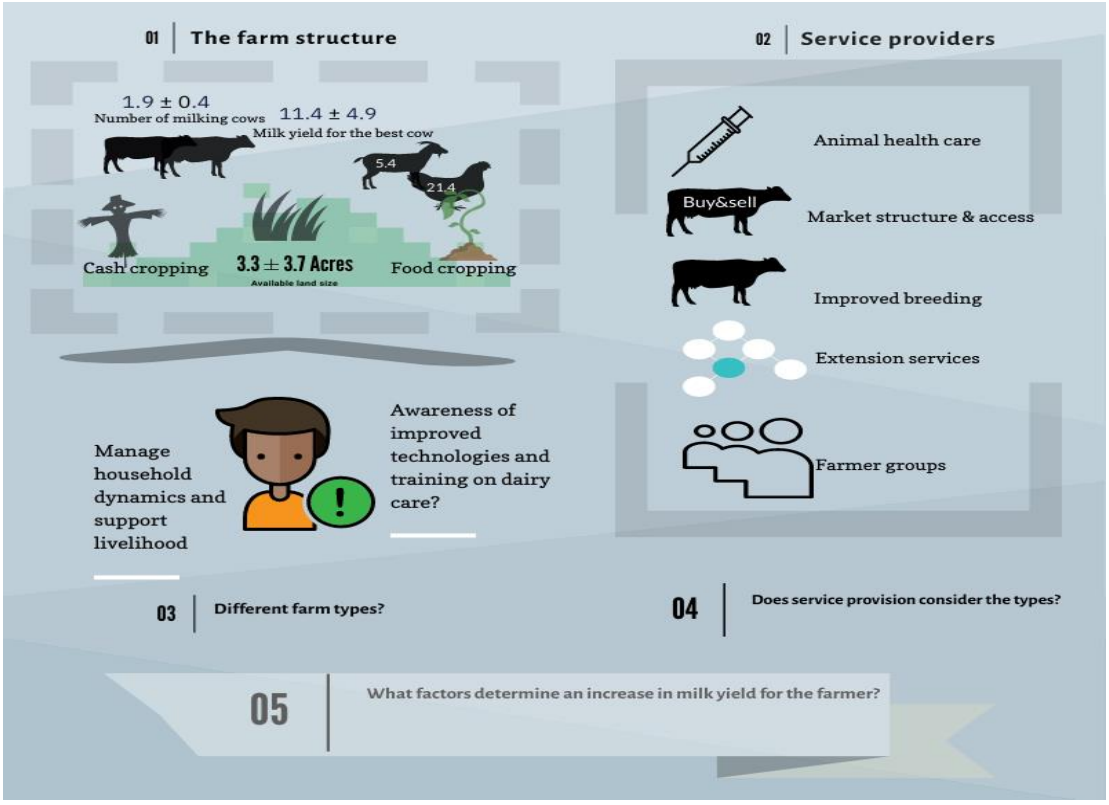


Figure 1: Overview of a smallholder dairy farm and its service providers

As the desire to model and simulate multiple entities with unlimited interactions grows, use of traditional object-oriented modelling becomes limited in terms of flexibility and scalability (Isoda, 2001). To mitigate this limitation, use of multi-agent modelling is proposed, in which case selected entities can be fully represented in one platform or environment, learn and progress as they interact with each other. Agent based modelling and simulation builds on object-oriented modelling given that the agents can be developed following the rules and constructs of object-oriented modelling. Based on a heterogeneous nature of entities, the agents' settlement, and unlimited interactions, use of multi agent-based modelling provides a multi-dimension approach to study the evolvement of such entities. In addition, various decision problems can be derived and implemented for the agents as demonstrated by Macmillan and Huang (2008) who modelled an agricultural society of producers and consumers.

The divergent characteristics of smallholder dairy farmers necessitates for a study to understand how this diversity influences their different production potentials. Different farm types evolve differently with regard to milk yield and sales, as such disaggregation is vital. In my research work, I characterized smallholder dairy farmers using machine learning models and association rules were studied for different farm types. The specific farm characteristics are represented in an agent-based model and simulated to show how such farms are expected to evolve over a period of time as farming practices change and they adopt best practices seen within better-yielding households.

The outcome of this research is an in-depth characterization of smallholder dairy farms based on cluster and associations rules analysis. Stable farm types that can be used to study smallholder dairy farmers' evolvement have been derived for Tanzania and Ethiopia. Importantly, factors that determine increase in milk yield for various farm typologies are proposed and their impact assessed through agent-based modelling and simulation. Given the various factors that when changed can result into increase in milk yield, the agent-based model provides an assessment of the most important factors that a farmer can consider based on the farmer's underlying features. Consequently, it becomes feasible for a farmer to predict the impact of changing certain farming practices and observe their significance prior to implementing them in the real world. Furthermore, the need for farmer social networks for knowledge sharing and best practices is described by a model that simulates farmer learning and adoption of best farming practices from each other, resulting into improved milk yield.

1.2 Statement of the problem

Current approaches being used to assist dairy farmers run their enterprises profitably are failing because characterization of the farms and critical associations between the farm types and the main factors that determine their evolvement are unknown. As a result, smallholder farmers are treated as one amorphous unit, leading to poor utilization of investments as well as demoralizing farmers from pursuing dairy farming. Disapproving that the smallholder dairy farmers are one amorphous unit; this study details an approach to disaggregate the farmers and establish important factors pertaining to the evolvement of the different categories of dairy farmers with respect to increase in milk yield. Therefore, this study was intended to shed light on what can be done to drive smallholder dairy farm transformation from low milk yield to mid and high yield.

1.3 Rationale of the study

Adequate research on how the smallholder farming systems may change in future is critical for poverty alleviation and food security (Herrero *et al*., 2014) especially in Africa where the smallholder farmers are the main food producers (Salami *et al*., 2010). Existing factors that determine smallholder dairy farms' evolvement must be explored to better predict the future of these small enterprises with the goal of improving milk yield (Swai *et al*., 2014). Several studies have attempted to model evolvement of smallholder farming systems by using traditional approaches such as statistical and linear programing models (French, Tyrer & Hirst, 2001; Herrero *et al*., 2014). However, real time modelling of complex and highly sophisticated systems (operating in longer time scales and incorporate complex human-environmental interactions), has resulted into higher analytical complexity (Nolan *et al*., 2009). Fortunately, Agent-Based Modelling and Simulation (ABMS) research is a proven method to integrate several multi-level systems' components into one modelling platform (Therond *et al*., 2014), making ABMS a feasible approach to model and simulate scenarios that would be complex laboratory experiments (Ghorbani, 2016), and human-environment interactions where systems' equilibrium is difficult to obtain (Nolan *et al*., 2009).

1.4 Objectives

1.4.1 General objective

The main objective of this research is to characterize, through computer modelling and simulation, the factors that determine smallholder dairy farm evolvement and commercialization, indicated by progressive increase in milk yield.

1.4.2 Specific objectives

This research intends to accomplish the following specific objectives:

(i) To characterize smallholder dairy farms and map out the key features that determine production environments (farm types).
(ii) To identify the determinants of farm evolvement and their association rules for the derived farm types.
(iii) To determine the differential influence of the evolvement determinants on milk yield maximization through agent-based modelling and simulations.

1.5 Research questions

This research intended to answer the following questions:

(i) How can cluster analysis be used to derive stable farm types and identify key features that determine production environments?

(ii) From the derived farm types based on cluster analysis, what are the main frequent patters that can be used as determinants for the farm types evolvement?

(iii) How can agent-based modelling and simulation be applied to establish the differential influence of the evolvement determinants identified through objective one and two?

1.6 Significance of the study

The need to characterize and study evolvement patterns for smallholder dairy farmers cannot be over emphasized as already highlighted in literature (Herrero *et al*., 2014; Swai *et al*., 2014). Developmental projects that bring in interventions for smallholder dairy systems need to be guided by thorough research on the farming systems. This research work informs policy makers, dairy boards, farmer groups and researchers on key factors that determine increase in milk yield for farmers: with selected case study sites in Tanzania and Ethiopia. Impact is envisaged for smallholder farmers who are led, informed or supported by various policy making boards, developmental partners, researchers and farmer groups. By helping the farmer understand the key factors that lead to higher milk yield in his/her production system, the common trial and error approach used for milk yield improvement is avoided.

Farm typology characterization commonly undertaken using clustering approaches, have been complemented in this research by providing a methodology to obtain robust clusters. Application of multiple clustering algorithms and supervised learning approaches to validate cluster robustness has been recommended and tested (Nyambo, Luhanga & Yonah, 2019a; Nyambo *et al*., 2019).

This research also presents an approach to study characteristics of smallholder farming systems through association rules mining. Use of frequent pattern has been demonstrated to be useful in characterizing farming systems based on majority practices (Nyambo, Luhanga & Yonah, 2019b). The presented approach complements the characterization of farmers based on cluster analysis. New features were identified and modelling through agent-based methodology, the features were proven significant. This research presents, for the first time use of agent-based modelling and simulation to study the differential influence of factors that determine increase

in milk yield for smallholder dairy farmers. An understanding on the differential influence of the determinants will influence farmers and policy makers to channel their focus and resources practices that have a higher likelihood to increase milk yield.

This study provides a robust approach that can be used by several stakeholders in the dairy industry including: researchers, dairy boards, developmental partners and government agencies that can be used in order to help farmers commercialize as they evolve from subsistence dairy farming to fully commercial enterprises. The use of agent-based modelling and simulation to study evolvement reduce the trial and error approaches for the farmers to increase milk yield, which consequently will improve the commercialization of the smallholder dairy farmers and improve their livelihood.

1.7 Delineation of the study

Use of agent-based modelling and simulation to study the evolvement of smallholder dairy farmers is presented with an emphasis on farmers' disaggregation. Formation of homogenous groups of farmers was done by using unsupervised learning algorithms, with a robust validation approach. To detail the approach for smallholder dairy farmers' characterization and evolvement modelling secondary datasets were used. Results from this study highlights key determinants for evolvement of smallholder dairy farmers. The methodology and results demonstrated in this research can be used by researchers rather than individual dairy farmers. This is because, development of mobile or web based user interface was beyond the scope of this research. Limitations based on datasets and computing resources are detailed below:

(i) Datasets

The research adopted datasets from the Program for Emerging Agricultural Research Leaders (PEARL) project which sampled emerging dairy sheds in four countries of East Africa namely Tanzania, Kenya, Uganda and Ethiopia. For Tanzania, six regions where surveyed based on a purposive sampling approach. As such not all smallholder dairy farmers were included in the study rather than those which the project had objectives for. For the same reason, only farmers keeping exotic or cross breed cattle were considered in the survey population. Therefore, this study does not represent the entire smallholder dairy population in Tanzania or Ethiopia and that its results and implications can be tailored to the study sites detailed in Section 3.1. However, this research results can be assimilated to other dairy population having similar baseline characteristics as the given study sites in the PEARL project with some level of bias.

Applicability of the findings of this research in pastoral communities is deemed irrelevant. In addition, study site differences within the countries were large. However, the study did not want to add this extra complexity in the modelling, therefore, locational differences were removed so that the productivity of the smallholder dairy farms could be studied in general.

The dataset from Tanzania indicated a high level of similarities particularly during cluster analysis. One reason for this could be poor data collection practices which resulted in records duplication with low level of differences. The impact of this could be observed in low rank correlation for the Tanzania clusters as reported in Section 3.1.2. If the data quality was improved, the rank correlation and clustering results could have been significantly altered. However, the methods and results described detail in-nutshell what can be achieved by using existing data to study the future improvements of dairy farms.

(ii) Computing resources

The agent-based model simulation cases were done with some computing challenges that resulted into failure to load complete output files. The used computing resources (i.e. Hewlett-Packard Intel core i7-5600U CPU @ 2.60GHz, 8GB RAM) with Microsoft excel, could not load an output file for farm type three for Ethiopia due to its large size (n=2689 producing 2689 x 731 rows of data in excel). As a result, the csv output could not be opened in full for analysis. This scenario was solved by having the farm type split in two parts, (a, with n=1344) and (b, with n=1345) to study the farmers' interactions based on scenario 1 and 2 as detailed in previous sections. Since the interactions are highly influenced by nearest neighbours, the split of the data could somehow reduce the chances of some members to meet influential peers who could lead to changes in their milk yield. It is tempting to imply that the simulation results for the farmer networks for farm type three could have some differences if the whole set was analysed at once.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Agent-based modelling and simulation has been applied in various fields such as in: - modelling the evolvement of agricultural societies (Macmillan & Huang, 2008), modelling growth and expansion of urban societies (Benenson, 2004; Arsanjani *et al*., 2013), social networks and supply chain management (Fox *et al*., 2001; Frayret *et al*., 2007) and diffusion research (Kiesling *et al*., 2012). The research work reported in this dissertation builds on the existing knowledge on agent-based models with a specific focus of maximizing milk yield in smallholder dairy farming systems. The study of the farmers' evolvement from low to mid to high milk yield was preceded by characterization to identify important factors that determine evolvement for various clusters of farmers (referred to as farm types). Thereafter, use of agent-based modelling and simulation was employed to study how the identified evolvement factors influence increase in milk yield.

2.2 Characterization of smallholder farmers

Characterization of smallholder farming systems refers to detailing the features of production environments. Through characterization, farming systems can better be understood and differentiated. Generally, characterization of farmers involves classifying farmers and farming systems into homogenous groups. The homogenous groups are also known as production clusters. The underlying goal of having the homogenous groups is for appropriate introduction of improved technologies and conversant policy support (Goswami *et al*., 2014) while avoiding aggregation of farmers as one amorphous entity. Mostly, in advanced analysis of farming practices, development of typologies is crucial to avoid aggregation. The sub-groups unveil existing variations among farmers or farm types and therefore, an improvement plan can be targeted to a particular group of farmers instead of considering them as one.

Various approaches are reported in previous researches concerning characterization of smallholder farmers; all of them ending up with development of some sub-groups of farmers. Characterization methods range from on farm participatory to advanced statistical and machine learning algorithms with increasing order of complexity. The mostly used algorithms in characterization can be grouped into deterministic and probabilistic approaches. In deterministic methods, with the same starting values and number of clusters, an algorithm will

produce the same type of results (Celebi & Kigravi, 2013). On the other hand, probabilistic methods may yield different groupings even without altering the starting values or number of clusters (Celebi & Kigravi, 2013). In general, the deterministic approaches use supervised learning while the probabilistic approaches use unsupervised learning. In unsupervised learning, an algorithm would self-group the data set according to the algorithm's settings.

Prior to the characterization done in this research, a detailed review of literature was done to study key methods used in previous researches (Nyambo, Luhanga & Yonah, 2019). The paper reports on three main approaches being used: deterministic (Van de Steeg *et al*., 2010; Berkhout *et al*., 2011; Gizaw *et al*., 2017) probabilistic (Salasya & Stoorvogel, 2010; Pelcat *et al*., 2015; Nazari *et al*., 2017) on farm and expert based methods (Musa, Peters & Ahmed, 2006; Nolan *et al*., 2009; Herrero *et al*., 2014). Reported review (Nyambo, Luhanga & Yonah 2019) recommended a combination of methods to properly characterize evolving systems such as the smallholder farmers.

2.2.1 Characterizing farmers for evolvement studies

Statistical models have been used to understand evolvement and dynamics of smallholder dairy farms. For example; French, Tyrer and Hirst (2001) used a statistical modelling approach to understand mortality, morbidity and productivity of smallholder dairy farms. In the study, Cox proportional – hazards model was used as part of survival analysis in statistics. The proportional hazards model relates the time taken for an event to happen to some covariates that can be associated to that quantity of time. Frequency and determinants of mortality, and demographic trends were used as their specific areas for modelling. Having the estimates of death rates in each cattle age category, the authors carried a simulation to find out the projected growth. The study assessed a number of factors affecting life span of cows (example age, sex and breed); and death rates were given based on seasons. However, an assumption that no other animal was introduced into the herd is contrary to evolvement of dairy farms in the real world. In the real world, survival of a herd is affected by other persistent factors directly or indirectly but with high impact such as availability of health services, feeding systems, breeding services and access to inputs. Therefore, the main drawback of the study done is that it did not consider the changing environmental and farming-system factors that affect herd survival. This draw back could be resulting from lack of data on her dynamics, that is influx and outflux of animals from the herd. Being an external factor to the animal parameters featured in the study, incoming

animals may have a significant impact on the survival of existing ones with respect to disease introduction and competition of resources.

In another study, spatial variation in environment and socio-economic conditions were considered to characterize farming systems that could be used to generalize systems in a big region. This technique is demonstrated by Van de Steeg *et al*. (2010) who used classification algorithm and regression models to characterize farming systems found in Kenya highlands. Parameters used for classification were: area under cultivation of food and cash crops, milk production, and usage of fertilizers. Classes of farms were formed and their variability were explained by location factors and household characteristics; by fitting them in a Logistic regression model. The authors' consideration for spatial parameters was influenced by gaps produced in other studies on land use and land cover changes which did not introduce spatial factors into the modelling. Although some studies which detail use of spatial data, land cover and human population existed, these parameters could not fit for a more localized and specific characterization such as the smallholder dairy farm. After having variable farming systems, given location and household characteristics the likelihood of finding a particular farming system in the area was estimated. Expert based classification was used to validate results given by the statistical method. A confusion matrix was used to compare the field validation data against the estimates done by the statistical model.

Characterizing smallholder farmers is crucial to better depict their evolvement, in addition to better service provision and policy support (Herrero *et al*., 2014). In another research, expert based classification of farmers was done by Herrero *et al*. (2014) using predefined factors. The results of the classification were validated by using a hierarchical cluster analysis same results were revealed. Household features were modelled by using a Linear Programing (LP) algorithm aiming at maximizing the farms' gross gain. The LP model was adapted for multi-time period modelling in which results from the annual optimization of the household gains were used as inputs for next year's run.

It is therefore acknowledged that, characterization can be done by the methods in the literature with some improvements proposed in Nyambo *et al*. (2019). Similarly, evolvement patterns can be studied as suggested in previous studies but, there is need to deal with the complexity and inclusion of all factors pertaining to the growth and development of smallholder dairy farmers. Presented methods for characterization of smallholder farmers have been done with a specific focus; covering certain parameters in production. This is observed as a huge limitation

in the studies since the entire heterogeneity of farming systems is not covered. Some studies highlight on the complexities involved in studying dynamic real world systems, one of them being the failure to capture real world heterogeneity (Nolan *et al*., 2009). One factor pertaining to such limitations is the lack of a framework for data collection that can guide the capturing of all parameters in production for a holistic characterization of farms. For the case of smallholder dairy farmers, evolvement is shown to be tied to key farming strategies that each individual farmer opts in order to maximize their milk yield. However, there could be a chain of decision rules that are related to the farmers' options but cannot be well captured in the studies due to lack of data. Availability of a data collection framework will result into a holistic data sets that can yield highly accurate models that are developed to reduce time and financial investments that farmers are facing in quest for increasing milk yield (Mbwambo *et al*., 2017).

2.3 Agent – based modelling of agricultural systems

In this section, Agent-based modelling of farming systems is reviewed. Modelling goals, approach and data used were the key informants towards understanding how previous research has been done in the area of livestock farming. Significant body of literature indicate that feeding systems (especially grazing) has been widely studied with respect to its effect on livestock productivity and upkeep of vegetation cover. Effects of various policy implementations to individual farmers is equally represented by use of agent-based modelling and simulation.

The work of Macmillan and Huang (2008) made use of agent-based modelling to model and simulate evolvement of a primitive agriculture society based on a single settlement with heterogeneous landscape that supports agriculture. Agents in this implementation evolved by having production and consumption plans; in which an agent would have inheritors if it produces enough to have surplus after selling and consuming. Agents in the setup needed to learn from their experiences to improve their planning. This knowledge acquisition and transfer (inheritance) presents how the real world evolves. The weakness here is that, the evolvement is presented to cover farmers only whilst, in the real world, the given scenario is affected by a number of actors including humans or systems that support agriculture. However, the authors recommend that inclusion of all entities as in the real world would result into a highly complex model.

In another study, an overview of computational modelling that was meant to address situations in which modelling assumptions are based on fixed neighbourhood conditions was presented (Nolan *et al*., 2009). Having more realistic and sophisticated models brings in the challenge of solving them analytically. This challenge calls in what the authors called computational economics in which, complex realistic and sophisticated models need to go through numeric optimization (rationalization) and use of simulation methods in which agents are highly heterogeneous and the whole system is out of equilibrium. Without use of methods such as agent-based modelling, demand and supply in agriculture-oriented systems may reach an equilibrium since all neighbourhood conditions are fixed. The reality is that, spatial relationships among producers and consumers may influence their demand and supply functions. Therefore, the challenge of model complexity resulting from heterogeneous nature of the environment can be mitigated by rationalization, where by, the agents will only adopt association rules which will increase their utility values (Nolan *et al*., 2009; Faliszewski & Rother, 2016).

An agent-based model simulation of long term climate-livestock and vegetation interactions on communal rangelands done by Hahn explains possible outcomes from overgrazing with climate playing an important role in vegetation growth (Hahn *et al*., 2005). The model was developed in order to find critical conditions that can occur in communal rangelands and to suggest other livestock management strategies to farmers if need be. Livestock events in the agent-based model were given as regression equations. To best reflect livestock growth in the real world, age-based categories were implemented for the livestock life cycles.

Similar work in rangeland management is reported by Fust and Schlecht (2018) where the model integrated movement and feeding metabolism of domesticated ruminants. Main goal in the model was to assess the potential of adaptive livestock production in a highly dynamic, heterogeneous and semi-arid rangeland. Forage selection was highly based on quality and spatial distribution. Water sources were also modelled as individual agents that are determined by climatic conditions. The livestock productivity was modelled in view of forage consumption, conversion of the forage into energy which defined herd fitness.

Productivity of the Ankole-Friesian cattle has been modelled in stochastic simulation model of Ankole pastoral production system by Mulindwa *et al*. (2011). A stochastic compartmental model was developed, and key components were: forage production, herd structure dynamics,

and gross margins. Milk production was a sub component in gross margin. Not basing on real data, model equations were used to define livestock production.

Understanding livestock farmers' behaviour and adaptation to various strategies has been well studied using real data. Work done by Schilling *et al*. (2012) described an agent-based model for pastoral farmers' decision and behaviour in response to changes in their operating environment. The main goal was to study production intensity based on the farmers' choices. In the model, a farmer can learn from fellow farmers and adopt those practices that are delivering better outcomes. At the end, it was identified whether farmers' networks have a big effect in small scale or in large scale productivity. Real farm data were used in the model development and simulations. Individual farmers modelling based on real data is also reported by Oudendag, Hoogendoorn and Jongeneel (2014), where response to policy changes was studied with respect to farming intensity.

Mack and Huber (2017) conducted a study on farm compliance costs and Nitrogen surplus reduction of mixed dairy farms under grassland-based feeding. Animal and land use activities were simulated under a scenario that all farmers accepted the grassland feeding system. Regression models were then used to predict the reduction of Nitrogen emissions for various marginal costs compliance.

Also, noted is that livestock grazing and feeding management is well researched with use of agent-based modelling. Some agent-based modelling works in farming have lacked validation based on real farm data. Although such models might have a wide range of applicability, challenges may arise when considerations are put on actual situations of the farmers and their dynamics. Farmers from the same place are not necessarily the same or facing similar constraints. Appreciating the distribution and categories of the farmers together with their attributes is important towards realistic evolvement models. The progresses reported by Schilling *et al*. (2012), Oudendag, Hoogendoorn and Jongeneel (2014) and Mack and Huber (2017) represents how well real farm data can be used. However, farmers' learning dynamics and adaptation to better practices is well presented in Schilling *et al*. (2012). With the goal of increasing milk yield, this research therefore studied how the farmers can learn from each other, adopt to better practices without infringing their social-economic status. In addition, agent-based modelling has been applied to study the effect of various farming practices in milk yield.

Unlike other computer modelling approaches that focus of systems behaviour, agent-based modelling focuses on individuals' behaviour and their effects to the system being studied (Shiflet & Shiflet, 2014). Generalization cannot be assumed in agent-based modelling since individual agent behaviour and attributes will have a great influence on model results. This imply that, in adoption of agent-based modelling for the case of dairy farming the differences such as in cattle breeds and production environments will have a high influence on the model accuracy and applicability. Therefore, a huge limitation towards adoption of agent-based modelling for dairy farming is the presence of multiple breeds of cattle with varying behaviours which may hinder a generalized study.

2.4 Related research on agent – based modelling

The agent-based model for supply chain management presents an example of agent-based modelling; specifically, multi-agents. Fox *et al*. (2001) defined an agent-oriented supply chain management in which different agents were defined to handle activities in the supply chain. Speech-act-based communication in the style of Knowledge Query and Manipulation Language (KQML) was used for the agents' communication language. Behaviours, activities and conversation states in the system were finite. The Markov Decision Process framework was used to develop conversation plans for each agent. The finite nature of states in the system allowed Discrete-state Markov decision process to be used for agents' decision making.

The agriculture sector is another field in which the use of agent-based modelling is succeeding. Macmillan and Huang (2008) presented agent-based modelling of agricultural societies in which the agents were scattered in a settlement and each had decision making capabilities to fulfil their production and consumption plans. The presented work relied and was built on a foundation made by Epstein and Axtell (1996), in which agriculture societies were gatherers and did not practice a settlement system. The research was advanced by Macmillan and Huang (2008) by introducing production and consumption in a single settlement.

Agent-Based Modelling (ABM) has played a significant role in integrated modelling of systems that make use of multiple models. An example of this is an integrated assessment modelling in land, water use and governance. Presented works by Therond *et al*. (2014) and Nolan *et al*. (2009) describe how various models could be integrated into one platform using ABM instead of separate modelling for the cases of water use and governance, for farming activities among others. MAELIA framework is a water use and governance platform

developed by using multi-agents modelling and simulation. The MAELIA framework was presented by Therond *et al*. (2014) and involved an integrated design which captured a number of separate models in water use and governance for river basins. In another research, Filatova (2013) shows that, through ABM, it is possible to define agents that produce effects on land such as farmers, instead of focusing on the effects only as done in traditional modelling. From the highlighted examples, the ability to integrate all the models and their respective parameters into one platform signifies the power and suitability of using ABM in modelling such systems.

2.5 Conclusion

The review of previous research on characterization of smallholder farming systems has resulted into identification of key research gaps as follows:

(i) Clustering inconsistency justifying unstable clusters when one clustering algorithm is used. Use of different clustering algorithm can yield different results; application of multiple clustering algorithms can help researchers identify the more robust clustering model.

(ii) Cluster validation approach. Since cluster analysis result into sub-grouping of datasets into homogenous units, a validation that these clusters can be used to predict productivity of farms is essential.

Previous studies on the use of agent-based modelling and simulation to study farming systems has also yielded several gaps that have been covered in this research. The identified gaps were:

(i) Use of real farm data in model development and simulations. Application of real data in simulation of the real-world scenarios may lower uncertainties that may arise when the models are subjected in real-world evaluation. With real data, models can also be developed in a more holistic manner covering important factors as portrayed in the real-world.

(ii) Inclusion of internal and external farmer dynamics, and farmer disaggregation. The production potential of a dairy farm may be hindered by service providers e.g. breeding, health and water sources to mention few, there is need to include these external factors in addition to the animal specific as highlighted in previous research works. Also, modelling the farms as separate farm types with different characteristics is important to avoid aggregation because not all farmers are facing similar constraints.

CHAPTER THREE

MATERIALS AND METHODS

The advancement and profitability of dairy farmers is projected to be fully dependent on Artificial Intelligence by 2067 (Britt *et al*., 2018). To achieve a significant level of farm profitability, previous and current data need to be studied to inform next stages in advancing the dairy sector. This research has applied artificial intelligence approaches to demonstrate a characterization and evolvement modelling of smallholder dairy farmers with a main goal of increasing milk yield. To better understand the evolvement of the smallholder dairy system, (that is how milk yield can be increased over time), three stages were considered in this research.

The first stage involved cluster analysis to derive and describe farm types. Cluster analysis is an unsupervised learning approach which groups entities into homogenous groups. The end results of a cluster analysis are groups of entities which have high intra-similarity or intactness. This approach was used in this research to categorize smallholder dairy farmers from a baseline data set into homogenous groups based on their farming practices and dairy farming intensity. The subsequent clusters of farmers were referred to as farm types.

The second stage involved use of an association rules mining algorithm to understand the practices of the majority of farmers within the farm types. Association rules mining is another approach that uses unsupervised learning to study frequent patterns and relationship among entities in a data set. In this research, the approach has been used to derive frequent items as well as identify relationship patterns among the variables involved within dairy farm types in order to have a robust characterization of smallholder dairy farm types. The robust characterization yielded important variables that should be considered in order to increase milk yield.

From the first and second stages, factors that determine evolvement (based on increasing milk yield) for the farmers within the farm types were identified. The third stage used Agent-Based Modelling and Simulation (ABMS) to study the influence of the evolvement determinants on actual milk yield realized as well as use of farmer peer to peer learning to increase milk yield. The agent-based models considered the farm types and the determinants for evolvement to study how can a farmer move from low yield to medium and high milk yield. Simulation scenarios were used to study the impact of each evolvement determinant. The developed agent-

based model had two main components: the generation of a knowledge rich society of smallholder dairy farmers, and improving milk yield. Figure 2 demonstrates the conceptual framework for the modelling.

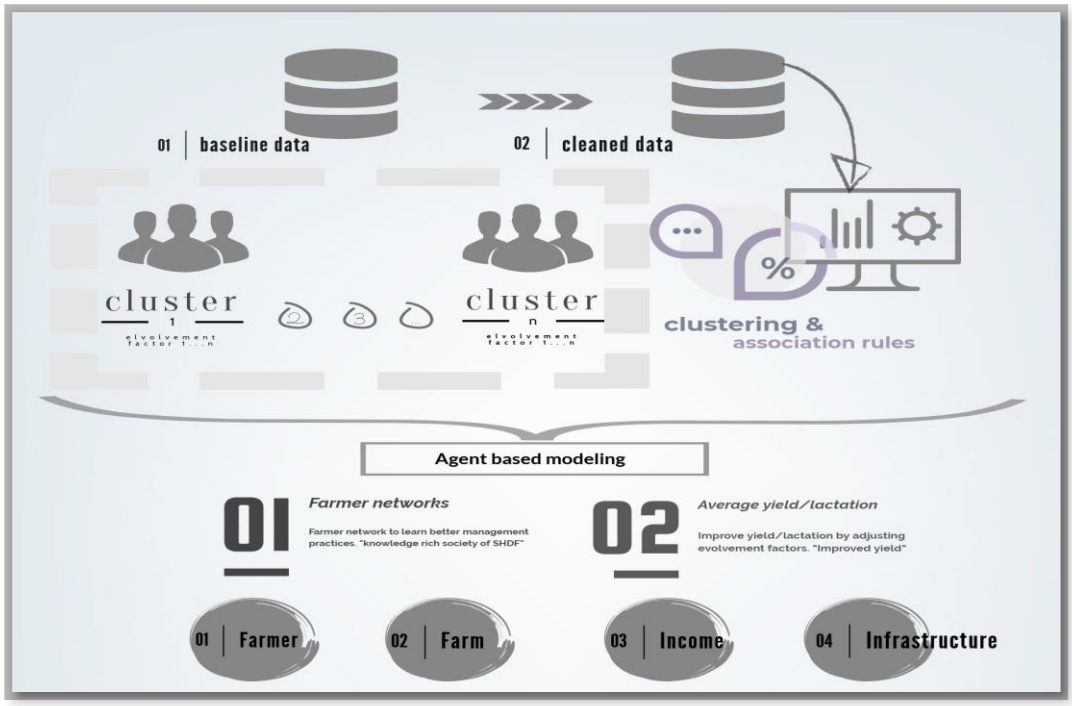


Figure 2: Data analysis conceptual framework

3.1 Data set preparation and features selection

The PEARL[1] Project collected data in four Eastern Africa countries (Ethiopia, Kenya, Tanzania and Uganda). The selection of Tanzania and Ethiopia as case studies for this Thesis was informed by both having nascent dairy value chains with low numbers of crossbred and pedigree dairy cattle. Despite Ethiopia being the top producer of milk in Africa, majority of the milk comes from indigenous cattle while Tanzania has less than one million crossbred and pedigree dairy cattle , which accounts for less than 2% of the total cattle population (Tanzania, 2016). Considering the fact that smallholder dairy systems across East Africa are similar, this study sought to establish evolvement factors attached to the two countries and for existing farm types within the countries to better inform policy making. The used methodology in this research can be applied in any case study other than Tanzania and Ethiopia. Nonetheless, other

[1] Program for Emerging Agricultural Research Leaders - Developing a framework for decision support tools to optimize smallholder dairy productivity in East Africa.

areas of application of the methodology can range from studying education systems, water resource management and health systems to mention a few.

