Modeling Farmers Intensi cation Decisions with a Bayesian Belief Network: The case of the Kilombero Floodplain in Tanzania

B.H. Gebrekidan;

University of Bonn, Institute of Food and Resource Economics, Germany

Corresponding author email: bisrat.gebrekidan@ilr.uni-bonn.de

Abstract:

Modeling farmers intensication decision requires a model that considers the dependencies between the perceived influences and their choices of intensication pathways, accounting uncertainties at the same time. A combination of data-driven Bayesian Belief Network (BBN) and Regression Tree is proposed in this paper. Data from 304 rural households in Kilombero Valley Floodplain in Tanzania is used to learn the structure and parameter of the model. The resulting BBN is able to drive the probabilities of intensication choices conditional on key market, biophysical and socio-economic characteristics of farm households.

Acknowledegment: This research was conducted under the GlobE Wetlands project, which is funded by the German Federal Ministry of Education and Research (FKZ: 031A250 A-H), with additional funding provided by the German Federal Ministry for Economic Cooperation and Development.

JEL Codes: D01, C11

#846

Modeling Farmers Intensification Decisions with a Bayesian Belief Network: The case of the Kilombero Floodplain in Tanzania

Abstract

Modeling farmers intensification decision requires a model that considers the dependencies between the perceived influences and their choices of intensification pathways, accounting uncertainties at the same time. A combination of data driven Bayesian Belief Network (BBN) and Regression Tree is proposed in this paper. Data from 304 rural households in Kilombero Valley Floodplain in Tanzania is used to learn the structure and parameter of the model. The resulting BBN is able to drive the probabilities of intensification choices conditional on key market, biophysical and socio-economic characteristics of farm households.

Contents

1. Introduction

Achieving food security while promoting sustainable development are at the top of priorities for the government of Tanzania (GOT). In fact, like most Sub Saharan African countries, the agricultural sector is the main stay of the country's economy, and a key driver for rural development. The sector still continues to provide employment to around 78 per cent of the total workforce and provides livelihoods to more than 70 percent of the population; it contributes to approximately 95 percent of the national food requirements; and it is the single largest contribution to gross domestic product and export accounting for about half of the total [Milder et al., 2013]. However, the sector remains largely subsistence, with population growth surpassing production growth, food self-sufficiency declining, and the numbers of malnourished people consequently rising.

The GOT vigorously pursues a policy that increases in agricultural production in the country could be an engine of economic growth, driven either by a shift to large scale commercial farms, or by improved productivity on smallholders through providing opportunity and access to resources. The idea highlighted an idea expressed on different policy statements and national visions including Kilimo Kwanza (Agriculture First)(2009), the Southern Agricultural Growth Corridor of Tanzania (SAGCOT) (SAGCOT 2010) and Big Results Now (2012)[Coulson, 2015].

One of these major focal areas for the Tanzanian government in its bid to transform towards a sustainable food basket and eradicate poverty is Kilombero valley floodplain wetland (KVFP). The low-lying plain with alluvial deposits is endowed with a productive natural resource base, fertile land, reliable water availability and extensive pastures [Bamford et al., 2010; Nindi et al., 2014a]. The floodplain is home to more than 500 thousand people[2012 Census] and provides essential raw material, income and nutrition benefits in form of crop production, fish, drinking water, forest products, and fuel wood for households [Mombo et al., 2011]. Like other floodplains, the KVFP also provides remarkably diverse array of ecosystem services including recreational facilities and aesthetic values [Kangalawe and Liwenga, 2005; Kato, 2007; Milder et al., 2013], opening further opportunities.

However, the supply of productive land is increasingly constrained due to intense competition between different groups; small holder farmers, migrating pastorals and agro-pastoral, large scale commercial ventures and Governmental and non governmental conservation groups [Bamford et al., 2010; Dinesen, 2016; Kato, 2007; Milder et al., 2013]. Mainly small-holder farmers are under considerable pressure: pressure to sell or lease their land to national and international investors; pressure to engage in national markets; and intensify production to meet their households food requirements [Kangalawe and Liwenga, 2005; Snyder and Cullen, 2014].

In the past, smallholders in the floodplain have enjoyed abundant land to increase their agricultural production by substantially increasing area under cultivation and expanding to marginal lands and bringing new wetland areas under cultivation [Bamford et al., 2010][see figure A.7]. This land use change in turn is usually associated with various negative environmental consequences, such as loss of habitat and above- and underground biodiversity [Jones et al., 2012]. However, with various pressures like surging population growth and in-migration into the valley, it is argued that agricultural intensification becomes the rule rather than an option [Binswanger and Pingali, 1988; Otsuka et al., 2013]. As Otsuka et al. [2013] argued "the tension caused by increasing scarcity of resource stimulates technological changes to save those resource and institutional support for those technologies also change". And the contemporary situation in KVFP is exhibiting such scenario. The need for increased agricultural production has recently led GOT and various NGOs to promote the use of optimized/high-quality inputs, adoption of new technologies or mechanization and value-chain development as means to increase productivity and closing yield gaps for generating sustainable and inclusive growth and reducing poverty [Agra, 2016].

The literature on agricultural intensification, its drivers, challenges and mechanisms are well established. One of the earliest and most dominant hypothesis of drivers of agricultural intensification are from Boserup [1965, 1981] and Lele and Stone [1989]; Ruthenberg et al. [1980], in which population growth results in increasing hardship in providing livelihoods, causing the farmers to opt for more intensive agriculture. This is reflected in increased land use intensities, e.g. shorter fallow periods to regenerate soil fertility and more frequent annual cropping and adoption of improved agricultural technologies [Jayne et al., 2014; Nin-Pratt, 2015; Okike et al., 2001; Otsuka et al., 2013].

At the end, the pressure to adapt to the changing environment and for production increase resides on the individual farmer's decision. Like most African rural farm population, farmers' livelihood strategies are usually in line with their agricultural production strategy. When faced with a decision to intensify their production, smallholders have multiple pathways or strategies they take and the choices of these strategies are usually conditioned on a number of socio-economic characteristics and the ecological context of their farming environment. According to Schelhas [1996], the choices that farmers make are not simply mechanistic responses to population density, economic conditions, and environmental factors. The level of acceptable production for a household is determined by needs to feed, clothe, house, and educate a family and to meet social obligations and they continuously have to adjust their farming strategy to cope with challenges they face. The underlying assumption is that the decision makers themselves are the experts on how they make the choices they make and their decision is influenced by their own knowledge and perception Darnhofer et al. [1997]

However, only little is known of the relative contribution of these factors on the choice of agricultural strategies of farmers in KVFP, an area particularly important for conservation and development objectives Milder et al. [2013].

It is in this light that further analysis of choices of intensification strategies in the floodplain is required. In order to understand the state of agricultural intensification and its role, it is important to understand the contexts in which farmers operate and to identify and understand different factors that influence their decision making. Our central focus is on the determinants and results of farmers decisions to uptake different paths of intensification and land management practices. We specifically focus on four land saving intensification strategies practiced in the study area, mainly 1) use of chemical fertilizers, 2) use of improved seed, 3) implementation of small-scale irrigation system, and 4) increasing frequency of planting. This is also helpful to efforts aiming to upscale these strategies and identifying indicators farmers use to prioritize strategies Leonard et al. [2011].

Our paper offers two novel contribution to the existing literature.First, we propose the possibility of using Bayesian belief network (BBN) modeling as an alternative tool to existing models. Different modeling tools have been proposed to understand the decision making and choices of individuals. Modeling paradigm ranging from , Probit models[Abay et al., 2016], Logistic regression models [Erenstein, 2006; Okike, 2001; Perz, 2003] and decision trees [Gladwin, 1980] are few to mention [see [Besley and Case, 1993] for modeling farmer adoption decision]. In this paper, A Bayesian belief network(BBN) approach was adopted as a modeling tool for identifying important factors in explaining probabilities of intensification choices. The methodological motivations behind this particular study are manifold. First, given all the uncertainties inherent in modeling farmers decision making due to our current understanding of the decision making process, the vulnerability of agriculture to random events such as changes in weather, uncertainty related to data and observation etc. ,we opted to use probability theory as our foundation to explicitly deal with uncertainty. And the Bayesian approach to uncertainty ensures that the system as a whole remains consistent and provides a way to apply the model to data [Koski and Noble, 2011]. As BBNs are joint probability distribution, uncertainty is propagated through the model and presented in the final results. Contrary to deterministic models, the probabilistic representation of knowledge in BBN prevents overconfidence in the strength of responses obtained by simulating changes in one or more variables of interest [Uusitalo, 2013]. Second, Unlike other 'black box' models, BBN provide generality and formalism of displaying relationships clearly and intuitively [Daly et al., 2011; Margaritis, 2003], making them amenable to analysis and modification by experts and stake-holders [Daly et al., 2011; Sun and Müller, 2013; Uusitalo, 2013]. At the same time incorporating the qualitative beliefs and attitudes of stake holders along with quantitative data. The other main advantage of BBN is updatability where BBN can learn from minimal data in data poor setting and the model parameters and structure can be updated as more data become available.

Second, by investigating how farm households make their intensification decisions when multiple pathways are available in KVFP, We highlighted factors driving the choice of alternative strategies in an area of high potential but ecologically sensitive landscape.

2. Material and Method

2.1 Bayesian Belief Network Modeling

BBN also known as Bayesian Net,Causal Probabilistic Network,Bayesian Network or simply Belief Network is a probabilistic graphical modeling tool that allows for knowledge representation and support for reasoning under uncertainty [Kjaerulff and Madsen, 2012; Korb and Nicholson, 2010; Pearl, 2011]. As other graphical models the nodes represent stochastic variables and the arcs represent direct causal dependencies based on process understanding, statistical, or other types of associations between the linked variables [Pollino et al., 2007]. More formally Bayesian network can be described as an acyclic directed graph (DAG) which defines a factorization of a joint probability distribution over the variables, where the factorization is given by the directed links of the DAG. More precisely, for a DAG, $\eth = (V, E)$, where V denotes a set of nodes and E a set of directed links (or edges) between pairs of the nodes, a joint probability distribution, $P(X_V)$, over the set of (typically discrete) variables X_v indexed by V can be factorized as

$$P(X_V) = \prod_{v \epsilon V} P(X_v | X_{pa(v)})$$

where $X_{pa(v)}$ is a set of parent nodes for variable X_v, for each node v an element of V [Kjaerulff and Madsen, 2012].

