Department of Economics

Working Paper Series

ISSN 1753 – 5816

Please cite this paper as:

Lietava, J & Morimoto, R (2019), “Regression tree analysis of soil fertility and agro-economic practices and the effects on yield in Tanzania”, SOAS Department of Economics Working Paper No. 218, London: SOAS University of London.

No. 218

Regression tree analysis of soil fertility and agro-economic practices and the effects on yield in Tanzania

by

Jan Lietava & Risa Morimoto

(February 2019)

SOAS University of London

Department of Economics
SOAS University of London
Thornhaugh Street, Russell Square, London WC1H 0XG, UK
Phone: + 44 (0)20 7898 4730
Fax: 020 7898 4759
E-mail: economics@soas.ac.uk
http://www.soas.ac.uk/economics/

Regression tree analysis of soil property and agro-economic practices and the effects on yield in Tanzania

Lietava, J & Morimoto, R*

Abstract

Food security and yield production have been extensively studied in regard to fertiliser response in the context of Sub-Saharan Africa. Soil fertility gradients within farms interact with other complex factors such as plot management, alongside bio-physical as well as socio-economic constraints, to create extensive heterogeneity in yield production. Hence, blanket recommendations regarding increasing inputs may not just lead to limited or no change in productivity but may affect long term sustainability. To try and explore the relationships between these factors, Classification and Regression Tree (CART) analysis was used to simplify the effect of plot management decisions. Data was from the focal plots of the 2016 and 2017 Tanzanian Agronomy Panel Survey (TAPS 2017; $n = 580$) with a range of yields and socio-economic contexts (2.10*tha*$^{-1}$ – 3.68*tha*$^{-1}$). Results suggested that in low resource fields, management factors are subservient to extreme soil degradation while in high resource fields good management such as optimal planting are needed for maximum predicted yield (4.48*tha*$^{-1}$*;* n = 51). Boundary line analysis was conducted, and maximum yield-nutrient response values calculated. The yield gap obtained suggested only up to *60%* of locally obtainable yield is reached, highlighting the necessary balance between intensity of resource use and good management for sustainable and sufficient crop production, especially in the case of extreme soil degradation.

Keywords: Regression tree analysis; soil property; Tanzania; agro-economic practice.

JEL classification: Q10, Q15, Q56, R15, R28.

[*] SOAS University of London. Address for correspondence: Department of Economics, SOAS University of London. Russell Square, London WC1H 0XG, UK. Tel: +44 207 898 4730. Email:rm36@soas.ac.uk

Introduction

Food security is a recurring problem within most of Sub-Saharan Africa, with a large majority of the continent depending on subsistence agriculture (Isinika et al., 2011). Crop yield is highly variable not only across countries, but within farms as small as $0.45ha^{-1}$, affected by both management factors as well as inherent soil properties (Giller et al., 2006: 11). However, these two factors are not independent, and interact with each other within a socio-economic context, with limitations such as low resource input availability and the opportunity cost of applied labour. Tanzania has a varied political economy history, with *70%* of its population relying on subsistence agriculture, mainly maize, using rainfall alongside little to no water management, and hence facing increasing vulnerability from climate change and precipitation variability (Lema and Majule, 2009: 206).

Extensive literature has covered the diverse response in terms of plot management to increasing uncertainty regarding crop yield in the context of the operational environment of small-holder farmers (Tittonell and Giller, 2013). With fertiliser usage in Tanzania below the Sub-Saharan average of $15.0kgha^{-1}$, at $8.0kgha^{-1}$, soil rehabilitation is costly and uncertain, and hence conservation management techniques such as no-tillage and manure application are essential (The World Bank, 2017; Isham, 2002).

Many methods have been used to estimate yield production, generally utilising Cobb-Douglas or transcendental logarithmic (translog) functional form. However, in the context of extensive links and feedback patterns within soil management, fertility and socio-economic factors, CART (Classification and Regression Tree) analysis offers a method without underlying parametric assumptions (Wiig et al., 2001; Ekbom et al., 2008). Hence, this paper aims to use CART analysis to explore these links, focusing on plot management techniques, within the institutional environment of Tanzania.

The first chapter explores the current literature, contextualising current limited input use in terms of slow technology uptake. Furthermore, the limit of blanket recommendations is high-lighted with varied plot management adaptation due to climate change, as well as extensive soil resource limitations. Chapter two explores the current institutional limitations, and the origin of high transaction costs which are currently present for smallholder farmers. This is followed by an explanation of the three main type of variables selected, focusing on management and conservation, soil properties and infrastructure and socio-economic variables. This is followed by a brief overview of the TAPS 2017 data, highlighting remnant wealth stratification, alongside diverse management responses as well as varied inherent soil properties. Chapter three shows the CART analysis results, demonstrating the importance of plot management techniques especially in low resource fields. Finally, chapter 4 discusses yield potential alongside socio-economic factors, concluding that the focus should be on attaining locally obtainable yield through soil property and plot management techniques.

1 Agricultural management and interactions

Agro-economic management and resource allocation: fertilisers, land use and technology change

Maize is relatively demanding of nutrients, with Nitrogen (*N*), Phosphorous (*P*) and Potassium (*K*) essential for increasing crop yield, more integral in areas where rainfall is limited or highly variable (Katinila et al., 1998; Tittonell et al., 2008). Productive assets are limited in Tanzanian agriculture, with *56%* of the planted area cultivated using the hand hoe, which is more emphasized for those farmers working on maize, with the percentage rising to *75%* (Katinila et al., 1998; Isinika et al.2011: 290-293). Furthermore, although the usage of fertiliser as an input increased by a factor of 5 in the period *2005 - 2010*, artificial fertiliser usage is very low, with only *21%* of maize farmers accessing and utilising this resource in the main season (Isinika et al., 2011: 296).

Several studies have looked at the factors affecting fertiliser adoption. *N* adoption is associated more intensely with decreasing land size, as well as with increasing levels of education (Nkonya et al., 1997; Cornia, 1985; Seyoum et al., 1998). However, the lack of adaptation of current technology has also been explored as a factor in the lack of the engagement of Tanzanian youth in agriculture combining with the ageing structure of institutional actors, such as extension and research personnel, potentially leading to worse conditions in terms of future technology adoption (Isinika et al., 2011). Furthermore, in the short-term, the uptake of chemical and inorganic fertilisers may be delayed due to the necessary amount of time required for farmers to collate and gather knowledge through local networks and extension agents (Seyoum et al., 1998; Isham, 2002; Katinila et al., 1998; Kaliba et al., 2000). Isham (2002) explore the process of technology adoption in Tanzania utilising three characteristics of social structures which positively affect diffusion: group homogeneity, participatory norms in the form of the level of interactive decision making and leadership heterogeneity (Isham, 2002: 41-42). This highlights the importance of focusing on the regional variety and diversity of social structure.

Climate change and adaptation

Vulnerability has been defined in many different contexts, but is used here to refer to a cross-section of social, economic and institutional factors and the ability of individuals to use ‘entitlements’ to adapt to risk (Kelly and Adger, 2000: 326). This can be summarised in the exposure-sensitivity chart in figure 1, which highlights the importance of adaptive capacity in terms of the level of mitigation, optimisation and maintenance of these response strategies in terms of risk-determination (Ionescu et al., 2009; Mongi et al., 2010).

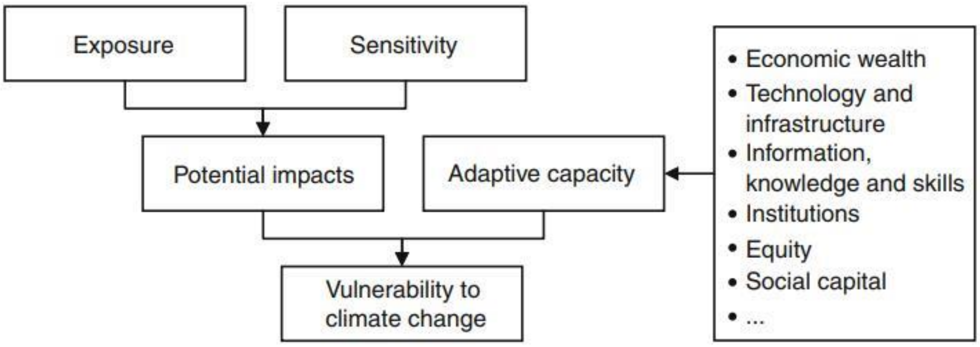


Figure 1: A conceptualisation of vulnerability in regard to climate change, with inputs of exposure, sensitivity and adaptation (Ionescu et al., 2009: 4).

Most of Tanzanian agriculture relies on rainfall, with only *2.7%* of the total planted land having irrigation (Isinika et al., 2011: 294). Furthermore, El-Niño Southern Oscillation (ENSO) influences rainfall, with more precipitation during warmer ENSO events (Rowhani et al., 2011: 452). The Northern areas face a bi-modal precipitation pattern, with rains between March and May, as well as October to December, with the rest of the country receiving rainfall in the main season from December to May (Rowhani et al., 2011: 450). The Northern areas face a bi-modal precipitation pattern, with rains between March and May, as well as October to December, with the rest of the country receiving rainfall in the main season from December to May (Rowhani et al., 2011: 450).

The effects of increasing variability due to climate change, in regard to agricultural productivity, can be subdivided into three main categories. First, unpredictable rainfall leading to changing onset and end of rains in tropical regions can have larger effects, especially due to the yield of maize being highly correlates to the length of the growing season (Mongi et al., 2010; Tabu et al., 2005; Croon et al., 1984; Lema and Majule, 2009). Secondly, increasing pests and diseases due to higher temperature and moisture lead to further risks to crops. Finally, increasing probability of drought can not only increase the chance of crop failure, especially in critical parts of the growing season, but can lead to declining soil fertility (Lema and Majule, 2009: 207). Furthermore, there is regional variation. Semi-arid sections of central and south-eastern Tanzania are more at risk due to low soil fertility and predicted decreasing soil moisture (Mongi et al., 2010; Morris, n.d.: 371-372).