3.1.1 Data preparation

Data was collected from June 2015 to June 2016 in Ethiopia and Tanzania. The total number of households surveyed was 3500 for Tanzania and 4679 for Ethiopia (Fig. 3). A digitized questionnaire was used in the data collection activity and were implemented by using Open Data Kit (Hartung, Lerer & Tseng, 2010). However, the data collection activity was not part of this research as the data sets were acquired as secondary data. Data quality checks included removal of outliers and erroneous data such as negative values, questionnaires whose total collection time was below a defined threshold (16 min), and data collected at night (survey start time beyond 7pm). Collected data included farm demographics, herd structure and cattle dynamics, feeding systems, diseases and health management, breeding services, milk production, marketing systems and farmer groups.

3.1.2 Feature selection

More than 500 features were available for each surveyed household. The data cleaning process trimmed the data sets to 3317 and 4394 records for Tanzania and Ethiopia equivalent to 94.7% and 93.9%, respectively (Fig. 4). From a total of the 500 variables (features) available for analysis, features selection was done to identify relevant variables for inclusion in the research.

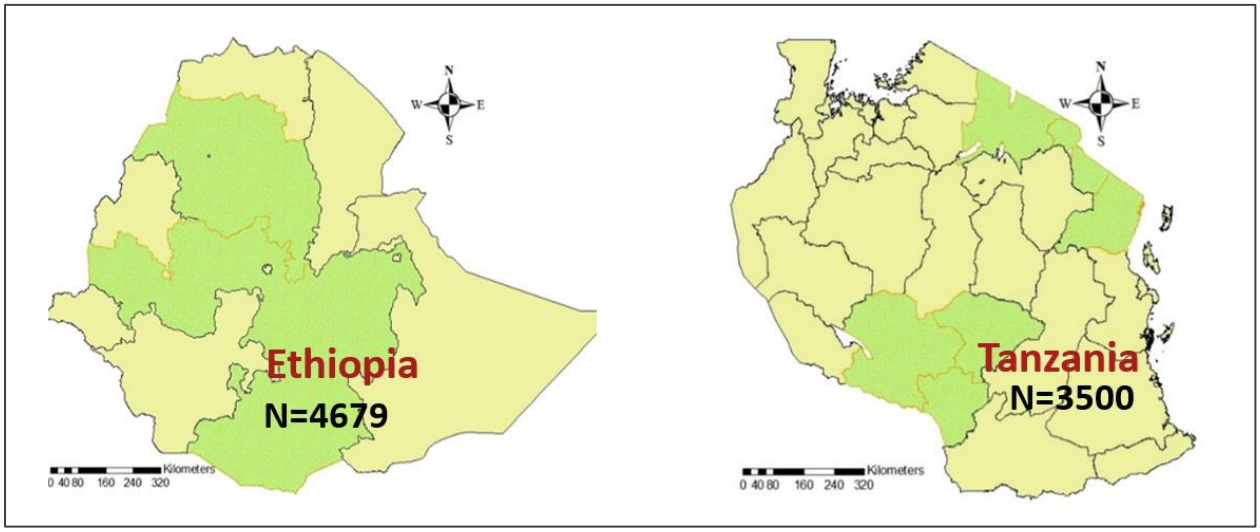


Figure 3: Data collection sites and number of households involved

An initial set of 46 features related to milk production, feeding systems, health, breeding, milk sales, extension services and farmer groups was identified in the two data sets as appropriate

for the research. In order to further identify the most unique features amongst the 46 variables, Principal Component Analysis (PCA) was undertaken to eliminate correlated variables. The top 21 features (based on the load score) with the lowest communality were then selected for further analysis. An additional 14 variables related to feeding systems and health management practices which are known to influence productivity in smallholder dairy farming were included based on expert domain knowledge. A total of 35 final features were included in the analysis.

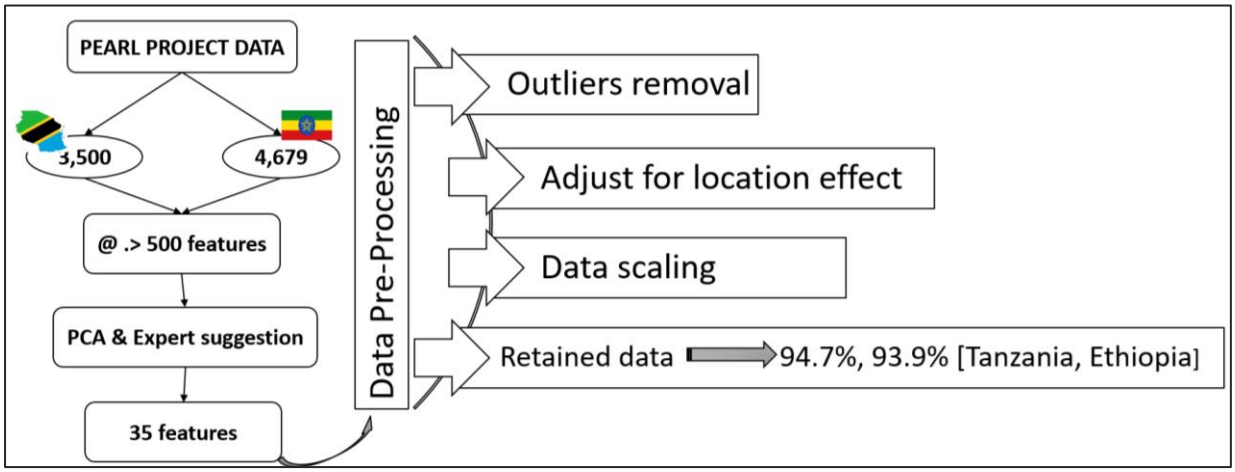


Figure 4: Data pre-processing steps

3.1.3 Data analysis

As a pre-requisite for clustering, missing values for continuous variables were identified and replaced with population means, while missing values for categorical values were replaced with the mode. To avoid bias causes by locational differences, the effect of location (study site) for each country was removed from the response variables by fitting a liner model $y = \mu + study_{site} + e$ (where y is the response variable, μ is the mean for the response variable, and e is the error term) and extracting adjusted values. As per clustering standards, quantitative variables were tested for normality and scaled to have a mean of zero and unit variance. Additionally, for each variable, outliers were identified as values above or below the bounds estimated using box plots. Outliers were removed to minimize bias and mis-clustering. Specifically, bias was minimized by applying the following inclusion criteria:

(i) The total number of cattle owned was restricted to a maximum of 50 per herd for Ethiopian farmers and a maximum of 30 per herd for Tanzanian farmers to reflect differential country-specific livestock densities, as well as to fit within the smallholder farmer definition (Guadu & Abebaw, 2016; Tanzania, 2016).

(ii) Some smallholder farmers held land holdings above 100 acres; all farmers with land holdings greater than 100 acres were removed. Smallholder definition was mostly based on cattle population.

(iii) The maximum amount of milk sold by smallholder farmers was restricted to 100 litres per day, based on expert domain knowledge of the herd sizes and yield per cow. Any additional milk was assumed to have been aggregated from other farmers for sale, and the farmers were excluded from the analysis.

(iv) It was assumed that, an extension officer could visit a farmer once each week. Any farmer who had more than 54 visits per year was considered an outlier.

The categorical data were converted into binary and numerical values (Huang, 1999; Huang, 1998; Kim *et al*., 2005).

3.2 Clustering

Three unsupervised learning algorithms, Fuzzy clustering, Self-Organizing Maps (SOM) and K-means, were used for cluster analysis. The combination of the approaches was meant to evaluate which one would provide more stable clusters.

The Self Organizing Maps (SOM) - Kohonen algorithm (Prayaga, 2001; Gelbard, Goldman & Spiegler, 2007) was used in clustering the data based on neural networks. Literature cites comparison of the SOM against hierarchical approaches (Chen & West, 1995; Mangiameli, 1996) whereby, the SOM is proved to be the best approach to cluster highly dimensional multivariate data sets with less/acceptable cluster dispersion (ratio of distance of nodes from the centroid and distance between centroids) and ability to produce accurate typologies as explained by Nazari *et al*. (2017) and Galluzzo (2015). The SOM algorithm calculates Euclidean distance by using Eqn. (1) and the best matching unit (BMU) satisfying Eqn. (2) (Galluzzo, 2015; Lähdesmäki, 2015).

$$Distance = \sqrt{\sum_{i=0}^{i=n}(vi - wi)^2} \qquad (1)$$

Where *v* and *w* are vectors in an n dimension Euclidean space relating to position of a member and neuron, respectively, and;

$$\forall n_i \in S: diff(n_{winnerweight}, v) \leq diff(n_{iweight}, v) \qquad (2)$$

Whereby, v is any new weight vector, $n_{winnerweight}$is the current weight of the winning neuron, and $n_{iweight}$ is a weight of any other i^{th} neuron on the map.

The K-means algorithm has been widely used in non-hierarchical clustering and characterizing smallholder dairy farms (Kuivanen *et al*., 2016; Dossa *et al*., 2011). Similar to SOMs, the algorithm uses Euclidean distance measures to estimate weights of data records. The algorithm is presented as Eqn. (3), with a segment of the Euclidean distance as in Eqn. (1).

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left|\left|x_i^j - c_j\right|\right|^2 \qquad (3)$$

Where $\left|\left|x_i^j - c_j\right|\right|^2$computes the Euclidean distance as in Equation 1; k = number of clusters, n= number of observations, j = minimum number of clusters, i= minimum number of observations, x_i = Euclidean vector for any i^{th} observation, c_j = cluster centre for any jth cluster.

Fuzzy analysis (fanny method) has been reported to have short convergence time and good measures for clusters separation (Basel & Nandi, 2015), which attracted its use in this study. Various methods based on fuzzy models have been used for cluster analysis (Salasya & Stoorvogel 2010; Gumma *et al*., 2011; Söderström *et al*., 2013; Journal *et al*., 2014). The fanny method adds a fuzzier and a membership value to the common K-means algorithm, Eqn. (3). In addition, the model uses the Dunn coefficient and a silhouette separation coefficient for assessing the solution fuzziness and inter-cluster cohesion, respectively. The general equation for fuzzy clustering (Bezdek, 1984) is given in Eqn. (4) and the Dunn definition of partitioning (Trauwaert, 1988) is given in (5).

$$J = \sum_{i=1}^{k}\sum_{j=1}^{n} U_{ij}^{m}\left|\left|x_i - c_j\right|\right|^2 \quad , \quad 1 \le m < \infty \qquad (4)$$

Where k = number of clusters, n = number of observations, i= minimum number of clusters, j= minimum number of observations, U_{ij}^{m} =membership coefficient, x_i = Euclidean vector for any i^{th} observation, c_j = cluster center for any jth cluster. Given Eqn. (4), the Dunn definition of partitioning is given by:

$$F_k(U) = \left(\frac{1}{n}\right)\sum_{i=1}^{k}\sum_{j=1}^{n} U_{ij}^{m} \qquad (5)$$

In the analysis, the number of groups (K) represented how many farm types (clusters) could be defined. Therefore, number of clusters were determined by using the Elbow method (where a bend or elbow in a graph shows a decline of within cluster sum of squares differences, as the number of clusters increase provides the best solution. Gap statistics and Silhouette separation coefficients were used to validate the results from the Elbow method (Kassambara, 2017). To the end, the Elbow method was found to be robust and was used for the rest of the analysis. The final clustering methods used were:

(i) Fanny for fuzzy clustering (Basel & Nandi, 2015)
(ii) SuperSOM with batch mode (Cottrell, Olteanu & Rossi, 2016)
(iii) Hartigan-Wong for K-means (Nidheesh, Abdul & Nazeer, 2017; Kazuaki, 2013).

Evaluation of the clustering algorithms was done by considering ranking consistency in the testing dataset, mean distance of observations from central nodes, mean silhouette separation coefficients. The effectiveness of the clusters in explaining variances in milk yield and sales was also sought as a means to validate the clusters fitness.

3.3 Association rules mining

The association rules mining was covered in the second objective where the target was to work on the derived clusters from the first objective to establish determinants for evolvement (milk yield) for the farmers in the two study cases. Use of association rules was sought to highlight on farming practices of the majority of the farmers in the derived clusters. Therefore, for each cluster of farmers, association rules mining was done to find frequent patterns.

Hence, mining for association rules proceeded with Apriori algorithm based on the following:

(i) Frequent items and patterns can be observed from the first parts of the data set and solutions are rare, so applying a depth-first search will require more time to generate all patterns.
(ii) The number of nodes in the data is finite even though the depth of the tree to be generated is infinite. Thus, depth-first search might fail to locate all children of nodes as it goes down.
(iii) The primary interest was to find out frequent patterns and then find out how the patterns are related by visualizing them. Considering the existing literature (Hunyadi, 2011;

Heaton, 2016) and the given assumptions, this research considered use of Apriori algorithm for association rules mining.

Data analysis was done in R software (Kabacoff, 2011). Four measures of rules interestingness were: support, count, confidence and lift. Support is a count of the number of times an item appears in the data set. Count is a number of observations in the data set supporting a particular association rule. Confidence is a measure of likelihood of occurrence of a rule; considering a rule [AUB] → C, confidence measure indicates the likelihood of this association rule by taking the ratio of the support of [[AUB] UC] to the support of [AUB] as represented in Eqn. (6). Lift is a measure of the deviation of the support of a whole rule from the support expected under independence, given the support of the antecedents and the consequent. From the example, $[AUB] \rightarrow C$, Lift is given by Eqn. (7).

Association rules were generated using the Arules package and visualized using the ArulesViz package, using graph and grouped matrix visualization (Hahsler & Chelluboina, 2011; Hahsler & Karpienko, 2017). In the graph and grouped matrix visualization, strong rules are indicated by higher lift values (strong colour intensity) and high support levels denoted by size of bubbles. Minimum support denotes the least number of times an item/pattern has appeared in the dataset, and the minimum confidence denotes the least likelihood of occurrence for the item in the Right Hand Side (RHS) upon occurrence of the item in the Left Hand side (LHS). During analysis, the support and confidence values were started at 0.1 and 0.5, respectively. These values were adjusted on each run to produce manageable number of rules (maximum 60) that can be easily visualized.

$$Confidence\ [AUB] \rightarrow C = \frac{Support\ [[AUB]UC]}{Support\ [AUB]} \quad (6)$$

Where; [AUB] is the antecedent and C is the consequent.

$$Lift = \frac{Support\ [[AUB]UC]}{Support\ [A].Support\ [B].Support\ [C]} \quad (7)$$

Where; $[AUB]$ is the antecedents and C is the consequent. An association rule is stronger if its lift value is high, meaning that; the frequent items are much stronger together than, when they are apart. Generally, good lift values must be greater or equal to 1.

3.3.1 Rationalization of association rules

The requirement to provide explanation is the main drive of rationalization. One can influence choice among possible alternatives to allow agents achieve their goals. As defined by Cherepanov *et al*. (2013), to rationalize a choice is to "find a subjectively appealing rationale that justifies that choice". Approaches to this can be either axiomatic or probabilistic. The first involves definition of desirable properties of rules then creation of a rule that will feature the desirable properties. In contrast, a probabilistic approach involves design of a rule that maximizes the probability of selecting the best choices. As stated by Elkind and Slinko (2016), objectively there is a best choice but as a result of errors in judgment, voters have different opinions. Likewise, in farming systems, specifically dairy farms, decisions to variants of management practices will be affected by the errors that agents make probably because of their prior knowledge. Therefore, the main question to ask is: *which among the existing practices are more likely (have a high probability) to maximize the agents' productivity in milk yield? Or which among the existing practices is milk yield more sensitive to?* A combination of the best practices will determine the probabilities of a farmer to either be in a low producer's stage or moving towards high producers, which is referred to as evolvement in this study. The factors that determine the probabilities of a farmer to increase its milk yield are therefore referred to as evolvement determinants. Figure 5 details two ideal primary clusters (low producers and high producers) whereby, a farmer in the low producers may be applying baseline practices found in the high producers so that its production may improve. This is supposed to be a gradual change done by changing and combining factors according to the found evolvement determinants.

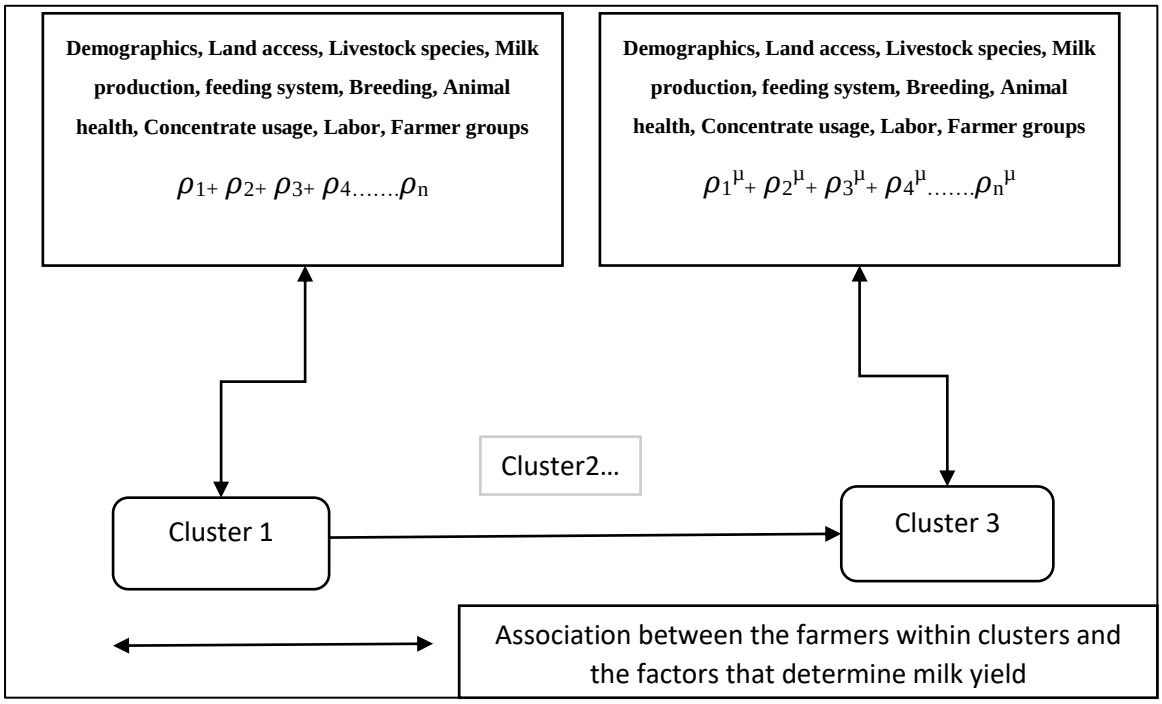


Figure 5: Moving from low to high milk yield by changing management practices

Considering Fig. 5, change of state (evolvement) for the farmer may be represented as: State1, state1.1, state1.2, state1.3…. state1.9 (state1 being a low producer and state1.9 being a high producer), for Fig. 5, state1 belongs to cluster1 while state1.9 belongs to cluster3.

Therefore:

$$\text{State1}|P = \rho_1 + \rho_2 + \rho_3 + \rho_4 \ldots\ldots \rho_n$$

$$\text{State1.1}|P = \rho_1^{\mu} + \rho_2 + \rho_3 + \rho_4 \ldots\ldots \rho_n$$

$$\text{State1.2}|P = \rho_1^{\mu} + \rho_2^{\mu} + \rho_3 + \rho_4 \ldots\ldots \rho_n$$

$$\text{State1.3}|P = \rho_1^{\mu} + \rho_2^{\mu} + \rho_3 + \rho_4^{\mu} + \rho_5^{\mu} \ldots\ldots \rho_n$$

$$\text{State1.9}|P = \rho_1^{\mu} + \rho_2^{\mu} + \rho_3^{\mu} + \rho_4^{\mu} + \rho_5^{\mu} \ldots\ldots \rho_n^{\mu}$$

For the clustering and association rules mining, SAS version 9.2 (SAS Institute Inc., Cary, NC, USA) and R software (Kabacoff, 2011) were used because of their powerful libraries which enabled scripting for data cleaning and implementation of the unsupervised learning algorithms on the datasets.

3.4 Agent – based modelling and simulation

Smallholder dairy farmers' characteristics from baseline data, determinants for milk yield derived through clustering and frequent pattern analysis, and literature review formed the key inputs to the agent-based modelling and simulation. Literature review provided dairy cows' input data (breed type, body weight, maximum milk level, energy requirements) and equations on how milk can be derived from Dry Matter Intake (DMI).

Modelling of the smallholder dairy farm takes a leaf from work done by Macmillan and Huang (2008) who modelled a primitive agricultural society, a society practicing settled agriculture with consumption plans, markets for their produce and evolves over time. The modelling was done with an assumption that all agents in the system are farmers practicing a settled agriculture. Therefore, these farmers were supposed to make various decisions for their evolvement and survival. For the smallholder dairy farms, an assumption is that, the farmers practice a crop-livestock farming and there are internal and external factors influencing their evolvement such as household characteristics, management practices and infrastructural services. For each agent, prior knowledge of what other agents in the system will do is unknown (Milch & Koller, 2000) so, decisions are made based on agents' knowledge base to maximize their milk yield.

Objective one and two of this research, provided key factors influencing milk yield for farmers in Tanzania and Ethiopia (referred to as evolvement determinants). Agent-based modelling proceeded by considering two questions:

(i) Can smallholder dairy farmers improve milk yield by learning good management practices from better milk producers within their farm types? It should be noted that, for Tanzania and Ethiopia, cluster analysis yielded various farm types together with key attributes of farmers within those farm types. The main task for the agent-based modelling was to study the possibilities of farmers learning from their peers the good practices that will help them improve milk yield, see Fig. 5.
(ii) How significant are the evolvement determinants (which were derived through clustering and association rules), can they be ranked in an order of significance?

To answer the above questions, two modelling perspectives were chosen: firstly, was to develop a knowledge rich society of smallholder dairy farmers through peer learning; and secondly was to study the effects of the evolvement determinants in milk yield (part 1 and 2,

respectively, see Fig. 2 for reference). Utilization of farmer networks to create a knowledge-rich society of smallholder dairy farmers has been adopted from Schilling *et al*. (2012) whose main goal was to study production intensity based on the farmers' choices. The authors demonstrated how a farmer can learn from fellow farmers and adopt those practices that are delivering better outcomes. At the end, it was identified whether farmers' networks have a big effect in small scale or in large scale productivity.

3.5 Agent – based model validation

Face validation approach was considered for all the models while involving several metrics as: correlation coefficients between simulated milk yield and actual milk yield from baseline data, models fitness by using total sum of square error, residual sum of square error and an r^2 that explained the proportion of variance in milk yield caused by the evolvement determinants. For the agent-based modelling, MARS (Multi – Agent Research and Simulation) platform and its associated cloud and local services (MARS LIFE and MARS DSL plugin on Eclipse IDE, respectively) were used. The model codes were written based on the MARS Design Specific Language (MARS DSL) implemented with Java and Python language constructs. With the MARS laboratory, modelling and simulation of multi-agents systems was done with highly flexible agents' abstraction, and a layered approach allowing the researcher to focus on one agent at a time (Christian *et al*., 2016). The choice for the MARS platform on model development and simulation was also attributed to the cloud-based model simulations, which created backups each time a model was run and saved the need to create local backups for each model run.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the main findings of the study with a discussion that elucidates the agreement and disagreement with previous research. The first and the second objective present main findings during characterization of smallholder dairy farmers in the selected case studies. While cluster analysis results were enriched with a cluster validation approach, association rules mining was enriched with frequent patterns that can be used to further characterize the derived farm types as per the cluster analysis. Results from the cluster and association rules analysis gave a conclusion on key determinants for increase in milk yield for the farmers in Ethiopia and Tanzania. Agent-based modelling and simulation then explained the differential influence of the determinants towards increase in milk yield.

4.2 Results

This sub-chapter describes the main findings of the study following the order of the research objectives and methodology. Subsequent activities were focused on identifying key evolvement determinants for smallholder dairy farmers in study cases. The determinants which were established from the first and second objectives, were further assessed and validated by using an agent-based model. The agent-based model established the differential influence of each determinant on milk yield maximization; main goal was to ease the farmers' trial and error attempts on deciding what management strategy to prioritize for increase in milk yield in limited resources. Moreover, the agent-based model was simulated to demonstrate the likelihood of farmers improving milk yield by learning better management strategies from their peers. The concept is hereby referred to as *creation of a knowledge rich society of smallholder dairy farmers.*

4.2.1 Features selected for the study

From features selection phase as detailed in Section 3.1, 35 features were selected as indicated in Table 1.

Table 1: Features used in the study

S/No	Feature Name	Type	Range
1	Exclusive grazing in dry season	Boolean	0(no) or 1(yes)
2	Exclusive grazing in rainy season	Boolean	0(no) or 1(yes)
3	Mainly grazing in dry season	Boolean	0(no) or 1(yes)
4	Mainly grazing in rainy season	Boolean	0(no) or 1(yes)
5	Mainly stall feed in dry season	Boolean	0(no) or 1(yes)
6	Mainly stall feed in rainy season	Boolean	0(no) or 1(yes)
7	Use of concentrates	Discrete	1 – 12 (months)
8	Watering frequency	Discrete	0 – 4
9	Distance to water source (km)	Continuous	0 – 15
10	Total land holding	Continuous	0 – 100
11	Area under cash cropping	Continuous	0 – 10
12	Area under food cropping	Continuous	0 – 83.25
13	Area under fodder production	Continuous	0 – 80
14	Area under grazing	Continuous	0 – 13
15	Number of employees	Discrete	1 – 10
16	Number of casual labours	Discrete	1 – 10
17	Vaccination frequency	Discrete	0 – 6
18	Deworming frequency	Discrete	0 – 5
19	Self-deworming service	Boolean	0(no) or 1(yes)
20	Membership in farmer groups	Discrete	0 – 5
21	Experience in dairy farming	Discrete	1 – 50
22	Years of schooling	Discrete	0 – 21
23	Preferred breeding method	Boolean	0 (bull) or 1(AI)
24	Distance to breeding service (km)	Continuous	0 – 100
25	Frequency of extension visits	Discrete	1 – 54
26	Herd size	Discrete	1 – 50
27	Number of milking cows	Discrete	1 – 20
28	Number of exotic cattle	Discrete	1 – 48
29	Number of sheep	Discrete	1 – 80
30	Peak milk production	Continuous	1 – 40
31	Amount of milk sold in bulk (Lt)	Continuous	1 – 100
32	Litres of milk sold (Lt)	Continuous	1 – 100
33	Distance to milk buyers (Km)	Continuous	1 – 37
34	Total crop sale	Continuous	0 – 21000 (Birr) 0 – 950000 (Tsh)
35	Distance to market (Km)	Continuous	0 – 8

4.2.2 Clustering: Households assignment into homogeneous clusters

A four-cluster solution was found to be optimal for the Ethiopia dataset and was used in the clustering models. The SOM and K-means algorithms clustered the Ethiopia dataset into four groups (Fig. 6), while the fuzzy model grouped the records into three clusters, with no members in the fourth cluster. Table 2 illustrate the cluster densities for each algorithm as applied on the Ethiopia data. For Tanzania, six clusters were well-defined based on the elbow method (Fig. 7) except for the fuzzy model. At K=6, the fuzzy model had highly fuzzy cluster memberships of 0.09 and 0.18 for each member. Such low membership values indicate an unstable cluster solution. Due to the low membership values for the households, the fuzzy model was discarded for the Tanzania data set. The cluster analysis proceeded with the K-means and Self-Organizing Maps (SOM) algorithms. Cluster densities associated with the six clusters are provided in Table 3.

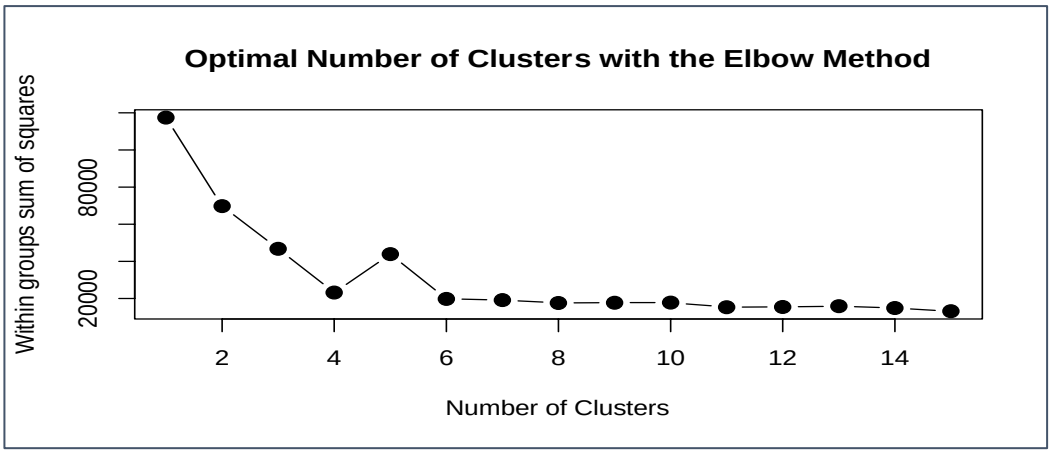


Figure 6: Graph showing four optimal clusters for the Ethiopia data set

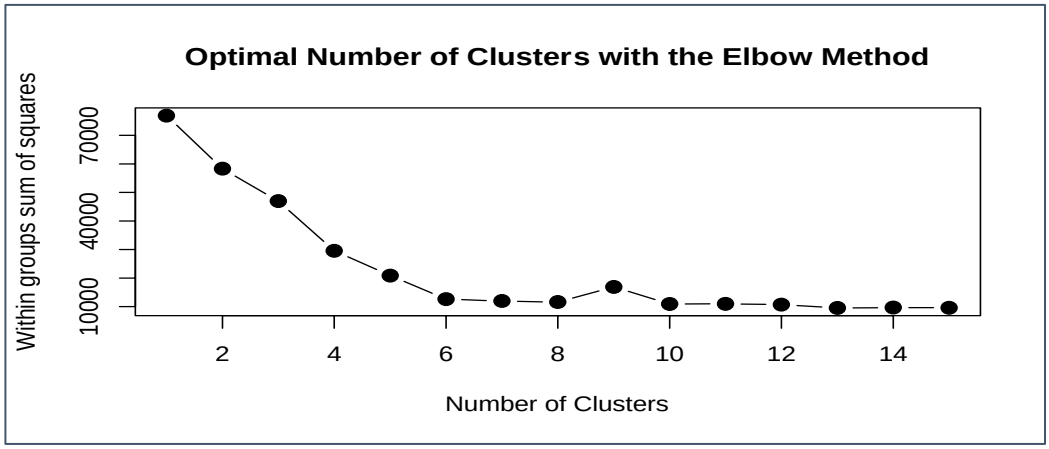


Figure 7: Graph showing six optimal clusters for the Tanzania data set

For the Ethiopia data, cluster densities given in Table 2 indicate there was one static cluster for both the K-means and SOM models (with the exact same number of members, 487). The number of members in the other clusters varied, indicating households being re-assigned to different clusters. Figures 8, 9 and 10 represent the cluster visualization for each algorithm in the Ethiopia dataset. Clusters derived using K-means were separated and had significant intra-cluster adhesion (Fig. 8), while Spatial distribution of SOM clusters (Fig. 9) indicated significant overlap between two of the 4 clusters (clusters in red). Clusters' densities for Tanzania are displayed in Table 3.