BBN uses Bayes theorem and probability calculus to represent a causal linkage between two connected stochastic variables. For instance $X \rightarrow Y$, where X directly influences Y, we need to drive the posterior probability distribution $P(X|Y = y)$ using the prior distribution $P(X)$ and the conditional probability distribution $P(Y|X)$.

Once a BBN is built it can be used to answer any question posed in a probabilistic form and can be answered correctly and with a level of confidence. These questions are usually restricted to determining the most likely hypothesis or, more specifically, the belief in, or probability of, each possible hypothesis, given the available observations or evidence [Daly et al., 2011]. Generally BBNs can be used to make two different types of inferences. The first is top-down inference[predictive tool], this involves finding the belief(probability) of query node(target node) being in certain state , given the other nodes(variables) are set to a certain values. Second, bottom-up inference[diagnostic tool], is finding the probability of the states of a given set of nodes that best explain why our target variable is set to certain value[Daly et al., 2011; Frank, 2015].

Designing a Bayesian belief network involves generally three steps that the modeler must undertake [Cain, 2001a; Choi et al., 2011; Korb and Nicholson, 2010; Marcot et al., 2006].(1)Variable selection and feature engineering , (2)Estimating the structure of the network and (3) Populating the network with CPTs.

BBNs emerged from Artificial Intelligence field and widely used in diverse domains, including medical field, environmental modeling and natural resources management and forecasting Daly et al. [2011]; Korb and Nicholson [2010]; Uusitalo [2013]. Although their application in farming system literature is limited, there are some application of BBN. For example Cain [2001b] used BBN to explore the determinants of crop yield. Sun and Müller [2013] combined Bayesian Belief network with opinion dynamics modeling and agent based model for simulating land-use decision making under the influence of payments for ecosystem services. The BBN was used to capture the choice of farmers weather to participate or not in payment for environmental service program. In similar work, Frayer et al. [2014] developed a BBN to analyze the proximate causes and underlying drivers of the decision to plant of trees on previous cropland. Aalders [2008] and Celio and Grêt-regamey [2016] built a Bayesian belief network to incorporate farmers choices of different land use options. Rasmussen et al. [2013] developed a large scale BBN tool for risk management in EU agriculture using Farmers Agricultural Data Network data (FADN). Pope and Gimblett [2017] used BBN in combination to agent based modeling to explore the different ranching strategies farmers choose under varying environmental conditions. When it comes to agricultural adoption literature, to the best of our knowledge, this study is the first to apply the Bayesian belief network to model farmers adoption decisions. And we argue that BBN can proved an alternative and/or additions to existing and well established statistical models. BBNs provides advantage of explicitly taking in to account uncertainty, integrate of wide range of input data including expert knowledge and easily adaptability of both the structure and dependencies between different influencing factors.

2.2 Data and Study Area

2.2.1 Study Area

The Kilombero Valley, located in the Ulanga and Kilombero Districts in southern Tanazania, forms one of the four principal sub-basins of the Rufiji River Basin and comprises a myriad of rivers and seasonally flooded marshes and swamps [Dinesen, 2016]. The seasonal change in water dynamic is huge and the plains sometimes becomes totally flooded during the wet season, while it dries up during the dry season with the exception of the rivers and river margins as well as the areas with permanent swamps and water bodies[Kato, 2007; Ntongani et al., 2014].

The Valley lies at the foot of the Great Escarpment of East Africa in the southern half of Tanzania, about 300 km from the coast[Kato, 2007; Nindi et al., 2014b]. It covers an area of about 11,600 km^2, with a total length of 250 km and width of up to 65 km. The elevation within the basin is about 300 m above sea level. Generally, the floodplain is humid with high temperatures ranging from 26°C to 32°C. The KVFP is typical fertile alluvial floodplain with loamy, clay, clay loamy and sandy soils and is an important source of nutrients and sediment [Milder et al., 2013; Nindi et al., 2014b].

The KVFP is of global , regional , national importance in terms of ecology and biodiversity. It comprises the kilombero Game controlled area which approximately 7000 km2 and kilombero valley ramser site which covers 7,0679 km^2 [Dinesen, 2016; Nindi et al., 2014b].

As one of Africas largest wetlands, the Kilombero Valley has a long history of productive activities, primary for farming [Kato, 2007; Rebelo et al., 2010]. And in resent year an increase in agricultural land use has been widespread and rapid [Jones et al., 2012]. Immigration into the valley has increased dramatically due to the perceived availability of high quality and cheap farmland. Conflicts between the pastoralists and farmers over land use is a chronic and widespread problem, which has resulted in injury and litigation disputes[MALF, 2015].

2.2.2 Data

We conducted a household survey in 21 villages in two Districts of the Kilombero Valley, Ulanga and Kilombero. In total 304 farm households were interviewed using a

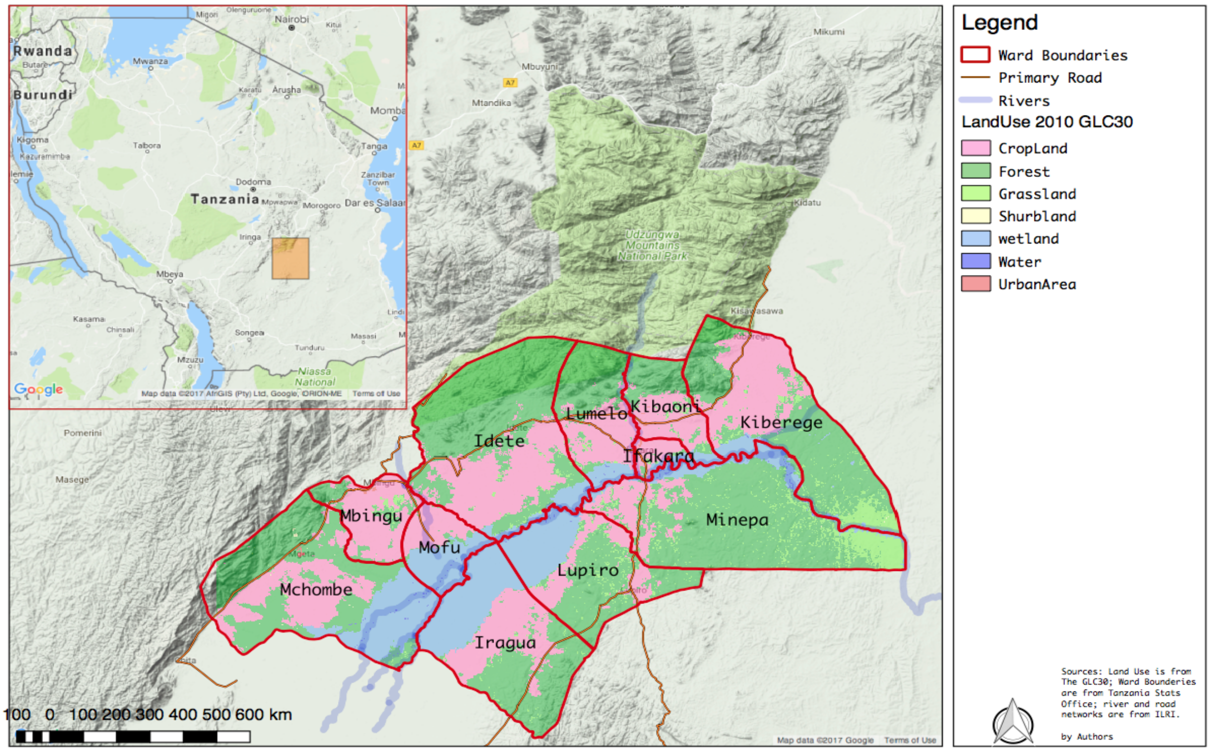


Figure 2.1: Study Area

standard questionnaire, giving their opinions upon a wide range of topics designed to discover the farming system in terms of resource availability and use, livelihood source.

The selection of households to be interviewed was based on a multi stage sampling strategy. In the first stage 12 wards were selected purposively based on the availability of floodplain farming. In the second stage 21 villages were selected randomly within the wards. In the final stage households were selected randomly from the list provided by each villages leader. The number of interviewees per village ranges from 5 in smaller villages to 15 in the biggest. A GIS coverage incorporating the land use map form GLC30 [Jun et al., 2014] and the administrative boundary and census data from Tanzania statistics office was use to estimate the boundaries and total population size in the study area.

2.2.3 Variable Selection and Feature Engineering

From the sample survey we selected those variables considered most relevant to determine the adoption of agricultural technologies pursued by particular households. These variables are selected based on a systemic review of the key features from the literature. A significant body of research on farmers intensification decision informs us about the processes driving intensification choices [Abay et al., 2016; Erenstein, 2006; Headey et al., 2014; Howley et al., 2012; Okike, 2001; Shriar, 2000, 2001]. Feder et al. [1985] reviewed the vast amount of literature on adoptions of different technologies in developing countries. Looking at the broader perspective, the adoption of agricultural technology depends on a range of personal, social, cultural, economic and ecological factors, as well as on the characteristics of the strategy itself. In order, to select the final subsets of variables, we used a decision tree algorithm and random forest algorithm to check feature importance.

The selected variables span a combination of Household-level factors such as households endowments(eg. Available income,off-farm income opportunities, access to credit,the size of the farm household, availability of both family and hired labor, the quantity and quality of land) , access to input and output markets which is expected to increase the use of purchased inputs and the capital intensity of agriculture by increasing the profitability and availability of such inputs. And variables representing the environment in which the farmer is operating agro-ecological characteristics, such as soil quality and the farming system also influence adoption choices. In addition a variables representing the farming system followed by a particular household and the crop choices also influences the decision [Okike, 2001].

Our interest variable (target node) intensification is a discrete node that contains seven states representing five strategies, one state to represent absence of intensification and the others state that captures the remaining combinations of strategies that are not observed in the data for this particular period. Here we can take the advantage of BBNs to updating the conditional probabilities once new data become available. The following table shows the description of 15 evidence nodes that are included in the final network. We used per capita income as a surrogate for a resource endowment and availability of capital. It comprises income from Agriculture(farming and fishing) ,income from non farm activities, income from land rental and brick making. We also included farmer type to represent the farming system followed by a particular household. A farm type variable is a typology created through Non parametric Multivariate Analysis to artificially stratify farmers in to clusters that are homogeneous according to their livelihood and land use. In order to capture the quality and hydrological characteristics of the farm, we generated a topographic wetness index using digital elevation model of our study area(slope and upslope contributing area). The index provides an indication of the relative wetness with in the catchment and highly correlated with soil moisture and ground level water [Sörensen et al., 2006]. Prices received by those who market their crop will have an effect on the crop choices they make and also the income they receive We included prices received by household for rice and maize(which is dependent on the distance from the market) as expected prices for the two crops. Due to lack of past price data, here we assume farmers received what they expected during the planting period. Our distance variable measure the distance from the farm to the nearest big market in km. Measuring the distance from the farm rather than homestead takes in to account the access to and cost of transport from the farm to either farmers home stead or the market(since most farms are located bottom valley quite far from home stead).