Studies have shown the awareness of smallholder farmers in regard to perceived change in rainfall and temperature especially regarding the shifting of the main season (Mongi et al., 2010; Lema and Majule, 2009). In response to unpredictable season length and start, transplanting, integration of livestock and splitting plots have all been used as techniques, alongside bi-modal planting and increasing labour input per hectare (Mongi et al., 2010; Morton, 2007). Furthermore, inter-cropping and substitution can also be seen as attempts to reduce risk, which are separate from coping mechanisms, such as employment diversification into charcoal production or intracommunity support (Morton, 2007). Hence, responses to climate change have to be flexible and compensate for both excess-water seasons and higher drought probability, such as through optimising soil organic matter to stabilise soil structure (Lema and Majule, 2009: 216-217). This highlights the importance of management techniques, as well as the level of diversity present in the responses of Tanzanian smallholder farmers.

Effects of soil on productivity and fertility management

Biophysical constraints in terms of soil properties are one of the largest factors affecting crop output variability in Sub-Saharan Africa, with high levels of soil degradation due to intensive farming practices, climate change as well as poor agro-economic management (Lungu et al., 1993; Kihara et al., 2016; Achieng et al., 2010; Lal, 2006). Different soil fertilities stemming from inherent soil properties, topographic position and the combination of agro-economic management techniques suggest that blanket recommendations are not satisfactory (Tabu et al., 2005; Ismail et al., 1994). Furthermore, nutrient balances and changing chemical properties can have varying impacts depending in adaptation strategies (Braun et al., 1997).

Soil organic matter (SOM) is a key factor, affecting the retention of water alongside chemical effects such as *pH* (Reeves, 1997; Braun et al., 1997: 18-21). Tabu et al. (2005) focused on smallholder farmers in Western Kenya, with a difference of *19%* in SOM between productive (4.3tha^{-1}) and unproductive (2.8tha^{-1}) niches. SOM has also been shown to be positively correlated with a transition from plow-till to conservation-till, manuring as well as improved fertility management (Lal, 2006: 201). With the majority of farmers still utilising the hand-hoe for tillage, this is a major potential area of improvement within Tanzanian agriculture.

On the other hand, in terms of socio-economic limitations, limited extension service currently present within the Tanzanian agricultural sector could potentially pose a limit on the transition from SOM-degradation to sustainable practices which would need to be based on solidifying the fact that soil resources are not unlimited. In the case of Tanzanian smallholder farmers, soil acidity as well as fertiliser cost are limitations. This is partially due to limited infrastructure and hence utilising on-farm resources is increasingly important (Achieng et al., 2010).

Impact of agro-economic practices and soil management

In the following chapters, this paper argues that focusing on achieving locally obtainable yields should take priority over increasing yield potential, due to input limitations as well as inherent soil properties. Furthermore, although management techniques can improve maize yield, through methods such as reducing planting delay and focusing on correct soil property management, they exist within the institutional and socio-economic constraints. Hence, other factors such as food security must be considered, especially in the context of climate change and risk management. Furthermore, the importance of CART analysis is demonstrated regarding its ability to explore non-parametric relationships between complex and inter-linked variables. There has been limited literature utilising this method in regards to yield and management exploration, and this paper aims to solidify and standardise this approach, in comparison to traditional parametric methods (Tittonell et al., 2008; Zheng et al., 2009).

2 Methodology

Tanzanian agriculture and its evolution

Structural Adjustment Programs and local political economy have influenced smallholder farmers through effects on inputs, infrastructure as well as direct involvement (Isinika et al., 2011; Putterman, 1995). In 1995, *13%* of arable land was utilised, while two-thirds of GDP originated from agriculture (Putterman, 1995: 312). Although this fell to *25.9%* by 2008, it is still the largest source of revenue, with *88%* of total agricultural area under small holders (Isinika et al., 2011; Skarstein, 2005).

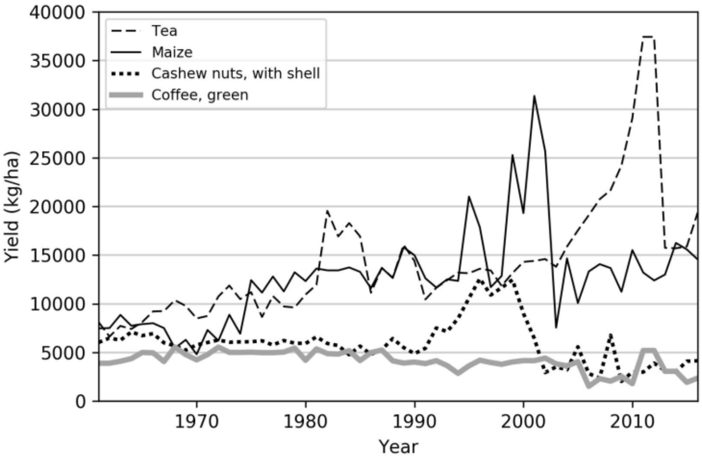


Figure 2: Data for four crops in Tanzania, extracted for 1961 to 2016 (Food and Agricultural Organisation of the United Nations, 2018*b*).

Pre-reform, agricultural stagnation and liberalisation

Pre-independence of Tanzania, commercial smallholder agriculture was successful and regulated state intervention (Ellis, 1982; Putterman, 1995: 312-313). Following "villagisation" through multi-purpose cooperatives, parastatal bodies were expected to mitigate transaction costs (Skarstein, 2005). However, a combination of institutional factors arising from uniform pricing led to reduction of official maize purchases from *224,000t* in 1978 - 1979 to *104,600t* in 1980 - 1981, as seen in figure 2 (Gibbon et al., 1993; Putterman, 1995: 313). Stagnation also arose from the cutting of ex-ternal aid due to a failure to meet International Monetary Fund conditions on structural adjustment, and the loss of local knowledge due to "villagisation", a process which was supported by interna-tional agencies, and lead to massive resettlement (Cooksey, 2011; Gibbon et al., 1993; Skarstein, 2005; Collier and Gunning, 1999).

Inter-regional barriers were removed by 1987, and a Structural Adjustment Program targeting input availability was launched in 1982. However, unfavourable exchange rates, low world market prices and removal of subsidies on inputs and transport all countered any positive potential effects from private sector competition following removal of parastatal bodies (Skarstein, 2005: 314). Alongside disappearing credit markets, this led to increasing wealth-stratification in favour of large farmsteads (Collier and Gunning, 1999; Gibbon et al., 1993).

Hence, farmers are increasingly diversifying for alternate income-sources while only relying on agriculture for subsistence. Decreasing fertiliser use (*10.5%* by 1997/1998) alongside increasing utilisation of land and population, combined with the effects of "villagisation" have lead to decreasing labour productivity (Gibbon et al., 1993). Even with the re-introduction of input subsidies in 2006, the presence of weak physical infrastructure means that not only does the remnant wealth-stratification remain, but higher transaction costs for smallholder farmers and lack of inputs make yield optimisation more difficult (Isinika et al., 2011).

Model form

Classification and Regression Tree (CART) Analysis

Stochastic frontier functions, such as translog and Cobb-douglas are commonly used to estimate inefficiency and deviance from the potential frontier, and applied to agricultural productivity (Umar et al., 2017; Ray, 1982; Pavelescu et al., 2011). Although they offer some flexibility regarding functional form, they have limitations. For example, Cobb-douglas assumes fixed elasticity amongst factor combinations while translog functions may require handling zero values and thus introducing bias (Ekbom et al., 2008; Heathfield and Wibe, 1987). This can lead to more complex interpretations and differences in policy conclusions just through transition of functional form as demonstrated by Umar et al. (2017).

CART is non-parametric and assumes no underlying distribution of the data (Tittonell et al., 2008; Heathfield and Wibe, 1987; De'ath and Fabricius, 2000). Prediction on continuous data is done by constructing a tree based on the homogeneity of the explanatory variable, where for every predictor all split combinations are considered and the sum of squares is optimised. For categorical variables there are $2^{k-1} - 1$ possibilities as opposed to $u - 1$ for numerical (De'ath and Fabricius, 2000: 3179). Impurity (heterogeneity) of a node is defined as follows:

$$I(A) = \sum_{i=1}^{C} f(p_i A) \qquad (1)$$

Where A is a node, C is the number of classes, p_{iA} is the proportion of A in class i and f is an impurity function, generally Gini index; $f(p) = p(1-p)$ (Therneau et al., 1997: 5-6). Splitting continues until a certain complexity parameter (α) is reached where the variance explained is lower than the cost of the next step, and hence a tree is generated.

Since only the rank of variables is important, a tree is invariant to transformations, and can handle missing variables through propagation based on characteristics of other values in the node. Though it can handle skewed and multi-modal data as well as categorical and continues variables, selection of variables can introduce bias (Tittonell et al., 2008; Tsien et al., 1998). On the other hand, in strongly linear situations it can under-perform compared to linear regression (De'ath and Fabricius, 2000: 3184). Crop management, soil fertility and bio-physical interactions create a complex network of growth-limiting interactions which would suggest this is not a limitation. On the other hand, CART can offer a flexible and adaptable technique to mapping soil fertility classes and management interactions in the context of limited or non-complete data (Zheng et al., 2009; Tittonell et al., 2008).