Table 2: Cluster densities for the Ethiopia data set

Cluster	K-means model	SOM model	Fuzzy model
1	342	487	2673
2	875	2084	411
3	2689	1217	1309
4	487	605	

Table 3: Cluster densities for the Tanzania data set

Cluster	K-means model	SOM model	Fuzzy model
1	811	1180	2506
2	452	952	811
3	374	203	
4	616	295	
5	372	516	
6	692	171	

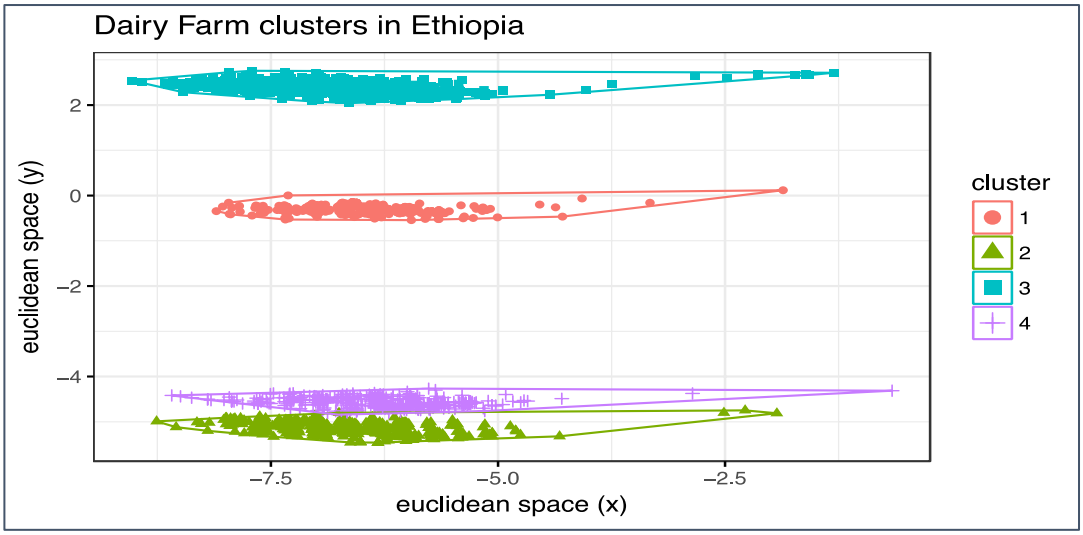


Figure 8: Household allocation using the K-means model for Ethiopia

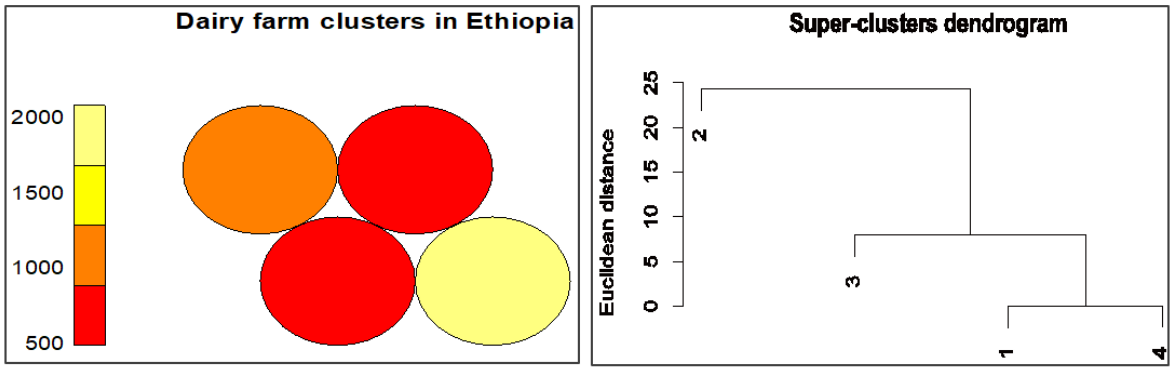


Figure 9a. Figure 9b.

Figure 9: Node counts for the SOM model for Ethiopia

Figure 9a and 9b is a heat map showing cluster densities and dendrogram, respectively, for the clusters' formation based on the SOM model. Figure 9a shows counts of households within clusters while Fig 9b indicates cluster relationship and separation. The numbers on the coloured plane indicate count of households in each cluster. Two clusters had an equal number of farmers (shown in red colour) and on the dendrogram these were categorized as clusters 1 and 4. These two clusters seemingly had few differentiating features since they originate from the same parent node. This phenomenon can also be observed in Fig. 8 for the K-means model (clusters 2 and 4). These clusters appear to be joined into one cluster in the fuzzy model (cluster 3 in Fig. 10). The fuzzy model resulted in 3 clusters, each with a significant number of outliers (Fig. 10). The outliers were more noticeable for cluster 2 than clusters 1 and 3.

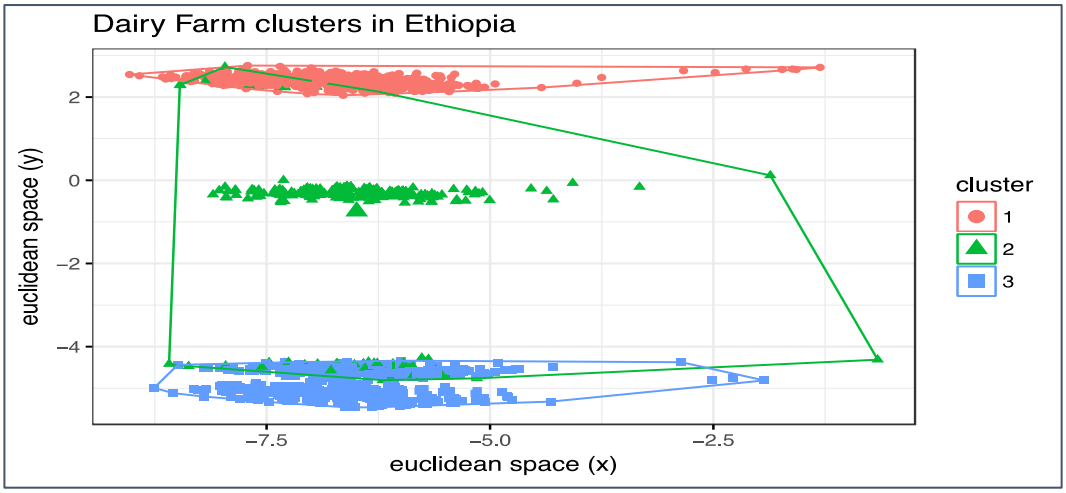


Figure 10: Household allocation using the fuzzy model for Ethiopia

Presence of the outliers and cluster overlap in the fuzzy model was supported by a low value of the Dunn coefficient (0.3014) which corresponds to a high level of fuzziness.

Based on the results obtained, the cluster composition parameters related to inter-cluster adhesion and intra-cluster cohesion indicated that, clusters from the k-means model were better separated (higher mean silhouette value) and more compact (lower mean distance from central node) than in the other models for Ethiopia (Table 4).

Table 4: Cluster composition parameters for Ethiopian households

Model	No. Clusters	Within sum of square	Mean distance from central nodes	Mean silhouette separation
K-means model	4	20758	0.74	0.66
SOM model	4	23178	0.92	0.51
Fuzzy model	3	21655	0.89	0.56

Based on the characteristics of the production clusters from K-means and Fuzzy models in Ethiopia, the K-means clusters were selected for further analysis in the research. Table 5 summarize the main characteristics of the Ethiopia's farm typologies founded on factors' loadings from the K-Means model. Results revealed some common patterns across all clusters which indicate some mutual characteristics of farmers in the country. For Ethiopia farmers, adherence to best health management practices especially deworming and vaccination together with cattle watering could be termed as general best practices. Summary statistics revealed that, at least 2910 farmers adhere to deworming and vaccination services whereby, a minimum of 52.28% of farmers in Addis Ababa do vaccinate their animals once per year ($p<0.001$); while

up to 44.07% deworm their animals twice a year in Asela Shed ($p<0.001$). Ethiopia farmers are constrained with small land holdings for crop farming resulting into low crop sales across majority of the clusters.

Table 5: Farm typologies based on the K-means cluster solution for Ethiopia

Cluster	Proportion	High loadings	Low loadings	Farm type
1	8%	Area under fodder production, grazing land, years of schooling	Number of milking cows, total cattle owned, number of exotics	Semi-intensive low production households
2	20%	Experience in dairy farming, cash cropping, frequency of visits by extension officers	Litres of milk sold, number of milking cows, peak milk production for the best cow	Commercially oriented mixed crop-dairy low production households
3	61%	Litres of milk sold, peak milk production for the best cow, total crop sales	Frequency of visits by extension officers, long distances to milk buyers, years of schooling	Commercially oriented mixed crop-dairy high production households
4	11%	Long distances to milk buyers, number of sheep, supplementary feeding	Food cropping, grazing land, long distances to breeding service providers	Market constrained, low production households

Clusters from the k-means model yielded four dairy farm types which are highly distinguished by milk yield and sales, and dairy herd size. *Semi-intensive low production households* are in cluster one forming 8% of the total households, they are typified by a mixed farming system with small herd sizes. *Commercially oriented mixed crop-dairy low production households* formed the second cluster and consisted of 20% of the households, typified by low dairy production in spite of high access to extension services. However, loading high on cash cropping indicates their dedication to commercial cropping rather than dairy. *Commercially oriented mixed crop-dairy high production households* formed the third cluster and had the highest number of households (61% of the total). This cluster represents dedicated farmers; whose performance needs to be enhanced by supplying them with extension support for

improved farm management. *Market constrained; low production households* formed the fourth cluster and consisted of 11% of the households. It is observed that, dairy market proximity can hinder farmers' motivation to dairy farming regardless of their closeness to improved breeding services.

For Tanzania, the mean silhouette separation coefficients were not significantly different (0.66 and 0.64 for K-means and SOM, respectively) as shown in Table 6. However, there was a tendency for the SOM to have better defined clusters given its lower within cluster sum of squares as well as lower mean distance from central node. The spatial distribution is illustrated in Figs. 11 and 12.

Table 6: Cluster composition parameters for Tanzania households

Model	No. Clusters	Within sum of square	Mean distance from central nodes	Mean silhouette separation
K-means model	6	12628	2.1	0.66
SOM model	6	11772	1.7	0.64

For the Tanzania clusters' separation and robustness can be observed through Figs. 11, 12. No significant difference could be observed with regards to the inter-cluster adhesion between K-means and SOM (Table 6).

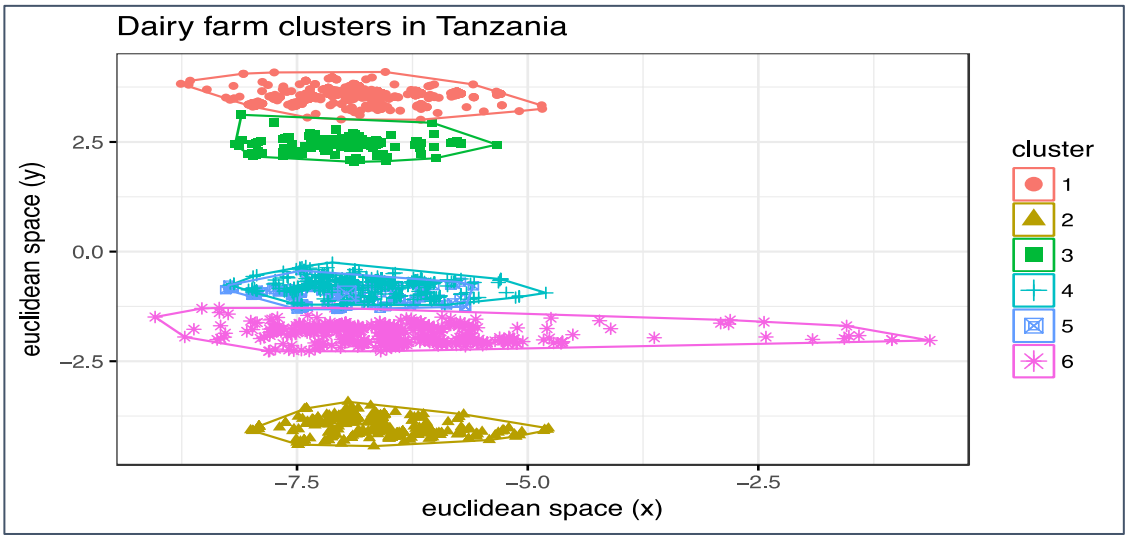


Figure 11: Household allocation using the K-means model for Tanzania

Figure 11 shows clusters visualization from the K-means model for Tanzania data set. Cluster 4 and 5 overlap and are in close proximity to cluster 6, indicating that they have few

differentiating characteristics. This overlapping is equally observed in the SOM model (Fig. 12).

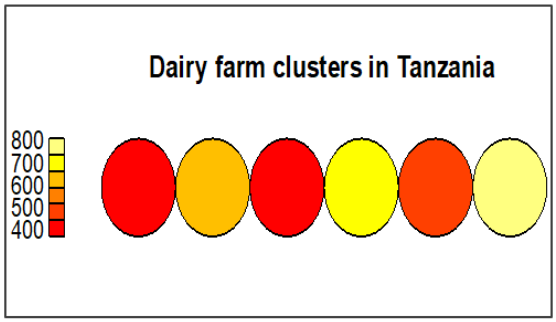


Figure 12a.

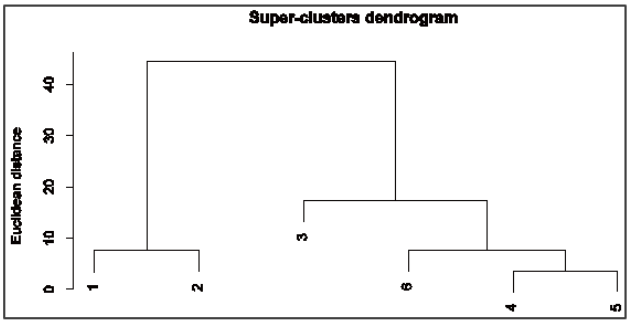


Figure 12b.

Figure 12: Node counts for the SOM model for Tanzania

The numbers on the coloured bar in Fig. 12a indicate densities of members in each cluster. There are only four well separated clusters based on density (from left: red, orange, yellow and light gold). However, the dendrogram (Fig. 12b) shows that three clusters, branching from the same node, which also are also seen as the overlapping clusters (clusters 4, 5, and 6) in the K-means plot (Fig. 11).

Clusters derived from the SOM model were characterized to identify similarities and differences, since SOM clustering results were more robust than the K-means clusters for the Tanzania case. Like in the Ethiopia case, factor loadings in the Tanzania clusters indicated that farmers had common characteristics significantly on higher loadings. Table 7 summarizes the characterization from SOM cluster solution for the Tanzania case. Results highlights that, generalizing features for Tanzania farmers were banked on adherence to best health management practices and cattle watering. Summary statistics revealed that, at least 2878 households indicated that they deworm and vaccinate their cattle. From that population, at least 37.42% deworm their cattle thrice per year ($p<0.0001$) and at least 82.13% vaccinate their animals once per year ($p<0.0001$).

Table 7: Farm topologies based on the SOM cluster solution for the Tanzania data set

Cluster	Proportion	High loadings	Low loadings	Farm type
1	36%	Vaccination frequency, frequency of watering, experience in dairy farming	Distance to buyers, total land size, area under fodder production	Non-commercial dairy production households
2	29%	Frequency of watering, area under fodder production, distance to breeding service providers	Area under grazing, years of schooling, sell of bulk milk	Small stock semi-intensive commercial dairy households with limited access to breeding services
3	6%	Frequency of watering, sell of bulk milk, vaccination frequency	Frequency of visits by extension officers, distance to breeding service providers, number of milking cows	commercially oriented and self-reliant high dairy production households
4	8.9%	Frequency of watering, vaccination frequency, area under fodder production	Area under grazing, years of schooling, sell of bulk milk	Semi-intensive commercially oriented medium production households
5	15%	Frequency of watering, litres of milk sold, vaccination frequency	Distance to markets, experience in dairy farming, total land size	Commercially oriented high production entrant households
6	5.1%	Distance to water sources, total cattle owned, frequency of watering	Total crop sales, distance to breeding service providers, experience in dairy farming	Commercially oriented mixed crop-dairy low production entrant households

Cluster one contained majority of the households (36%) consisting of *non-commercial dairy production households*. Cluster one farmers were characterized with high vaccination and watering frequency, many years of experience in dairy farming, small distances to milk buyers, small land sizes and small lands under fodder production. Cluster two which had 29% of the sampled households consisted of *Semi-intensive commercially oriented medium production households with limited access to breeding services*, and characterized with high frequencies of cattle watering, large areas under fodder production, long distances to improved breeding service providers, low grazing lands, few years of formal education, and low sales of bulk milk. Cluster three contained 6% of the households and consisted the *commercially oriented and self-reliant high dairy production households*. Cluster three was characterized with high frequencies of cattle watering and vaccination, high sales of bulk milk, low frequency of visits by extension officers, small distances to breeding service providers and small number of milking cows. Cluster four consisted of 8.9% of the households who categorized as *Semi-intensive commercially oriented medium production households*. Cluster four was characterized with high frequencies of watering and vaccination, large areas under fodder production, small grazing lands, and few years of schooling and low sales of bulk milk. 15% of the households formed cluster five which was categorized as *Commercially oriented high production entrant households*. Key characteristics of the farmers in cluster five were: high frequencies of watering, vaccination and milk sales, small distances to market, low experience in dairy farming and small land sizes. Cluster six comprised of 5.1% of the households who were categorized as *Commercially oriented mixed crop-dairy low production entrant households*. Cluster six was characterized with: long distances to water source, large herd sizes, low crop sales, short distances to breeding service providers and low experience in dairy farming.

4.2.3 Cluster validation: Membership re-ranking

Spearman ranking correlation was used to study the levels of household re-location for the training and testing data sets. Generally, the clustering models applied to the Ethiopia dataset indicated low membership re-location. Table 8 summarizes the results for Ethiopia where, despite a lower Akaike Information Criteria (AIC) estimate, the fuzzy model had the highest number of members re-allocated to other clusters (32%) compared to the K-means and SOM. The high correlation coefficients for SOM and K-Means indicate lower re-allocation of cluster members. In contrast, results from Tanzania indicated high re-ranking of cluster membership between training and testing datasets (Table 9).

Table 8: Cluster membership re-allocation for the Ethiopia cluster models

Model	AIC	Residual deviance	Ranking accuracy (r)
K-means model	102	2.7e^-2	0.85
SOM model	102	2.8e^-2	-0.88
Fuzzy model	68.09	9.35e^-2	0.68

Table 9: Cluster membership re-allocation for the Tanzania cluster models

Model	AIC	Residual deviance	Ranking accuracy (r)
K-means model	200	0.001	-0.21
SOM model	200	0.006	0.39

(i) Cluster robustness and fitness

In order to assess whether the clusters defined by the various algorithms reflect differences in production characteristics between households, we evaluated the variance accounted for by these cluster on milk yield and sales. For Ethiopia, total variance was 1.015 and 0.988 for milk yield and sales, respectively, while in Tanzania, the total variance was 1.076 and 1.09 for milk yield and sales, respectively. The differences between residual variances for two linear models (Eqns. 8 vs. 9 for Ethiopia and Eqns. 10 vs. 11 for Tanzania) were significant ($p < 0.00001$). Results show that for Ethiopia data, the fuzzy model clusters accounted for 89% and 70% of the total variance in milk yield and milk sales, respectively. On the other hand, the K-means clusters accounted for 71% and 65% of the total variation in milk yield and milk sales, respectively. Tables 10 - 11 summarize the proportion of variances accounted for by the clusters for each clustering model.

$$y_i = x_e * \gamma_e + c_e + e_e \quad (8)$$

$$y_i = x_e * \gamma_e + e_e \quad (9)$$

$$y_i = x_t * \gamma_t + l_t + \sigma_t + c_t + e_t \quad (10)$$

$$y_i = x_t * \gamma_t + l_t + \sigma_t + e_t \quad (11)$$

Where, for the Ethiopia models: c_e is cluster of production, e_e is the error term, x_e is experience in dairy farming, and γ_e is years of schooling. For the Tanzania models: c_t is cluster of production, e_t is the error term, x_t is experience in dairy farming, γ_t is years of schooling, l_t is total land size and σ_t is area under fodder production.

Table 10: Proportion of variance accounted for by cluster of production in Ethiopia

	Fitted model	**Total Variance***	**Residual variance**	**-2log likelihood**	**P value**	**Variance**
K-means		Milk yield				
	Model without cluster	1.015	0.239	1867.4	<0.00001	73%
	Model with cluster		0.977	3718.4		
		Milk sales				
	Model without cluster	0.988	0.222	1770.1	<0.00001	54%
	Model with cluster		0.76	3388.6		
SOM		Milk yield				
	Model without cluster	1.015	0.283	2091.8	<0.00001	68%
	Model with cluster		0.977	3718.4		
		Milk sales				
	Model without cluster	0.988	0.258	1969.8	<0.00001	51%
	Model with cluster		0.76	3388.6		
Fuzzy		Milk yield				
	Model without cluster	1.015	0.074	337	<0.00001	89%
	Model with cluster		0.977	3718.4		
		Milk sales				
	Model without cluster	0.988	0.073	319.4	<0.00001	70%
	Model with cluster		0.76	3388.6		

*For the data that was scaled to have unit variance and mean of zero

Table 11: Proportion of variances accounted for by cluster of production in Tanzania

	Fitted model	Total variance*	Residual Variance	-2log likelihood	P value	Variance
K-means		Milk yield				
	Model without cluster	1.076	0.0027	-2981	<0.00001	71%
	Model with cluster		0.771	2584.2		
		Milk sales				
	Model without cluster	1.09	0.018	-1084.3	<0.00001	65%
	Model with cluster		0.723	2520		
SOM		Milk yield				
	Model without cluster	1.076	0.294	1633	<0.00001	44%
	Model with cluster		0.771	2584.2		
		Milk sales				
	Model without cluster	1.09	0.228	1381.6	<0.00001	45%
	Model with cluster		0.723	2520.2		

* For the data that was scaled to have unit variance and mean of zero.

The need to further investigate the farm type characteristics was determined in order to observe whether there were associations among variables within the farm types. Ultimate goal was to have robust characterization for the farm types to better inform agent-based modelling and simulation of the important determinants for increase in milk yield. To this end, association rules mining based on the Apriori algorithm (as detailed in Section 3.2) was undertaken.

4.2.4 Association rules mining

Clustering creates homogeneous groups and show their characteristics based on high or low factor loadings. Sometimes there are hidden attributes that cannot be identified through clustering and can only be identified by frequent pattern analysis and association mining. For example, frequent pattern analysis can reveal that majority of farmers in a farm type water their cattle twice a day. This attribute when given in a cluster solution will just indicate a high factor loading without the details on frequency, number of farmers who have shown that attribute and

even the confidence that the attribute is likely to be found when farmers of similar practices are studied. Therefore, through frequent pattern and association rules mining, support and confidence of the characteristics is given and this can be used in future studies for farmers bearing similar practices.

The association rules were derived based on different values for minimum support and confidence to capture manageable number of rules (<100). In this section, outstanding associations for the farm types have been presented by using graph of rules and grouped matrix, size of bubbles indicates support level and colour intensity indicate lift values. The graph of rules highlights key associations among the frequent items while, the grouped matrix highlights significance of the frequent items by showing number of rules that the item has formed in combination with other frequent item sets.

(i) Ethiopia farm types

A total of four farm types were determined for Ethiopia. These are described below.

Farm type one (semi-intensive low production households)

Farmers in farm type one practice stall feeding system and have medium milk yield and sales. High lift values (>1) were observed in rules involving: medium milk production for best cows, low amount of milk reserved for home consumption, no formal training in dairy care and handling, medium amount of milk sold, less than one-acre land holdings, no responses in tick control frequencies, use of concentrates limited to three months, and zero distance to water sources (Figs. 13 & 14). Clustering results had high loadings on area under fodder production, grazing land and years of schooling. Small herd sizes were also attributed to this farm type during clustering. Given that farmers in type are characterized with medium milk yield, practice stall feeding, have grazing lands and maintain small herd sizes, their production system is assumed to be semi-intensive low production households. Therefore, it can be demonstrated at 65% confidence level that, farmers in type one from Ethiopia have low dairy production for subsistence. It can be ascertained that, at least 65% of the times that semi-intensive feeding, small land holding, medium yield and small herd sizes are found together for farmers in Ethiopia, the farmer would be a semi-intensive low producer for subsistence.

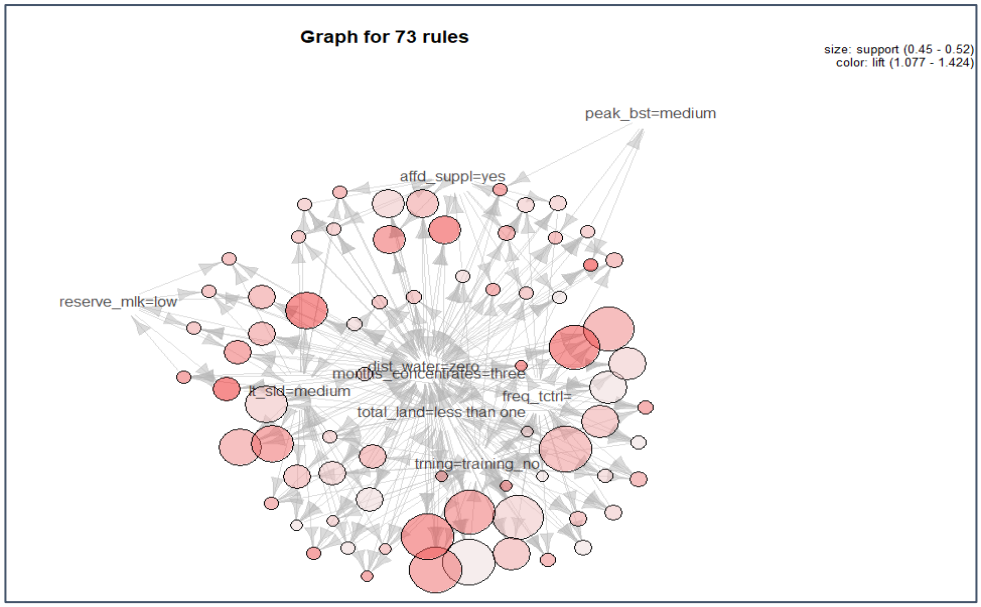


Figure 13: Graph of rules for farm type one in Ethiopia

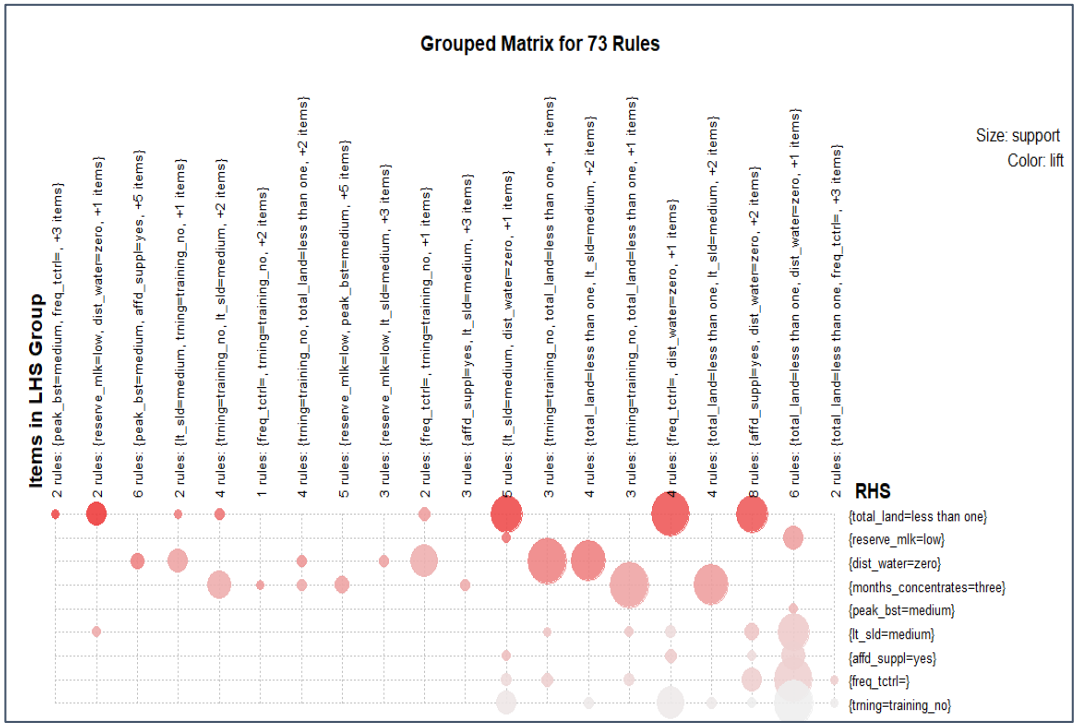


Figure 14: Grouped rules for farm type one in Ethiopia

Farm type two (commercially oriented mixed crop-dairy low production households)

Farmers in farm type two have medium milk yield, small herd sizes and water their cattle once daily. High lift values were observed in rules involving Artificial Insemination (AI) for breeding, medium peak milk production, no use of purchased fodder, two milking cows and

preference on cross breeds (Figs. 15 & 16). Clustering results indicated high experience in dairy farming, cash cropping, frequency of visits by extension officers, low milk yield and sales, and small number of milking cows. Given that the farmers practice improved breeding, have adequate extension service, are experienced in dairy production, and yet have low milk yield and sales their production system is assumed to be commercially oriented mixed crop-dairy low production households. It can be demonstrated, at least at 60% confidence level, farmers in type two have been practicing dairy farming for subsistence. At least 60% of the times that cash cropping, high experience in dairy farming, extension services, medium milk yield, small herd sizes and use of AI are seen together, the farmer would be a commercial oriented low producer.