Although BBNs are capable of handling continuous nodes, the existing software tools are limited in terms of capability of including continuous variable as a node. Hence, we discretized all continuous variables to different discrete bins. There a number of ways to discretize continuous variables. We partition our continuous variables using a heuristic method called equal frequencies where the variable is transformed to K equal lengths or width [Clarke and Barton, 2000; Nojavan A. et al., 2017].

Variable	Description and Measurement
ageOfTheHouseholdHead	age of the household head expressed in years
SizeOfTheHousehold	total number of individuals living in the household
SizeOfTheCropLand	total size of farm in Hectares
LabourInManDays	Labour in man days available
ShareOfHiredLabour	share of the hired labour
FarmerType	farmer typology based on livelihood and landuse
CommercializationIndex	share of output sold [a combined index of all crops]
DistanceFromTheBigestMarket	Distance in Km from the main input and output market
Income	total income per individual
PercentOfNonFarmIncome	share of income from non agricultural activities
CreditAccess	whether the household has access to credit services
Topographic Wetness Index	surrogate for biophysical characteristics of the plot[highly correlated with soil moisture]
ricePrice	Expected Rice Price
maizePrice	Expected maize Price
CropChoice	farmer choice of crops to plant
ChoiceOfIntensificationStrategy	choice of intensification strategy

A descriptive analysis of our variables in terms of distribution and correlation between variables are provided in the appendix.

3. Result

3.1 Structure Learning

Once the variables are selected and feature engineering is done[discretising continuous variables, define the nodes and their states] the next step is to learn the structure of the network that encodes the interdependencies between variables. There are two different approaches to build the structure of a BBN. (1) Learning through knowledge engineering from experts and literature or theory (2) learning from empirical data. In this study we used the second approach , learning the BBN directly from the data. However,following [Sun and Müller, 2013] we augmented our approach based on theory and conceptual intuition. In cases where the links are mathematically correct but intuitively not acceptable either we reverse or remove the arc.

There exist different classes of algorithms for learning the structure of a Bayesian network from a data[For detailed explanation of learning algorithms see Koller and Friedman [2009] Part III and chapter seven of [Nielsen and Jensen, 2009]].Generally there are two different classes of algorithms for learning from a data:-

- Constraint-based structure learning: These approaches view a Bayesian structure learning network as a representation of independences. Using some statistical tests (such as chi-squared or mutual information),the approach try to test for conditional dependence and independence in the data and use these relationships as constraints to construct a BN [Koller and Friedman, 2009; Neapolitan et al., 2004].

- Score-based structure learning: Score-based methods is an optimization-based search approach that considers a Bayesian network as specifying a statistical model and produce a potential of candidate Bayesian networks, calculate a score for each candidate, and return a candidate of highest score [Kjaerulff and Madsen, 2012; Nielsen and Jensen, 2009].

For our study we adopted constraint based structure learning called tree-augmented naïve Bayesian (TAN) network[Friedman et al., 1997]. TAN models are a restricted family of Bayesian networks in which the class variable has no parents and each attribute has as parents the class variable and at most one other attribute [Cerquides and

De Màntaras, 2003]. Tree augmented naive Bayes is a semi-naive Bayesian Learning method which has an advantage over the popular naïve Bayesian . It relaxes the naive Bayes attribute independence assumption by employing a tree structure and choose the tree that maximizes the likelihood of the training data[Friedman and Koller, 2003; NorsysSoftwareCorp, 2016; Zheng and Webb, 2010]. According to [Friedman et al., 1997], learning the structure of a network using TAN embodies a good trade-off between quality of estimation of correlation between predictors and the computational complexity. And the learning procedure is guaranteed to find an optimal TAN structure

The structure was developed with the application and a Java API version of Netica (5.4) [NorsysSoftwareCorp, 2016]. Netica provides a number of simplifying tasks for the modeler including high visual capability to display the the network and advanced algorithms to learn the structure and parameters of the network. Using the Java API for construction of BBNs provides an advantage in terms of transparency, reproducibility and easily integration with other modeling tools of interest.

3.2 Parameter Learning

As in the structure learning of BN, there are several possible ways of generating estimates for the conditional probabilities (CPTs). In this study, the probabilities were derived from survey data using maximum likelihood estimation since learning these parameters from observed data provide some level of objective probabilities rather than completely subjective probabilities and can make the computational requirement easier [Kocabas and Dragicevic, 2013]. Netica provides three algorithms to parameterize the CPTs from data: Count learning, expectation-maximization (EM) and gradient descent [Frank, 2015; NorsysSoftwareCorp, 2016]. In this study we explored both Count learning algorithm and expectation maximization to populate the CPTs from the data. Count learning also called Multinomial Parameterization (Spiegelhalter and Lauritzen method) is the simplest and the most widely used [Korb and Nicholson, 2010; NorsysSoftwareCorp, 2016]. When learning using count learning , the net starts from state of ignorance meaning each node's states starts as uniform distribution, and for each instances of data, we identify which state the node takes and update the distribution to the corresponding parent instantiation [Korb and Nicholson, 2010]. On the other hand, EM searches for maximum likelihood estimates even when there are missing values in the data. The algorithm works through two steps, In the first step (expectation step or E-step), we compute expected counts from the net and in the second step(M-step), we treat the expected counts like complete data, and compute the maximum likelihood estimates from them. We typically repeat this iterative process until it no longer improves the likelihood[Choi et al., 2011]. Although Count learning is the simplest and true Bayesian learning [NorsysSoftwareCorp, 2016], our final network is based on EM learning as it is more robust and provide better calibration to our data.

The final network is presented in figure 3.1. The network provides the prior probabilities for all the variables in the network. Looking at our target node, while around 62% of our sample households did not intensify their production, 38% of the households have adopted one or more of the intensification strategy. 12% of the households have used improved seed variates, 8% are planting in both short and long rainy season, 7% use chemical fertilizers , 6.73% use irrigation and chemical fertilizer combined , while only 3.7% use irrigation. In terms of crop choices Rice is the dominant crop chosen by the households.Around 42% of the farmers plant rice as a mono-crop and 27% produce rice with combination of maize and 7.7% produces rice with combinations of maize and vegetables. Majority of the households are small scale rice based farmers (60%), 30% are small scale farmers that diversifier in terms of their land use and 10% represent large scale ago-pastoralists who practiced crop production and livestock keeping. 70% of the farmers own less than 3 hectares of crop land and 18% between 3 and 6 hectares. And around 12 % owns more than 6 hectares. We can also see from the BBN that farmers generally participate in output market, 41% of the farmers marketed between 30 and 60% of their output to the market and 34% sold more than 60 percent of their output.

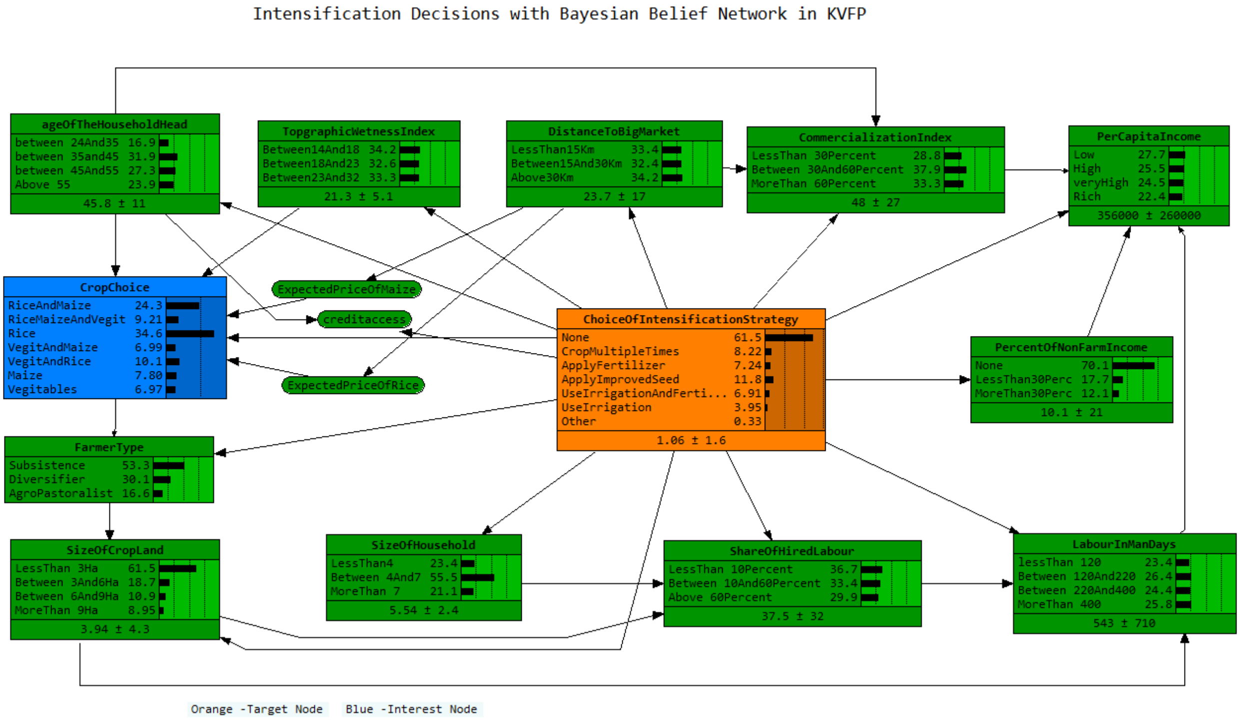


Figure 3.1: The resulting BBN Using Estimation Maximization Algorithm

3.3 Validation of the BBN

An important aspect of constructing a Bayesian network is validation. Validation is key to ensuring the high quality of the model. There are a number of ways to check the validity of the constructed BBN both quantitative and qualitative. In case of qualitative one can check the validity using expert opinion [Celio and Grêt-regamey, 2016; Frank, 2015]. The quantitative validation used a test data set to check the validity of predictions on our target variable. We validate our BBN using to quantitative validation techniques. we first performed five fold cross validation by partitioning our data in to five disjoint sub sets and do the iterative validation for five folds. In order to take in to account unbalanced nature of our target node intensification strategy and ensure that all the states are equally represented in the split, we used stratified cross validation tool from Caret [Kuhn, 2008]. Based on the avarage performance across the folds, the error rate of our model was 45%.