Data processing and principal component analysis (PCA)

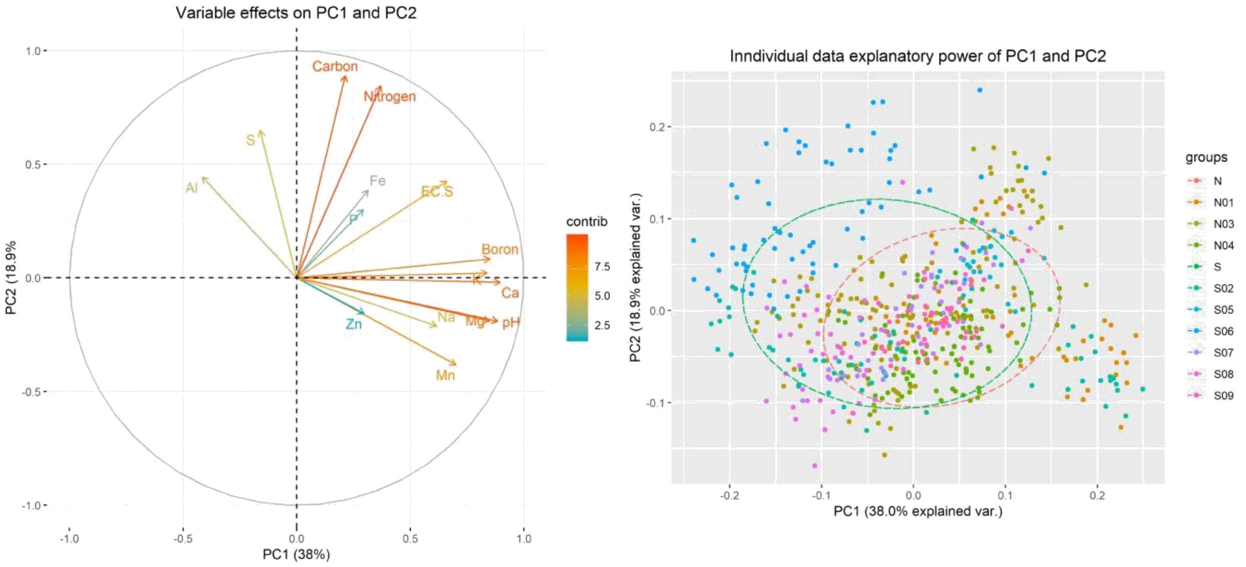
Data collection and processing

Data processing was done using the *R* package (R Core Team, 2013). Furthermore regression trees were constructed using "*RPART* ", but the "*partykit*" package was also used (Therneau and Atkinson, 2018; Zeileis et al., 2008). Plotting was generally done using "*ggplot2*" and a multitude of helper functions were utilised such as "*raster*", "*dplyr*" and "*sp*". (Wickham, 2016; Wickham et al., 2018; Hijmans, 2018; Bivand et al., 2013).

Soils data used were from the TZAPS 2017 survey and collected at 0 - 20*cm* (top) as well as 20 - 50*cm* (bottom) at four locations through geo-referenced at each field and analysed by mid-infrared (IR) and x-ray fluorescence (XRF) methods (Chamberlin et al., 2018). To simplify and assess the complex inter-relationships of the top soils data, principle component analysis was utilised to try and minimise the number of indicators for the CART analysis (Hammer et al., 1990; Tittonell et al., 2008). The goal of PCA is to identify correlated subsets within data sets with the goal of attempting to understand inter-relationships between them while minimising the number of variables but maintaining explanatory power (Hammer et al., 1990: 91-93). The data was normalised by using *ln* and PCA was then applied (*n* = 570) and the results visualised using ”factoextra” and ”ggbiplot” *R* packages (Vu, 2011; Kassambara and Mundt, 2017).

Soil PCA analysis

Figure 3 demonstrates the contribution of the top 10 principal components (PC). PC1 and PC2 explained *38%* and *18.8%* of the variation respectively. Furthermore, figure 4a and 4b show the main components of PC1 and PC2. For the former, *Ca*, *pH*, *Boron*, *M g* and *K* have approximately *15%* each, while PC2 has more than *70%* contribution from *Carbon*, *N itrogen* and *S*. Figure 3 is an orthogonal plot of the variable contributions to the first two principal components. Figure 3 (a) also shows the inter-relationship between variables in regards to correlation while (b) shows the contribution across regions with probability ellipses for both regions relatively similar. A strong positive relationship is observed between *Carbon* and *Nitrogen* (r^2 = 0.90) as well as between *pH* and *Ca* (r^2 = 0.72) as well as some other ions. The first 8 principal components were selected to be used in modelling yield response, and together explained *89%* in the soil property variations.



(a) **(b)**

Figure 3: (a) shows the orthogonal plots of variables and their contribution on PC1 and PC2 while (b) highlights the individual contribution of each site. Ellipses for north and south are within 1 σ.

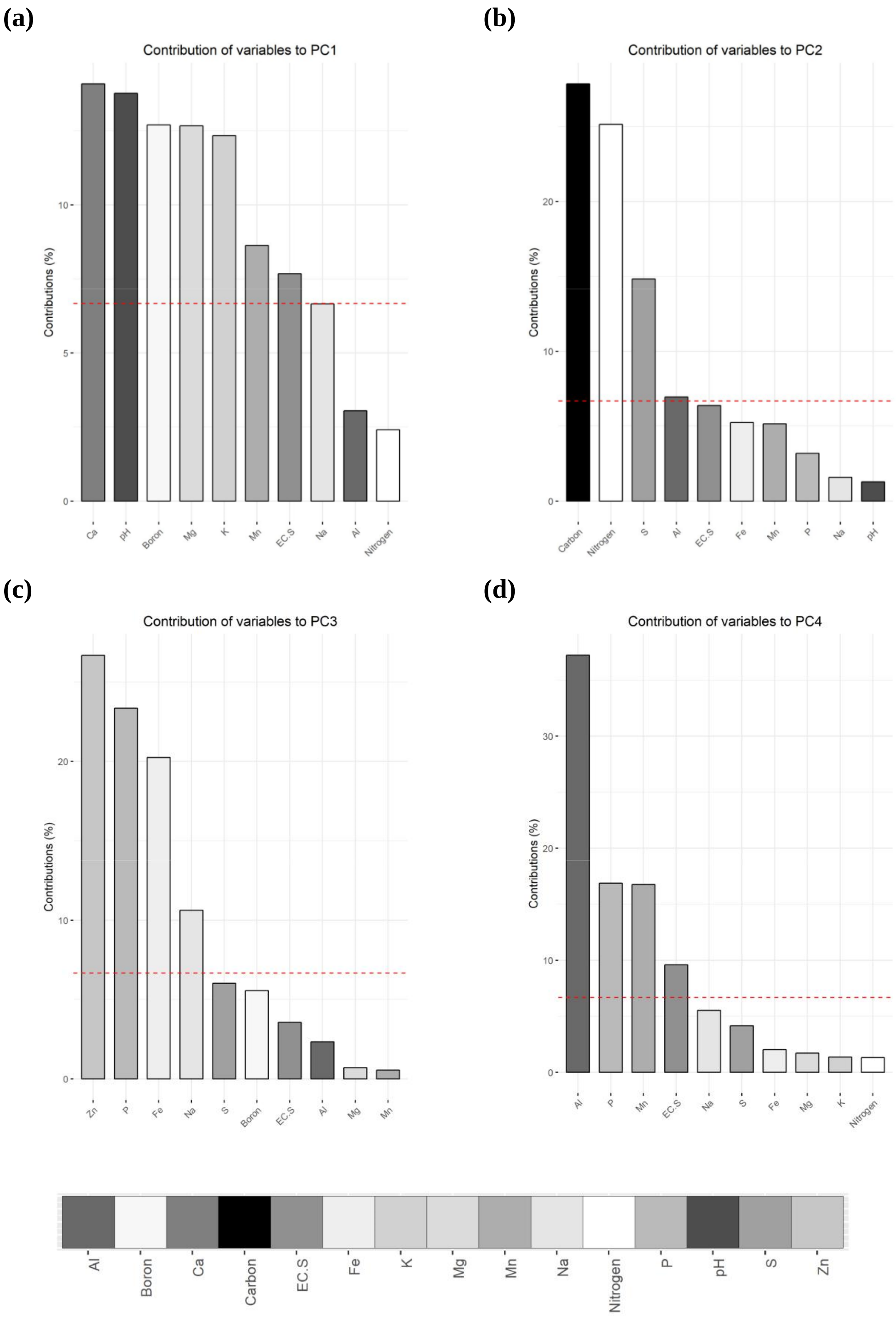


Figure 4: Variable contribution to *P C*1-4.

Choice and definition of variables

Three models were created - the first with yield dependant only on general management and site factors, with the second adding infrastructure proxies to explore transaction cost interaction regarding access to markets and inputs. Finally, soil properties were added to see how bio-physical constraints changed the interaction. The variables used were as follows:

Output (*Q*): the output measures, obtained through averaging three crop samples taken on the focal plot during the harvest and normalised as kilogram per hectare.

General site (*G*) : Rainfall (*R*) was included to count as a site proxy variable across the regions and account for the precipitation differences and extracted for the growing season, November to July for 2017 while for 2016, yearly data was used (Funk et al., 2015). Self-selected fertility of the plots by the farmers (*F*) and labour input (*L*) were also included.

General site variables (*G*) for regression tree:

$$G[R, F, L] \tag{2}$$

Management and conservation (*M*): resource intensity usage is one of the largest factors affecting yield variability, compromising of expenditure on improved seed, tillage systems, fertilisers, herbicides as well as other inputs such as labour. Due to multicollinearity, input expenditure per hectare (K_i) was chosen out of fertiliser expenditure and household assets. This was due to the fact that input expenditure per hectare performed significantly better than house hold assets ($r^2 = 0.05$ over 500 cross validation runs) and would include expenditure on other management when normalised through including labour inputs (*L*) as compared to just a fertiliser variable. Seed expenditure per hectare (K_s) was included as a separate input to try and account for hybrid seed usage. Planting density (M_{pden}), calculated as plants per m^2 from the ridge and planting distance was also included. Instead of herbicide data which was not available, recorded disease (M_d) for the focal plot was used alongside a weeding count for *N*-interaction (M_w).

Late planting due to drought stress can reduce yield by up to *20%*, and hence planting delay (M_{pdel}) was used to account for this as a management technique (Yonah et al., 2006). The bimodal rainfall pattern of Northern Tanzania as well as rainfall variability and soil moisture interaction make it difficult to estimate optimal planting date, but a crop planting calculator was utilised and earliest date picked for both zones, which was the start of December for South regions and beginning of November for North (Food and Agricultural Organisation of the United Nations, 2018*a*). An irrigation variable (M_i) was included to account for potential variability adaptation, but only *7%* of focal plots used this feature.

Soil conservation techniques can affect the plot fertility, soil moisture retention and run off, as well as soil degradation (Kangalawe et al., 2008; Lopes, 1980; Ekbom et al., 2008). Farmyard manure (M_m) has been extensively studied in regards to improving degraded soil condition and *pH* effects, alongside *N*-fixing legume inter-cropping and rotation, M_{ic} and M_r respectively (Lungu et al., 1993; Achieng et al., 2010). Forms of tillage as a conservation practice regarding topsoil restoration and acidity effects, as well as improvement of soil organic carbon, are commonly utilised (Ismail et al., 1994). Hence, maximum tillage level (M_t) was included as a measure of soil pro-cessing every season, alongside slash and burn in the last year, terracing, crop residues, ridging and contour bunds ($M_{c1\ 5}$).