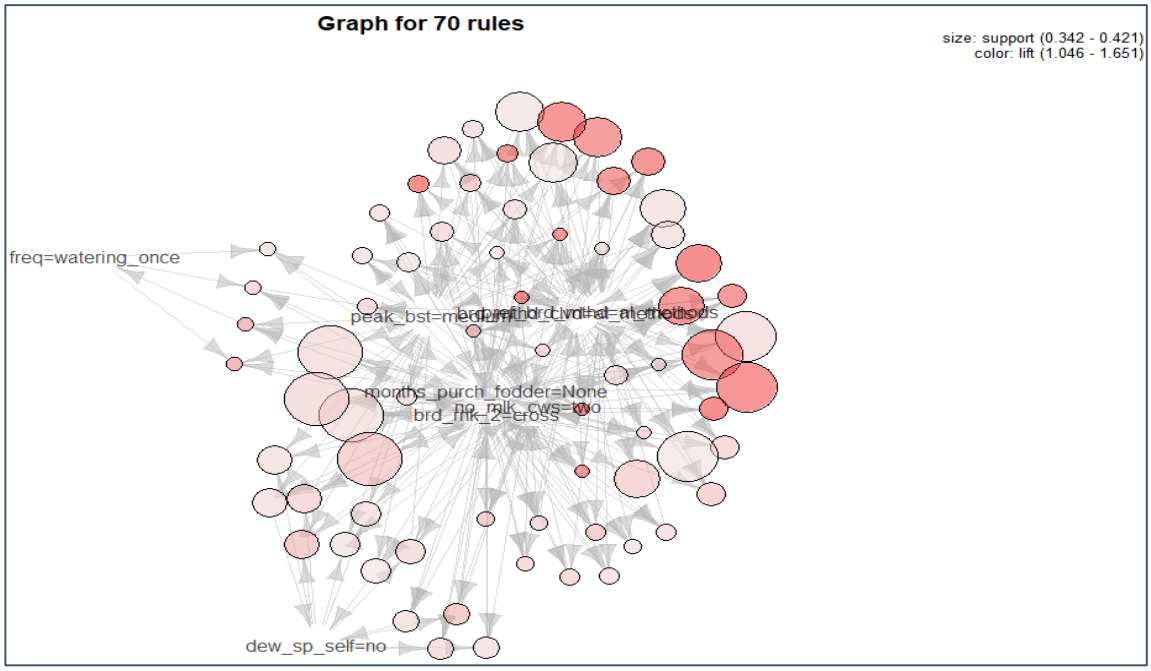


Figure 15: Graph of rules for farm type two in Ethiopia

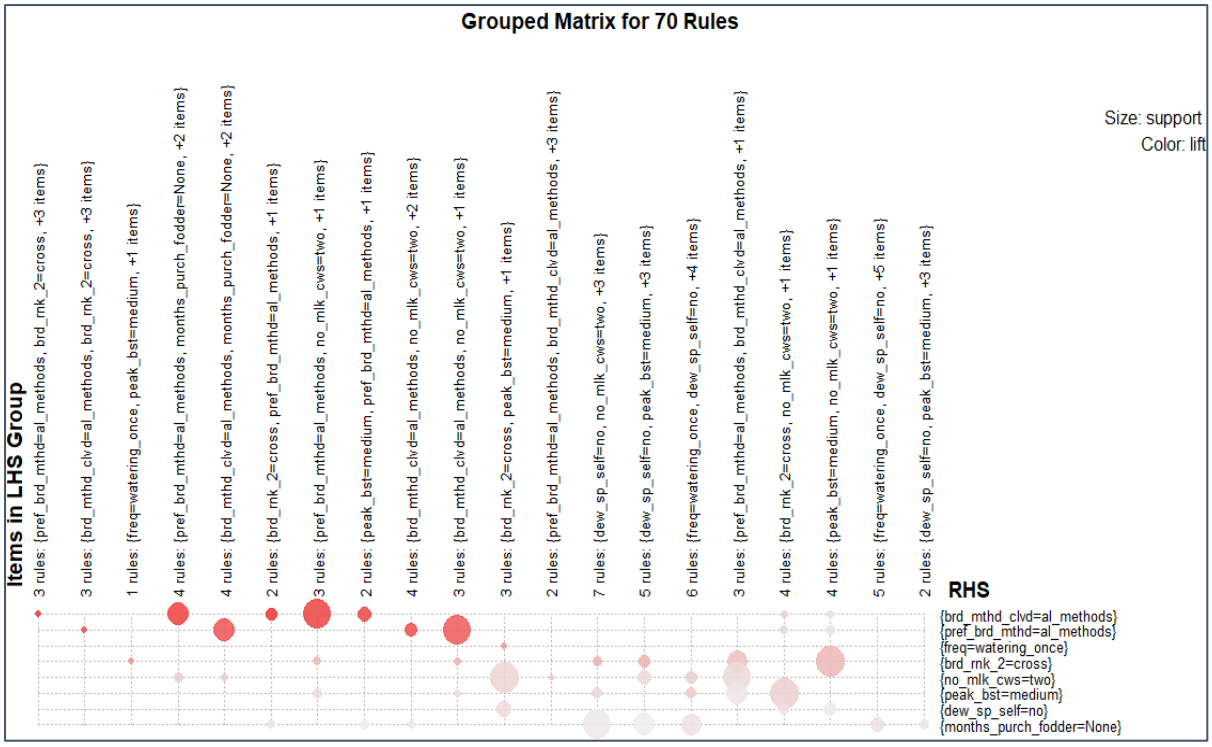


Figure 16: Grouped rules for farm type two in Ethiopia

Farm type three (commercially oriented mixed crop-dairy high production households)

This farm type covers the commercially oriented mixed crop-dairy high production households. On cluster-based characterization, this farm type loaded high on litres of milk sold, peak production and total crop sale while low loadings were on frequency of visits by extension officers, years of schooling and distance to milk buyers. Affirmatively, this farm type performs better than other types in Ethiopia. From the association rules in farm type three, high lift values were observed in rules involving: choice of breeding methods (Although majority of farmers preferred bull breeding services, there was a category of farmers who preferred use of artificial insemination at 94% confidence and higher lift, see Fig. 17), medium milk production and sales, ownership of two milking cows and landholdings of between one to three acres, three vaccination frequencies and preference on Holstein-Friesian cattle breed. It can be demonstrated at 75% confidence that farmers in farm type three practice dairy for commercial purposes regardless of low extension services. As such, at least 75% of the times that medium milk yield, low extension services, 3 times vaccination per year, use of AI breeding and preference in Holstein-Friesian breed are seen together, the farmer would be a high commercial dairy producer.

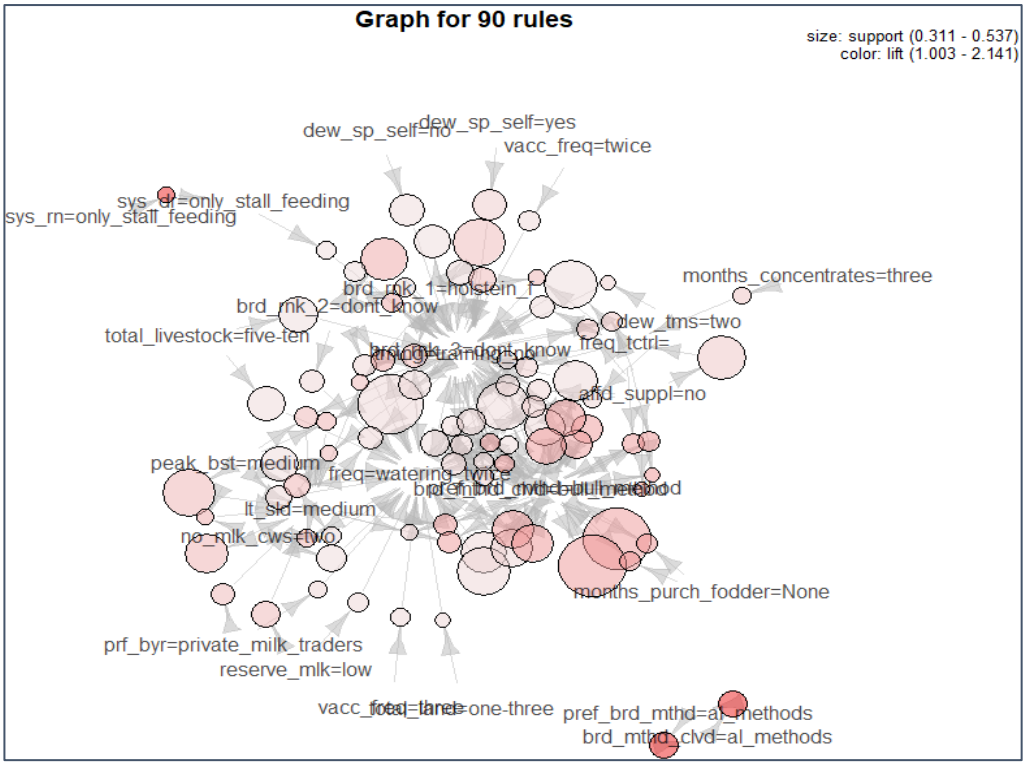


Figure 17: Graph of rules for farm type three in Ethiopia

Farm type four (market constrained; low production households)

Farm type four includes market constrained low production dairy households. In clustering this farm type loaded high on distance to milk buyers, number of sheep and supplementary feeding; while loading low in food cropping, grazing land, and had short distances to breeding service providers. High lift values with low support were observed in rules including: preference in AI breeding method, lack of preferred milk buyers (Fig. 18). Other rules with high lift and support included: medium milk yield, low sales, lack of vaccination and use of concentrate three times a year. The production system in farm type four is assumed to be market constrained at 82% confidence level. Figure 6 highlights strong associations among the small number of milking cows, low milk sales and lack of preferred buyers. It can be affirmed that, at least 75% of the times that, preferred breeding method is AI, zero distances to water sources, medium milk yield and low sales, and lack of preferred buyer are seen together, the farmer would be market constrained low producer.

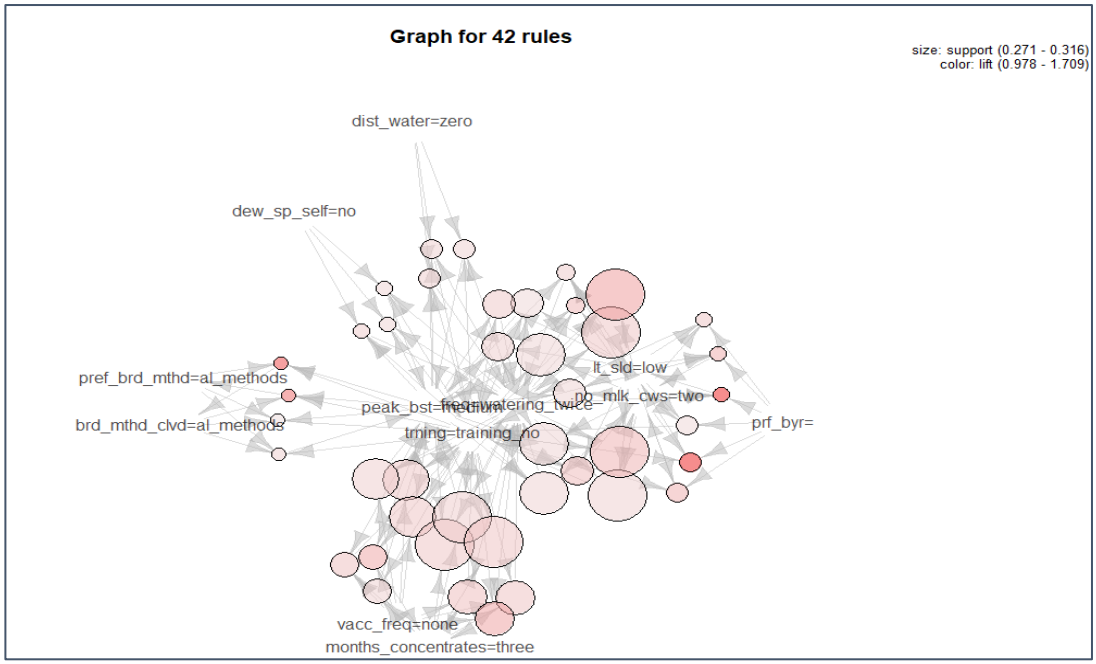


Figure 18: Graph of rules for farm type four in Ethiopia

Results revealed some common patterns across all farm types in Ethiopia, which indicate some mutual characteristics of farmers in the country. The farmers, adherence to best health management practices especially deworming and vaccination together with cattle watering could be termed as general best practices. Summary statistics revealed that, at least 2910 farmers adhere to deworming and vaccination services whereby, a minimum of 52.28% of farmers in Addis Ababa do vaccinate their animals once per year ($p<0.001$); while up to 44.07% deworm their animals twice a year in Asela Shed ($p<0.001$). Ethiopia farmers are constrained with small land holdings for crop farming resulting into low crop sales across majority of the clusters.

(ii) Tanzania farm types

A total of six farm types were determined for Tanzania. These are described below.

Farm type one (non-commercial dairy production households)

Based on clustering results, this type of dairy farms consists of farmers who have been practicing dairy farming for a long time (loading high on experience in dairy farming). Investigating the farm type by using association rules revealed that, farmers in this group practice the traditional dairy farming based on stall feeding, few numbers of milking animals (1 – 3), prefer bull method for breeding and they do not have land for fodder production. These experienced dairy farmers have not attended any formal training and receive very few visits

for extension support (1 – 9 visits per year). Figure 19 indicates strong rules (high support and lift values) being concentrated on stall feeding system. Visualizing the rules by using a grouped matrix in Fig. 20, shows that the feeding system is frequent in both, Left-Hand Side (LHS) and Right-Hand Side (RHS).

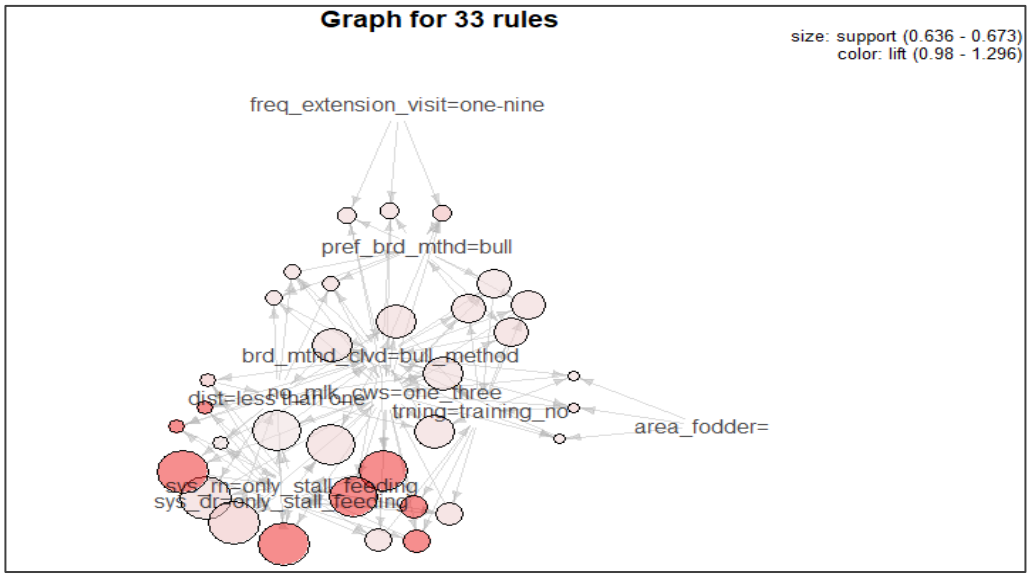


Figure 19: Graph of rules for farm-type one in Tanzania

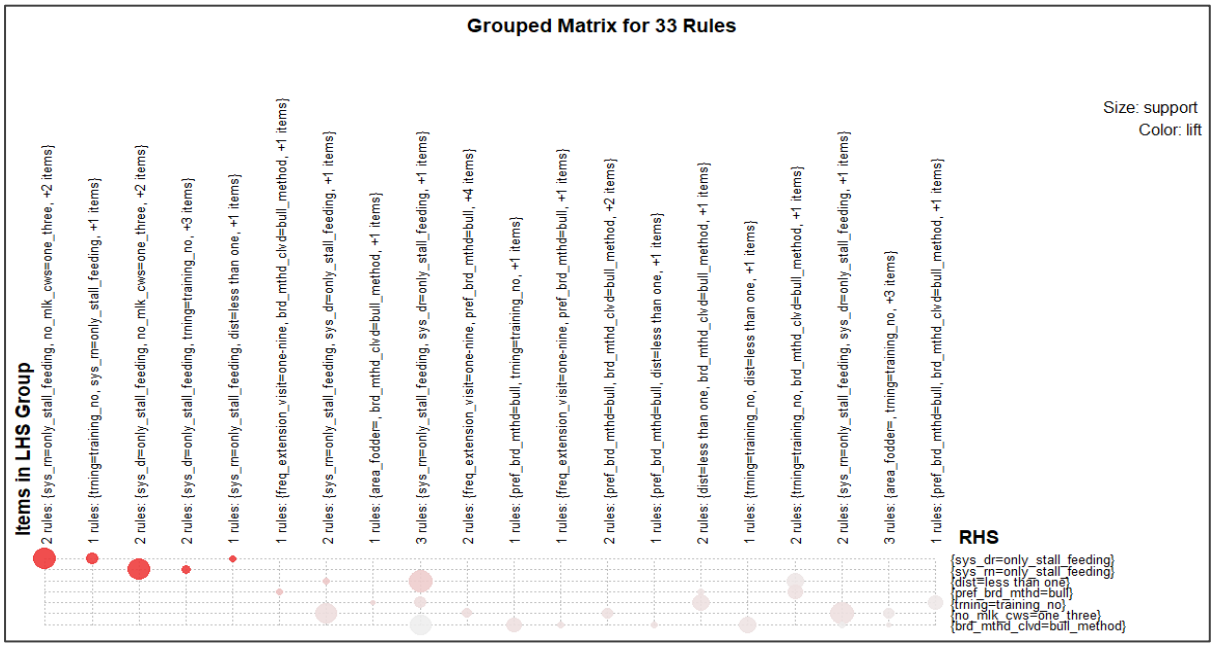


Figure 20: Group of rules for farm-type one in Tanzania

Characteristics of the smallholder dairy farmers in farm type one was generated with good quality measures in terms of support, confidence and lift. High lift values (>1) were observed

in rules involving number of milking cows 1 – 3, lack of training, stall feeding, preference and use of bull breeding and lack of areas for fodder production. Clustering results indicated low loading on the distance to milk buyers. Given that these farmers maintain a small number of milking cows, they lack formal training, and practice stall feeding with no areas available for fodder production, the level of their commercial orientation is assumed to be low. These results indicate with high confidence (> 85%) that more than 63% of the farmers practice subsistence dairy farming. Consequently, the results confirm that when attributes related to lack of training, stall feeding and no areas for fodder production are characteristic of a farmer, that farmer is likely to be a subsistence farmer.

Farm type two (semi-intensive commercially oriented medium production households with limited access to breeding services)

Clustering results characterized farm-type two as semi-intensive commercial farming with medium production. Farmers in this group plant fodder but also utilize small grazing lands. Association rules indicate their preference to bull breeding, which is associated to long distance to improved breeding service providers. High lift values (>1.2) were observed in rules involving: stall feeding system, preference and use of bull breeding, average milk production, lack of farm labourers, no use of purchased fodder, short distance to water source (< 1 km) and 1 – 3 milking cows. Clustering results indicated low amounts for bulk milk sales, presence of areas for fodder production and small grazing lands. A graph of rules further indicates that lack of farm labourers and absence of purchased fodder are associated with stall feeding and preference to bull breeding. Further associations among the features are given in Figs. 21 and 22. Given that these farmers have average milk production and low sales of bulk milk with a mixed feeding system (dominated by stall feeding), their production system is semi-intensive and commercially oriented. Therefore, the results indicate with high accuracy (88%) that, 57% of farmers in farm type two have a medium commercial orientation. The results further imply that, at least 88% of the times that attributes: stall feeding with areas for fodder production, small grazing lands, lack of farm labours, no use of purchased fodder and presence of 1 – 3 milking cows are a farmer's characteristics the farmer would be a small stock semi-intensive commercial dairy farmer.

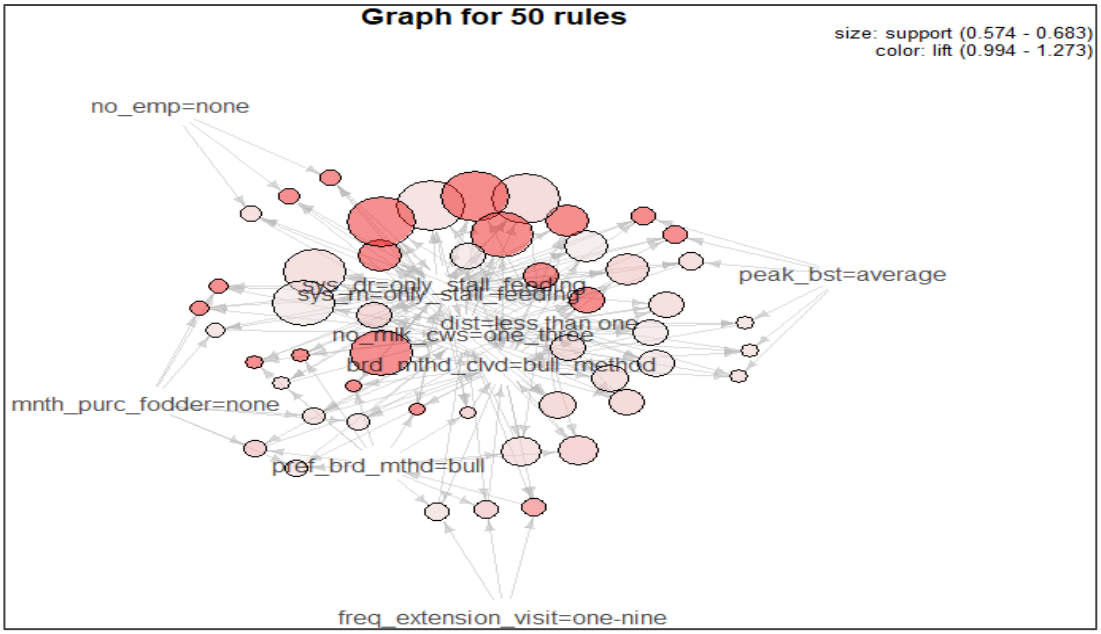


Figure 21: Graph of rules for farm-type two in Tanzania

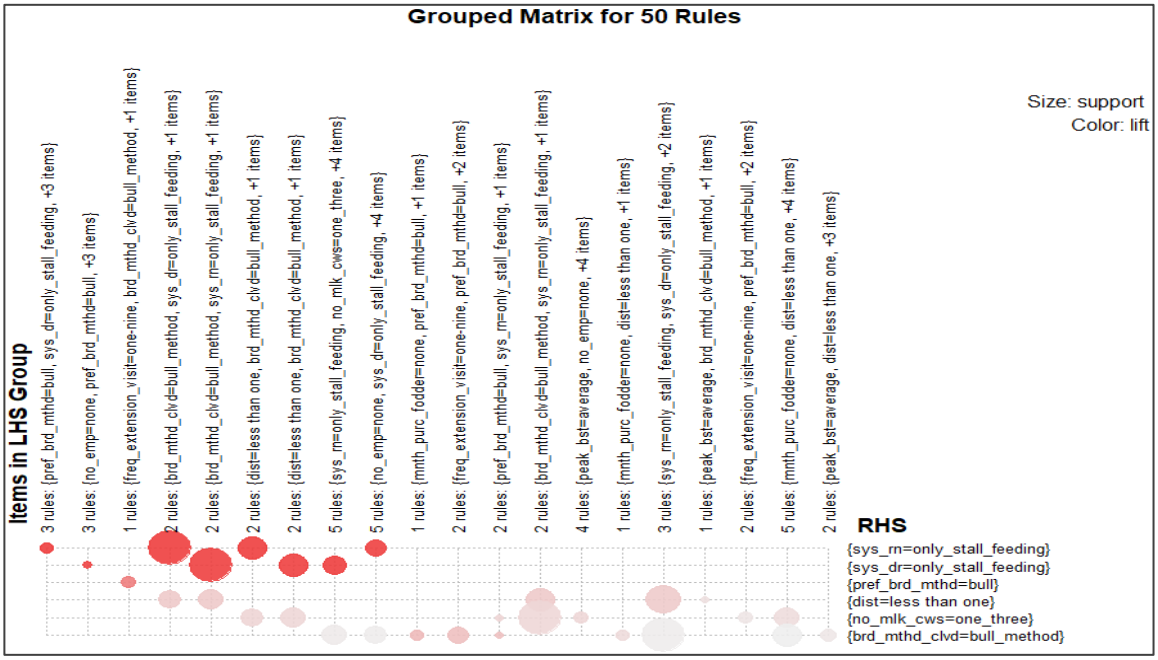


Figure 22: Group of rules for farm-type two in Tanzania

Farm type three (commercially oriented and self-reliant high dairy production households)

From the clustering results, this farm type has the high dairy production households with high loadings on the sale of bulk milk together with vaccination frequencies, and low loading on frequency of extension visits and distance to improved breeding services. High lift values (>1) were observed for association rules covering: stall feeding, preference and use of bull breeding,

vaccination frequency if two/year, lack of farm labour and no membership into farm groups. Figure 23 shows preference to bull breeding being associated to two times vaccination/year and non-membership into farmer groups. Farmers in farm type three are located in the same region as National Artificial Insemination Centre (NAIC).

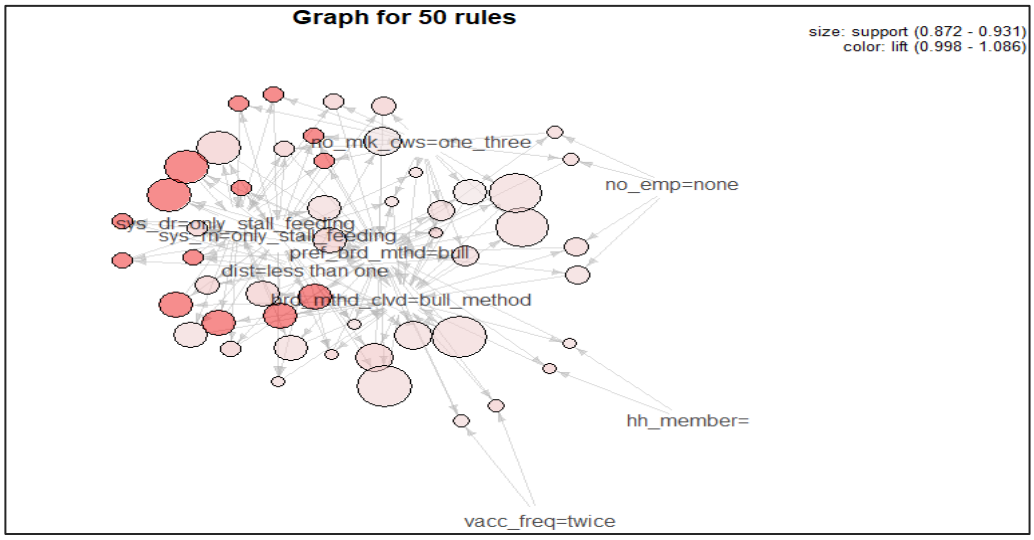


Figure 23: Graph of rules for farm-type three in Tanzania

A grouped matrix for the rules in farm type three shows that only feeding system, breeding method and the number of milking cows have appeared in both RHS and LHS, indicating their frequency pattern and hence importance. Attributes such as lack of farm employees, lack of farmer groups, the frequency of vaccination and shorter distance to water sources (less than a kilometre) are concentrated on the LHS as antecedents of the rules (Fig. 24). Given that these farmers have high amounts of milk sold in bulk, vaccinate their cattle at least twice a year, receive few or no visits from extension officers and do not belong to farmer groups; their production system can be termed to be self-reliant with respect to extension services and with high milk production. As such, it can be stated at 97% confidence level that, 87% of farmers in farm type three practice commercial dairy farming. Consequently, at least 87% of the times that the attributes: stall feeding system, preference to bull breeding, less/no visits from extension officers, vaccination twice/year, no membership in farmer groups and high amounts of bulk milk sale are a farmer's characteristics, the farmer would be a self-reliant high production dairy farmer.

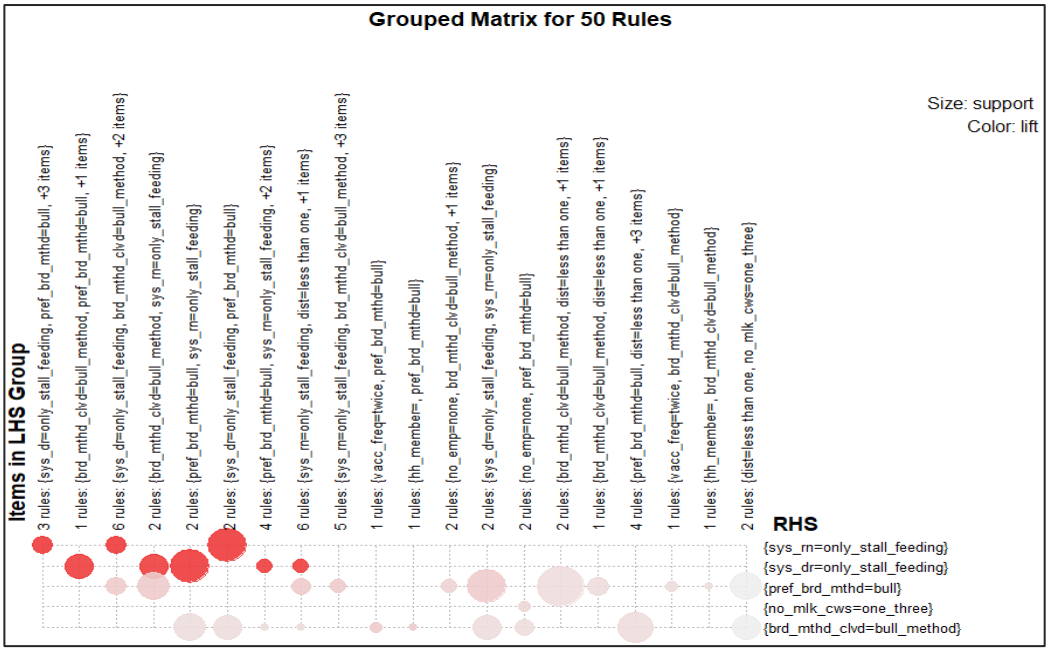


Figure 24: Group of rules for farm-type three in Tanzania

Farm type four (semi-intensive commercially oriented medium production households)

Cluster analysis characterized farm type four with large areas under fodder production and some small areas available for grazing. This farm type identifies as a small stock system based on a low loading for sale of bulk milk in cluster attributes. Association rules revealed high lift values (>1) in rules covering: stall feeding, preference and use of bull breeding, lack of farm employees and maintenance of 1 – 3 milking cows. Clustering results have provided unique features of farm type four, more than what the frequent items could reveal. Milk production and sales coefficients did not appear as frequent items for this group. However, 64% of the membership had average milk production, 68% had below average milk sales while 59% had below average milk reserved for home consumption. Associations and clustering of the variables appearing in high lift rules are given in Figs. 25 & 26. From the unique attributes given by the clustering results, a lowly educated farmer practicing a semi-intensive feeding system with average yield and low sales of bulk milk can be identified as an average producer with a low commercial orientation. Lack of training is assumed to be a complementing factor to the low commercial orientation for farm type four.