TABLE 3.1: Sensitivity of 'ChoiceOfIntensificationStrategy' to a finding at another node

Variable	Variance Reduction(%)	Entropy Reduction(%)
CropChoice	7.97	5.87
DistanceToBigMarket	6.26	3.77
PerCapitaIncome	2.01	2.05
FarmerType	1.97	2.01
ShareOfHiredLabour	1.42	1.72
TopgraphicWetnessIndex	1.36	1.64
SizeOfHousehold	1.18	1.62
CommercializationIndex	1.18	2.02
SizeOfCropLand	1.08	1.3
PercentOfNonFarmIncome	0.519	2.32
ExpectedPriceOfRice	0.284	0.225
LabourInManDays	0.262	1.42
ageOfTheHouseholdHead	0.209	2.19
ExpectedPriceOfMaize	0.0259	0.0158
creditaccess	0.00893	0.96

3.4 Sensitivity analysis

Since the final output of BBN is dependent on a priori assigned probabilities, we used sensitivity analysis to measure changes in probabilities of target node when there are changes in critical input parameters [Pollino et al., 2007].The sensitivity analysis based on Bayesian Network also serves as an aid to identify the significant and informative variables that affect our target variable intensification strategy [Sun and Müller, 2013]. Since the input parameters required for the sensitivity analysis contains discrete values, Entropy Reduction (Mutual Information) method is used here to determine the sensitivity of the BBN model's output to variation in a particular input parameter.The entropy reduction method works by computing the expected reduction in entropy of the target node due to finding at another node F . It is calculated as [Marcot et al., 2006; NorsysSoftwareCorp, 2016; Pearl, 1988]:

$$I = H(Q) - H(Q|F) = \sum_{q} \sum_{f} \frac{P(q,f) \log_2 [P(q,f]}{P(q)P(f)}$$

where $H(Q)$ and $H(Q|F)$ are the entropy of Q before and after any new findings respectively.

Table 3.1 shows the output of the sensitivity of intensification strategy to influencing variables. The result indicate that crop choices of the farmer is having the greatest influence in choice of intensification strategy with 5.87 reduction in entropy, followed by the distance from the nearest market with 3.8% variance reduction.In addition, Per capita income,share of income from non farm activity, age, Farmer type ,share of hired labor and topographic wetness index also influence the variation in choice of intensification strategy. All other variables have less than 1% reduction in entropy.

Although the sensitivity analysis based on a measure of entropy provides us an interesting insight regarding the main influencing factors for the choice of an intensification strategy, it does not tell us much how the probability of each strategy is influenced by the factors included in our model. In this study we made a further effort by conducting a global sensitivity analysis using a combination of Design of Experiment (DEO) with meta-modeling approach. In order to generate sample configurations of the evidence nodes we used Nearly Orthogonal Latin Hypercubes(NOLH) [Sanchez, 2005] that covers the parameter space of our evidence nodes. After creating the sample points, we provide the values to the network as evidence and we recorded the probabilities of each strategy for each sample point. To determine the effect of the different nodes on the variation of the probabilities of each strategy , we followed a meta modeling approach, and applied a regression tree model for each of the strategies [Coutts and Yokomizo, 2014]. The regression tree modeling approach has the advantage of automatically incorporating higher level interactions and of dealing with nonlinearity, they make very few assumptions about the structure of the data, and they are robust to outliers and implicitly handle variable selection [Coutts and Yokomizo, 2014; Kuhn and Johnson, 2013]. In addition, it provides an analogy for an easy rule induction from the results. The regression trees were implemented using Scikit-learn a machine learning tool in Python [Pedregosa et al., 2011]. The resulting regression tree [Fig A.1 A.2 A.3 A.4 A.5] and feature importance Fig A.6 for each strategy is provided in the annex.

The feature importance from the regression tree [Fig A.6] revels that variation in probabilities of choosing cropping multiple times is captured by variation in total labor available during the year, commercialization index, topographic wetness index , income and distance to central market. The variations in the probabilities of fertilizer application are also affected by topographic wetness index, if the farmer is diversifier, age, commercialization and distance to the market. On the other application of improved seeds is influenced by share of non farm income, age, household size , distance to the market and farm size. The probability of use of irrigation and fertilizer application is affected by distance to the market , farm size , share of non farm income and topographic wetness index. The variation in probabilities of use of irrigation is affected by variation in topographic wetness index , non farm income , farm size, if the farmer is of type substance and availability of labor.

4. Discussion and Conclusion

The present study was designed to explore the different factors that affects the choices of intensification strategies by considering a diversity of pathways that are tailored to the agro-ecological potential and production systems of floodplains. By building a data driven BBN we tried to explore factors that affect the choices of intensification strategy in KVFP. The probabilistic representation of BBN allow us to establish dependencies between hypothesized factors and intensification strategy while taking in to account uncertainties.

Based on our sample size, around 38% of farmers have intensified their production according to our identified sets of strategies. Given the urge for increase in agriculture production from the existing farm land we observe there is still room for intensifying more. The results from the sensitivity analysis also revealed range of factors influencing for farmers to choose one strategy over the others. In other words, we are able to identify under which social, economic and environmental conditions a particular intensification strategy is adopted.

In our study area, the choice of planted crops is the most import influencing factor for the choice of intensification strategy. Given that rice is the main crop produced in the area, the variations are more or less dependent on the mixed cropping of rice with maize and vegetable. Income and availability of off-farm income also further modify the choice of intensification strategy. Availability of income is a surrogate for farmers endowment and their ability to invest additional resources required for adopting the strategies.

Distance from the farm to nearest market and commercialization also has a strong influence on the choice of intensification strategy. Access to market has an effect both in terms of access to key inputs and also access to the output market and significantly affect intensification[Erenstein, 2006]. Although the availability of both family and hired labor is a crucial determinant of intensification choice, the sensitivity show a moderate connection.

The global sensitivity analysis provides the factors leading to the choice of a particular intensification strategy. In general, from the meta-modeling analysis we observe that each strategy is influenced differently by factors under consideration. Although the variations in the probabilities of strategies are influenced by a common set of variables, the magnitude and order of the effects are different across the strategies.

Closer inspection of the regression trees shows, farmers whose field is located in relatively wetter areas, who sale less than 66% of their output and located relatively far from the market and with higher market participation will have higher (33%) probability to choose cropping multiple times a year. In case of applying fertilizer, farmers who own relatively bigger land size and located in wetter areas will have on average 27% probability to apply fertilizer. The highest probability (28%) for Adoption of improved seed variety as strategy is found for households with less than 4 household members, located less than 22 km from the central market and less than 10 hectares of land. Looking at the characteristics of farmers who chooses to combine small-scale irrigation with fertilizer application, farmers who receives more than 38% of non-farm income with farm size greater than 7 hectares and located less 22 km from the market have an average probability of 22%. As expected Irrigation use is affected by labor availability, farm type and topographic wetness index of the farm. The highest probability for use of irrigation (5%) is for farmers with labor more than 109 man-days, a type of small-scale subsistence and topographic wetness index of less than 28.

There are some limitations to our study and one should interpret our findings in light of these limitations. Our limitations mainly arises from limited data set both for training and validation of our BBN. In addition , our data also lacks combinations of intensification strategies that might be choices for the farmer. The BBN presented in this study is static in nature and does not take in to account the dynamics over time. Discretization of the continuous variables also will result in loss of information from the data and the choice of the method of discretization also might affect the structure and parameters of the BBN. Although our objective is on land augmenting intensification strategies adopted in our study area , farmers might have other strategies that they might uptake which are not included in our research.