General conservation and management variables for the regression tree:

$$M[K_i, K_s, M_t, M_{c1-5}, M_w, M_d, M_{pden}, M_{pdel}, M_{ic}, M_r, M_m] \tag{3}$$

Soil properties (*S*): soils data from the top 20*cm* (S_{t20}) of the soil were utilised, with non-correlated properties used for the regression tree, while slope (S_s) was also included as a physical aspect of soil properties due to the relationship between some soil variables and erosion (Lal, 2006).

Soil property variables can be summarised as follows:

$$S[S_{t20}, S_s] \tag{4}$$

Infrastructure and socio-economic factors (*I*): transaction costs regarding market and information access are a limitation, with an interaction between input resources and accessibility. This would also affect general management and hence distance to market (I_m), and fertiliser transport cost (I_f) were used as proxies. To test for the positive correlation observed between education and extension service retention and technology uptake, the education of the household head was used H_h.

Infrastructure and socio-economic variables for regression tree:

$$I[I_m, I_f, H_h] \tag{5}$$

Hence, the three models can be summarised as follows:

Model 1:

$$Q = f(M + G) \tag{6}$$

Model 2:

$$Q = f(M + I + G) \tag{7}$$

Model 3:

$$Q = f(M + I + S + G) \tag{8}$$

Study area and household statistics

The TAPS conducted in 2016 and 2017 covered a range of bio-physical and socio-economic environments and included 580 households (Chamberlin et al., 2018). The focal plots data used in the dissertation were from nine regions, split into two main zones; North and South. The five South regions experience a uni-modal rainfall pattern with a mean season precipitation of around *500mm* (figure 7 (a)). On the other hand, the three North areas receive *950mm* in a bi-modal environment. The nine areas cover a range of socio-economic and bio-physical backgrounds, with an average yield of 3.0tha^{-1} for the North and 2.1tha^{-1} for the South. Furthermore, a range of different practices is utilised - farmyard manure is much more prominently used as an input in the northern areas of Tanzania. However, the usage of a hand-hoe is still the most common form of tillage which highlights the low technological level, with irrigation remaining an issue at only *7%* of plots utilising this feature.

Resource use intensity is one of the most important factors affecting yield productivity and soil degradation effects (Lopes, 1980; Reeves, 1997). As expected, labour is one of the largest sources if inputs in the sampled plots, with a mean of 40 days per year. The average input expenditure per hectare on the focal plot is 50,700 TSh which is equivalent to approximately $22 - fertiliser usage and seed expenditure are both low demonstrating the continued limited resource usage, emphasized with the *66%* disease incidence rate. Inter-cropping is commonly used with *55%* of focal plots recording some form of inter-crop. Maize is the most common crop grown, alongside beans, and sunflowers are often used for extra money as a cash crop, sometimes in the short season of the bimodal areas.

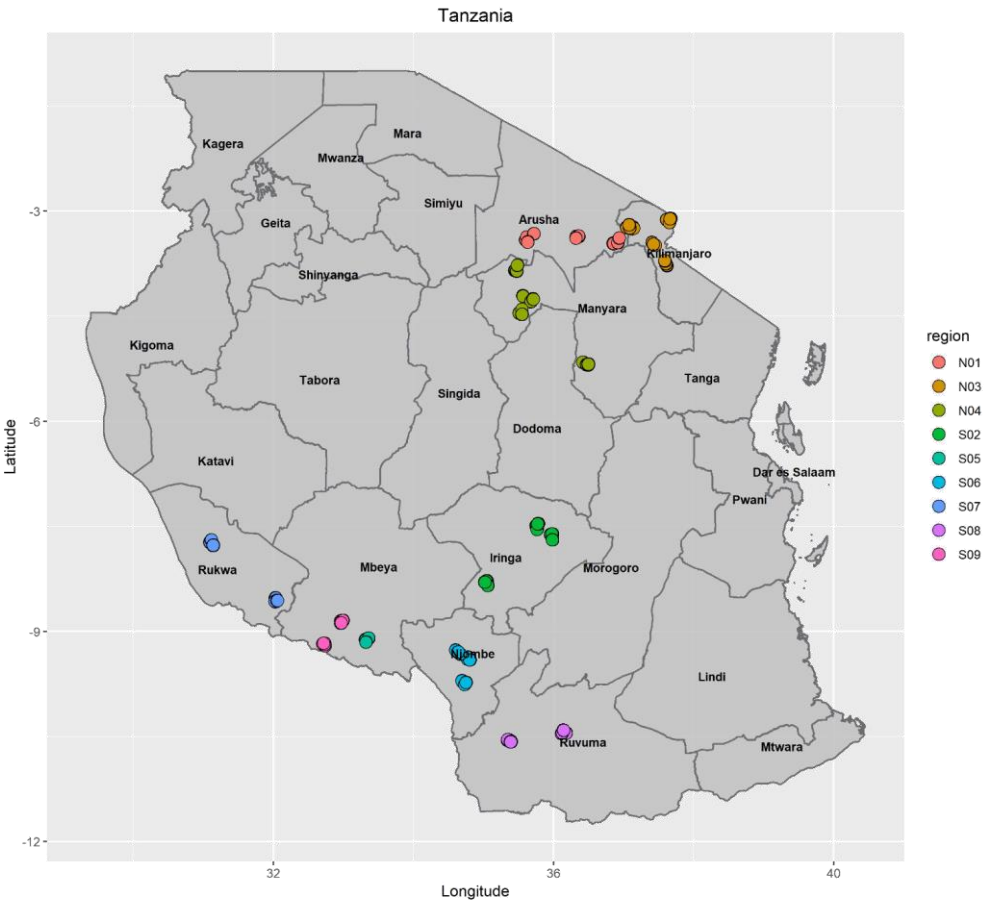


Figure 5: The individual focal plots as well as their respective regions with borders from *Global Administrative Areas* (2012).

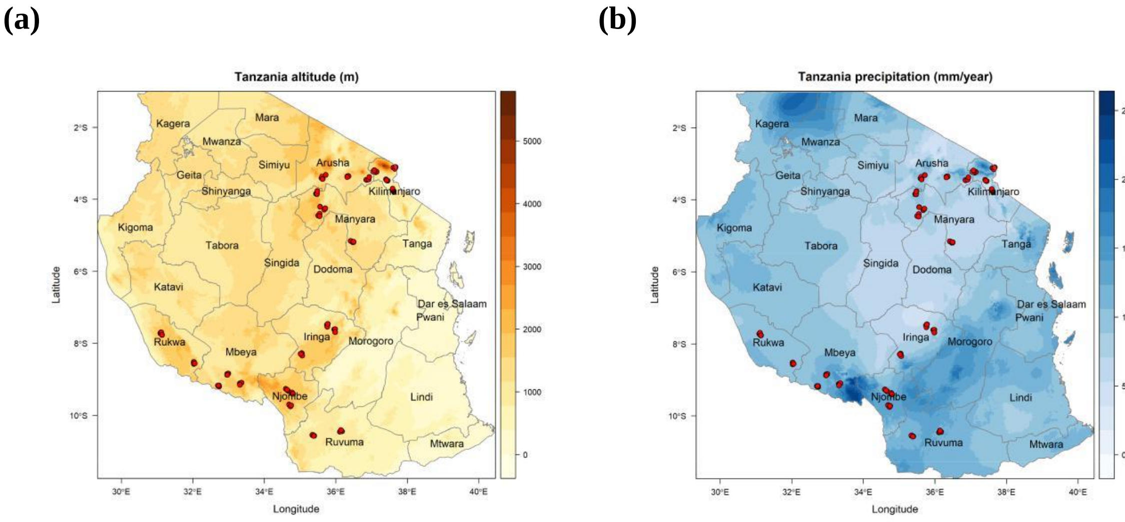


Figure 6: Variation in altitude and rainfall across Tanzania with data from Fick and Hijmans (2017) and borders from *Global Administrative Areas* (2012).

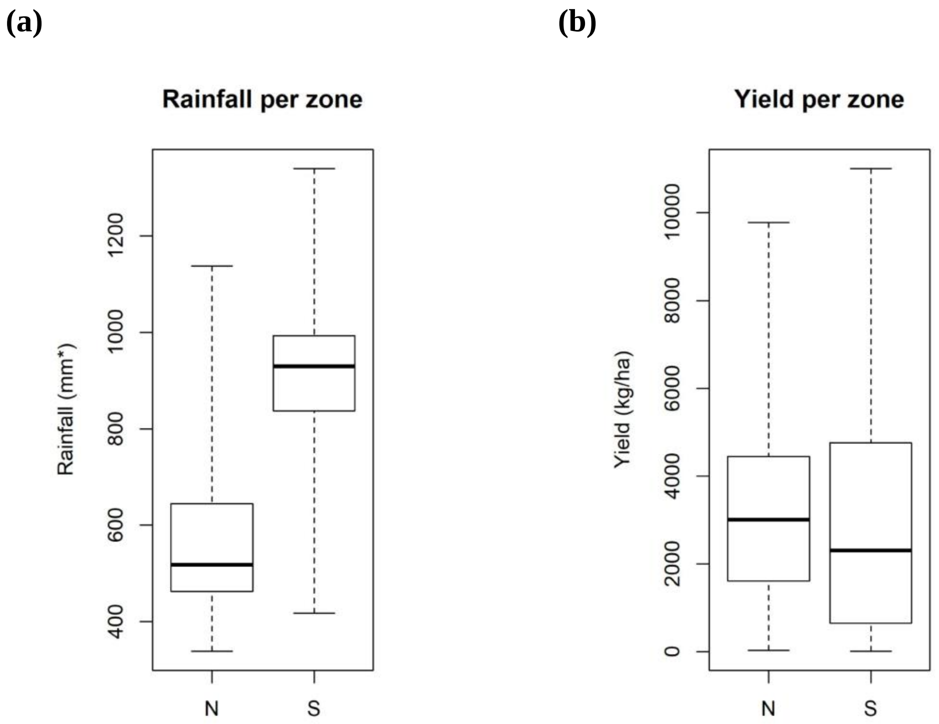


Figure 7: Boxplots showing the variation across zones for both rainfall and yield.