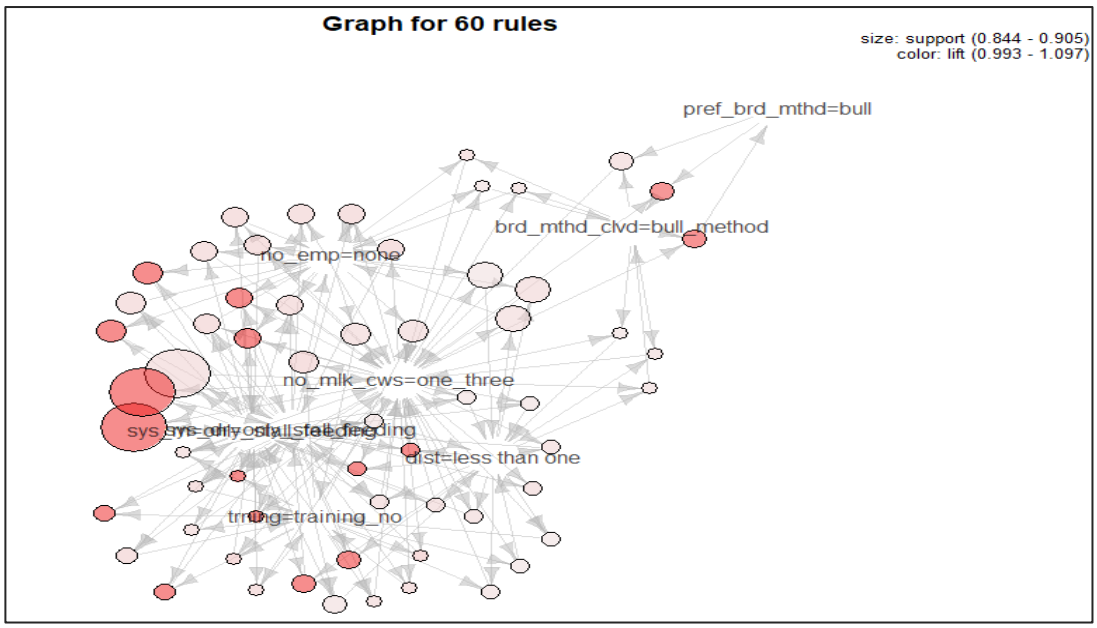


Figure 25: Graph of rules for farm-type four in Tanzania

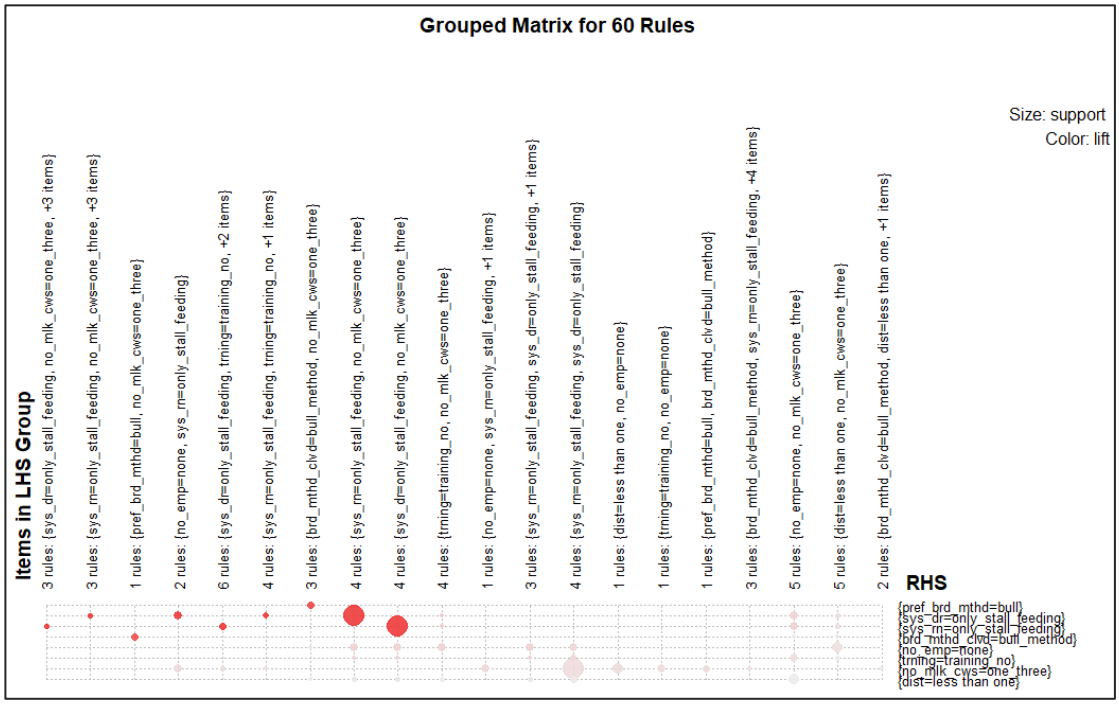


Figure 26: Group of rules for farm-type four in Tanzania

Farm type five (commercially oriented high production entrant households)

Clustering results revealed this farm type with high production, low experience and with limited land sizes. High lift values (>1) were revealed in rules covering: stall feeding system, lack of farmer groups, distance to market 1 – 5 km, lack of formal training in dairy care and

reason for choice of stall feeding being insufficient land. From the frequent pattern analysis, the intensification system appears to be contributed insufficient land (Fig. 27). Although patterns show that, the number of milking animals in this farm type is one to three, commercial orientation of entrant dairy farmers is demonstrated by their high loadings on litres of milk sold and the short distance to markets (1 – 5 km) as given by the cluster analysis. Being close to formal markets, commercial orientation for farm type five appears to be on formal traders rather than neighbours who buy on retail.

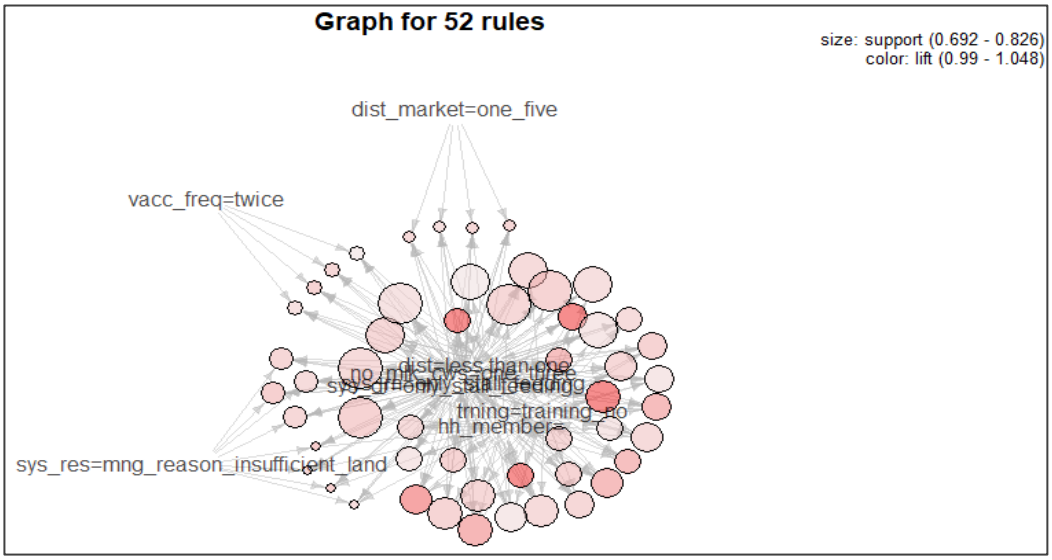


Figure 27: Graph of rules for farm-type five

Lack of membership in farmer groups, lack of formal training in the dairy care and intensive feeding have appeared as strong frequent patterns in both RHS and LHS as shown on the clusters of rules in Fig. 28. It can be demonstrated at least at 69% of entrant farmers who have high amounts of milk sales in formal markets, practice an intensive feeding in small land sizes, have no formal training in dairy care and do not belong to farmer groups, are commercial oriented high production farmers. As such, at least 88% of the times that the attributes: few years of experience in dairy, high milk sales in formal market, intensive feeding, small land size, lack of training in dairy care and lack of farmer groups, are found together the farmer would be a commercial oriented high production entrant.

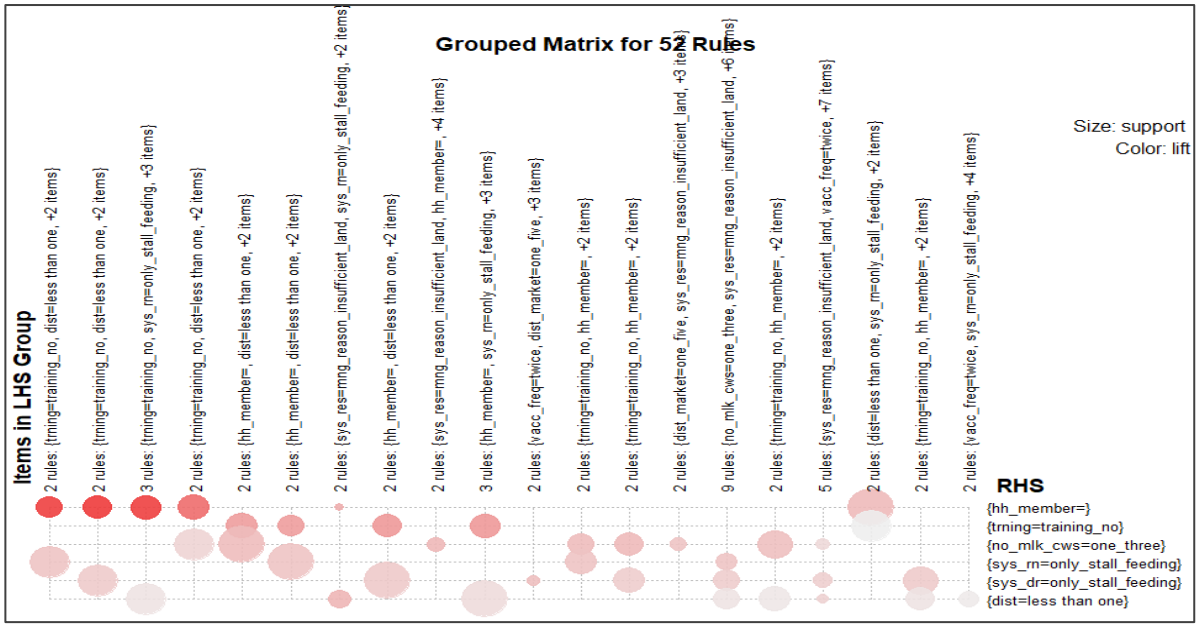


Figure 28: Group of rules for farm-type five in Tanzania

Farm type six (commercially oriented mixed crop-dairy low production entrant households)

Farm type six consist of commercial entrants who load high on a number of cattle owned, walk long distances to water sources, located near service providers for improved breeding and have few years of experience in dairy farming. Lift values from association rules were high (>1) in rules covering: land below average, number of milking cows being 1 – 3, preference to Artificial Insemination (AI), stall feeding system, lack of formal training in dairy care and lack of farmer groups. Loading low on the distance to breeding service providers, association rules have indicated that, this group of farmers prefer and use Artificial Insemination for breeding. Stall feeding is the default system during rainy and dry seasons associated with small pieces of land assumed to be used for crop farming (low loading on total crop sale in clustering), see Fig. 29.

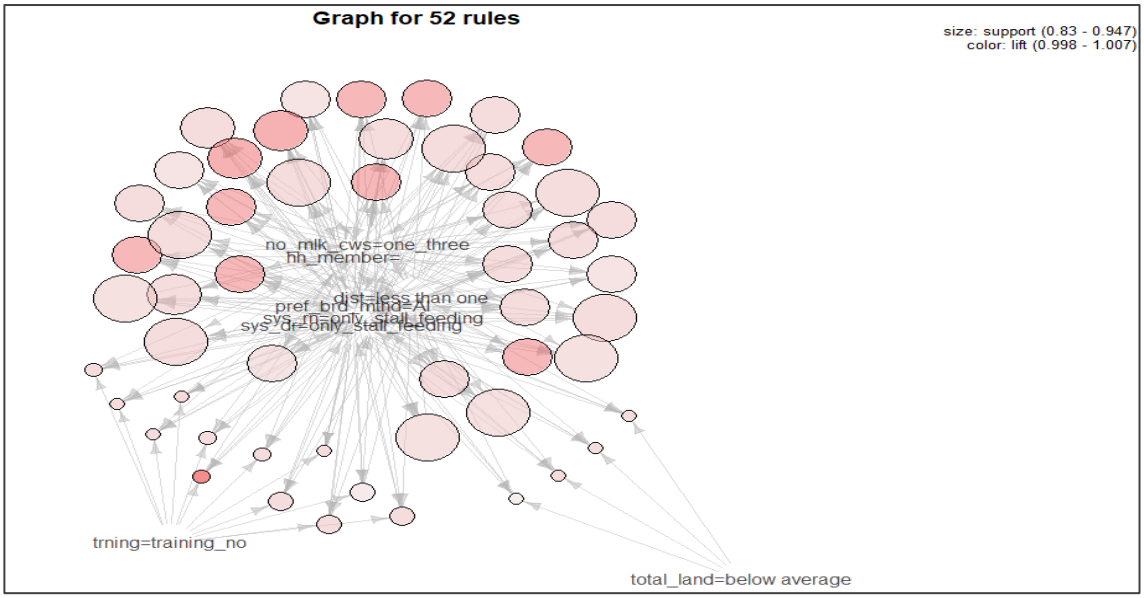


Figure 29: Graph of rules for farm-type six in Tanzania

Although cluster loadings indicate long distances to water sources, association rules reveal that, majority of farmers in type six walk less than a kilometre to water source. Rules clustering for farm-type six can be observed further in Fig. 30.

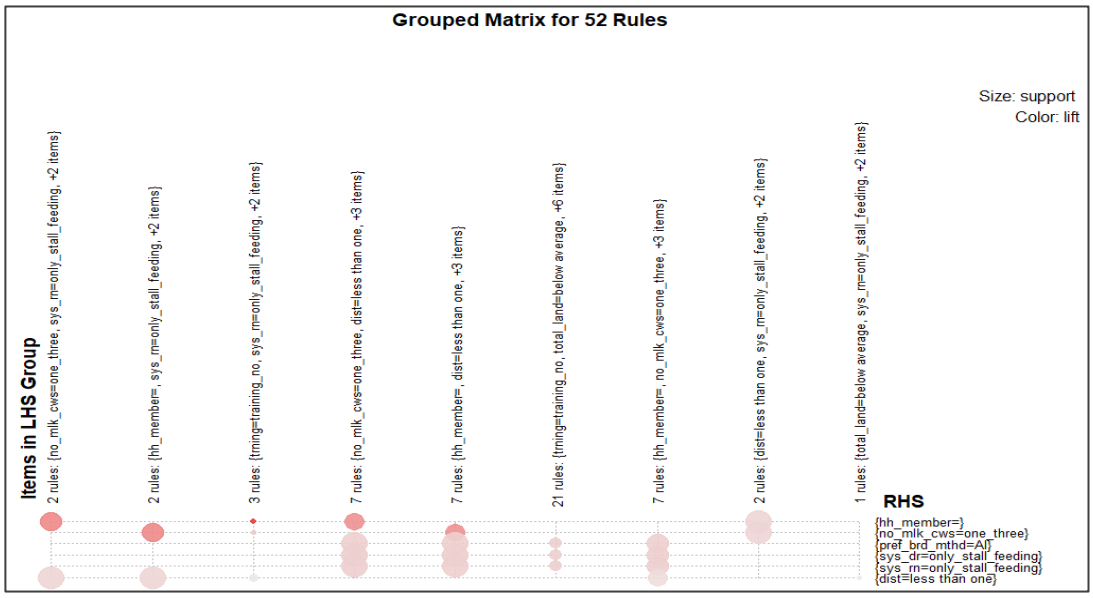


Figure 30: Group of rules for farm-type six in Tanzania

Thirty rules from farm type six were generated at a 100% confidence level. The remaining rules (22) ranged between 95% - 98% confidence level for at least 83% of the farmers. These quality measures imply that the rules can be used to generalize the characteristics of farm type six. It can be demonstrated that, at least 83% of farmers in type six are commercial entrants in

mixed crop-dairy farming system. Equally, at least 95% of the times that dairy entrant farmers have low crop sales, small land sizes, prefer Artificial Insemination (AI), practice stall feeding system, lack formal training in dairy care and lack farmer groups; the farmer would be commercial with low crop-dairy production.

Clusters factor loadings in the Tanzania data set indicated that farmers had commonalities across different production systems for attributes with high loadings, while differences among the clusters mostly came from attributes with low loadings. Common features for Tanzania farmers were; adherence to best health management practices and cattle watering. Summary statistics revealed that, at least 87% of 3317 households indicated that they deworm and vaccinate their cattle. From that population, at least 37.42% deworm their cattle thrice per year ($p<0.0001$) and at least 82.13% vaccinate their animals once per year ($p<0.0001$). Low milk production and sales can be marked as a mutual challenge for Tanzania farmers except for cluster three. Association rules indicate stall feeding is dominant (at least 90% confidence) and is used even in regions where fodder production is not seen.

4.2.5 Evolvement determinants

To this end, structure of the dairy farm types considered the clusters attributes and the associations among the cluster variables. There are characteristics that supports development of the farm enterprises and others limiting improvements in the same. The limiting factors are regarded as areas that can be focused on to help the farmers improve yield and sales. Altogether, these are termed as the farms' evolvement determinants. The determinants have been grouped into four categories following a decision framework for smallholder dairy farmers (Mwanga *et al*., 2019). The four categories are labelled as: farm characteristics, farmer characteristics, farm income and institutional settings.

(i) Evolvement determinants for Ethiopia

Farm characteristics

Farm type one was characterized with available areas for fodder production, grazing lands, small herd sizes (for milk cows and total cattle), and low numbers of exotic cattle. Stall feeding was dominant in farm type one, lack of tick control strategies and fed concentrates at least 3 months a year. Farm type two was characterized with 2 milking cows, low watering frequencies (once daily), lack of use of purchased fodder, preference in AI breeding and cross breed types,

and lack of self-deworming. Farm type three was characterized with two times watering daily, intensive stall feeding, preference in bull breeding and Holstein-Friesian breed, minority of farmers preferred AI with 94% confidence, three times vaccination in a year and owned 1 – 3 Ha of land. Farm type four had high number of sheep, perform supplementary feeding, owned 2 milking cows and watered twice daily, self-deworming service, low food cropping, small grazing lands, preferred use of AI breeding.

Farmer characteristics

Farm type one was characterized with high number of years in formal schooling and lack of formal training on dairy care. Farm type two had high number of years of experience in dairy farming. Farm type three was characterized with low number of years of formal schooling and lacked formal training on dairy care. Farm type four was characterized with lack of formal training on dairy care.

Income characteristics

Farm type one was characterized with medium milk yield and sales and low milk reserves for home consumption. Farm type two was characterized with high amounts of cash cropping, low amounts of milk yield and sales. Farm type three was characterized with medium milk yield and sales, and high amounts of total crop sales. Farm type four was characterized with lack of preferred milk buyers and low milk sales.

Infrastructure characteristics

Farm type one was characterized with zero distance to water sources. Farm type two had high frequencies of visits from extension officers, while, farm type three had low frequencies of visits and short distances to milk buyers. Farm type four had high distances to milk buyers and short distances to breeding service providers and to water sources.

(ii) Evolvement determinants for Tanzania

Farm characteristics

Farm type one was characterized with two times vaccination per year and cattle watering, large land sizes, available areas for fodder production, stall feeding system, 1-3 milking cows and bull breeding. Type two was characterized with high frequencies of cattle watering, available areas for fodder production, small grazing land, use of bull breeding, no employed labours and

no use of purchased fodder. Type three was characterized with high frequencies of watering and cattle vaccination, 1 – 3 milking cows, bull breeding, no employed labours and vaccinate cattle twice per year. Type four was characterized with high frequencies of watering, vaccinate twice per year, available areas for fodder production, small grazing lands, use of bull breeding, no employed labours, and 1 – 3 milking cows and stall-feeding system. Type five was characterized with high frequencies of watering and vaccinate twice per year, small land sizes, and 1 – 3 milking cows and stall-feeding system because of insufficient land. Type six was characterized with large herd sizes, high frequencies of watering, small land sizes, milking cows 1 – 3, artificial insemination for breeding and stall-feeding system.

Farmer characteristics

Farm type one was characterized with many years of experience in dairy farming and no training in dairy care. Type two was characterized with few years of formal schooling. Type three was characterized with non-membership in farmer groups. Farm type four was characterized with few years of formal schooling and lack of formal training in dairy care. Type five was characterized with few years of experience in dairy farming, no formal training in dairy care and non-membership in farmer groups. Type six was characterized with few years of experience in dairy farming, no formal training and non-membership in farmer groups.

Income characteristics

Income parameters were not represented in farm type one. Type two was characterized with low sales of bulk milk, average milk yield and regular sales. Type three was characterized with high amount of milk sold in bulk implying their high commercial orientation. Type four was characterized with low sales of bulk milk implying regular sales. Type five was characterized with high amounts of regular milk sales. Type six was characterized with low crop sales implying their crop-dairy low production.

Infrastructure characteristics

Farm type one was characterized with small distances to milk buyers and water sources, and received 1 – 9 visits from extension officers per year. Type two was characterized with long distances to improved breeding service providers and received 1 – 9 visits from extension officers per year. Type three was characterized with low frequencies of visit from extension officers per year and short distances to improved breeding service providers. Type four was characterized with low distances to water sources. Type five was characterized with short

distance to formal markets (1 – 5 km). Farm type six was characterized with long distances to water sources and short distances to improved breeding service providers.

The clustering results yielded cluster attributes which pertain to the farmers within different farm types. Association rules mining enriched the characteristics of the farm types by providing hidden attributes described by identification of frequent sets of occurrences of specific activities and features. The results from clustering and association rules highlights the main factors pertaining to the smallholder dairy farmers milk yield and commercialization within the various farm types. Based on the analyses, factors that can determine improvement in milk yield and commercialization for the six farm types in Tanzania and four farm types in Ethiopia are generally given in Table 12.

Agent based modelling considered these evolvement determinants to understand their influence in annual milk yield and in peer to peer learning where, the farmer agents interacted by exchanging their practices based on the determinants. Simulations for the milk yield scenario assumed a one-year cow lactation period while, the farmer networks scenario assumed two years' interactions.

Table 12: Evolvement determinants for Tanzania and Ethiopia farm types

	Farm	Farmer	Income	Infrastructure
Tanzania	Frequency for vaccination	Years of experience	Retail milk sale	Distance to buyers
	Frequency for watering	Years of schooling	Bulk milk sale	Distance to water sources
	Total land	Formal training	Total crop sale	Extension visits
	Land for fodder	Membership groups		Distance to breeding
	Stall feeding in rainy			Distance to market
	Stall feeding in dry			
	Grazing in rainy			
	Grazing in dry			
	Milking cows			
	Herd size			
	Employed labour			
	Purchased fodder			
	Breeding method			
Ethiopia	Land for fodder	Years of experience	Retail milk sale	Distance to buyers
	Grazing land	Years of schooling	Total crop sales	Distance to water
	Herd size	Formal training	Preferred buyer	Extension visits
	Milking cows			Distance to breeding
	Number of exotic			
	Stall feeding in rainy			
	Stall feeding in dry			
	Months of using concentrates			
	Frequency for watering			
	Months of purchasing fodder			
	Breeding method			
	Breed type			
	Self-deworming			
	Frequency for vaccination			
	Total land			
	Supplementary feeding			
	Cash cropping			
	Food cropping			

4.2.6 Resulting agent – based model design

The documentation of the design of the agent-based models has been done according to the guideline provided in (Grimm *et al.*, 2010). As identified in Section 3.4, two models were designed according to the resulting evolvement determinants listed in Table12.

(i) Average Milk Yield Model

Design concepts

- Agent cow: The agent has an amount of energy required daily to achieve its production potential. The only source of feeds for the agent is assumed to be grass. Amount of grass eaten by the agent is converted to energy. Daily energy requirements and rates of conversion of grass into energy were obtained from literature (Sulabo, 2011; Dairy, 2015). The agent's energy use were allocated for movement, maintenance, resting, milk production and gain in weight (Moran, 2005). The model design considered the effects of infectious diseases that can affect milk yield. In order to minimize complexity, a random distribution of clinical mastitis was adopted (Østergaard, Sørensen & Kristensen, 2000). By calculating the total probability of catching the infection, the agent cow eating and drinking rates were determined.
- Agent Observer: This agent controls changes in the layers, which in turn affects the behaviour of the cow and farmer agents. The observer agent controls grass growth rate, and the distribution of mastitis base probability in rainy season and reduces in dry season.
- Agent farmer: This is the farm manager, and is responsible to fill feeds and water troughs for the cows. This agent has input data set that details how the farm is managed. A change on the input data for the farmer has an impact on the farm productivity including feeding pattern and annual milk yield from the cow agents. It is the farmer's set of choices and strategies that is altered to observe changes in the milk yield. The alteration is done by adjusting values of the evolvement determinants given in Table 12.
- Adaptation: The cow agent decision room had to decide when to eat and drink based on energy, hunger and thirsty values. The cow could move to the feeds and water troughs every time, either its energy value was below the minimum requirement, hungry or thirsty. Observing its food value, current energy level, and hydration in

respect to amount of food eaten. Since these values were stored in the cow's memory, their alteration alerted an action by the cow.

- Objectives: The farmers' main objective was to increase amount of milk which is measured in kilograms per year. The changes in daily milk yield was affected by the choices of management strategies for the determinants indicated in Table 12. The cows' main objective was to increase total energy value measured in MJ to achieve daily requirements and milk yield.
- Interactions: Direct interactions were between individual cow and biomass layer, individual cow and feeds and water trough layer, observer and biomass layer, farmer and feed and water trough layer, and observer and mastitis layer. Indirect interactions were between individual cows competing for grass and water as they strive to reach to their daily objective.
- Stochasticity: Growth of biomass, feed and water trough re-filling and distribution of probability for infectious mastitis in the environment.
- Observations: At each simulation time, observations were made on the cow agent (milk level, body weight, probability of infection), grass layer (increase in biomass), feeds and water trough layer (increase in feeds and water level), and the mastitis layer (random growth and decline of infection probabilities).

Initialization

- Cow agent: for each cow agent, the following were the initial values as gathered from existing research (Sulabo, 2011; Dairy, 2015):
    - Location: randomly distributed in 45*45 grid cells
    - Instances: 5
    - Energy (MJ) = 0
    - Milk level (kg) = 0.5
    - Probability of infection = ((0.16*Mhistory)/5) + ((0.16*parity)/5) + (0.16*(1/breed))
    - Body weight (kg) = initial value given for each breed in an input data file
    - Breed = fixed breed type given for each agent in an input data file
    - Mastitis history = initial value given for each agent in an input data file
    - Parity = initial value given for each agent in an input data file

- Farmer agent: the farmer agent was randomly distributed in a 45*45 grid cells, and a set of values for each evolvement determinant (as indicated in Table 12) was given as an input file
- Observer agent: For the observer agent, initial values were a random distribution in a 45*45 grid cells, and one instance.

Model layers

The implementation of layers was used to have highly flexible model interactions and improving focus on individual agents (Christian *et al*., 2016). Figure 31 shows the organization of three independent layers which the agents interact with, and each other.

- Grass layer: Each cell in a 100*100 grid contained 4kg of grass at initialization and increases/reduces randomly during simulation.
- Feeds and water trough: 4*9 cells for feeds and 9*9 cells for water out of 100*100 grid. The farmer agent randomly increased values of the cells from 0 (at initialization) depending on management practices given in the farmer input file.
- Mastitis base probability: Each cell in a 100*100 grid had 0 base probability at initialization and increases or reduces randomly during simulation.

Input data

Input data was generated for the model layers, cow and farmer agents. The raster files for the layers were in .asc format and each cell carried a value of 4 for the grass raster, and 0 for the feeds and water trough, and the mastitis raster files.

- Grass raster file (100*100)
- Feeds and water trough raster file (100*100)
- Mastitis probability raster file (100*100)
- Cow agent input csv file
- Farmer agent input csv file

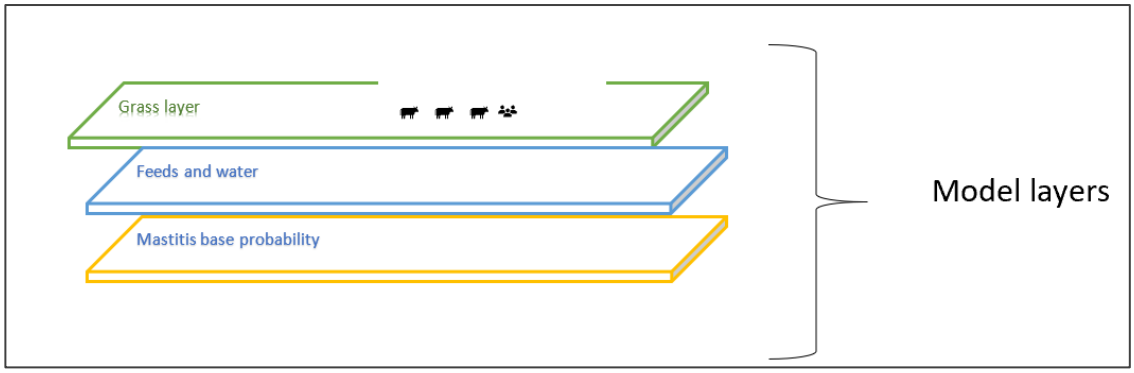


Figure 31: Milk yield model layers

Milk yield

Unlike in the farmer networks modelling where milk yield was derived by a regression model, in this milk yield model the cows are left to produce milk as an effect of their eating patterns, energy requirements and as affected by the farmer's decisions. The following mathematical expressions (Eqns. 12 – 17) have been used to calculate the daily energy requirements to milk yield and increase in body weight (for individual cow agents).

- Energy requirements (equation adapted from (Bruinenberg, Zom & Valk, 2002))

$$EM = 6.9\left(42.4B_w^{0.75} + 442M_y\right)\left(1 + M_y - 15\right)0.00165 \quad (12)$$

Where, EM is maximum daily energy requirement, B_w is body weight, M_y is maximum milk yield.

- Probability of infection

$$P(I) = P_e + 0.0032a + 0.0032b + \frac{0.016}{c} \quad (13)$$

Where, P(I) is total probability of infection, P_e is probability of infection from the external environment, a is mastitis history, b is parity, c is breed type.

- Conversion of grass matter and water to milk yield

$$y = 0.2f + 0.0625w \quad (14)$$

Where, y is milk yield, f is food value and w is water intake.

- Energy use

$$E_u = (10 + 0.1B_w) + 7.1y + 20 \quad (15)$$

Where, E_u is daily energy use, B_w is body weight and y is milk yield. The constants 10 and 20 have been used for the assumptions that 10MJ is a standard amount of energy required for body maintenance for the cows (the total amount will change depending on the cow's body weight), and 20MJ is a standard amount of energy required for movements for the cows.

- Excess energy

$$E_e = EM - E_u \quad (16)$$

- Weight gain

$$B_w = \frac{E_e}{44} \quad (17)$$

Simulation scenarios

- To understand the effect of the evolvement determinants (Table 12) in annual milk yield.
- To identify among the categories of evolvement determinants (farm, farmer, income and infrastructure characteristics), which category have a higher likelihood of improving milk yield.

(ii) Farmer networks model

From the cluster analysis, dairy farmers were placed into homogenous groups also known as farm types. Therefore, the farmer networks model that intended to simulate the peer-to-peer farmer learning considered the homogenous groups of farmers detailed in Section 4.2.2 The model design concepts and relationships are further detailed below.

Design concepts

According to the clustered datasets in Section 4.2.2, each individual farmer had a set of strategies that are followed in dairy farming, and for the purpose of learning new strategies there was a second set of strategies that the farmer had adapted from its peers. Therefore, two relationships were important for the model:

- $F \times S$ Where, F is a set of farmers and S is a set of farming strategies. A corresponding function is: $F \rightarrow 2^S$ i.e. for every farmer f in F returns a set of strategies (non-empty) that f follows.
- $F \times S'$ Where, F is a set of farmers and S' is a set of strategies that are adapted. A corresponding function is: $F \rightarrow 2^{S'}$ i.e. for every farmer f in F returns a set of strategies (potentially empty) that f as adapted from its peers.

With the prescribed relationships and conceptual model, the general pattern of farmer processes was identified and simulated. For each simulation time, the farmer agent acted in the following steps:

(a) Establish an understanding of what farm type it belongs
(b) Build awareness of the farm type characteristics
(c) Assess own milk yield against the farm type average
(d) Evaluate the need and possibility of adopting new strategies from farm type standards
(e) Produce milk after adopting new strategies
(f) Perform a self-evaluation by comparing its milk yield with the farm type average milk yield
(g) If self-milk yield is still below the farm type average, repeat steps (b) - (f)
(h) Interact with nearby farmers and study their milk yield, is it higher than self-milk yield?
(i) If (h) = TRUE, compare self-strategies to those of neighbour
(j) Decide whether to adopt a new strategy based on social economic status
(k) If strategy-changed = TRUE, evaluate self-milk yield after adoption of a new strategy
(l) Repeat (h) – (k), until there is no neighbour producing better that self-milk yield with different strategy.

Milk yield

Milk yield in the model adopted a regression modelling approach. Mack and Huber (2017) demonstrated a similar approach where regression models were used to predict the reduction of Nitrogen emissions for various marginal costs compliance. Since evolvement factors were

analysed separately for Tanzania and Ethiopia, two models were fitted as given in Eqns. (18) and (19), respectively.

- Tanzania

  Tanzania farm type had a total of 25 determinants as summarised in the model below.

$$milk_{yield} = -\beta - \sum_{i=1}^{12} x_i + \sum_{i=13}^{25} x_i + \varepsilon \qquad (18)$$

Where: β is the intercept and x_1 to x_{25} are the evolvement determinants for Tanzania (Table 12), ε is the error term. For every determinant there was an associated coefficient that was obtained through the regression, it was therefore considered as the weight value for that determinant in the agent-based model (Appendix 3).

- Ethiopia

  Ethiopia farm types had a total of 28 determinants as summarised in the model below.

$$milk_{yield} = \beta - \sum_{i=1}^{12} x_i + \sum_{i=13}^{28} x_i + \varepsilon \qquad (19)$$

Where: β is the intercept and x_1 to x_{28} are the evolvement determinants for Ethiopia (Table 12), ε is the error term. For every determinant there was an associated coefficient that was obtained through regression, it was therefore considered as the weight value for that determinant in the agent-based model (Appendix 3).

Simulation scenarios

- Farmer learning from peers was modelled to happen without knowledge of farm type characteristics.
- Farmer learning from neighbours was modelled to occur only after the farmer has knowledge of their farm type characteristics.