A. Annex

Figure A.1: Fertilizer Application

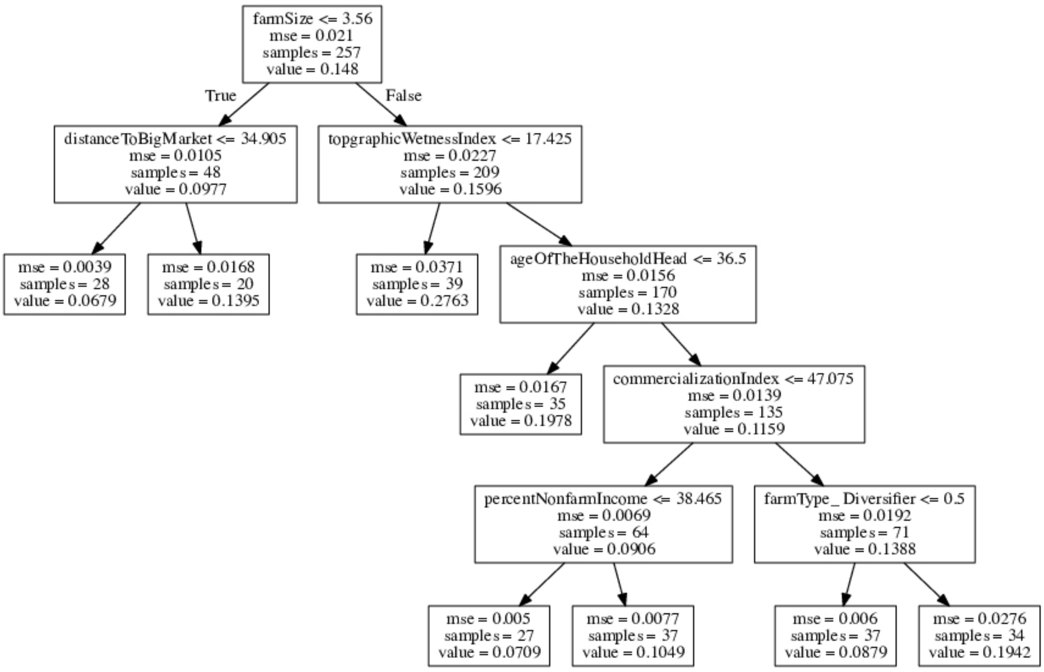


Figure A.2: Crop Multiple times

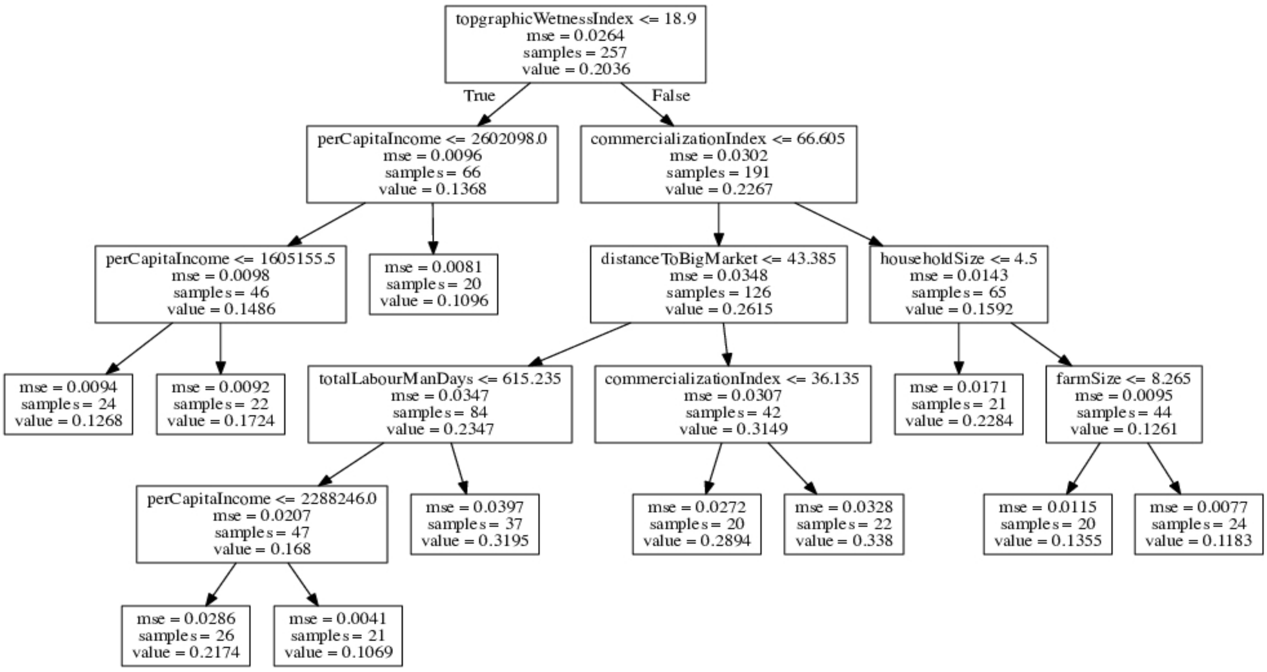


Figure A.3: Improved Seed

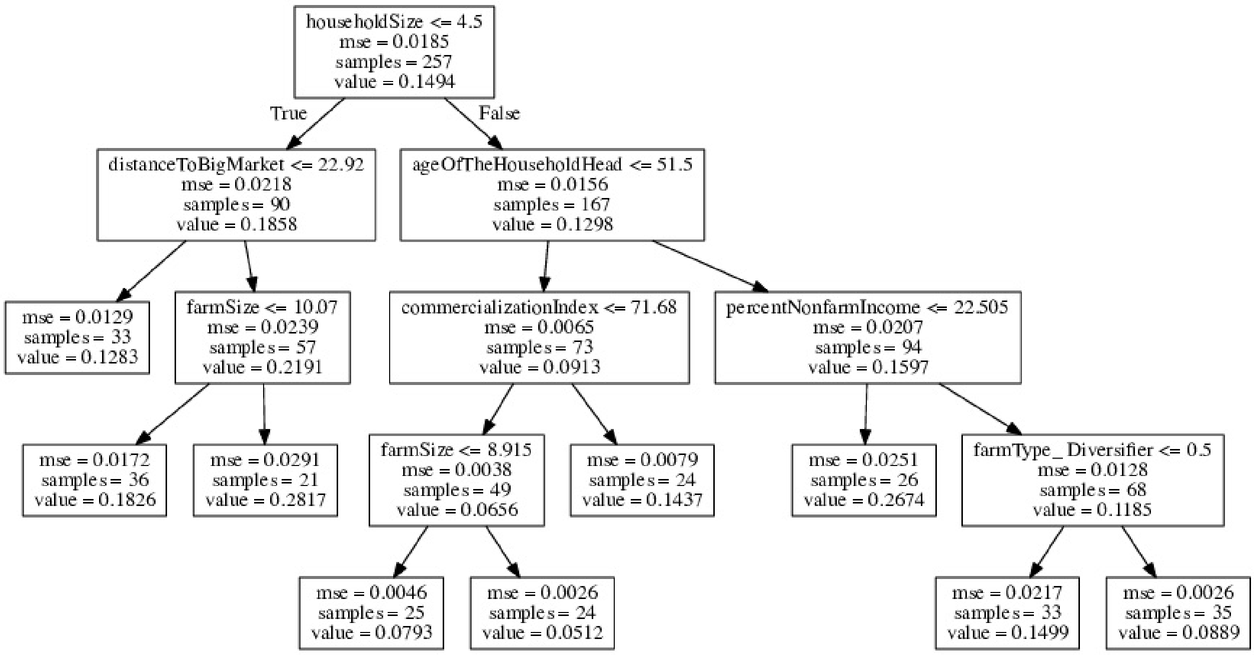


Figure A.4: Irrigation and Fertilizer

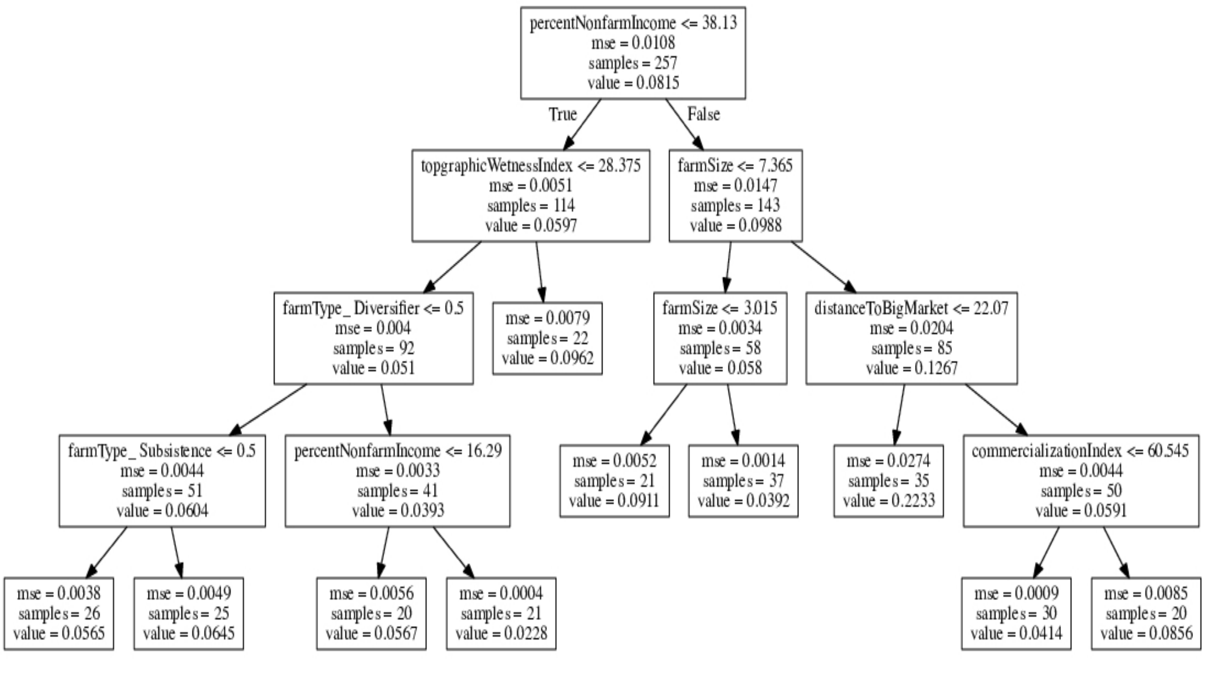


FIGURE A.5: Irrigation Use