Land scarcity on a national scale is not an issue, but fragmentation is more common with fertile land being scarce, with a mean fertility rating of 2 (moderate) across all sites. The average size of the focal plot is only 2.13*ha*, compared to 3.34*ha* for the mean total farm area not only highlighting the level of subdivision of land but also the reliance on the main plot to deliver food. The mean time taken to get to the focal plot from the homestead is 20 minutes, and potentially, along with fragmentation, a remnant of the villagisation process. Fields are generally planted on moderately sloping soils (mean = 2), but often higher slopes are utilised resulting in a reduction of yield.

Finally, *87%* of all households have a male household head, with an average maximum education of around 7 years. Family sizes are large with an average of nearly 6 members per household, as well as an average age of around 39 years.

Soil properties

Tanzania has a range of soil types from volcanic soils in the Northern highlands, as well as gneiss soils, to sandy soils of the coast line with red soils in the central plateau area (Food and Agricultural Organisation of the United Nations, 2012). However, the interaction between soil type, its properties and yield is complex and varies with a lot of factors such as conservation practices and input use (Kangalawe et al., 2008). Although soils are generally acidic (pH_{mean} = 6.11), there is some variation across regions especially with SOM. This is especially the case of K and P, where regional variations are larger than the North to South differences. This is expected from the fact that soil properties (top 20*cm*) are liable to change over relatively short term periods of time with different practices. Furthermore, soil organic content is generally below *1%* by weight for all regions, except for Mbeya and Njombe which are considerably better.

3 Regression results

Models were ran with a number of cross validations between 20 and 100. A common issue with using regression tree analysis, especially in the case of data that is not normally distributed, is over-fitting. Often, the bias-variance trade-off is ignored regarding the complexity of the tree and explanatory power is purely optimised. Hence, two techniques were utilised to minimise cross-validation error. First, constraints were placed on the selection of the splitting nodes with a minimum number of 30 items required for a split. Hence, the cross-validation error was minimised. This was followed by pruning, where the minimum value of the cross-validation results for the splits was selected and any other branches past this point were cut.

Split	**Complexity Parameter (cp)**		
	2017		*2016*
	Model 1	***Model 3***	***Model 1***
1	0.051	0.0551	0.0785
2	0.0365	0.0457	0.0410
3	0.0296	0.0365	0.0214
4	0.0214	0.0214	0.0148
5	0.0191	0.0201	0.0145
6	0.0148	0.0153	0.0096
7	0.0070	0.0102	-
8	0.0001	0.0001	-
Final rel. error	0.810	0.778	0.751

Table 1: Complexity parameter (α) reduction over time number of splits for each of the models.

Variable	**Unit**
Input resources	10^5 TSh ha^{-1}
Fertiliser*	0 = no, 1 = yes
Fertility*	1 - 4
Distance to plot	Minutes
Delay*	Weeks
Seed input*	TSh ha^{-1}
Manure	0 = no, 1 = yes
Labour	Days per year
Carbon	% by weight
Fe*	mg kg^{-1}
P	mg kg^{-1}
K	mg kg^{-1}
pH*	log H^+

Table 2: Variables selected from all three models for 2016 and 2017. The * denotes variables selected for 2016 by CART analysis, which follow the same units (but of 10^3 magnitude, and on a farm level not per hectare).

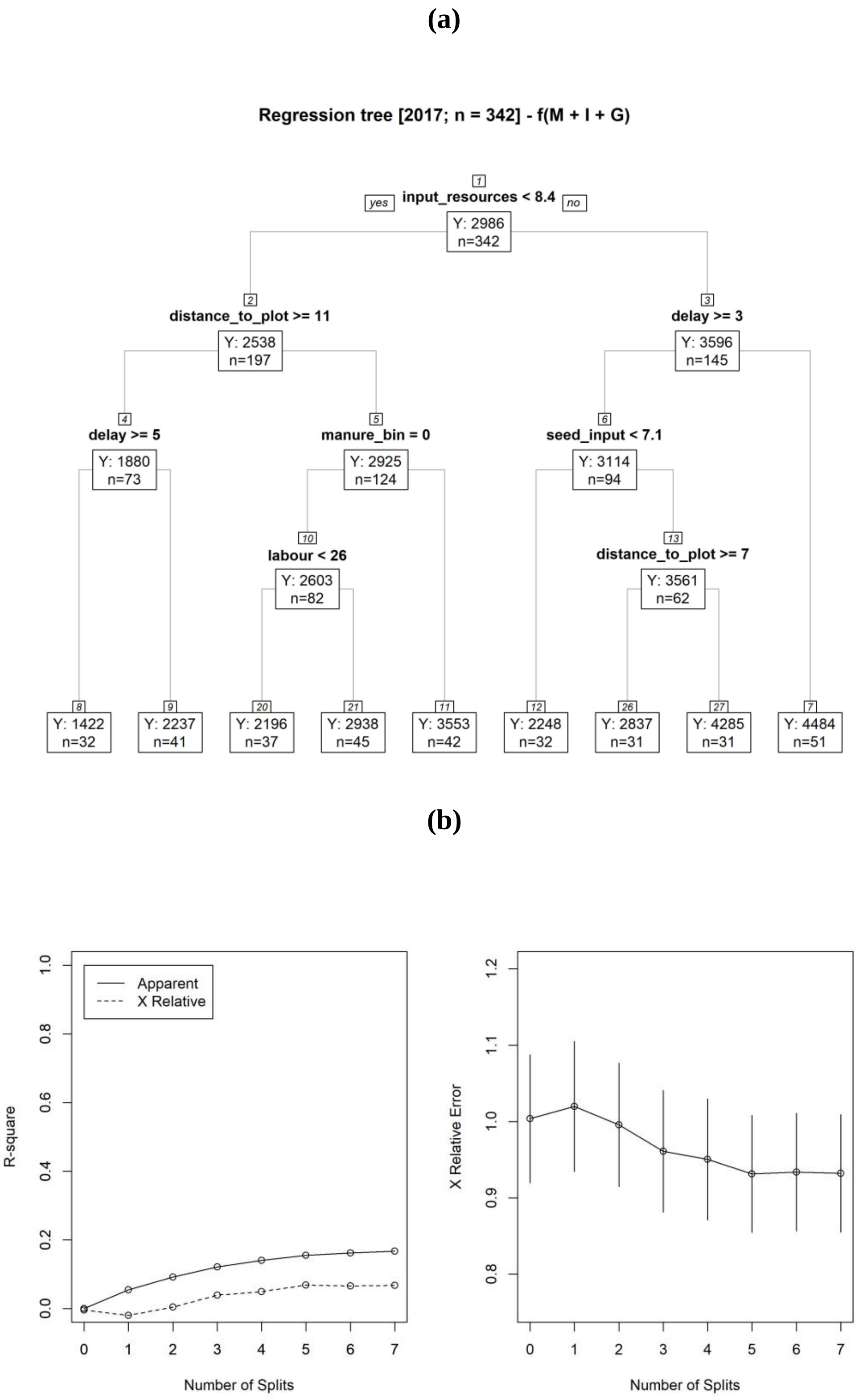


Figure 8: Regression tree from function of management and socio-economic factors for 2017 with *n* denoting count and *Y* the average yield. See table 2 for explanation of variables.

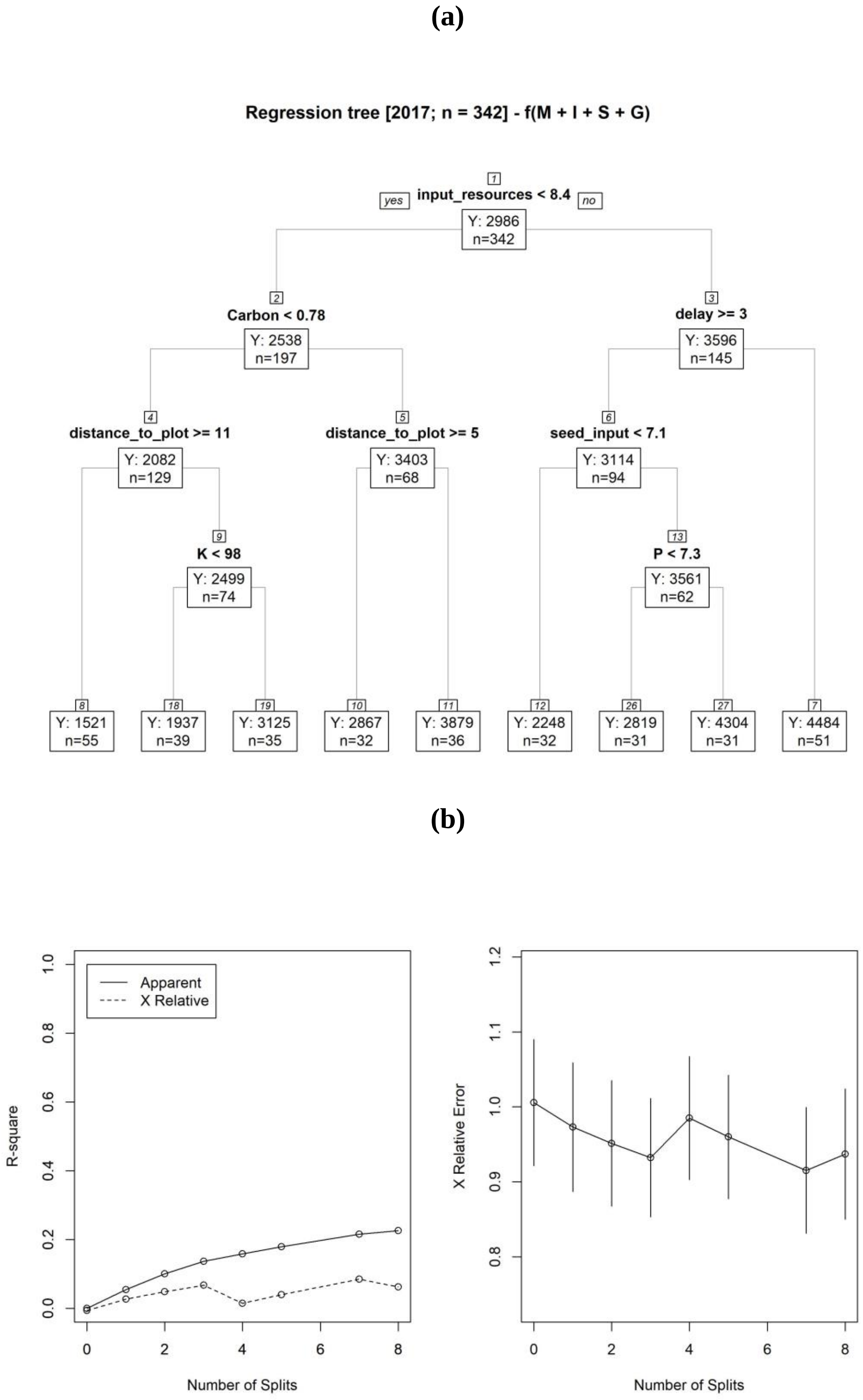
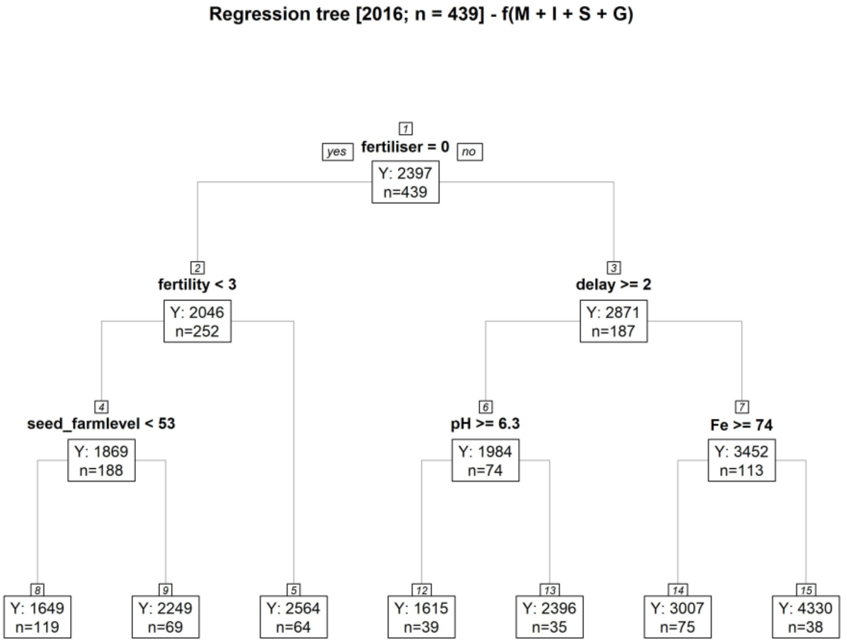