4.2.7 Agent – based modelling and simulation results

Two models were developed: 1) a model to study the impact of the identified evolvement determinants towards increase in milk yield, and 2) a model to study the influence of farmer networks for sharing best strategies to increase milk yield based on the identified determinants.

Therefore, it was important that the second model was developed after the first one so as to put the evolvement determinants in an order of their influence in increasing milk yield.

Averages for milk yield in Ethiopia were shown to be higher than those of Tanzania as per the datasets and previous research highlighted in Section 1. For that reason, it was necessary to study the evolvement determinants based on good practices just to observe which determinants would reveal higher milk yield. Therefore, the simulations for Ethiopia milk yield model considered only the good strategies which were labelled with "*base +*", indicated with blue line in radar plots. Actual values for the determinants were extracted from the dataset.

In order to observe the impact of the determinants for the Tanzania case study, the model was simulated with two sets of users i.e. those who had good strategies and those with bad strategies. For example, if a farmer watered the cows more than two times a day it was considered as a good strategy and the simulation was labelled "*base +*". If a farmer watered the cows less than two times a day, it was considered as a bad strategy and the simulation was labelled "*new -*". This approach was applied to all the individual determinants, and actual values for the determinants were extracted from the datasets that have been used in this research for Tanzania. For both Ethiopia and Tanzania, number of cows was five for all simulations, and an individual farmer was considered as the farm manager.

(i) Impact of the evolvement determinants towards higher milk yield for Ethiopia

Income characteristics and commercial orientation was shown to have a significant influence on milk yield for Ethiopia. A farmer who sold at least 10 litres of milk was likely to produce up to 7.38 litres above the average yield in actual data, leading to 21 litres (Fig. 32). Similarly, farmers who had high crop sales had a higher likelihood of producing up to 7.38 litres of milk above the average in actual data per day. For a farmer who preferred to sell milk to private milk traders, an increase of at least 6.38 litres of milk above the average from actual data was observed. The determinants on income characteristics, that is milk, crop sales and buyer preference when fitted together had an r^2 of 0.93.

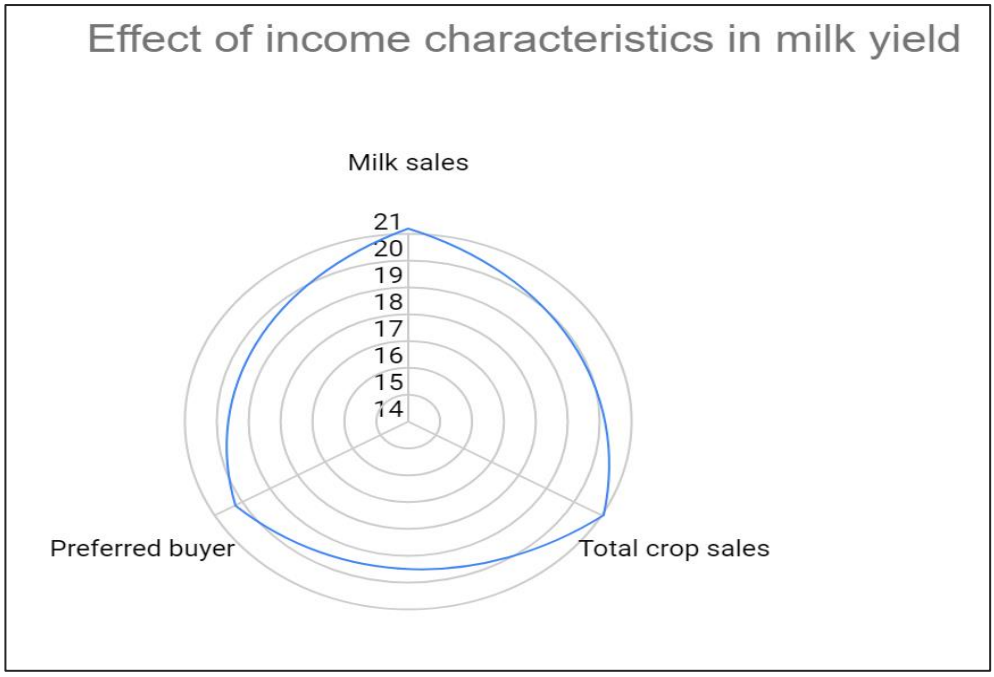


Figure 32: Effect of income characteristics in milk yield for Ethiopia

Farm characteristics ranked second with an r^2 of 0.93 when all features shown in Fig. 33 were combined. Specifically, cash cropping, and area for fodder production, use of AI breeding, watering frequency and stall feeding in dry season have been shown to have a positive impact on milk yield. However, best practices in supplementation, deworming, available land sizes, and stall feeding in rainy season have been shown to yield at least 4.38 litres above the average yield in actual data, giving a total of at least 18 litres of milk (Fig. 33).

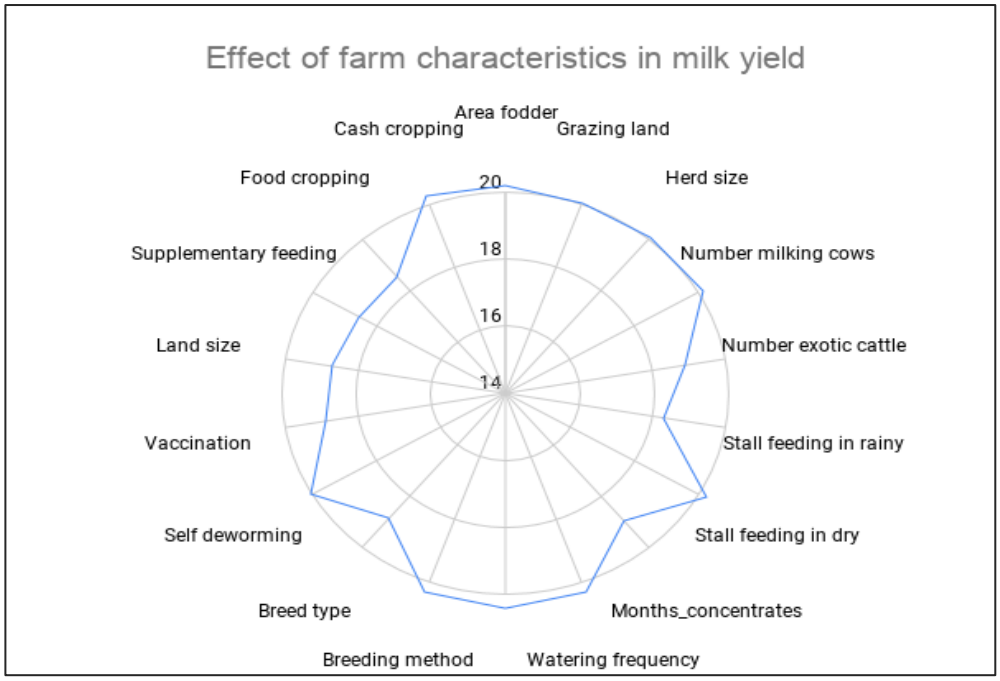


Figure 33: Effect of farm characteristics on milk yield for Ethiopia

Infrastructure characteristics ranked third for Ethiopia with an r^2 of 0.93 when the determinants were combined. Shorter distances to AI breeding service providers were shown to cause an increase of 6.38 above the average for actual values totalling to milk yield of up to 20 litres, while other determinants causing milk yield of at least 18 litres (Fig.34).

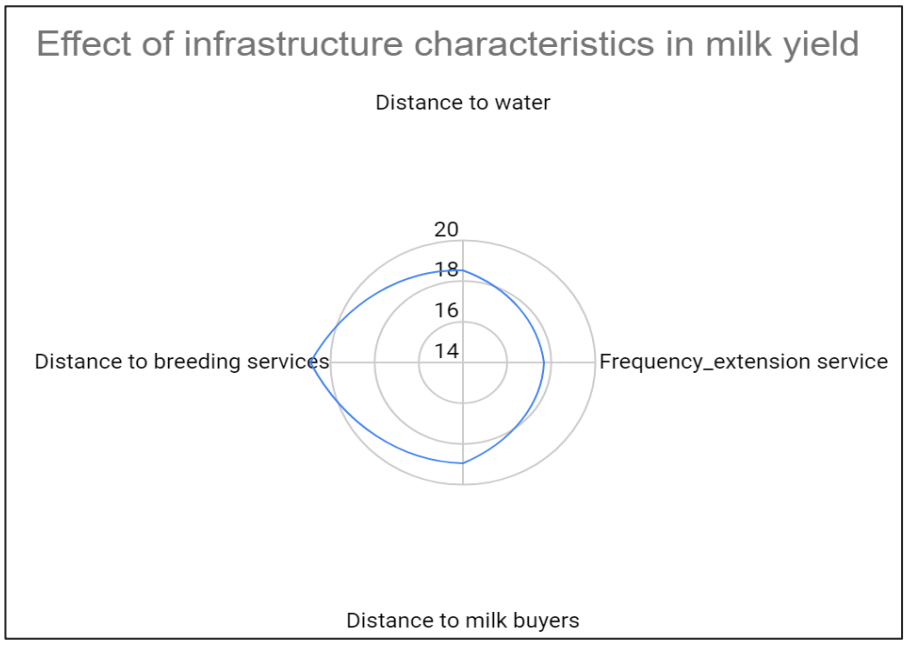


Figure 34: Effect of infrastructure characteristics in milk yield for Ethiopia

Experienced farmers were shown to have higher milk yield for Ethiopia based on farmer characteristics, which ranked fourth with an r^2 of 0.93. At least 16 litres of milk could be achieved for trained farmers and those who had at least completed a primary school level education (Fig.35). This is equivalent to an increase of 5.38 litres of milk above the average actuals.

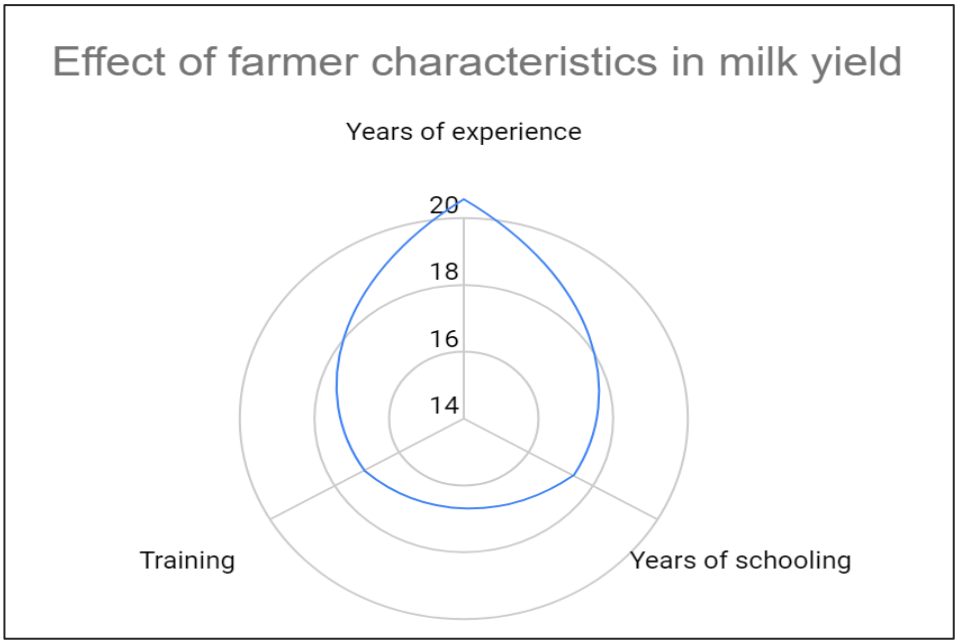


Figure 35: Effect of farmer characteristics in milk yield for Ethiopia

The fitting of the Ethiopia's evolvement determinants all had an r2 of 0.93 which imply that, the fitted evolvement determinants accounts to up to 93% of the variance in milk yield.

(ii) Impact of the evolvement determinants towards higher milk yield for Tanzania

Simulation results indicated that farmers who use purchased fodder get an increase in their average milk yield. That is, lack of purchased fodder led to 28% drop in milk yield from 22 litres to 16 litres. An increase on the levels of employed labours and watering also resulted into increase in milk yield, while an increase in herd size, number of milking cows, use of grazing and total land caused up to 11.77% decrease in milk yield (Fig. 36). Other farm characteristics as vaccination and choice between stall feeding and mixed did not make a significant change in milk yield. The farm characteristics significantly explained the variances in milk yield, with an r^2 of 0.96 when combined.

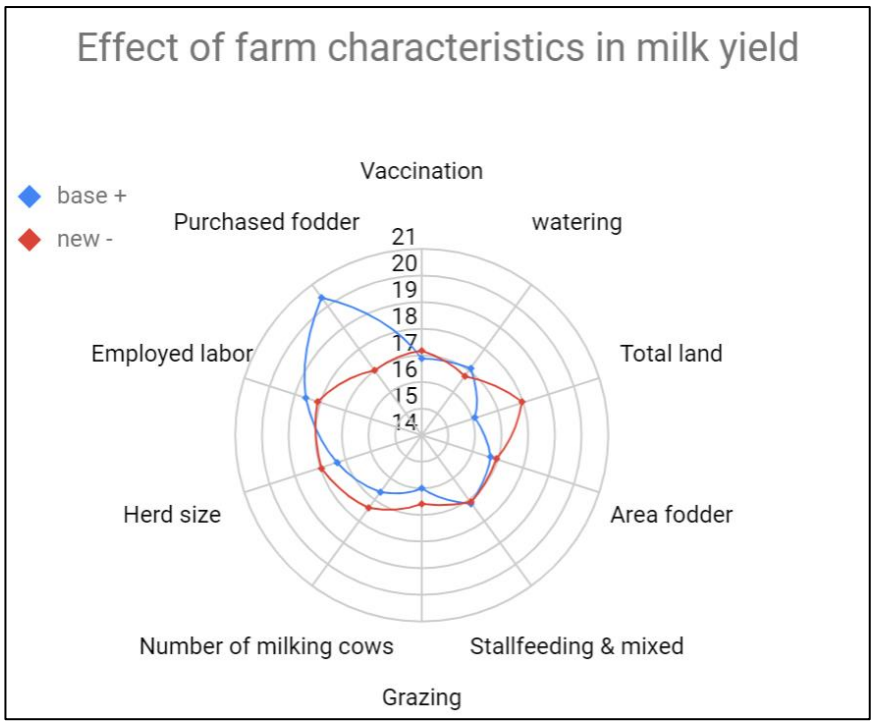


Figure 36: Effect of farm characteristics in milk yield for Tanzania

Experience and formal training in dairy care have been shown to have a positive impact in milk yield, yielding up to 18 litres of milk, equivalent to an increase of 4.3 litres above the average yield in actual values. There was a slight difference in milk yield based on years of formal schooling and membership in farmer groups (Fig. 37). The farmer characteristics alone could explain a significant level of variance in the milk yield ($r^2 = 0.94$).

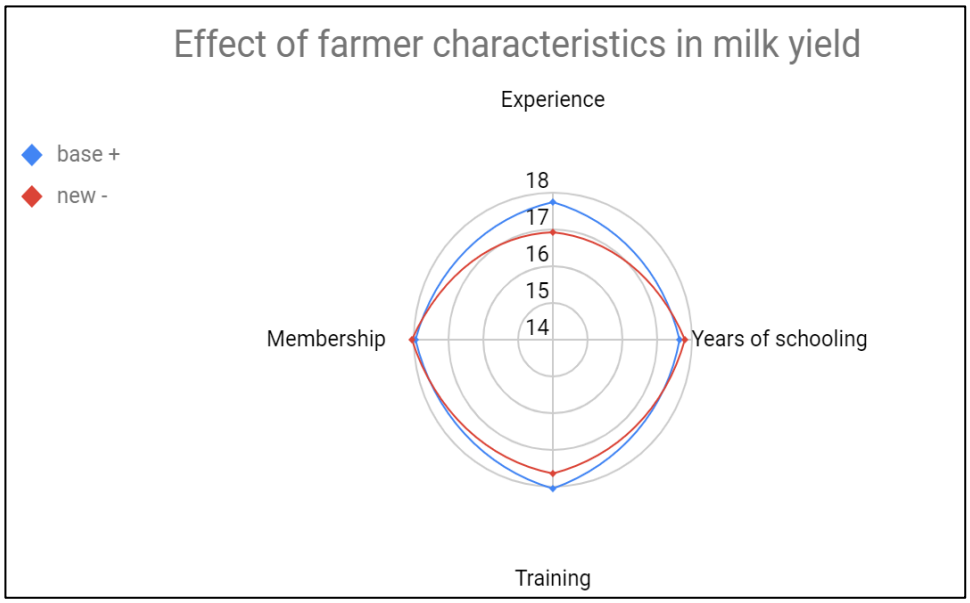


Figure 37: Effect of farmer characteristics in milk yield for Tanzania

Regular sell of milk was associated with an increase of 4.8 litres of milk. Sale of bulk milk indicated a drop of 4.5% from farmers who had less sale of bulk milk. Figure 38 indicates a drop of milk yield from 18 litres to approximately 17.3 litres for farmers who sale milk in bulk. When the three characteristics were fitted, an r^2 of 0.93 was obtained as a proportion of the variance explained by the income characteristics.

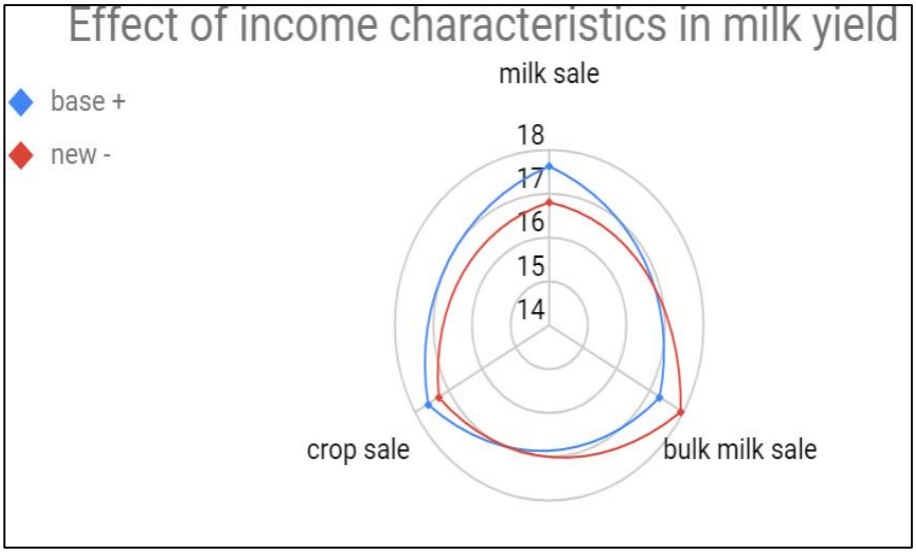


Figure 38: Effect of income characteristics in milk yield for Tanzania

Farmer located far from improved breeding services demonstrated an increase of 5.8 litres above the average for the actual data, giving milk yield of 18.5 litres than those living near service providers (16.5 litres). On the other hand, short distances to milk buyers, water sources, and at least 9 extension visits per year resulted to an increase of at least 4.3 litres of milk above the average for actual yield (Fig. 39). Moreover, farmers living near formal markets had lower milk yield than their peers by at least 0.2 litres. Considering the effect of sales given in Fig. 38, that more milk sales implied more milk yield; it is becoming evident that farmers do not sell their milk in formal markets. This result attests the effect of buyer preference demonstrated for Ethiopia case where, preference to private milk traders and individual customers highlighted more yield. When all the infrastructure characteristics were fitted, an r^2 of 0.93 was obtained.

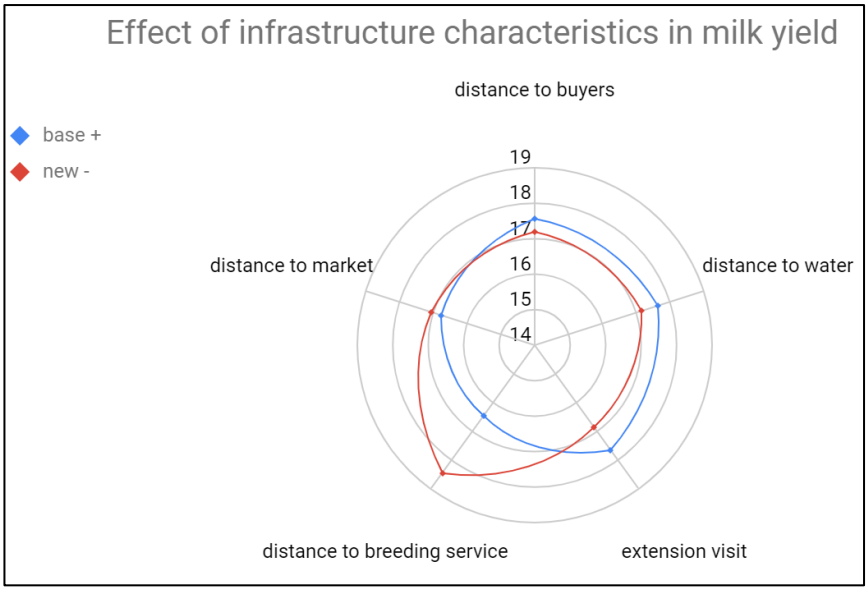


Figure 39: Effect of infrastructure characteristics in milk yield for Tanzania

The simulation results for the milk yield model indicated the differential influence of each evolvement determinant towards increase in milk yield for the Ethiopia and Tanzania case studies. The differences in milk yield, that is, litres of milk above or below the average for actual yield was considered in ordering the evolvement determinants in addition to the r^2 obtained in each model run. Consequently, for Ethiopia; income, farm, infrastructure then farmer characteristics was the order of importance. For Tanzania; farm, farmer, income then infrastructure characteristics was the order of importance. Table 12 lists the evolvement determinants.

Two scenarios were run for a model that simulated increase in milk yield by the farmers' adaptation to better strategies, based on the identified determinants and their influence in milk yield. The scenarios were:

(i) The farmers adapted to better farming strategies by learning from their neighbours without knowledge on the characteristics of their farm types.
(ii) The farmers adapted to better farming strategies after having knowledge of their farm type characteristics and spent 145 days adapting the farm type characteristics.

In each simulation, the milk yield was compared to the real-world dataset for each farm type and the correlation (r) for each farm type was calculated.

(iii) Knowledge sharing through peer-to-peer farmers' learning for Ethiopia

Simulation of farm type three had computational limitations due to its size (n=2689 producing 2689 x 731 rows of data in excel) which resulted into the farm type being split it to two halves: farm type three (a) and (b) as shown in Fig. 40 and 41. Simulated milk yield was correlated to the actual milk yield from actual data. The highest prediction accuracy was 0.64 and the lowest was 0.051. The simulation of farm type four in Ethiopia for the two scenarios, appeared to have produced milk yield values that were not correlated with actual milk yield (r = 0.051).

On average, more than 87% of the farmers from Ethiopia had learned and adapted new practices within the first 290 days (Fig. 40), when they were not aware of their farm type characteristics. Results highlight a drop of that proportion (to 34%) in the second scenario when the farmers were first trained on the farm types characteristics (Fig. 41).

Figure 40 indicates for all the farm types; the first 290 days were significant in milk increase even without awareness of the farm types characteristics. Also, the proportion of farmers who didn't learn a new practice to improve their milk yield was very low (0.23%).

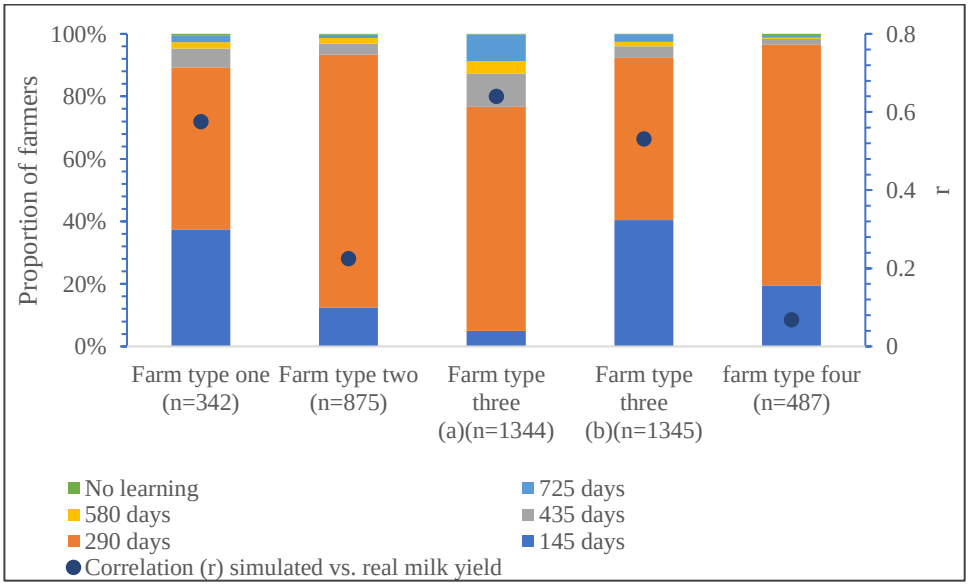


Figure 40: Farmer learning based on scenario 1 for Ethiopia

When the farmer agents were led to first understand and adopt to the farm types characteristics, the proportion of learned farmers within 290 days of simulation for all farm types dropped to 34.30% but for farm type three (b) which appeared to have not changed its pattern from scenario one. As a result, some farmers' first increase in milk yield were observed at least after 435 days of simulation. There were no significant differences in the correlation coefficients

between the two scenarios for the Ethiopia farm types. The proportion of farmers who didn't adopt to a practice to increase their milk yield was doubled from the first scenario and became 0.4%. Observing the simulated milk yield for the two scenarios, the averages for all farm types were insignificantly different. That is, the two scenarios produced the same end results at the end of simulation time. Milk yield averages based on scenario 1 were: 20.94, 20.7, 23.3, 20.79, 20.2 and based on scenario 2 they were: 20.99, 20.73, 18.17, 21.54, 20.3 for farm types one, two three (a), three (b), and four respectively.

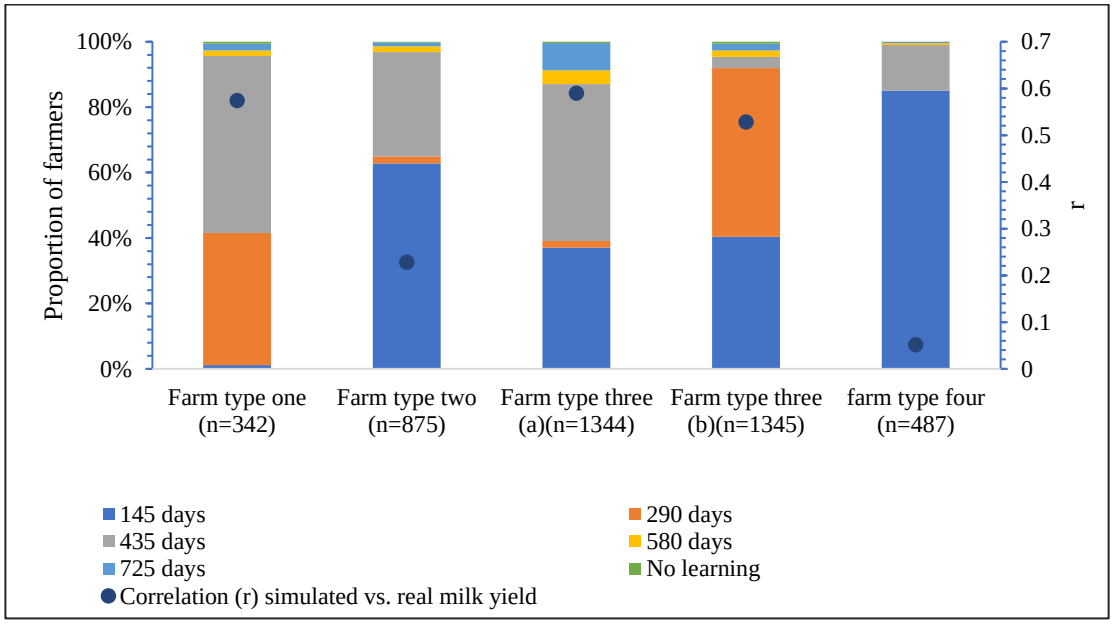


Figure 41: Farmer learning based on scenario 2 for Ethiopia

(iv) Knowledge sharing through peer-to-peer farmers' learning for Tanzania

Under the first scenario, majority of the farmers in Tanzania adopted better farming practices and increased their milk yield in the first 145 days of simulation except for farm type three where only 41% of the farmers adopted improved practices in the period (Fig. 42). A significant portion of farmers (up to 15% for farm type one and 8.23% from the entire population) could not improve on milk yield at the end of simulation. Farm type five had the highest prediction accuracy (0.6), while the remaining farm types had at least 0.45 prediction accuracy and 0.3 for farm type three.

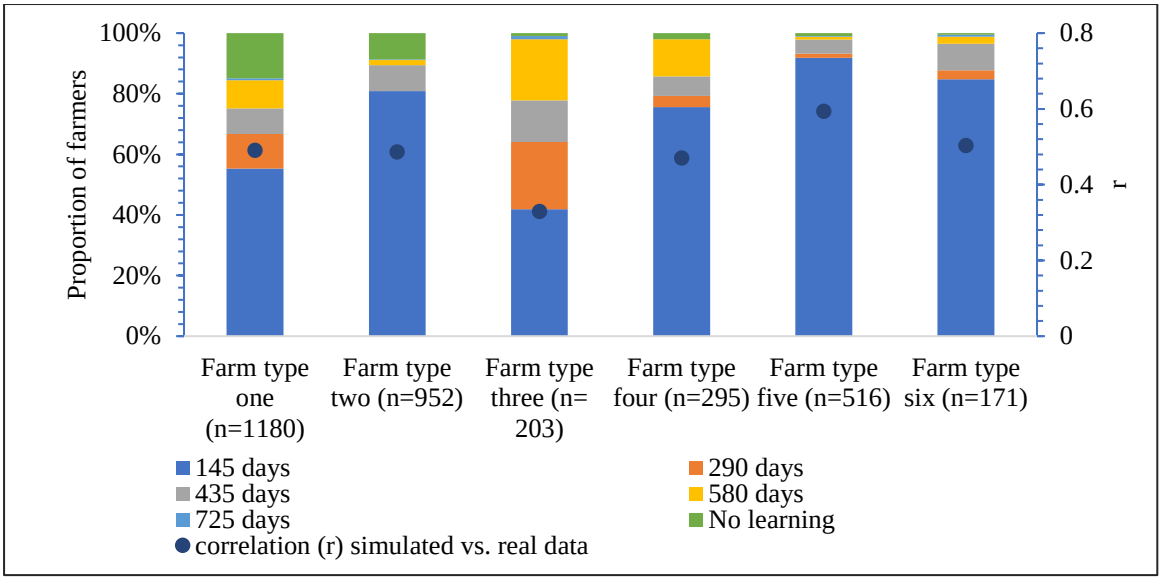


Figure 42: Farmer learning based on scenario 1 for Tanzania

Simulation results for the second scenario indicated that majority of the farmers (95.12%) across all farm types adapted to better farming practices and improved milk yield in the first 290 days (Fig. 43). There were no differences in the prediction accuracy between the simulated milk yield and actual yield for the two scenarios. It is important to note that the farm type characteristics were developed from frequent pattern analysis and therefore the minority were the ones expected to adapt to the farm type characteristics in the first 145 days. Therefore, the proportion of farmers who adapted to better practices in the first 145 days was lower than in the first scenario. As such, majority of the farmers under the second scenario had improved their milk yield on the 290 day of simulation, and the proportion of those farmers who took longer time to improve under scenario one was reduced significantly. The proportion of farmers who could not improve their milk yield to the end of simulation was reduced for farm type one from 15% under the first scenario to 4% under the second scenario, and from 8.23% to 3.28% for the entire population. Notably, 98.3% and 99% of farmers for farm type four and five had improved their milk yield under the second scenario during the first 290 days of simulation.