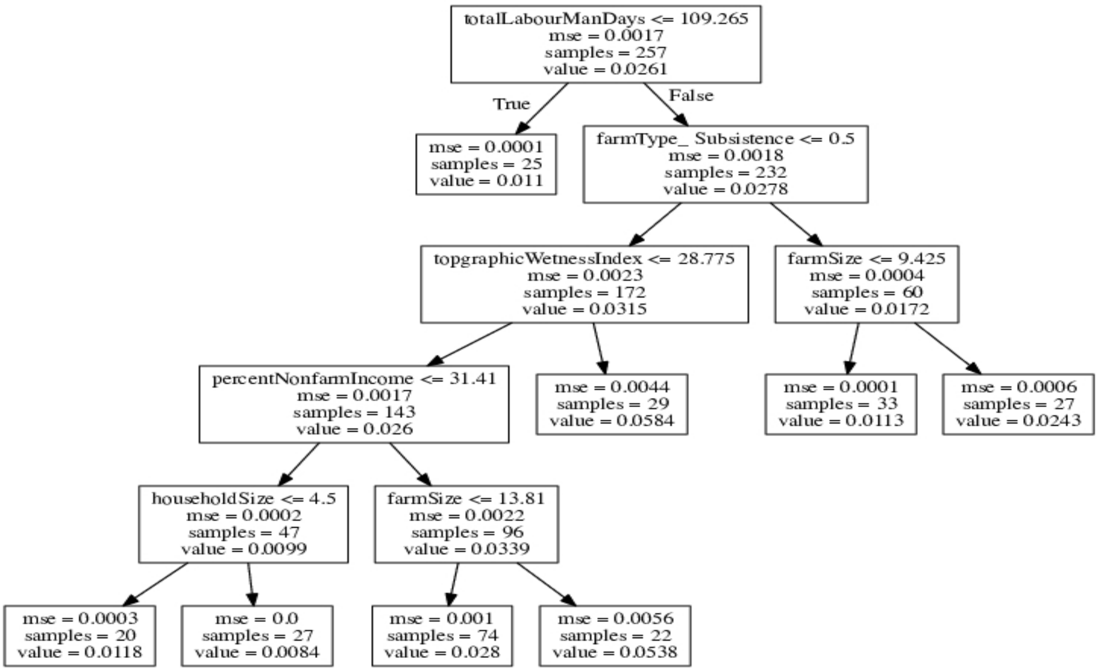
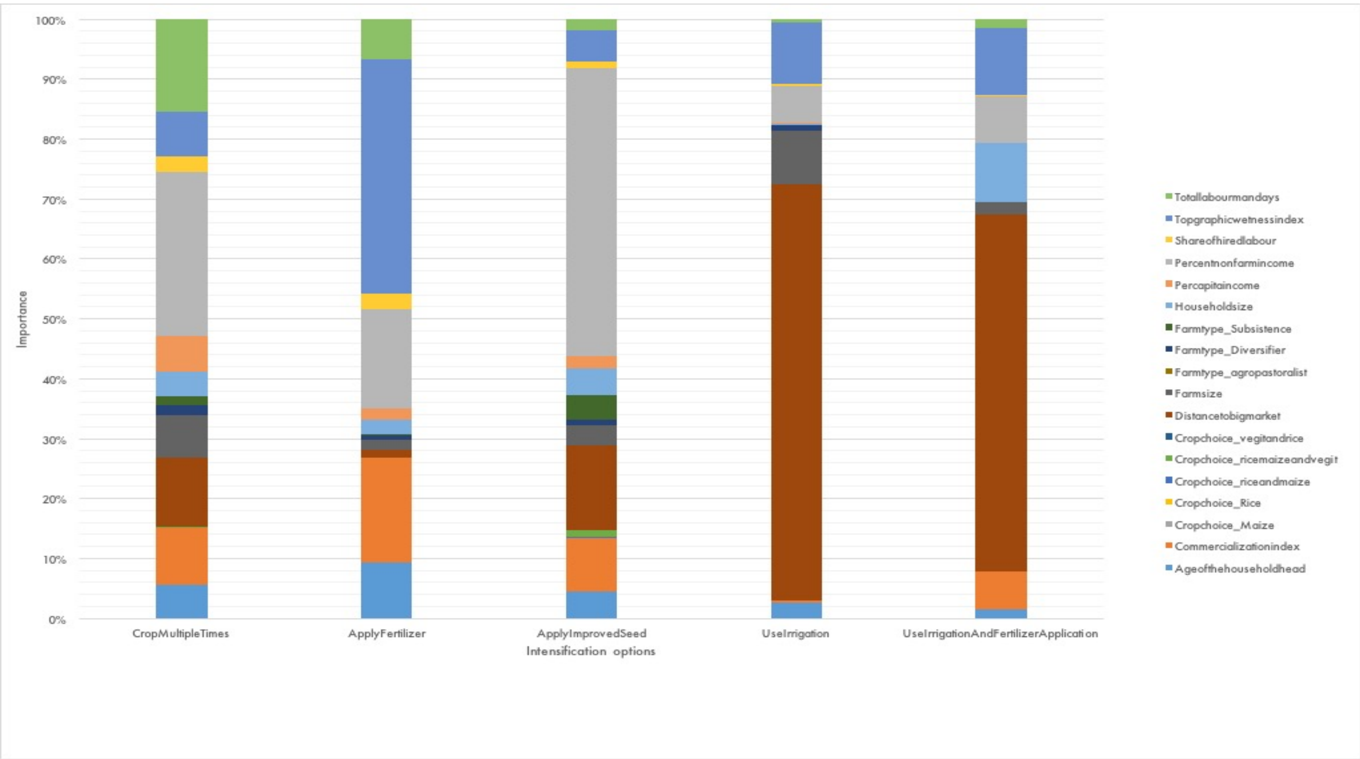


FIGURE A.6: Variable Importance



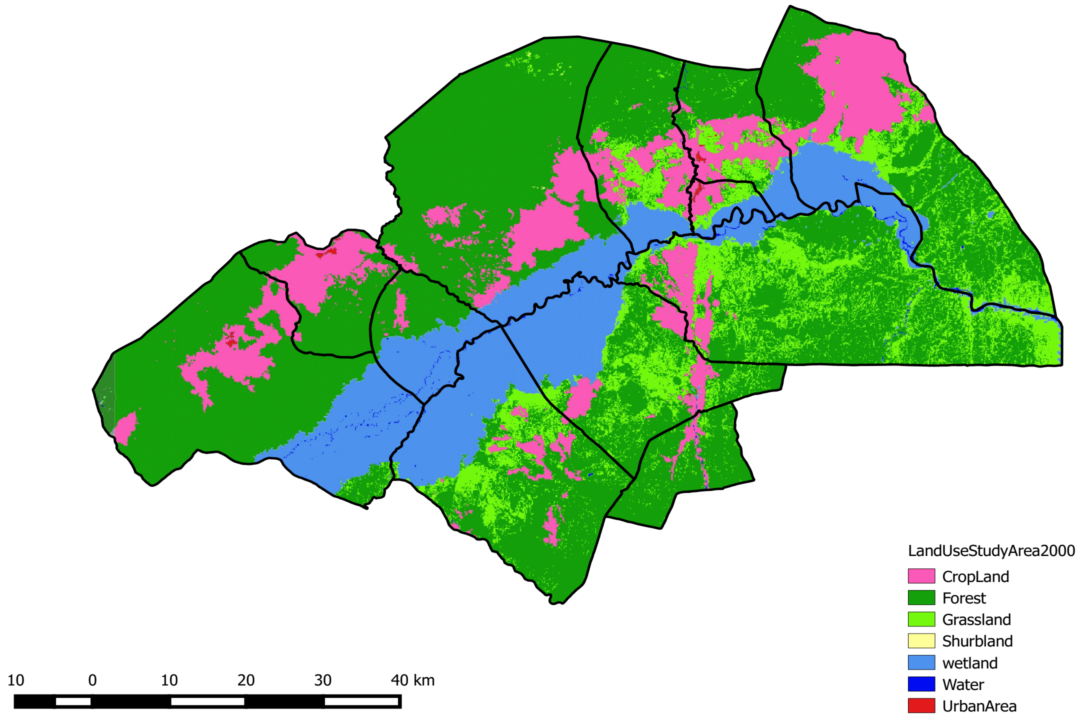


FIGURE A.7: LandUse for 2000

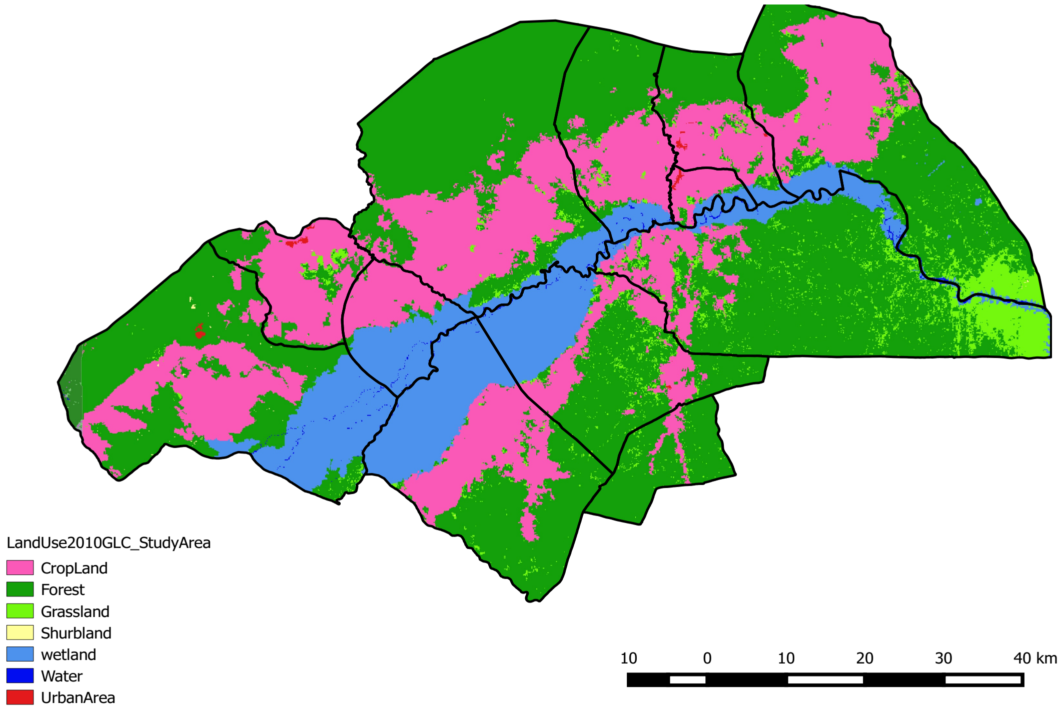


FIGURE A.8: LandUse for 2010

[Jun et al., 2014]

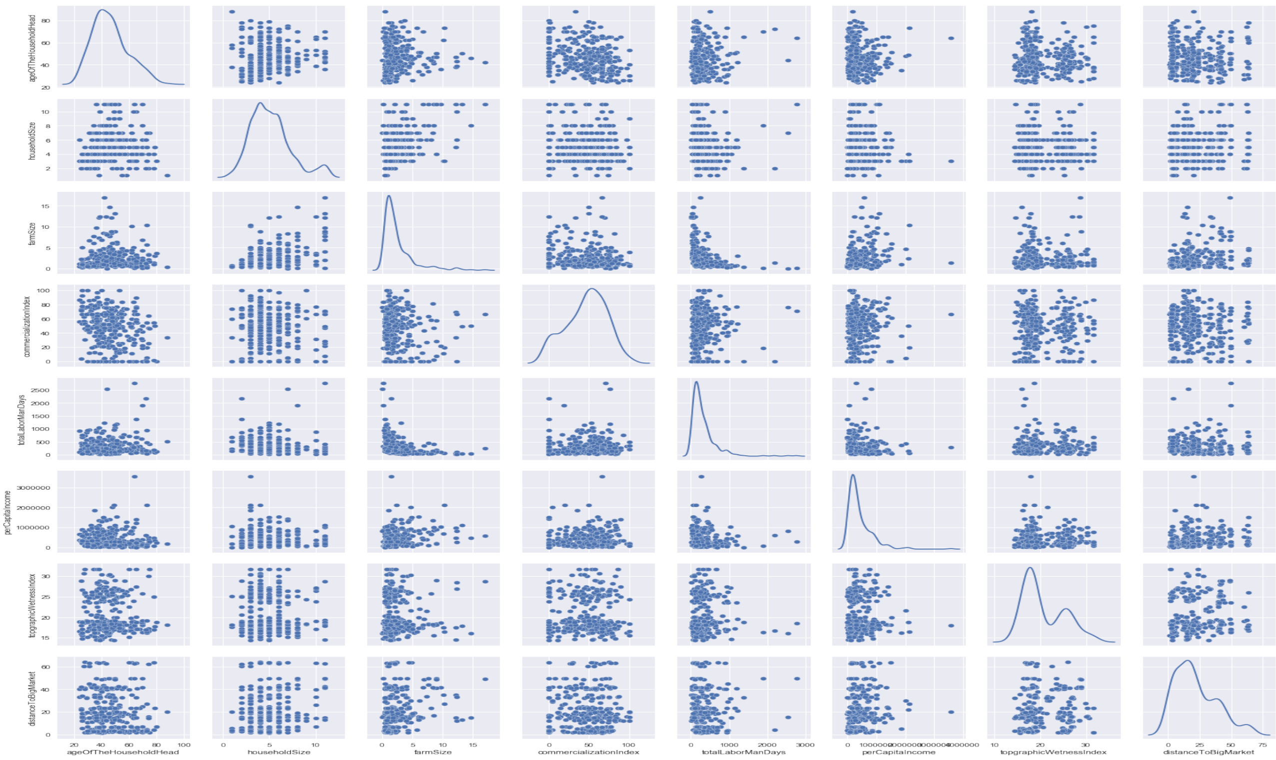


FIGURE A.9: paired combinations of continuous variables.