Figure 9: Regression tree from function of management, socio-economic and soil factors for 2017 with *n* denoting count and *Y* the average yield. See table 2 for explanation of variables.

(a)



(b)

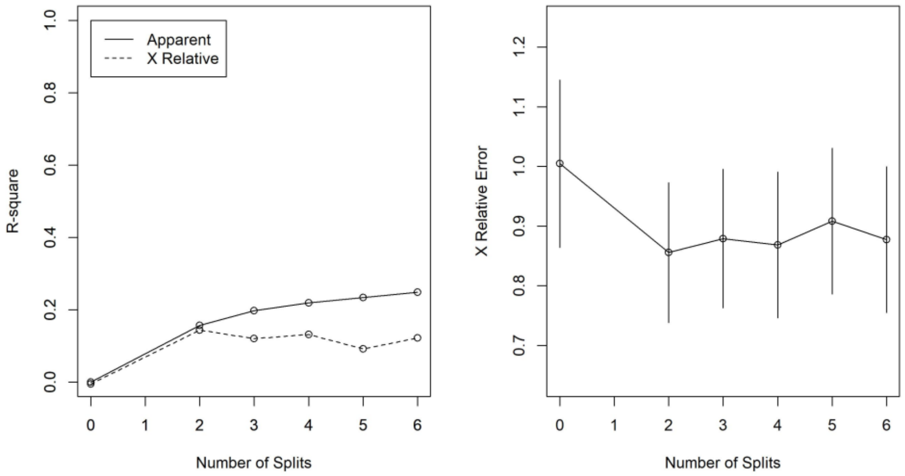


Figure 10: Regression tree from function of management, socio-economic and soil factors for 2016 with *n* denoting count and *Y* the average yield. See table 2 for explanation of variables.

Models 1 and 2 for 2017

The regression tree (figure 8) constructed from general management and site variables had 8 splitting nodes (R^2 = 0.19). Input resource intensity explained the most variance, splitting yields based on low (2.5*tha*$^{-1}$) and high (3.6*tha*$^{-1}$). In low resource fields, distance to the plot decided was second most important, deciding between fields with lowest yields (1.4*tha*$^{-1}$) and above average production (3.6*tha*$^{-1}$; *n* = 42). This is in line with the expectation from ease of access to fields and land fragmentation, supported by the fact that variance between the two extremes was affected by inputs of labour (planting on time) and manure application (Tittonell et al., 2007; Giller et al., 2006). High resource fields managed highest yield with on-time planting (4.5*tha*$^{-1}$) while seed expenditure and distance to the plot affected how much of the reduction in late planting was offset. The regression tree did not select any of the socio-economic or infrastructure variables as having more explanatory power so model 2 was identical.

Model 3 for 2017

Model 3 (figure 9a) improved with added soil properties (R^2 = 0.22). The first split remained the same as before, with resource use intensity being most important. In the high resource section, planting delay remained the most important, followed by seed expenditure, while the *P* threshold at around 7*mgkg*$^{-1}$, consistent with other literature, differentiated between close to maximum yield (4.3*tha*$^{-1}$) and sub-average yield (2.8*tha*$^{-1}$) (Vanlauwe et al., 2006). However, soil degradation became the most important split in lower-input fields, where sufficiently fertile fields with *C > 0.78%* close to the homestead had only a *13%* reduction of maximum yield. Distance to the plot and *K* deficiency decided lowest yields.

Model 1 for 2016

The model from the 2016 (figure 10a) data had the highest R^2 (0.75) with the lowest number of splits (table 1 and figure 10). However, planting density was not available (and planting stand density was highly correlated with yield r^2 = 0.28 and possibly implemented later on) while farm level expenditures were used instead. Fertiliser split the plots relatively symmetrically, with non-fertilised (2.0*tha*$^{-1}$) performing a lot worse than fertilised fields (2.9*tha*$^{-1}$). On non-fertilised fields, even with highest perceived fertility, yield remained close to average (2.6*tha*$^{-1}$). Interestingly, even with fertiliser improper planting management constrained maximum yield to only 2.4*tha*$^{-1}$ (*n* = 35), which is expected with large reductions in growing season length and compared to 4.3*tha*$^{-1}$) the maximum model yield, also very similar to 2017 (Mahoo et al., 1999).

4 Discussion and conclusion

None of the models utilised site proxy variables, suggesting that that the soil property and management variables had the most explanatory power, within the constraints set on the regression tree complexity (figures 9a and 10a). Unless specified otherwise, the 2017 model 3 (figure 9a) will be used for interpreting the data. Generally, the split of sites varied across all terminal nodes, except for three specific classes. In the high resource, high-yield terminal nodes (TN7-TN9) the separation was relatively homogeneous across North and South (table 3: $n_N = 23$; $n_S = 8$; $n_N = 26$; $n_S = 5$; $n_N = 4$; $n_S = 48$, respectively). This is consistent with the fact that the bi-modal North sites, which experience two growing seasons, had a larger variation of planting dates as opposed to the uni-modal South leading to a larger delay variable value. This is especially exacerbated by increasing variability due to changing rainfall patterns (Yonah et al., 2006; Mahoo et al., 1999).

Zone	Region name	Region code	*n*	Q (tha^{-1})	Number of objects per each terminal node								
					8 (TN1)	18 (TN2)	19 (TN3)	10 (TN4)	11 (TN5)	12 (TN6)	26 (TN7)	27 (TN8)	7 (TN9)
North			179	3.09 ± 1.91	24	21	27	20	24	10	23	26	4
	Arusha	N01	59	3.05 ± 1.82	10	0	10	10	15	3	9	2	0
	Kilimanjaro	N03	63	3.68 ± 2.04	3	5	5	3	7	7	10	23	0
	Manyara	N04	57	2.49 ± 1.71	11	16	12	7	2	0	4	1	4
South			162	2.87 ± 2.53	31	18	8	11	12	22	8	5	47
	Iringa	S02	50	2.10 ± 1.76	13	14	4	1	0	7	2	3	6
	Njombe	S06	24	2.52 ± 2.60	0	2	0	1	5	4	1	0	11
	Rukwa	S07	29	2.97 ± 2.57	7	0	3	8	6	0	2	0	3
	Ruvuma	S08	37	3.65 ± 2.88	8	2	0	0	0	9	3	2	13
	Songwe	S09	22	3.66 ± 2.87	3	0	1	1	1	2	0	0	14

Table 3: Breakdown of objects within each region for the terminal nodes. The distribution of North-South regions is especially highlighted.

Secondly, TN2 (*n* = 39) was mainly composed of sites from Manyara and Iringa (*n* = 16; *n* = 14) where fields were very close with highly degraded soils with lowest soil carbon content (0.48) as well as highest acidity (5.63) out of all the terminal nodes (table 3). This is possibly explained by a cyclical process - the lack of investment is potentially due to the fact that a large amount of inputs are necessary to restore the soil quality after a critical threshold is passed, and hence farmers prioritise other fields which are deemed more fertile (Tittonell and Giller, 2013; Giller et al., 2006: 20-21). This is further supported by the fact that TN2 has the lowest labour input out of all terminal nodes (table 4).