The main differences between scenario 1 and 2 for Tanzania is the proportion of farmers who improved milk yield within one year of peer to peer interactions, and the proportion of farmers who could not improve their milk yield at the end of the simulation period. Figure 44 detail the differences between the scenarios by considering the proportion of farmers who improved their milk yield within a period of 290 days (within 1 year of peer to peer interactions). In addition, the averages for simulated milk yield for the farm types are as follows: based on the first

scenario they were, 15.35, 14.49, 14.92, 13.58, 14.16, 14.63 and based on the second scenario they were, 17.45, 17.59, 18.42, 16.78, 16.66, 18.53 for farm types one, two, three, four, five and six, respectively.

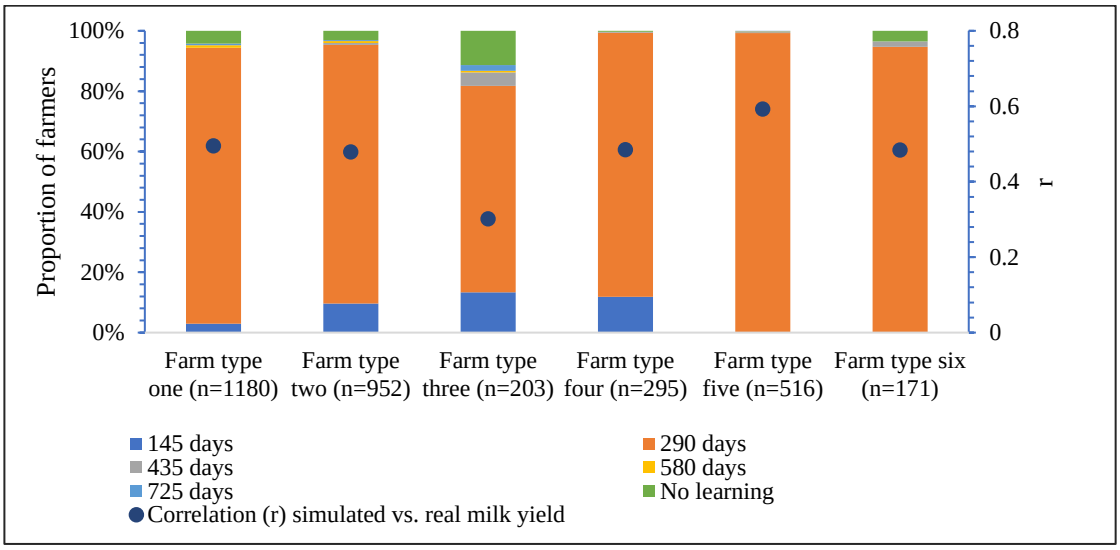


Figure 43: Farmer learning based on scenario 2 for Tanzania

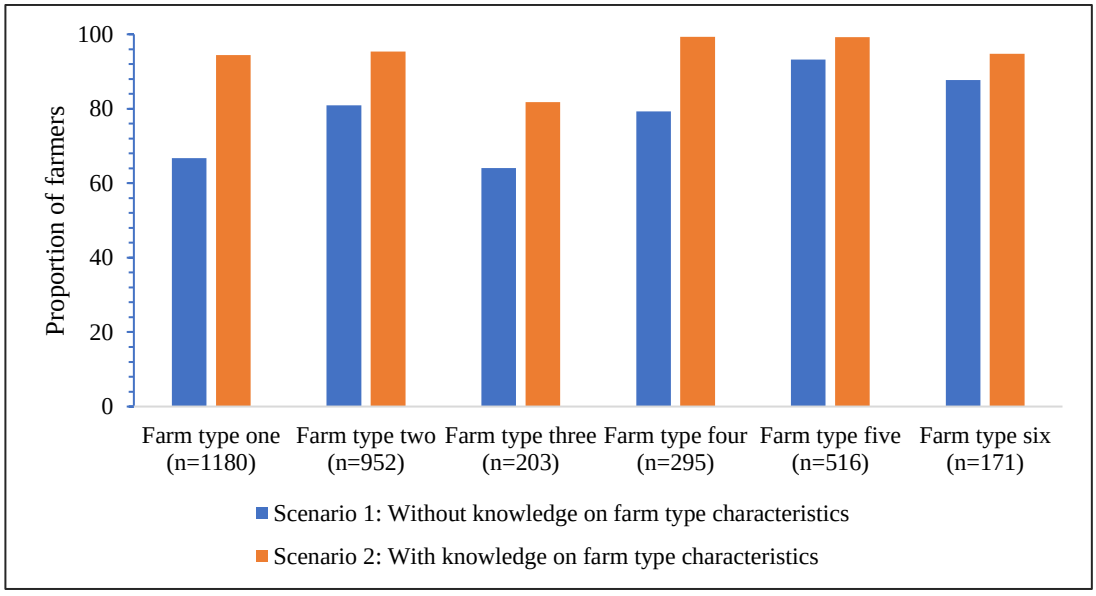


Figure 44: Differences between scenario 1 and 2 for Tanzania within 290 days

An outstanding difference between the Tanzania and Ethiopia simulations is, Tanzania increase in milk yield is sensitive to the farm type characteristics while the increase for the same in Ethiopia is not sensitive to farm type characteristics. Farmers in Tanzania could improve their average milk yield much better on the second scenario as compared to the first scenario. In general, there was a significant increase in average milk yield when the identified determinants

were observed during simulation of the agent-based models; from baseline data average milk yield of 12.7 ± 4.89 and 13.62 ± 4.47 to simulated milk yield average of 17.57 ± 0.72 and 20.34 ± 1.16 for Tanzania and Ethiopia, respectively.

4.3 Discussion

The main objective of this study was to identify factors that determine evolvement of smallholder dairy farmers while considering disaggregation of the farmers through agent based modelling. As highlighted in the introduction, smallholder dairy farmers have differing milk yield, and thus, the influential factors for their evolvement will differ. Therefore, disaggregating farmers into different groups, can allow more targeted approaches for evolvement to be identified. For the farmers to apply targeted approaches, the influence of the approaches towards increase in milk yield should equally be established. Therefore, in this study differential influence of the factors that influence increase in milk yield was investigated. The main advantage of establishing the differential influence of the factors, is the ability of the farmer to be able to prioritize resources and focus on the factors which have the highest influence.

To achieve this aim, we used unsupervised models to cluster farmers into types and to mine frequent patterns associated with each cluster/farm type. The farm type characteristics and the frequent patterns formed four categories of evolvement determinants; farm, farmer, income and infrastructure determinants. For each farm type, agent-based modelling and simulation was then applied to study the influence of the identified evolvement determinants. While this research presents a combination of methods to solve the ever-existing challenge of unmet dairy demands, it equally demonstrates that other case studies that involve dynamic human to human, human to environment interactions can be explored using machine learning and agent-based modelling.

4.3.1 Cluster – based characterization

In this study, unsupervised learning models were used to characterize smallholder farmers despite previous study claiming that they lack consistency and are highly unpredictable (Gelbard *et al.*, 2007). However, the use of the unsupervised models in this study has been done with a robust cluster validation approach, ensuring stable clusters formation and clusters that can be used in evolvement and prediction of smallholder dairy productivity.

Three commonly used algorithms for clustering farming households, namely: K-means, Fuzzy and SOM were chosen for comparison. However, unique to this study, a set of validation criteria to assess the robustness of the defined clusters was proposed. This approach is seldom used for similar studies, as reviewed in chapter two. The application of the three algorithms revealed differences in their performance based on data type and structure. In cases where observations were highly identical, soft clustering (Fuzzy model) failed to converge and categorize the records into appropriate six clusters as observed in the Tanzania cluster analysis. The main indicator of the failure was the fuzzy model lack of convergence even after many iterations. The Fuzzy model allocated households into 3 clusters despite four clusters being determined as appropriate, for the Ethiopia dataset. It would appear that the fuzzy model could be best suited to situations where data is highly heterogeneous. Otherwise it doesn't lend itself well to robust cluster identification.

In previous studies, Balakrishnan (1994) compared K-means and SOM algorithms in cluster identification within specific criterion of intra-cluster similarity and inter-cluster differences. In addition, the dataset had known cluster solutions; so, the only target was to find out performance differences between the two algorithms. Results indicated that the K-means algorithm had good perfomance over the SOM algorithm. Mingoti and Lima (2006) compared K-means and SOM models' performance by using smallholders' farm data. Results indicated that K-means was more robust. In this study, the SOM performed poorly compared to the fuzzy and K-means for the Ethiopia dataset having higher within cluster dispersion, as well as lower separation between clusters. For the Tanzania dataset, the SOM performed similarly as the K-means algorithm. Results from our study show that, the performance of SOM is concordant with that of Nazari *et al*. (2017) who characterized dryland farming systems. In contrast to observations by Mingoti and Lima (2006), the fuzzy model used in their study failed spectacularly for both datasets. This reinforces observations by Xu (2005) who concluded that the performance of clustering algorithms is subject to the nature of data and area of application. More studies need to be undertaken to see how the fuzzy algorithm can be best adapted to datasets with high level of homogeneity.

Based on the results from Ethiopia, where all the models could be evaluated, it would seem that model choice depends on the problem that needs to be solved. For a clustering problem, where the intention is to obtain robust membership allocation, then the K-means algorithm would be the most appropriate, to ensure maximal homogeneity within clusters. The use of this

model would minimize re-ranking when applying the model to new datasets without need for new learning. However, in the event that clusters are to be used in prediction models, the fuzzy algorithm would be the best for cluster definition owing to the prediction accuracy being highest. Cluster analysis ended up with six farm types for Tanzania and four farm types for Ethiopia. The characteristics of the farm types were given based on cluster factor loadings but more details of the farm types were further studied by using association rules mining as discussed in the next section.

4.3.2 Association rules mining

Smallholder farms have been characterized in previous studies using statistical and unsupervised machine learning algorithms with the main goal of understanding the nature of homogenous farm types within regions (Gizaw *et al*., 2017; Paas & Groot, 2017; Kuivanen *et al*., 2016) for appropriate introduction of policy, technology and extension support (Goswami *et al*., 2014). However, reliance on clustering results has various limitations, such as use of factor loadings only to understand cluster attributes in addition to limited explanation on reasons behind the homogenous groups formation (Maciejewska & Pempkowiak, 2015). In this research, association rules mining has been used to demonstrate frequent items and patterns within clustered smallholder dairy baseline data. Results from this analysis are key inputs in understanding evolvement of smallholder dairy farms and chances to predict their performance based on application of the identified determinants for increase in milk yield.

Use of association rules mining to characterize clustered dataset followed a similar approach as the one reported by Robardet, Bruno and Jean (2002) who characterized patient medical records. Other examples of characterization based on association rules have been in the commerce domains (Kamsu, Rigal & Mauget, 2013; Ramos *et al*., 2015; Suchacka & Chodak, 2017).

Results from the association rules mining indicate important attributes of farmers within the studied farm types that could not be discovered through clustering. In agreement with Robardet, Bruno and Jean (2002), use of association rules has provided a detailed explanation behind formation of the farm types and their prevailing characteristics. The farm types characteristics are therefore, referred to as determinants for increase in milk yield.

Validation of selected association rules was conducted on the basis of confidence and lift values as recommended by Bazaldua, Baker and Pedro (2014). The use of lift measure has been rated

by domain experts (Bazaldua, Baker & Pedro, 2014) as the most likely to identify the most interesting rules. However, for large biological datasets use of lift has been overtaken by custom validation approaches (Mallik, Mukhopadhyay & Maulik, 2015). Maximum number of records in the clusters used in this study was 2689, making rules filtering not as difficult as in large databases. The use of clusters and association rules mining provided a robust characterization approach for the six and four farm types in Tanzania and Ethiopia, respectively. Conclusively, twenty-five and twenty-eight evolvement determinants were found for the case of Tanzania and Ethiopia, respectively. The influence of the identified determinants was studied by using agent-based modelling and simulation as discussed in the next section.

4.3.3 Agent – based modelling and simulation

The term evolvement as used in this research refers to a gradual increase in milk yield across production systems to improve the livelihoods of smallholder dairy farmers. As such, the timelines assumed in the evolvement is one lactation period (365 days) for simulations that investigated the influence of the evolvement determinants and two years (730 days) for the simulations that investigated the peer to peer interactions for farmers. The agent-based models have highlighted the uniqueness of the farm types for Tanzania and Ethiopia, and also demonstrate that Tanzania evolvement case is different from Ethiopia: strengthening the disaggregation concept as highlighted in the problem statement.

The simulation results have concluded that in order of priority, if farmers from Ethiopia case study have to increase their milk yield based on the identified evolvement determinants then; income, farm, infrastructure and farmer characteristics need to be considered, in the given order, as given in Figs. 32 – 35. Equally, if farmers from Tanzania have to increase their milk yield then; farm, farmer, income and infrastructure characteristics have to be considered as given in Figs. 36 – 39. The agent-based modelling revealed income and farm-based characteristics influenced an increase of up to 7.58 litres above the average (13.62 ± 4.47) for Ethiopia. For Tanzania, farm and farmer-based characteristics influenced an increase of up to 7.72 litres of milk above the average (12.7 ± 4.89). The identified factors could predict up to 96% and 93% of the variances in milk yield for Tanzania and Ethiopia, respectively. Generally, there was an increase in milk yield based on the identified evolvement determinants; from baseline data average milk yield of 12.7 ± 4.89 and 13.62 ± 4.47 to simulated milk yield average of 17.57 ± 0.72 and 20.34 ± 1.16 for Tanzania and Ethiopia, respectively. The increments realized in this study are at least 60.7% for Tanzania and 55.6% for Ethiopia above the average

milk yield in the baseline data. Evaluating the agent-based models in real-world scenarios will strengthen the assurance that the identified determinants can move smallholder dairy farmers from low to higher milk yield.

Therefore, in cases of limited resources farmers should not go through a trial process to identify a working strategy. To our knowledge, this is the first study that has explored evolvement determinants for farmers in Ethiopia and Tanzania by using unsupervised learning approaches and agent-based modelling and simulation. The reported experiments which are based on artificial intelligence, complement the projections reported for the development of the dairy sector by 2067 as studied in previous research (Britt *et al*., 2018).

Other studies have presented the use of agent-based modelling to explore increase in dairy production and livestock management in general. However, the concept of disaggregation (farm types) and association rules to understand determining factors has not been done in previous research. It is worthwhile to compare this research to other researches on using agent-based modelling and simulations to study livestock management systems especially feeding such as: Schilling *et al*. (2012) and Oudendag, Hoogendoorn and Jongeneel (2014) who studied farmers adaptation to strategies and policies to increase dairy productivity, Mack and Huber (2017) who studied farmers' compliance cost and Nitrogen surplus reduction, and Fust and Schlecht (2018) who studied rangeland management for livestock feeding.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

This chapter concludes on the reported research, key findings, strengths, limitations and recommendations. For the strengths and limitations, key areas of considerations have been highlighted for researchers who will adopt the methodology in similar or related study cases. Recommendations highlights next research directions for extension of the reported study or adoption of the findings in policies and other working documents.

5.1 Conclusions

The reported study meant to identify and study key factors that determine increase in milk yield for smallholder dairy farmers. By using unsupervised learning models, the study presented a characterization approach for smallholder dairy farmers and propose farm types or production clusters for farmers in Tanzania and Ethiopia. With the farm types, are the identified determinants for increase in milk yield which were studied and evaluated by using agent-based modelling and simulation. To this end, the study has outlined the important determinants that can help smallholder dairy farmers from the study locations increase their annual milk yield. In addition, the study has demonstrated the usefulness of farmer networks/groups for peer to peer knowledge sharing to create a knowledge rich society of farmers (targeting to learn new management practices that influence increase in milk yield).

The use of unsupervised learning algorithms in characterization of farming systems is dominated by the use of K-means, hierarchical or Wards, Principal Component Analysis, Self-Organizing Maps, Fuzzy, and Naïve Bayes algorithms. Previous research has indicated use of one algorithm during characterization, leaving behind clusters validation requirement. The presented study identified the need for validation and applied three clustering algorithms in characterization. Based on the divergent performance of the three algorithms evaluated, it is evident that despite similar information being available for study populations, uniqueness of the data from each study site provided an over-riding influence on cluster robustness and prediction accuracy. It is therefore concluded that, application of multiple clustering algorithms result into selection of the best performing algorithm for each case study and stable clusters are guaranteed.

Further clusters validation approach demonstrated the use of regression analysis to establish predictive power of derived clusters. This approach is unique to the reported study and inform

on generation of farm types that can be used to predict production trends. Presented findings highlights that the clusters generated can be used to predict up to 89% of the variances in milk yield and 70% of the variances in milk sales.

From cluster analysis and association rules for frequent patterns, twenty-eight and twenty-five evolvement determinants were identified for Ethiopia and Tanzania, respectively. Use of agent-based modelling and simulation established the differential influence of the evolvement determinants in milk yield maximization. The identified determinants could predict up to 93% and 96% of the variances in milk yield for Ethiopia and Tanzania, respectively. Generally, there was an increase in milk yield based on the identified evolvement determinants; from baseline data average milk yield of 12.7 ± 4.89 and 13.62 ± 4.47 to simulated milk yield average of 17.57 ± 0.72 and 20.34 ± 1.16 for Tanzania and Ethiopia, respectively.

The methods used in this research carry all the weight of its strength. Firstly, the availability of baseline data from study sites in Tanzania and Ethiopia. For a PhD research to have generated such huge amount of data from real smallholder dairy farmers, significant financial and time resources were to be invested. This research kicked off as soon as the proposal was approved because datasets were available.

The clustering approach proposed and used in this research presents a novel way of handling unsupervised learning models in studying evolving real world systems such as the smallholder dairy farming. The ultimate goal of clustering was to obtain clusters that could be used in prediction models, so the need to have robust allocation was inevitable.

Characterization of the smallholder dairy farm types presented in this research details an approach of utilizing frequent pattern analysis to reveal hidden and underrepresented attributes. Based on confidence levels, the frequent pattern analysis could reveal important attributes which though having low representation, their influence could not be ignored. As a result, the characteristics highlighted in the discussion section have been thoroughly assessed and confirmed.

The factors that determine milk yield were studied and ranked prior to the study of farmer networks. Therefore, in the farmer networks that farmer agents interacted and exchanged their knowledge only for the factors that were identified to influence higher milk yield. This approach does validate the proposed determinants for increase in milk yield and that if they

can be adopted by farmers increase in milk yield will be observed. Therefore, this study presents the following contributions to the body of knowledge:

(i) A clusters validation approach that ensures formation of stable types and ones that can be used in prediction studies.
(ii) An approach to study characteristics of smallholder farming systems through cluster analysis and frequent patterns analysis.
(iii) Use of agent-based modelling and simulation with real data to study the differential influence of the determinants for higher milk yield while maintaining disaggregation.

5.2 Recommendations

Structure, cleanness and originality of data sets used in machine learning is important for accuracy of final results. The Tanzania data used in this study yielded challenges in cluster analysis including failure of the fuzzy algorithm to form appropriate number of clusters. Heterogeneity of the records can be questioned since failure of the fuzzy algorithm is attributed to the fact that majority of the records were highly homogeneous. This scenario can be due to collection of data from groups rather than individual surveys. More research is recommended on how clustering can be done for the case of group surveys or when the records of data are suspiciously homogeneous.

Evaluation of the agent-based models in real-world scenarios is recommended to confirm the agent-based models' accuracies. Although the models were developed based on real farm data sets, personalized feedbacks from the farmers and long-term evaluation of at least six months (half lactation period) would yield more accurate results. In addition, development of the agent-based model User Interface (UI) is recommended to enable dairy farmers assess personal evolvement trends and permute strategies that can lead them to increase in milk yield.

This research has outlined key determinants for increase in milk yield for Tanzania and Ethiopia under respective farm types. It is recommended that; dairy development partners and research institutions consider the determinants and their order of importance as given in this research for best results of various intervention programs. While long term development programs (example Public Private partnership for Artificial Insemination Delivery (PAID) and Africa Dairy Genetic Gains (ADGG)) are currently underway, resource allocation from the farmers' side should consider the identified evolvement determinants in order to boost milk yield and strengthen the farmers' desire in dairy farming.

Development of policy briefs as working documents for various dairy development stakeholders would ensure adequate dissemination of knowledge created in this research. The policy briefs and working documents can equally be used by farmers or in farmer groups. Expansion of the study in terms of geographic locations is highly recommended while taking into account the study limitations presented above. The study sites used in this research were adapted from the PEARL project as detailed in the methodology section, it is anticipated that other locations where dairy farming is practiced are bounded to similar challenges of low milk yield.

REFERENCES

Allan, K. (2019). Top ten milk producing countries in Africa, Kenya will surprise you. *Daily Active Kenya*. Retrieved May 30, 2019, from http://dailyactive.info/2019/03/05/top-ten-milk-producing-countries-in-africa-kenya-will-surprise-you/.

Armstrong, D. J. (2006). The quarks of object-oriented development. *Communications of the ACM*, 49(2), 123-128.

Arsanjani, J. J., Helbich, M., & Noronha V. E. (2013). Spatiotemporal simulation of urban growth patterns using agent-based modelling: the case of Tehran. *Cities*, 32, 33-42.

Basel, A. R. F., & Nandi, A. K. (2015). Fuzzy Clustering. In *Integrative Cluster Analysis in Bioinformatics*.

Bazaldua, D. A. L., Baker, R. S., & San Pedro, M. O. Z. (2014). Comparing Expert and Metric-Based Assessments of Association Rule Interestingness. *Proceedings of the 7th International Conference on Educational Data Mining*, 1(917), 44–51.

Benenson, I. (2004). Agent-based modelling: From individual residential choice to urban residential dynamics. *Spatially integrated social science, examples in best practice*, 42(6-7), 67-95.

Berkhout, E. D., Schipper, R. A., Keulen, H. V., & Coulibaly, O. (2011). Heterogeneity in farmers' production decisions and its impact on soil nutrient use: Results and implications from northern Nigeria. *Agricultural Systems*, *104*(1), 63–74. https://doi.org/ 10.1016/j.agsy.2010.09.006.

Bezdek, J. C. (1984). FCM: The Fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2), 191–203.

Britt, J. H., Cushman, R. A., Dechow, C. D., Dobson, H., Humblot, P., Hutjens, M. F., & Jones, G. A. (2018). Invited review : Learning from the future — A vision for dairy farms and cows in 2067. *Journal of Dairy Science*, 101(5), 3722–3741. https://doi.org /10.3168 /jds.2017-14025.

Bruinenberg, M. H., Zom, R. L., & Valk, H. (2002). Energy evaluation of fresh grass in the diets of lactating dairy cows. *Netherlands Journal of Agricultural Science*, *50*, 67–81.

Celebi, E. M., & Kigravi, A. H. (2013). Deterministic Initialization of the K-Means Algorithm Using Hierarchical Clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, *26*(7).

Cherepanov, V., Feddersen, T., & Sandroni, A. (2013). Rationalization. *Theoretical Economics*, 8(3), 775-800.

Chen, S. K., & West, D. (1995). The Comparative Ability of Self-organizing Neural Networks to Define Cluster Structure. *Omega*, *23*(3), 271–279, https://doi.org/10.1016/0305-0483(95)00011-C.

Christian, H., Dalski, J., Thiel-clemen, T., & Grundmann, L. (2016). Modelling & Simulation as a Service with the Massive Multi-Agent System MARS, *Spring Simulation Conference 2016, https://*10.22360/SpringSim.2016.ADS.001.

Cook, R. (2015). World Cattle Inventory: *Ranking of Countries (FAO)*. Retrieved May 30, 2019, from https://www.drovers.com/article/world-cattle-inventory-ranking-countries-fao.

Cortez, A. J., Groot, J. C., Massiotti, R. D. A., Scholberg, J. M., Aguayo, D. V. M., Tittonell, P., & Rossing, W. A. (2014). Resource use efficiency and farm productivity gaps of smallholder dairy farming in North-west Michoacán, Mexico. *Agricultural Systems*, *126*, 15-24.

Cottrell M., Olteanu M., Rossi, F. V. N. (2016). Theoretical and Applied Aspects of the Self-Organizing Maps. In *Advances in Self-Organizing Maps and Learning Vector Quantization.* (pp. 3–26).

Dairy, A. (2015). The eternal question: “how much milk can you get from grass?” *AHDB Dairy*, Retrieved August 3, 2018, from https://dairy.ahdb.org.uk/news/technical-articles/june-2015/how-much-milk-can-you-get-from-grass/#.W2QcI9gzbOQ.

Dossa, L. H., Abdulkadir, A., Amadou, H., Sangare, S., & Schlecht, E. (2011). Landscape and Urban Planning Exploring the diversity of urban and peri-urban agricultural systems in Sudano-Sahelian West Africa : An attempt towards a regional typology. *Landscape and Urban Planning*, 102(3), 197–206. https://doi.org/10.1016/j.landurbplan.2011.04.005.

Elkind E., & Slinko A. (2016). Rationalizations of Voting Rules, *Chapter 8 in Hand Book of Computational Social Choice* 2016, pp. 169-195, Published by Cambridge University Press, ISBN 978-1-107-06043-2, New York, USA.

Epstein, J. M., & Axtell, R. (1996). Growing artificial societies: social science from the bottom up. Brookings Institution Press.

FAO. (2010). Status and Prospects for Smallholder Milk Production A Global Perspective.

FAO. (2014). FAOSTAT.

Faliszewski, P., & Rothe, J. (2016). Control and Bribery in Voting. *Chapter 7 in Hand Book of Computational Social Choice 2016*, 146-168, Published by Cambridge University Press, ISBN 978-1-107-06043-2, New York, USA.

Filatova, T., Verburg, P. H., Parker, D. C., & Stannard, C. A. (2013). Spatial agent-based models for socio-ecological systems: challenges and prospects. *Environmental modelling & software*, 45, 1-7.

Fox, M. S., Barbuceanu, M., & Teigen, R. (2001). Agent-oriented supply-chain management. *In Information-Based Manufacturing*, 81-104.

Frayret, J. M., D'Amours, S., Rousseau, A., Harvey, S., & Gaudreault, J. (2007). Agent-based supply-chain planning in the forest products industry. *International Journal of Flexible Manufacturing Systems*, 19(4), 358-391.

French, N. P., Tyrer, J., & Hirst, W. M. (2001). Smallholder dairy farming in the Chikwaka communal land, Zimbabwe: birth, death and demographic trends. *Preventive Veterinary Medicine*, 48(2), 101–112.

Fust, P., & Schlecht, E. (2018). Integrating spatio-temporal variation in resource availability and herbivore movements into rangeland management: RaMDry—An agent-based model on livestock feeding ecology in a dynamic, heterogeneous, semi-arid environment. *Ecological Modelling*, *369*, 13–41. https://doi.org/10.1016/j.ecolmodel.2017.10.017.

Galluzzo, N. (2015). Technical and economic efficiency analysis on Italian smallholder family farms using Farm Accountancy Data Network dataset. *Studies in Agricultural Economics*, 117, 35–42.

Gelbard, R., Goldman, O., & Spiegler, I. (2007). Investigating diversity of clustering methods : An empirical comparison. *Data and Knowledge Engineering*, 63(1), 155–166. https://doi.org/10.1016/j.datak.2007.01.002.

Ghorbani, A. (2016). Managing the commons : a simple model of the emergence of institutions through collective action. *International Journal of the Commons*, 10(1), 200–219. https://doi.org/10.18352/ijc.606.

Gizaw, S., Abera, M., Muluye, M., Aliy, M., & Alemayehu, K. (2017). Validating the Classification of Smallholder Dairy Farming Systems Based on Herd Genetic Structure and Access to Breeding Services. *Agricultural Sciences*, 2017(8), 545–558. https://doi.org/10.4236/as.2017.87041.

Gondro, C., Okeyo, A. M., Ojango, J., Rao, J., Dessie, T., Rege, J. E. O., & Gibson, J. P. (2018). Genetic diversity of the indigenous cattle of Kenya , Uganda , Ethiopia and Tanzania using high-density SNP data Results and Discussion. *In Proceedings of the World Congress on Genetics Applied to Livestock Production*, Volume Electronic Poster Session - Genetic Gain - Strategies for Local Breeds, 2, 777.

Goswami, R., Chatterjee, S., & Prasad, B. (2014). Farm types and their economic characterization in complex agro-ecosystems for informed extension intervention : study from coastal West Bengal India. *Agricultural and Food Economics*, *2*(5), 1–24. https://doi.org/10.1186/s40100-014-0005-2.

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010). The ODD protocol: A review and first update. *Ecological Modelling*, *221*(23), 2760–2768. https://doi.org/10.1016/j.ecolmodel.2010.08.019.

Guadu, T., & Abebaw, M. (2016). Challenges , Opportunities and Prospects of Dairy Farming in Ethiopia : A Review. *World Journal of Dairy & Food Sciences*, 11(1), 1–9. https://doi.org/10.5829/idosi.wjdfs.2016.11.1.10140.

Gumma, M. K., Thenkabail, P. S., Hideto, F., Nelson, A., Dheeravath, V., Busia, D., & Rala, A. (2011). Mapping irrigated areas of Ghana using fusion of 30m and 250m resolution remote-sensing data. *Remote Sensing*, *3*(4), 816–835.

Lähdesmäki, H. T. V. (2015). Self-organizing and missing values in SOM and GTM. *Neurocomputing*, *147*, 60–70.

Hahn, B. D., Richardson, F. D., Hoffman, M. T., Roberts, R., Todd, S. W., & Carrick, P. J. (2005). A simulation model of long-term climate, livestock and vegetation interactions on communal rangelands in the semi-arid Succulent Karoo, Namaqualand, South Africa. *Ecological Modelling*, *183*(2–3), 211–230. https://doi.org/10.1016/j.ecolmodel.2004.07.028.

Hahsler, M., & Chelluboina, S. (2011). Visualizing Association Rules: Introduction to the R-extension Package arulesViz. Retrieved August 27, 2018, from http://www.comp.nus.edu.sg/~zhanghao/project/visualization/[2010]arulesViz.pdf.

Hahsler, M., & Karpienko, R. (2017). Visualizing association rules in hierarchical groups. *Journal of Business Economics*, 87(3), 317–335. https://doi.org/10.1007/s11573-016-0822-8.

Hartung, C., Lerer, A., & Tseng, C. (2010). Open Data Kit : Tools to Build Information Services for Developing Regions. *ICTD 10 Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, Article 18, 12. https://doi.org/10.1145/2369220.2369236.

Heaton, J. (2016). Comparing dataset characteristics that favor the Apriori, Eclat or FP-Growth frequent itemset mining algorithms. *Conference Proceedings - IEEE Southeastcon*, *2016-July*. https://doi.org/10.1109/SECON.2016.7506659.

Herrero, M., Thornton, P. K., Bernués, A., Baltenweck, I., Vervoort, J., Van de Steeg, J., … Staal, S. J. (2014). Exploring future changes in smallholder farming systems by linking socio-economic scenarios with regional and household models. *Global Environmental Change*, *24*(1), 165–182. https://doi.org/10.1016/j.gloenvcha.2013.12.008.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3), 283-304.

Huang, Z. (1999). A fuzzy k-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems*, 7(4), 446-452.

Hunyadi, D. (2011). Performance comparison of Apriori and FP-Growth algorithms in generating association rules. *Proceedings of the European Computing Conference*, 376–381. Retrieved May 30, 2019 from http://www.wseas.us/e-library/conferences/2011/Paris/ECC/ECC-64.pdf.

Isoda, S. (2001). Object-oriented real-world modelling revisited. *Journal of Systems and Software*, 59(2), 153-162.

Journal, W. A., Ecology, A., Cirad, D. S., & Board, C. (2014). Mapping Fertilizer Recommendations for Cocoa Production in Ghana Using Soil Diagnostic and GIS Tools Mapping Fertilizer Recommendations for Cocoa Production in Ghana Using Soil Diagnostic and GIS Tools. *West African Journal of Applied Ecology*, 17, 97-108 .

Kabacoff, R. I. (2011). R in action: *Data analysis and graphics with R*.

Kamsu-Foguem, B., Rigal, F., & Mauget, F. (2013). Mining association rules for the quality improvement of the production process. *Expert Systems with Applications*, *40*(4), 1034–1045. https://doi.org/10.1016/j.eswa.2012.08.039.