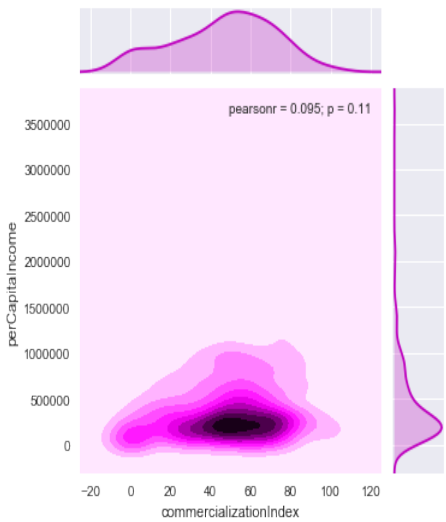


FIGURE A.10: Correlation between commercialization and Income

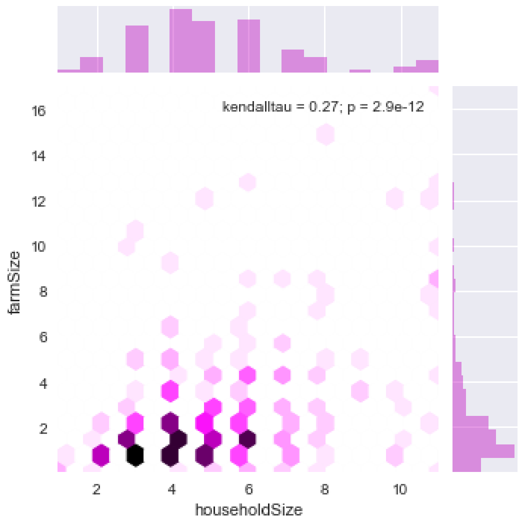


FIGURE A.11: Correlation between family Size and Farm size

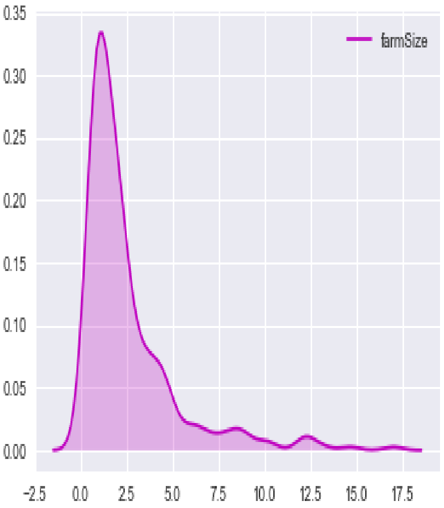


FIGURE A.12: Distribution of farm size

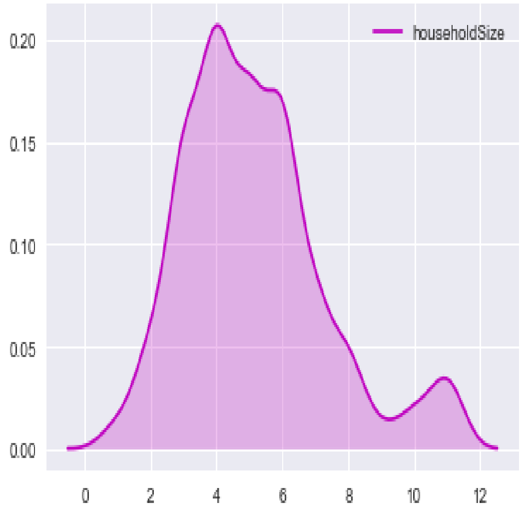


FIGURE A.14: Distribution of Household Size

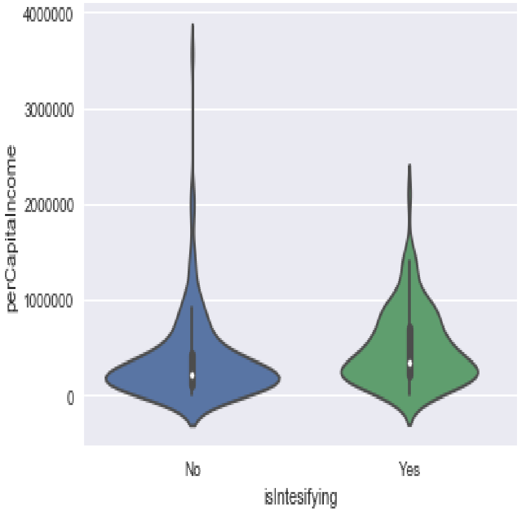


FIGURE A.13: violin plot between income and intensification

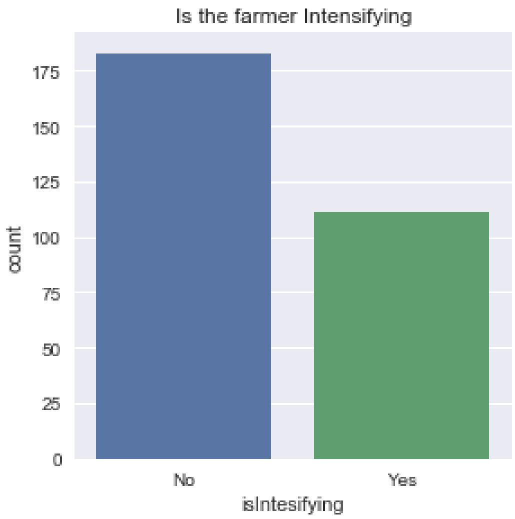
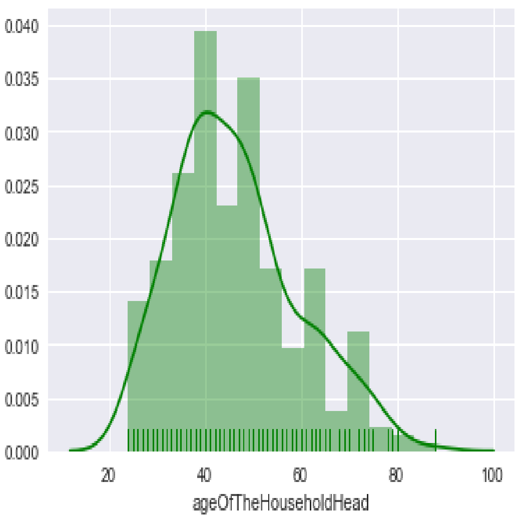


FIGURE A.15: Proportion of Intensifying



Bibliography

Aalders, I. (2008). Modeling Land-Use Decision Behavior with Bayesian Belief Networks. *Ecology and Society 13*(1).

Abay, K. A., G. Berhane, A. S. Taffesse, B. Koru, K. Abay, et al. (2016). Understanding farmers' technology adoption decisions: Input complementarity and heterogeneity. *Ethiopia Strategy Support Program Working Paper 82*.

Agra (2016). Africa agriculture status report:progress towards agricultural transformation in africa. Technical report, Green Revolution in Africa (AGRA).

Bamford, A., D. Ferrol-Schulte, and H. Smith (2010). The status of the ruipa corridor between the selous game reserve and the udzungwa mountains. *Report to Frontier Tanzania, Dar es Salaam, Tanzania*.

Besley, T. and A. Case (1993). Modeling technology adoption in developing countries. *The American Economic Review 83*(2), 396–402.

Binswanger, H. and P. Pingali (1988). Technological priorities for farming in sub-saharan africa. *The World Bank Research Observer 3*(1), 81–98.

Boserup, E. (1965). The condition of agricultural growth.

Boserup, E. (1981). Population and technological change: A study of long-term trends.

Cain, J. (2001a). Improvements in Natural Resources. *44*(0).

Cain, J. (2001b). *Planning improvements in natural resource management. guidelines for using bayesian networks to support the planning and management of development programmes in the water sector and beyond*. Centre for Ecology and Hydrology.

Celio, E. and A. Grêt-regamey (2016). Understanding farmers ' influence on land-use change using a participatory Bayesian network approach in a pre-Alpine region in Switzerland. *Journal of Environmental Planning and Management 59*(11), 2079–2101.

Cerquides, J. and R. L. De Màntaras (2003). Tractable bayesian learning of tree augmented naive bayes models. In *ICML*, pp. 75–82.

Choi, A., L. Zheng, A. Darwiche, and O. Mengshoel (2011). A tutorial on bayesian networks for system health management. *Machine Learning and Knowledge Discovery for Engineering Systems Health Management 10*(1), 1–29.

Clarke, E. J. and B. A. Barton (2000). Entropy and mdl discretization of continuous variables for bayesian belief networks. *International Journal of Intelligent Systems 15*(1), 61–92.

Coulson, A. (2015). Small-scale and large-scale agriculture: Tanzanian experiences. In M. Stahl (Ed.), *In Looking Back, Looking Ahead – Land, Agriculture and Society in East Africa*, pp. 44–73. Nordic Africa Institute.

Coutts, S. R. and H. Yokomizo (2014). Meta-models as a straightforward approach to the sensitivity analysis of complex models. *Population Ecology 56*(1), 7–19.

Daly, R., Q. Shen, and S. Aitken (2011). Learning bayesian networks: approaches and issues. *The knowledge engineering review 26*(2), 99–157.