Split	*n*	Node (*TN*)	Fertility*	Delay (weeks)	Pld. density (plts. m^2)	Labour (days/year)	Distance (minutes)	Plot area (ha)	C (%)	N (%)	P (*mgkg*$^{-1}$)	K (*mgkg*$^{-1}$)	pH (log H^+)	Intercrop*	Irrigation*	Rainfall (mm/season)
Low resource use																
C <0.78																
Distant	55	8 (1)	2.00	4.45	2.97	39.7	44.2	1.21	0.51	0.037	6.55	122.3	6.29	0.62	0	626
Close (K < 98)	39	18 (2)	2.00	2.67	2.79	38.0	1.74	0.85	0.48	0.029	5.17	54.0	5.63	0.77	0.077	648
Close (K > 98)	35	19 (3)	2.17	5.00	2.89	43.5	3.60	0.99	0.59	0.039	6.54	188.2	6.49	0.71	0.11	531
C >0.78																
Distant	32	10 (4)	2.38	4.13	2.92	52.9	19.6	0.97	1.23	0.073	9.67	189.7	6.44	0.59	0.063	658
Close	36	11 (5)	2.14	4.83	2.91	42.0	1.11	0.87	1.25	0.071	10.4	150.9	6.31	0.53	0.028	703
High resource use																
Delayed planting																
Low seed input	32	12 (6)	1.81	5.19	3.33	28.7	26.6	0.46	0.86	0.048	8.60	117.1	6.18	0.53	0.031	799
High seed input (P < 7.3)	31	26 (7)	1.94	7.06	2.63	36.8	18.5	0.66	0.72	0.038	4.67	133.0	6.08	0.68	0	693
High seed input (P > 7.3)	31	27 (8)	1.97	6.58	2.65	34.4	13.1	0.57	1.27	0.072	14.5	137.9	6.31	0.55	0.16	773
On-time planting	51	7 (9)	1.88	0.59	3.762	45.4	23.2	0.66	0.69	0.040	6.13	74.4	5.87	0.24	0	885
P <			0.0011	0.001	0.0039	0.32	0.001	0.0018	0.001	0.001	0.001	0.001	0.001	0.001	0.0068	0.001

Table 4: Breakdown of variables for each terminal node. Dummy variables for presence (0 = no, 1 = yes) are highlighted by *.

Furthermore, the highest average yield recorded (4.5*tha*$^{-1}$) is similar to yields classified as high (5.67*tha*$^{-1}$) across East and Southern Africa (Tittonell and Giller, 2013: 86). On the other hand, the lowest average yield node of 1.5*tha*$^{-1}$ is above the Tanzanian average of 1.46*tha*$^{-1}$ (Food and Agricultural Organisation of the United Nations, 2018*b*). This is potentially due to the model settings, minimum of 30 objects per node, as well as the 236 missing observations.

Variation in yield response

Lack of inputs, especially regarding chemical fertilisers, as well as soil fertility management are seen as the biggest constraints in terms of yield potential in Sub-Saharan Africa (Giller et al., 2006; Tittonell et al., 2007; van Vugt and Franke, 2018). The full model (model 3; figure 9a) obtained with all variables for 2017 only corroborates this in specific contexts. Within high resource input fields however, general management such as the planting date (reduction of *31%*, see figure 9a) becomes the limiting factor in regards to reaching yield potential, which is especially problematic with increasing variation in the start of the growing season (Yonah et al., 2006). Furthermore, even with delayed planting, improved seed usage as well as soil inputs can offset the delay effect leading to similar yields. The model selected a value of 7.3$mgkg^{-1}$ (TN8) for P which is similar to the thresh-hold plateau of yield response obtained by Vanlauwe et al. (2006: 178) at 8$mgkg^{-1}$. This would suggest that with high seed expenditure and fertiliser input (*84%* of objects within TN8), near maximum yield is still obtainable, which is supported by an only *4%* yield reduction from terminal node 9.

On the other hand, within the Tanzanian context, low resource fields are possibly degraded to the extent that the lack of chemical fertiliser and carbon limit yield potential more than management techniques. Fertiliser response has been shown to be variable and constrained in some contexts depending on the level of degradation, with N and P being the limiting factors (Kihara et al., 2016). Interestingly, the model selected C content over Nitrogen. Though clay content in soil organic matter allows for a building up of N-mineral stacks, and is hence highly correlated (r^2 = 0.81), this also suggests that the soil physical aspects play an important role, with regards to water retention as well as ridge formation (Giller et al., 2006: 20).

With fields closest to the homestead, this can be offset by application of manure and inter-cropping with N-fixing legumes, as opposed to the fields further away which are generally of even worse soil quality and are often not prioritised due to their perceived lack of fertility (Giller et al., 2006). Furthermore, this has been shown to be the case on smaller farms and plots where resource intensity usage is higher. This increase in input intensity with decreasing land area is corroborated in the model with the average plot areas of all higher resource input terminal nodes (TN6-9) smaller than TN1-6. This is possibly explained by increased experimentation with fertiliser volume with increasing land leading to inefficiency, as well as the lack of choice regarding prioritising more fertile fields, especially in the case of Tanzania with high land fragmentation (Nkonya et al., 1997; Cornia, 1985).

Soil fertility gradients and general management

Another common issue is inefficiency created from soil fertility gradients within the farm. Factors such as slope and inherent parent material influence the fertility, but they also interact with management factors through chemical fertiliser use, tillage and especially farmyard manure application which has long standing effects on soil organic content. Due to the proximity of neighbouring plots as well as large differences in asset wealth and livestock ownership, some studies have highlighted the transfer of nutrients towards large herd owners through grazing, leading to a decreasing fertility for smallholders (Achard and Banoin, 2003: 187). Especially in areas of Tanzania like Iringa, where free-roaming cattle are only present illegally, manure usage is highly dependent on the amount of income available to the farmer (Giller et al., 2006; Kangalawe et al., 2008). However, using non-roaming livestock for manure was relatively common within the survey - yet the highest incidents of manure usage (TN2 and TN5) were also plots closes to the homestead, most likely home gardens at less than two minutes walk (table 3).

Furthermore, crop yield heterogeneity in the full model differentiates close and further fields as the third most important factor for low resource usage. Planting density as well as planting time are generally better managed due to the better efficient labour application, which is already often limited (Giller et al., 2006: 18). Figure 11 shows the terminal node variation of both planting density (a) and fertility (b). Interestingly, the average perceived fertility of all plots within the high resource input nodes is lower than the low input nodes, suggesting farmers within the sample were proactive to some extent - inputs were not applied predominantly on the pre-determined concept of plot fertility as opposed to research carried out in other parts of Eastern Africa (Tittonell et al., 2008). This is further supported by the presence of irrigation in *11%* of all plots in TN3 which has the lowest rainfall per season at *531mm*, we as well as *N*-fixing legume inter-cropping being utilised in more than *65%* of fields in three lowest Nitrogen nodes, TN2, TN3 and TN7 (table 3). However, with the level of fragmentation present, the choice of resource focus could be constrained to certain fields due to food security, which would seem to be the case with the large distances to the plots in high resource terminal nodes (TN6-9).

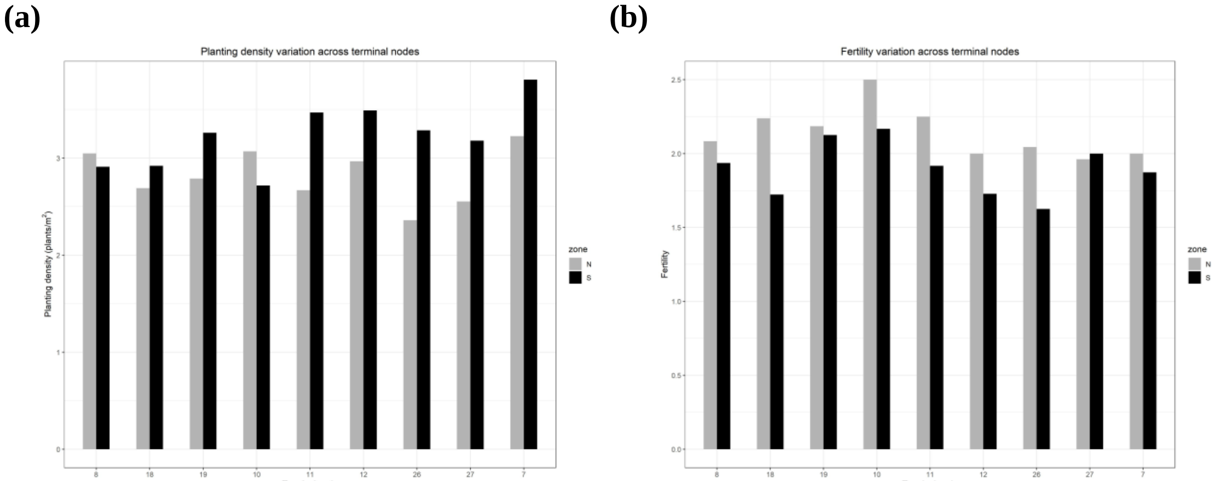


Figure 11: Zonal variation of yield with terminal nodes in regards to (a) planting density and (b) fertility.

The perceived fertility of the plot is often different to the actual potential yield. Boundary line simulations have been used to predict the maximum yield potential from soil properties and rainfall and can highlight areas of limitation (Tittonell et al., 2008, 2007; Tittonell and Giller, 2013; Giller et al., 2006; Fermont et al., 2009). To standardise the process, and following a similar method as utilised by van Vugt and Franke (2018), boundary lines were plotted by splitting soil property variables into equal width bins ($n_{bin} = 20 - 50$) and selecting the 'mean + 2x STDEV' (*95%* interval) for each bin to estimate local yield potential. A second order polynomial was fitted through the boundary points.

Management factors such as planting density and on-time planting could play a bigger role in plots such as TN9 where figure 12 suggests that yield is maximised within nutrient constraints. This is especially the case since soil properties are heterogeneous across regions - soil nutrients do not fall of rapidly as distance from the homestead increases (figure 14). Furthermore, defining $K < 50 mgkg^{-1}$ as the critical point for yield limitation, only fields in TN2 are close, which is supported by figure 13 which shows the nodes close to yield potential. However, most fields (like those in TN10) have the potential to transition to higher yield with better management and without heavy soil inputs as opposed to extremely carbon-degraded plots (TN7).