Karmella, B., & Dolecheck, J. (2015). Assessment of the technical and economic potential of automated estrus detection technologies for dairy cattle. *UK College of agriculture food and environment, Department of animal and food science*. Retrieved May 30, 2019, from https://afs.ca.uky.edu/dairy/extension/around-the-world.

Kassambara, A. (2017). The Elbow Method. *In Practical Guide to Introduction in R: Unsupervised Learning*, 101.

Kazuaki, K. (2013). Experiment of Document Clustering by Triple-pass Leader-follower Algorithm without Any Information on Threshold of Similarity. *IPSJ SIG Technical Report*, *2013*(23), 1–6.

Kiesling, E., Günther, M., Stummer, C., & Wakolbinger, L. M. (2012). Agent-based simulation of innovation diffusion: A review. *Central European Journal of Operations Research*, 20(2), 183-230.

Kim, D. W., Lee, K., Lee, D., & Lee, K. H. (2005). A k-populations algorithm for clustering categorical data. *Pattern recognition*, 38(7), 1131-1134.

Kuivanen, K. S., Michalscheck, M., Descheemaeker, K., & Adjei-nsiah, S. (2016). A comparison of statistical and participatory clustering of smallholder farming systems e A case study in Northern Ghana. *Journal of Rural Studies*, *45*, 184–198. https://doi.org/10.1016/j.jrurstud.2016.03.015.

Kurwijila, L. R. (2001). Evolution of dairy policies for smallholder production and marketing in Tanzania. In D. R. and W. Thorpe (Ed.). *Smallholder dairy production and marketing-Opportunities and constraints*, 298–316.

Larman, C. (2005). Applying UML and patterns: an introduction to object-oriented analysis and design and iterative development. *Pearson Education India*.

Otte, J., Costales, A., Dijkman, J., Pica-Ciamarra, U., Robinson, T., Ahuja, V., & Roland-Holst, D. (2012). Livestock sector development for poverty reduction: an economic and policy perspective. *Food and Agriculture Organization of the United Nations (FAO)*.

Maciejewska, A., & Pempkowiak, J. (2015). DOC and POC in the southern Baltic Sea. Part II - Evaluation of factors affecting organic matter concentrations using multivariate statistical methods. *Oceanologia*, *57*(2), 168–176. https://doi.org/10.1016/j.oceano.2014.11.003.

Macmillan, W., & Huang, H. Q. (2008). An agent-based simulation model of a primitive agricultural society. *Geoforum,* 39(2), 643-658.

Mack, G., & Huber, R. (2017). On-farm compliance costs and N surplus reduction of mixed dairy farms under grassland-based feeding systems. *Agricultural Systems*, 154, 34–44. https://doi.org/10.1016/j.agsy.2017.03.003.

Maleko, D., Msalya, G., Mwilawa, A., & Pasape, L. (2018). Smallholder dairy cattle feeding technologies and practices in Tanzania : failures , successes , challenges and prospects for sustainability Smallholder dairy cattle feeding technologies and practices in Tanzania. *International Journal of Agricultural Sustainability*, 1–13. https://doi.org/10.1080/14735903.2018.1440474.

Mallik, S., Mukhopadhyay, A., & Maulik, U. (2015). RANWAR: rank-based weighted association rule mining from gene expression and methylation data. IEEE Transactions on Nanobioscience, 14(1), 59–66. https://doi.org/10.1109/TNB.2014.2359494.

Mangiameli, P. (1996). A comparison of SOM neural network and hierarchical clustering methods. *European Journal of Operational Research*, 93(2), 402 - 417.

Mbwambo, N., Nigussie, K., & Stapleton, J. (2017). Dairy development in the Tanzanian. *Livestock master plan*, 1–4.

Mingoti, S. A., & Lima, J. O. (2006). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research*, 174(3), 1742–1759. https://doi.org/10.1016/j.ejor.2005.03.039.

Milch, B., & Koller, D. (2000). Probabilistic models for agents' beliefs and decisions. *In Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, 389-396). Morgan Kaufmann Publishers Inc.

Moran, J. (2005). Nutrient requirements of dairy cows. *In Tropical dairy farming: feeding management for small holder dairy farmers in the humid tropics.* (pp. 312). Landlinks Press.

Mujibi, F. D. N., Rao, J., Agaba, M., & Nyambo, D. (2019). Performance Evaluation of Highly Admixed Tanzanian Smallholder Dairy Cattle Using SNP Derived Kinship Matrix. *Frontiers in Genetics*, 10(April), 375. https://doi.org/10.3389/fgene.2019.00375.

Mulindwa, H., Galukande, E., Wurzinger, M., Ojango, J., Okeyo, A. M., & Sölkner, J. (2011). Stochastic simulation model of Ankole pastoral production system: Model development and evaluation. *Ecological Modelling*, *222*(20–22), 3692–3700. https://doi.org/10.1016/j.ecolmodel.2011.08.027.

Musa, L. M., Peters, K. J., & Ahmed, M. A. (2006). On farm characterization of Butana and Kenana cattle breed production systems in Sudan On farm characterization of Butana and Kenana cattle breed production systems in Sudan. *Livestock Research for Rural Development*, 18(12).

Mwanga, G., Mujibi, F. D., Yonah, Z., & Chagunda, M. G. (2019). Multi-country investigation of factors influencing breeding decisions by smallholder dairy farmers in sub-Saharan Africa. *Tropical Animal Health and Production*, 2019(51), 395–409.

Nidheesh, N., & Nazeer, K.A., (2017). An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data. *Computers in Biology and Medicine*, (91), 213–221.

Nazari, M., Razzaghi, F., Khalili, D., & Zarandi, S. M. T. (2017). Regionalization of dryland farming potential as influenced by droughts in western Iran. *International Journal of Plant Production*, *11*(April).

Nolan, J., Parker, D., Kooten, G. C. V., & Berger, T. (2009). An Overview of Computational Modelling in Agricultural and Resource Economics. *Canadian Journal of Agriculture Economics*, 57(4), 417–429.

Nyambo, D. G., Luhanga, E. T., & Yonah, Z. O. (2019a). A Review of Characterization Approaches for Smallholder Farmers: Towards Predictive Farm Typologies. *The Sci entific World Journal*. 2019, 9.

Nyambo, D. G., Luhanga, E. T., & Yonah, Z. O. (2019b). Characteristics of smallholder dairy farms by association rules mining based on apriori algorithm. *International Journal of Society Systems Science.* 11(2), 99–118.

Nyambo, D., Luhanga, E., Yonah, Z., & Mujibi, F. N. (2019). Application of Multiple Unsupervised Models to Validate Clusters Robustness in Characterizing Smallholder Dairy Farmers. *The Scientific World Journal*, *2019*, 12. https://doi.org/https://doi.org/10.1155/2019/1020521.

Østergaard, S., Sørensen, J. T., & Kristensen, A. R. (2000). A Stochastic Model Simulating the Feeding-Health-Production Complex in a Dairy Herd. *Journal of Dairy Science*, *83*(4), 721–733. https://doi.org/10.3168/jds.S0022-0302(00)74934-4.

Oudendag, D., Hoogendoorn, M., & Jongeneel, R. (2014). Agent-based modelling of farming behavior: A case study for milk quota abolishment. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence)*, *8481LNAI* (part 1), 11–20. https://doi.org/10.1007/978-3-319-07455-9_2.

Paas, W., & Groot, J. C. J. (2017). Creating adaptive farm typologies using Naive Bayesian classification. *Information Processing in Agriculture*. https://doi.org/10.1016/j.inpa.2017.05.005.

Pelcat, Y., McConkey, B., Basnyat, P., Lafond, G., & Moulin, A. (2015). In field management zone delineation from remote sensing imagery in-field management zone delineation from Remote Sensing. *Indian Head Agricultural Research Foundation*.

Prayaga, L. (2001). Self-Organising Mapping (SOM) networks with SAS e-miner.

Ramos, S., Duarte, J. M., Duarte, F. J., & Vale, Z. (2015). A data-mining-based methodology to support MV electricity customers' characterization. *Energy and Buildings*, *91*, 16–25. https://doi.org/10.1016/j.enbuild.2015.01.035.

Robardet, C., Bruno, C., & Jean, F. J. (2002). Characterization of unsupervised clusters with the simplest association rules : application for child's meningitis. *Intelligent Data Analysis, In proceed*, 61–65.

Salasya, B., & Stoorvogel, J. (2010). Fuzzy classification for farm household characterization. *Outlook on Agriculture*, 39(1), 57–63.

Salami, A., Kamara, A. B., & Brixiova, Z. (2010). Smallholder agriculture in East Africa: Trends, constraints and opportunities. *Tunis: African Development Bank*.

Schilling, C., Kaye-Blake, W., Post, E., & Rains, S. (2012). The importance of farmer behaviour: an application of Desktop MAS, a multi-agent system model for rural New Zealand communities. *Nelson, New Zealand: New Zealand Agricultural and Resource Economics Society*. Retrieved from http://researcharchive.lincoln.ac.nz/handle/10182/5200.

Shiflet, A. B., & Shiflet, G. W. (2014). An introduction to agent - based modelling for undergraduates. *Procedia Computer Science*, 29, 1392–1402. https://doi.org/10.1016/j.procs.2014.05.126.

Smith, J., Sones, K., Grace, D., MacMillan, S., Tarawali, S., & Herrero, M. (2013). Beyond milk, meat, and eggs: Role of livestock in food and nutrition security. *Animal Frontiers*, 3(1), 6-13.

Söderström, M., Isendahl, C., Eriksson, J., Araújo, S. R., Rebellato, L., Schaan, D. P., & Stenborg, P. (2013). Using proximal soil sensors and fuzzy classification for mapping Amazonian Dark Earths. *Agricultural and Food Science*, *22*(4), 380–389.

Suchacka, G., & Chodak, G. (2017). Using association rules to assess purchase probability in online stores. *Information Systems and E-Business Management*, *15*(3), 751–780. https://doi.org/10.1007/s10257-016-0329-4.

Sulabo, R. C. (2011). Energy Evaluation. *Lipids*.

Swai, E. S., Mollel, P., & Malima, A. (2014). Some factors associated with poor reproductive performance in smallholder dairy cows: the case of Hai and Meru districts, northern Tanzania. *Livestock Research for Rural Development*, 6(26).

Tanzania. (2016). Tanzania Livestock Modernization Initiative. *Ministry of Livestock and Fisheries Development.*

Therond, O., Sibertin-blanc, C., Lardy, R., Gaudou, B., Sauvage, S., Taillandier, P., … Mazzega, P. (2014). Integrated modelling of social-ecological systems : The MAELIA high-resolution multi-agent platform to deal with water scarcity problems. *7th International Environmental Modelling and Software Society (IEMSs 2014)*, (2004).

Thorpe, H. G., & Muriuki, W. (2001). Smallholder dairy production and marketing in eastern and southern Africa: Regional synthesis. In W. Rangnekar, D. and Thorpe (Ed.), *Smallholder dairy production and marketing—Opportunities and constraints* (pp. 185–198).

Trauwaert, E. (1988). On the meaningg of dun partition coefficient for fuzzy clusters. *Fuzzy Sets and Systems*, *25*, 217–242.

Van de Steeg, J. A., Verburg, P. H., Baltenweck, I., & Staal, S. J. (2010). Characterization of the spatial distribution of farming systems in the Kenyan Highlands. *Applied Geography*, 30(2), 239–253. https://doi.org/10.1016/j.apgeog.2009.05.005.

Wodajo, W. A., & Ponnusamy, K. (2015). Challenges of generation and dissemination of improved dairy practices : Evidence of literature from Ethiopia. *Global Science Research Journals*, *3*(7), 270–277.

Xu, R. (2005). Survey of clustering algorithms for MANET. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. https://doi.org/10.1109/TNN.2005.845141.

APPENDICES

Appendix 1: Milk yield model pseudocode

```
model Milk yield from individual cows for one lactation period (365 days)
use Mars
layer Grassland
Raster layers definition
raster-layer GrassRaster as biomass
        def EatGrass(cell : Tuple<integer, integer>, amount : real)
                        var graze = biomass.Reduce(cell.Item1, cell.Item2, amount)
                        var growth = biomass.Increase(cell.Item1, cell.Item2, amount)
raster-layer feed_trough as feed
        def eatgrass (cell : Tuple<integer, integer>, amount : real)
                        var fillfeed = feed.Increase(cell.Item1, cell.Item2, amount)
                        var eat = feed.Reduce(cell.Item1, cell.Item2, amount)
                        if (fillfeed > 0) cow eats
raster-layer water_trough as water
        def drinkwater (cell : Tuple<integer, integer>, amount : real)
                        var fill = water.Increase(cell.Item1, cell.Item2, amount)
                        var drink = water.Reduce(cell.Item1, cell.Item2, amount)
                        if (fill > 0) cow drinks
raster-layer Mastitis as mastitis
        def infection(cell : Tuple<integer, integer>, amount : real)
                        var low = mastitis.Reduce(cell.Item1, cell.Item2, amount)
                        var high = mastitis.Increase(cell.Item1, cell.Item2, amount)
                        if(low >= 0) infection probability
agent Observer on Grassland
Begin
        tick (for each simulation time)
                var cells = random(45*45)
                        while(cells > 0)
                        increase the biomass for the specified grid of cells then reduce grid count
                        cells--
            if (simtime >= 180)
        var cellsred = random (45*45)
                while (cellsred > 0)
```

```
                increase biomass at lower rate during dry season then reduce grid
                cellsred--
    var infect = random (45*45) occupied grid cells that can carry infections
        while (infect > 0)
                println "Access to biomass raster at (" + xcor + ", " + ycor +")"
                mastitis.Increase(random(45), random(45), random(2)) // Mastitis infection base probability (0.275) rounded to 0.3*10 - with random generation from 3,2,1, then reduce grid count
                infect--
    println "Current simulation tick " + simtime
    if (simtime >= 180)
        var infectred = random (45*45)
        while (infectred > 0)
                println "Access to biomass raster at (" + xcor + ", " + ycor + ")
                mastitis.Reduce(random(45), random(45), random(3)) reduce the infection probability during dry season, then reduce grid count
                infectred--
End.
agent farmer on Grassland layer, definition of evolvement determinants
  var herdsize : integer = 5
        var feeds : real = 0
        external var area_fodder: real
        external var area_grazing: real
        external var hhh_exp: integer
        external var hhh_yr_sch: integer
        external var land_foodcrops: real
        external var land_cashcrops: real
        external var total_land: real
        external var Totalctl_ownd: integer
        external var no_exotics: integer
        external var no_mlk_cws: integer
        external var lt_sld: real
        external var dist_byr: real
        external var mainly_stallfeed_rn: integer
        external var mainly_stallfeed_dr: integer
        external var freq_water: integer
        external var dist: real
        external var pref_brd_mthd: integer
```

```
        external var sp_dist: real
        external var dew_sp_self: integer
        external var vacc_freq: integer
        external var tms_vist: integer
        external var trning: integer
        external var breed_type: integer
        external var months_purch_fodder: integer
        external var affd_suppl: integer
        external var months_concentrates: integer
        external var prf_byr: integer
        external var Total_cropsale: real

        initialize agent farmer
                pos at #(random(45), random(45))
        Tick
Begin fit evolvement determinants
                var watereff = freq_water*0.52
                var farmerbase = (trning*0.034) + (hhh_yr_sch*0.071) + (hhh_exp*0.01)
                var income = (prf_byr*-0.0084)+(lt_sld*0.22)+(Total_cropsale*-0.0000052)
        var infrustructure = (dist_byr*0.0027)+(tms_vist*0.012)+(dist*     0.067)+(sp_dist*0.015)
                var farm = (area_fodder*0.078)+(area_grazing*-0.095)+(affd_suppl*-
                0.75)+(land_foodcrops*0.0077)+(land_cashcrops*0.29)+(Totalctl_ownd*-
                0.013)+(no_mlk_cws*-0.24)+(no_exotics*-0.019)+(mainly_stallfeed_rn*0.072)+
                (mainly_stallfeed_dr*-0.74)+(breed_type*-
                0.059)+(pref_brd_mthd*0.83)+(dew_sp_self*0.046)
        +(total_land*0.024)+(months_concentrates*-0.023)+(months_purch_fodder*-0.0061)

                        var sink = random(9*9)
while(water.Increase(random(4), random(4), random(2)) < sink)
        water.Increase(random(4),random(4), (random(4) /**  + watereff*/))
        println "Access to sink raster at (" + xcor + ", " + ycor + ")"
        sink--
      if (simtime >= 180)
        var sinkred = random (9*9)
        while (water.Increase(random(4), random(4), random(2)) < sinkred)
Introduce watering strategies into watering frequency
                water.Increase(random(4), random(4), (random(2) + watereff))
```

```
println "Access to sink raster at (" + xcor + ", " + ycor + ")"
        sinkred--
var feeds = random(4*9)
Introduce evolvement determinants into feeding strategies
        while(feed.Increase(random(4),random(9), random(5)) < 5)
feed.Increase(random(4),random(9), (random(5)+ infrustructure +income + farmerbase + farm))
println "Access to feeds raster at (" + xcor + ", " + ycor + ")"
feeds--
if (simtime <= 180)
var feedred = random (4*9)
while (feed.Increase(random(4), random(9), random(3)) < 5)
        feed.Increase(random(4), random(9), (random(2)+infrustructure + income + farmerbase + farm))
println "Access to feeds raster at (" + xcor + ", " + ycor + ")"
        feedred--
def RandomMoveCutGrass() => move to #(xcor + random(20)-1, ycor + random(2)-1)
def RandomMove() => move to #(xcor + random(2)-1, ycor + random(2)-1)
End.
agent cow on Grassland, definition of cow attributes for metabolism and milk yield
observe external var cow_id: integer
var IsAlive : bool = true
var IsHealthy : bool = true
var grazing : bool = true
var energy : real = 0
var maintanance : real
var energy_milk : real
var energy_rest : real
var energy_left : real
var Energy_grass : real
external observe var bodyweight: real
external var milkMax : real
var maxenergy = (6.9* (42.4 * bodyweight**0.75 + 442*milkMax)* (1+(milkMax-15)*0.00165)) as real
external var waterreq : real
var milklevel : real = 0.5
var milklevel1 : real
```

```
        var milklevel2 : real
        observe var milkyield : real = 0
        var Hunger : integer
        var eat : real
        var quenched : real = 0
        var water_res: real = 0
    var foodvalue : real = 0
        external var parity : integer
        external var Mhistory: integer
        external var breed : integer
        observe var prob_env : real
        observe var prob_sick : real = 0
        var gain : real
        var feedscore: real
        var breedscore : real
        var healthscore : real
    var Rule : string = ""
    var observe sstot : real
    var observe ssres : real
        initialize cow
            pos at #(random(15), random(15))
            println "Current position (" + xcor + ", " + ycor + ")"
        tick
Begin
            energy lost for each cow tick
            energy = energy - (1 + random(2))
                if (energy  < 50) IsHealthy = false
                if (energy  < 1) IsAlive = false
            var diff = (maxenergy - energy )
            var rel = (diff / ((maxenergy) as real))
            var hunger = (rel * 100)
            Hunger = hunger as integer
Calculate self-probability of infection
                    Probability from the environment =
            (Math::Abs(mastitis.infection(#(random(15), random(15)), random(3))))/100
                Probability from self = prob_env + ((0.16*Mhistory)/5) +
            ((0.16*parity)/5) + (0.16*(1/breed))/10
```

```
Estimate desire to eat based on health status
                if (foodvalue < 20 and (energy < maxenergy) and (prob_sick < 0.5))
                                                eat = Math::Abs(feed.eatgrass(#(random(10),
                        random(15)), random(2)))
Conversion of grass into energy
                 foodvalue = foodvalue + eat
                 Energy_grass = 16.3 * foodvalue
                 energy = energy + Energy_grass
                RandomMove()
if the cow probs of mastitis > 0.5 feed intake is reduced by 50%
get drinking water whenever food intake is or above 20
        if ((foodvalue >= 20) and (water_res < waterreq) and (prob_sick < 0.5))
         move to water trough
         quenched = quenched + Math::Abs(water.drinkwater(#(random(4), random(4)),
random(10)))
if the cow probs of mastitis > 0.5 water intake is reduced by 50%
milk from the water (1.8kg water = 0.45kg milk)
       milklevel1 =  (((0.45*quenched)/1.8)*0.25)
       water_res = quenched - ((0.45*quenched)/1.8)
reduce maintenance energy
       maintenance = (10 + ((10/100)*bodyweight))
1.25 to produce 1 liter of milk
       milklevel2 = (foodvalue/1.25) * 0.25
energy used to produce the milk
       energy_milk = milklevel2 * 7.1
       energy_left = energy - energy_milk - maintenance - 1 - (1 + random(5))
energy used while resting
       energy_rest = energy_left - 20
energy required to put one kg more weight
      gain = ((energy_left-energy_rest)/44)
      milkyield = milkyield + milklevel1 + milklevel2
To obtain the variance explained by the evolvement determinants on milk yield (Sum of square
differences)
        sstot = (milkMax - 17.59) * (milkMax - 17.59)
        ssres = (milkMax - milklevel2) * (milkMax - milklevel2)
cow vaccination
        if (prob_sick >= 0.45){
```

```
var farmers = nearest farmer
move to farmers
if (farmers.Getvaccination() > 0){
prob_sick = prob_sick/farmers.Getvaccination()
```

End.

Appendix 2: Farm Networks Model Pseudocode

```
model Farm networks, peer-to-peer learning of new strategies to increase milk yield
use Mars
layer Grassland
This is the farmer agents
agent farm on Grassland, definition of evolvement determinants (table 12)
        observe external var Farm_id : integer = 0
        external var dist_market : real = 0.0
        external var hhh_yr_sch : real = 0.0
        external var hhh_exp : real = 0.0
        observe external var peak_bst : real = 0.0
   external var lt_sld : real = 0.0
   external var bulk_sale : real = 0.0
        external var hh_member : real = 0
        external var freq_water : integer = 0
        external var dist : real = 0.0
        external var ex_grazing_dry : integer = 0
        external var ex_grazing_rn : integer = 0
        external var mainly_grazing_rn : integer = 0
        external var mainly_grazing_dr : integer = 0
        external var mainly_stall_dry : integer = 0
        external var mainly_stall_rn : integer = 0
        external var pref_brd_mthd : integer = 0
        external var sp_dist : real = 0
        external var dew_tms : integer = 0
        external var vacc_freq : integer = 0
        external var no_mlk_cws : integer = 0
        external var total_cattle : integer = 0
        external var total_land : real = 0.0
        external var area_fodder : real = 0.0
        external var dist_byr : real = 0.0
        external var dew_sp_self : integer = 0
        external var trning_days : integer = 0
        external var breed_type : integer
        external var mnths_cropresidue : integer = 0
        external var mnths_concentra : integer = 0
        external var mnth_purc_fodder : integer = 0
        external var freq_extension_visit : integer = 0
        external var no_emp : integer = 0
        external var affd_suppl : string
        external var cluster : integer = 0
        observe var predicted_milk : real = 0.0
```

observe var sstot: **real** = 0.0
observe var ssres: **real** = 0.0
observe var count_interact: **integer** = 0
var isAlive : **bool** = **true**

Initialize the farm households in random locations within specified grid limits

tick

Begin

First self-assessment: agent farmer is trying to predict its yield based on it's evolvement determinants' values – coefficients for each determinant, intercept and error term as indicated in Section 4.1.4 is shown (Given for Tanzania case study)

$$milk_{yield} = -\beta - \sum_{i=1}^{12} x_i + \sum_{i=13}^{25} x_i + \varepsilon \qquad (8)$$

predicted_milk = -0.0389 - 0.14*vacc_freq + 0.37*freq_water -0.06*total_land - 0.21*area_fodder -0.27*mainly_stall_dry -0.27*mainly_stall_rn
-0.08*no_mlk_cws -0.35*pref_brd_mthd +0.12*mainly_grazing_dr +0.12*mainly_grazing_rn -0.88*no_emp -0.03*mnth_purc_fodder +0.05*total_cattle
+0.03*hhh_exp -0.03*trning_days +0.17*hhh_yr_sch +0.91*hh_member - 0.0001*bulk_sale +0.27*lt_sld +0.48*mnths_cropresidue -0.002*dist_byr
+0.05*dist +0.005*freq_extension_visit +0.02*sp_dist +0.05*dist_market +3.089

After one month

If yield is low than the farm type average, the agents try to adjust their, determinants' values based on the farm type characteristics, then begin to learn from their peers after attaining the farm type milk yield standard.

Adapt farm type characteristics first

if ((cluster === 1)
Define determinants and values for farm type one
else if (cluster ===2))
Define determinants and values for farm type two
else if (cluster ===3)
Define determinants and values for farm type six
else if (cluster ===4)
Define determinants and values for farm type six
else if (cluster ===5)
Define determinants and values for farm type five
else if (cluster ===6)
Define determinants and values for farm type six

then predicts again, in each tick this procedure is executed until the agent yield is greater than that of the farm type average

Fit prediction model with updated strategies

After working on the farm type characteristics now start exploring for higher producers than self

if (**simtime** >= 146)
var nearestfarm = **nearest** farm
move to nearestfarm
if (nearestfarm.**Getpredicted_milk** > predicted_milk)
if (vacc_freq < nearestfarm.**Getvacc**())
vacc_freq = nearestfarm.**Getvacc**()
Update interaction count

```
                    if (freq_water < nearestfarm.Getfreq())
                            freq_water = nearestfarm.Getfreq()
                            Update interaction count
                    if (mainly_stall_dry > nearestfarm.Getstall())
                            mainly_stall_dry = nearestfarm.Getstall()
                            Update interaction count
                    if (mainly_stall_rn > nearestfarm.Getstall2())
                            mainly_stall_rn = nearestfarm.Getstall2()
                            Update interaction count
                    if (pref_brd_mthd > nearestfarm.Getbreeding())
                            pref_brd_mthd = nearestfarm.Getbreeding()
                            Update interaction count
                    if (mainly_grazing_dr < nearestfarm.Getgrazing())
                            mainly_grazing_dr = nearestfarm.Getgrazing()
                            Update interaction count
                    if (mainly_grazing_rn < nearestfarm.Getgrazing2())
                            mainly_grazing_rn = nearestfarm.Getgrazing2()
                            Update interaction count
                    if (no_emp > nearestfarm.Getlabor())
                            no_emp = nearestfarm.Getlabor()
                            Update interaction count
                    if (mnth_purc_fodder > nearestfarm.Getpurchfodder())
                            mnth_purc_fodder = nearestfarm.Getpurchfodder()
                            Update interaction count
Fit prediction model with updated strategies
                    if (simtime >291)
                if ((peak_bst <= 20.00) or (predicted_milk <= 17.65))
                    var nearestfarm = nearest farm
                  move to nearestfarm
                  if (nearestfarm.Getpredicted_milk > predicted_milk)
                    if (mnths_cropresidue > nearestfarm.Getcropresidue())
                            mnths_cropresidue = nearestfarm.Getcropresidue()
                            Update interaction count
Fit prediction model with updated strategies
After working on the income characteristics move to infrastructure characteristics
        if (simtime > 291){
                if ((peak_bst <= 20.00) or (predicted_milk <= 20.00)){
                    var nearestfarm = nearest farm
                  move to nearestfarm
                  if (nearestfarm.Getpredicted_milk > predicted_milk){
                  if (freq_extension_visit < nearestfarm.Getextension()){
                            freq_extension_visit = nearestfarm.Getextension()
                            Update interaction count
Fit prediction model with updated strategies
                if (simtime >= 436){
                if ((peak_bst <= 20.00) or (predicted_milk <= 20.00)){
                  var nearestfarm = nearest farm
                  move to nearestfarm
                  if (nearestfarm.Getpredicted_milk > predicted_milk){
                  if (trning_days < nearestfarm.Getstraining()){
                            trning_days = nearestfarm.Getstraining()
                            Update interaction count
                   if (hh_member < nearestfarm.Getmember())
                            hh_member = nearestfarm.Getmember()
                            Update interaction count
```

```
Fit prediction model with updated strategies
                    if (simtime >= 581)
                        var nearestfarm = nearest farm
                      move to nearestfarm and interact about better strategies than self
                      if (nearestfarm.Getpredicted_milk() > predicted_milk)
                        if (vacc_freq < nearestfarm.Getvacc())
                            vacc_freq = nearestfarm.Getvacc()
                            Update interaction count
                        if (freq_water < nearestfarm.Getfreq())
                            freq_water = nearestfarm.Getfreq()
                            Update interaction count
                        if (mainly_stall_dry > nearestfarm.Getstall())
                            mainly_stall_dry = nearestfarm.Getstall()
                            Update interaction count
                        if (mainly_stall_rn > nearestfarm.Getstall2())
                            mainly_stall_rn = nearestfarm.Getstall2()
                            Update interaction count
                        if (pref_brd_mthd > nearestfarm.Getbreeding())
                            pref_brd_mthd = nearestfarm.Getbreeding()
                            Update interaction count
                        if (mainly_grazing_dr < nearestfarm.Getgrazing())
                            mainly_grazing_dr = nearestfarm.Getgrazing()
                            Update interaction count
                        if (mainly_grazing_rn < nearestfarm.Getgrazing2())
                            mainly_grazing_rn = nearestfarm.Getgrazing2()
                            Update interaction count
                        if (no_emp > nearestfarm.Getlabor())
                            no_emp = nearestfarm.Getlabor()
                            Update interaction count
                        if (mnth_purc_fodder > nearestfarm.Getpurchfodder())
                            mnth_purc_fodder = nearestfarm.Getpurchfodder()
                            Update interaction count
                      if (mnths_cropresidue > nearestfarm.Getcropresidue())
                            mnths_cropresidue = nearestfarm.Getcropresidue()
                            Update interaction count
                         if (freq_extension_visit < nearestfarm.Getextension())
                            freq_extension_visit = nearestfarm.Getextension()
                            Update interaction count
                         if (trning_days < nearestfarm.Getstraining())
                            trning_days = nearestfarm.Getstraining()
                            Update interaction count
                         if (hh_member < nearestfarm.Getmember())
                            hh_member = nearestfarm.Getmember()
                            Update interaction count
Fit prediction model with updated strategies
Sum of square differences to calculate the proportion of variance explained by the update of strategies
        sstot = (peak_bst - 12.72) * (peak_bst - 12.72)
        ssres = (peak_bst - predicted_milk) * (peak_bst - predicted_milk
Return to your peers and give access for all the values for the evolvement determinants

End.
```

Appendix 3: Coefficients and determinants used in equations 18 and 19

For Tanzania, predicted_milk = -0.0389 - 0.14*vacc_freq + 0.37*freq_water -0.06*total_land -0.21*area_fodder -0.27*mainly_stall_dry -0.27*mainly_stall_rn
-0.08*no_mlk_cws -0.35*pref_brd_mthd +0.12*mainly_grazing_dr +0.12*mainly_grazing_rn -0.88*no_emp -0.03*mnth_purc_fodder +0.05*total_cattle
+0.03*hhh_exp -0.03*trning_days +0.17*hhh_yr_sch +0.91*hh_member -0.0001*bulk_sale +0.27*lt_sld +0.48*mnths_cropresidue -0.002*dist_byr
+0.05*dist +0.005*freq_extension_visit +0.02*sp_dist +0.05*dist_market +3.089

For Ethiopia predicted_milk = 9.09+0.078*area_fodder-0.095*area_grazing-0.013*Totalctl_ownd-0.24*no_mlk_cws-0.019*no_exotics+0.072*mainly_stallfeed_rn
-0.74*mainly_stallfeed_dr-0.023*months_concentrates+0.52*freq_water-
0.0061*months_purch_fodder+0.83*pref_brd_mthd-0.059* breed_type
+0.046*dew_sp_self+0.075*vacc_freq+0.024*total_landsize-
0.75*affd_suppl+0.0077*land_foodcrops+0.29*land_cashcrops+0.071*hhh_yr_sch
+0.034*trning+0.01*hhh_exp+0.22*lt_sld-0.0000052*Total_cropsale-0.0084*prf_byr-
0.067*dist+0.012*tms_vist+0.0027*dist_byr+0.015*sp_dist + 3.87