Darnhofer, I., R. Gretzmacher, and W. Schneeberger (1997). Cart for oats-vetch adoption in the ethiopian highlands. *Die Bodenkultur 271*(48), 4.

Dinesen, L. (2016). *Kilombero Valley Floodplain (Tanzania)*, pp. 1–8. Dordrecht: Springer Netherlands.

Erenstein, O. (2006). Intensification or extensification? factors affecting technology use in peri-urban lowlands along an agro-ecological gradient in west africa. *Agricultural systems 90*(1), 132–158.

Feder, G., R. E. Just, and D. Zilberman (1985). Adoption of agricultural innovations in developing countries: A survey. *Economic development and cultural change 33*(2), 255–298.

Frank, S. K. (2015). Expert-based bayesian network modeling for environmental management. Frankfurt Hydrology Paper 11, Institute of Physical Geography, Goethe University Frankfurt.

Frayer, J., Z. Sun, D. Müller, D. K. Munroe, and J. Xu (2014). Analyzing the drivers of tree planting in yunnan, china, with bayesian networks. *Land Use Policy 36*, 248–258.

Friedman, N., D. Geiger, and M. Goldszmidt (1997). Bayesian network classifiers. *Machine learning 29*(2-3), 131–163.

Friedman, N. and D. Koller (2003). Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine learning 50*(1), 95–125.

Gladwin, C. H. (1980). A theory of real-life choice: Applications to agricultural decisions. *Agricultural decision making: Anthropological contributions to rural development*, 45–85.

Headey, D., M. Dereje, and A. S. Taffesse (2014). Land constraints and agricultural intensification in Ethiopia: A village-level analysis of high-potential areas. *Food Policy 48*, 129–141.

Howley, P., C. O. Donoghue, and K. Heanue (2012). Factors Affecting Farmers' Adoption of Agricultural Innovations: A Panel Data Analysis of the Use of Artificial Insemination among Dairy Farmers in Ireland. *Journal of Agricultural Science 4*(6), 171–179.

Jayne, T., J. Chamberlin, and D. D. Headey (2014). Land pressures, the evolution of farming systems, and development strategies in africa: A synthesis. *Food Policy 48*, 1 – 17.

Jones, T., A. J. Bamford, D. Ferrol-Schulte, P. Hieronimo, N. McWilliam, and F. Rovero (2012). Vanishing wildlife corridors and options for restoration: A case study from tanzania. *Tropical Conservation Science 5*(4), 463–474.

Jun, C., Y. Ban, and S. Li (2014). China: Open access to earth land-cover map. *Nature 514*(7523), 434–434.

Kangalawe, R. Y. and E. T. Liwenga (2005). Livelihoods in the wetlands of kilombero valley in tanzania: Opportunities and challenges to integrated water resource management. *Physics and Chemistry of the Earth, Parts A/B/C 30*(11), 968–975.

Kato, F. (2007). Development of a major rice cultivation area in the kilombero valley, tanzania. *African Study Monographs* (36), 3–18.

Kjaerulff, U. B. and A. L. Madsen (2012). *Bayesian networks and influence diagrams: a guide to construction and analysis*, Volume 22. Springer Science & Business Media.

Kocabas, V. and S. Dragicevic (2013). Bayesian networks and agent-based modeling approach for urban land-use and population density change: A BNAS model. *Journal of Geographical Systems 15*(4), 403–426.

Koller, D. and N. Friedman (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Korb, K. B. and A. E. Nicholson (2010). *Bayesian Artificial Intelligence, Second Edition* (2nd ed.). Boca Raton, FL, USA: CRC Press, Inc.

Koski, T. and J. Noble (2011). *Bayesian networks: an introduction*, Volume 924. John Wiley & Sons.

Kuhn, M. (2008). Caret package. *Journal of Statistical Software 28*(5), 1–26.

Kuhn, M. and K. Johnson (2013). *Applied predictive modeling*, Volume 810. Springer.

Lele, U. J. and S. W. Stone (1989). Population, pressure, the environment, and agricultural intensification: variations on the boserup hypothesis. *MADIA discussion paper (USA)*.

Leonard, S. H., G. D. Deane, and M. P. Gutmann (2011). Household and farm transitions in environmental context. *Population and environment 32*(4), 287–317.

MALF (2015). Tanzania livelihood baseline profile: Kilombero-ulanga-lusewa paddy, maize, and cassava livelihood zone. Zone Discription 43, Ministry of Agriculture Livestock and Fisheries: Agriculture Department.

Marcot, B. G., J. D. Steventon, G. D. Sutherland, and R. K. McCann (2006). Guidelines for developing and updating bayesian belief networks applied to ecological modeling and conservation. *Canadian Journal of Forest Research 36*(12), 3063–3074.

Margaritis, D. (2003). Learning bayesian network model structure from data. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE.

Milder, J., L. Buck, and A. Hart (2013). Applying an agriculture green growth approach in the sagcot clusters: Challenges and opportunities in kilombero, ihemi and mbarali. Technical report, SAGCOT Centre, Dar es Salaam.

Milder, J., L. Buck, A. Hart, S. Scherr, and S. Shames (2013). A framework for agriculture green growth: Greenprint for the southern agricultural growth corridor of tanzania (sagcot). *Dar es Salaam: SAGCOT Centre*.

Mombo, F., S. Speelman, G. V. Huylenbroeck, J. Hella, and S. Moe (2011). Ratification of the Ramsar convention and sustainable wetlands management : Situation analysis of the Kilombero Valley wetlands in Tanzania. *Journal of Agricultural Extension and Rural Development 3*(9), 153–164.

Neapolitan, R. E. et al. (2004). *Learning bayesian networks*, Volume 38. Pearson Prentice Hall Upper Saddle River, NJ.

Nielsen, T. D. and F. V. Jensen (2009). *Bayesian networks and decision graphs*. Springer Science & Business Media.

Nin-Pratt, A. (2015). Agricultural intensification in africa. *IFPRI Discussion Paper 01433*.

Nindi, S. J., H. Maliti, S. Bakari, H. Kija, and M. Machoke (2014a). Conflicts over land and water resources in the kilombero valley floodplain, tanzania. *African Study Monographs 50*, 173–190.

Nindi, S. J., H. Maliti, S. Bakari, H. Kija, and M. Machoke (2014b). Conflicts Over Land and Water Resources in the Kilombero Valley Floodplain, Tanzania. *African Study Monographs 50*(October), 173–190.

Nojavan A., F., S. S. Qian, and C. A. Stow (2017). Comparative analysis of discretization methods in Bayesian networks. *Environmental Modelling and Software 87*, 64–71.

NorsysSoftwareCorp (2016). *Netica-J Manual*. http://www.norsys.com/netica-j/docs/NeticaJ_Man.pdf,Retrieved on 10/09/2016.

Ntongani, W. A., P. K. Munishi, S. R. More, and J. J. Kashaigili (2014). Local knowledge on the influence of land use/cover changes and conservation threats on avian community in the kilombero wetlands, tanzania. *Open Journal of Ecology 4*(12), 723.

Okike, I. (2001). *Agricultural intensification and efficiency in the West African savannahs: Evidence from northern Nigeria*. Number 33. ILRI (aka ILCA and ILRAD).

Okike, I., M. A. Jabbar, V. Manyong, J. W. Smith, J. A. Akinwumi, and S. K. Ehui (2001). Agricultural intensification and efficiency in the West African savannahs: evidence from northern Nigeria. *Socio-economics and Policy Research Working Paper - International Livestock Research Institute* (33).

Otsuka, K., F. Place, et al. (2013). Evolutionary changes in land tenure and agricultural intensification in sub-saharan africa. Disscussion Paper 13-22.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Pearl, J. (2011). Causality: Models, reasoning and inference. *International Statistical Review 79*(2), 272–301.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12*(Oct), 2825–2830.

Perz, S. G. (2003). Social Determinants and Land Use Correlates of Agricultural Technology Adoption in a Forest Frontier : A Case Study in the Brazilian Amazon Author (s): Stephen G . Perz Published by : Springer Stable URL : http://www.jstor.org/stable/4603460 Accessed :. *Human Ecology 31*(1), 133–165.

Pollino, C. A., O. Woodberry, A. Nicholson, K. Korb, and B. T. Hart (2007). Parameterisation and evaluation of a bayesian network for use in an ecological risk assessment. *Environmental Modelling & Software 22*(8), 1140–1152.

Pope, A. and R. Gimblett (2017). Linking participatory, bayesian, and agent-based modeling techniques to simulate coupled natural-human system: A case study with ranchers in sonora, mexico. In *Environmental Modeling with Stakeholders*, pp. 345–359. Springer.

Rasmussen, S., A. L. Madsen, M. Lund, et al. (2013). Bayesian network as a modelling tool for risk management in agriculture.

Rebelo, L. M., M. P. McCartney, and C. M. Finlayson (2010). Wetlands of Sub-Saharan Africa: Distribution and contribution of agriculture to livelihoods. *Wetlands Ecology and Management 18*(5), 557–572.

Ruthenberg, H. et al. (1980). *Farming systems in the tropics.* Number Ed. 3. Oxford University Press.

Sanchez, S. M. (2005). Nolhdesigns spreadsheet. *Software available online via http://diana. cs. nps. navy. mil/SeedLab*, 1430–1439.

Schelhas, J. (1996). Land use choice and change: intensification and diversification in the lowland tropics of costa rica. *Human organization 55*(3), 298–306.

Shriar, A. (2000). Agricultural intensity and its measurement in frontier regions. *Agroforestry systems 49*(3), 301–318.

Shriar, A. J. (2001). The dynamics of agricultural intensification and resource conservation in the buffer zone of the maya biosphere reserve, petén, guatemala. *Human ecology 29*(1), 27–48.

Snyder, K. A. and B. Cullen (2014). Implications of sustainable agricultural intensification for family farming in africa: Anthropological perspectives.

Sörensen, R., U. Zinko, and J. Seibert (2006). On the calculation of the topographic wetness index: evaluation of different methods based on field observations. *Hydrology and Earth System Sciences Discussions 10*(1), 101–112.

Sun, Z. and D. Müller (2013). A framework for modeling payments for ecosystem services with agent-based models, Bayesian belief networks and opinion dynamics models. *Environmental Modelling and Software 45*, 15–28.

Uusitalo, L. (2013). Advantages and challenges of Bayesian networks in environmental modeling networks in environmental modelling. (May 2007).

Zheng, F. and G. I. Webb (2010). *Tree Augmented Naive Bayes*, pp. 990–991. Boston, MA: Springer US.