Nitrogen leeching has been extensively studied as a limitation in large parts of degraded soils of Sub-Saharan Africa (Tittonell and Giller, 2013; Kihara et al., 2016). Practices such as vegetation clearance disturb soil physical properties and lead to decreasing carbon rates, which can inter-act with management forms such as tillage and inputs. Furthermore, mineral fertiliser and crop residues alone are not necessarily enough to maintain carbon content. Often an organic-inorganic fertiliser trade off is present, where farmers may believe substitutability, while "quick-acting" fer-tilisers often do not build up carbon content sufficiently (Giller et al., 2006: 18-19). This is possibly the case for TN9, where plots are nearly all fertilised, and may be close to yield potential due to good management, but are facing increasing soil, and specifically, *N* degradation.

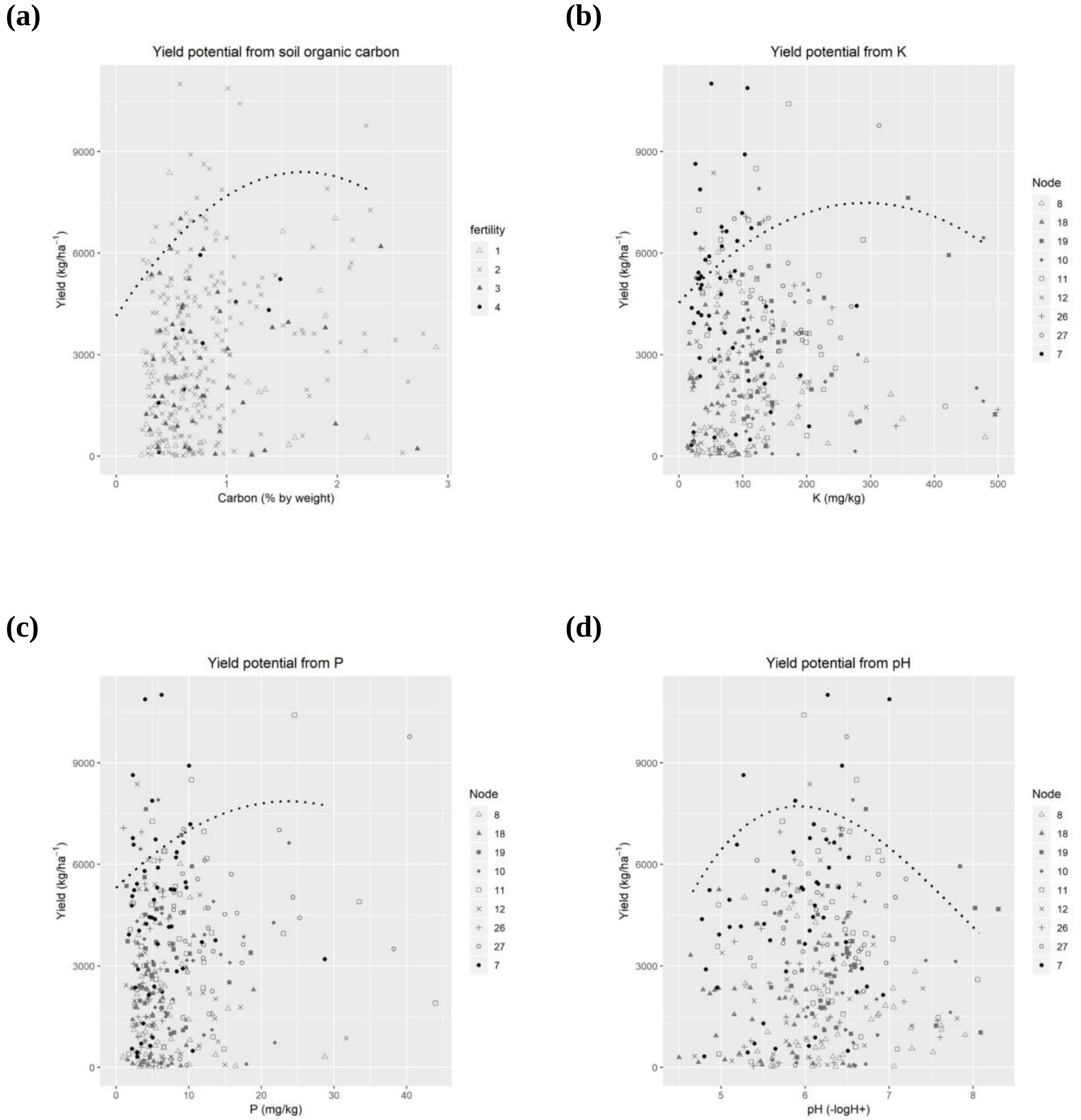


Figure 12: (a) shows the fertility rating against yield and carbon content of the soil while (b), (c) and (d) show terminal node objects plotted against their soil property values and yield. The dotted line shows the yield potential.

Limitations, socio-economic context and food security

Although management has been highlighted to affect soil properties, and explain more variation regarding yield output, there are inherent costs. For example, increasing planting density alongside planting within the optimal range may be desired, but come at the cost of increasing labour input which can be a limitation if hired labour is required (Giller et al., 2006). Increasing variability in rainfall and temperature suggest that it is increasingly difficult to estimate the optimal planting

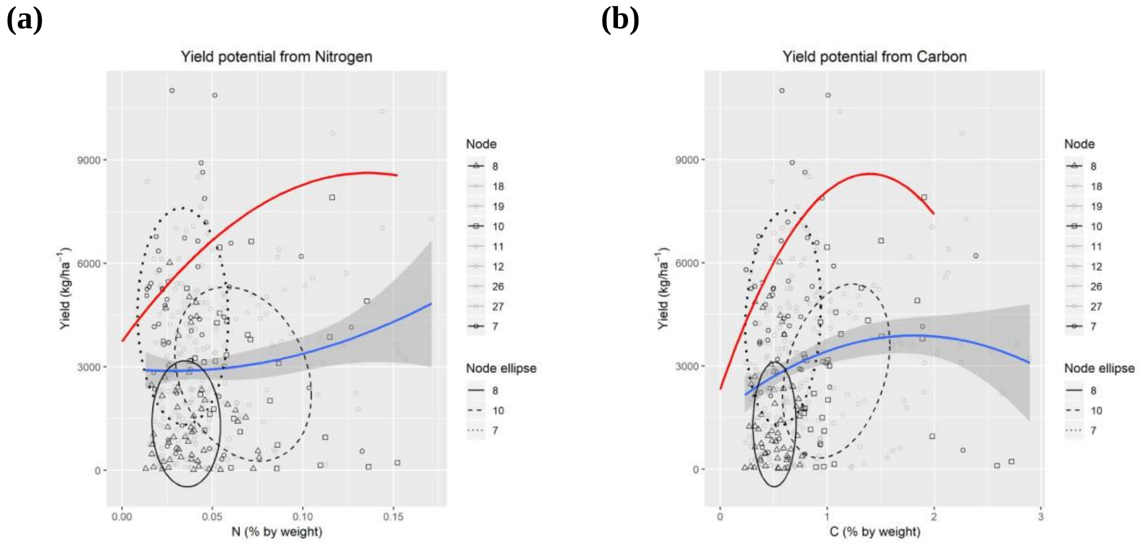


Figure 13: Nitrogen (a) and Carbon (b) yield variation with ellipses (level = 0.65) for selected nodes. The red line denotes the yield potential.

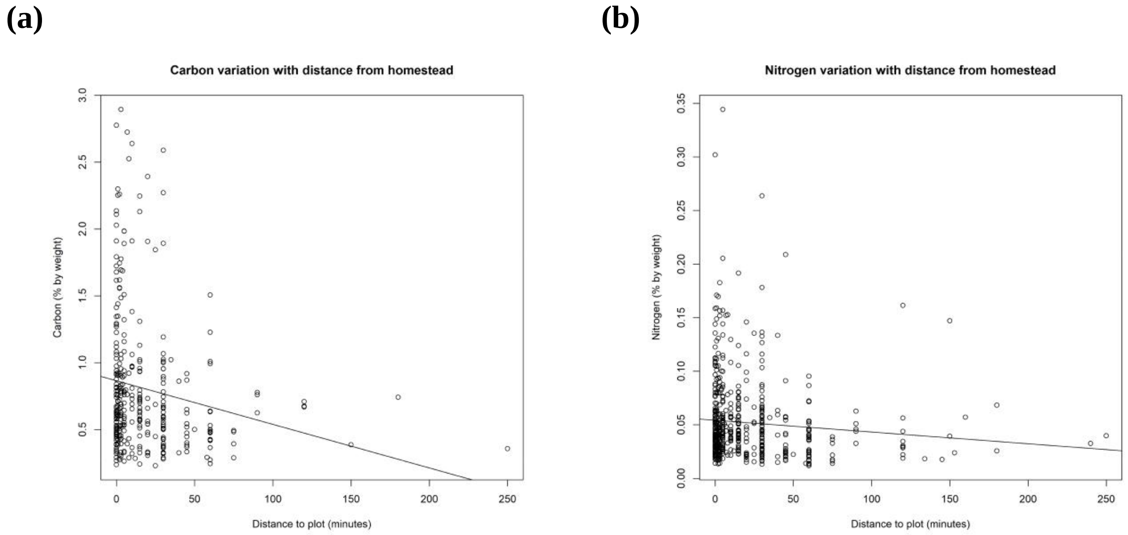


Figure 14: Carbon (a) and Nitrogen (b) variation with distance from homestead.

time, but increased variance can lead to crop failure and hence food insecurity (Mahoo et al., 1999; Yonah et al., 2006). Seed expenditure was also demonstrated to have a large impact in overcoming rainfall variability due to delay in planting. However, this is something that farmers are aware of, and often a conscious choice is to prioritise local varieties which may offer food security advantages rather than yield through possible early harvesting or improved storage characteristics of the strain (Tittonell and Giller, 2013: 88-89).

Furthermore, soil heterogeneity and soil degradation have a stratifying effect regarding yield, and hence food security. Not only does rehabilitation of soils require extensive manure and N-inputs, but a consistent application is required over a period of time (Tittonell and Giller, 2013: 82-83). This is especially problematic in the context of Tanzania where infrastructure difficulties and fertiliser procurement are variable and often difficult (Smaling and Braun, 1996; Harrington and Grace, 1998). Resource allocation decisions and opportunity costs are decided on a farm level, which limits conclusions drawn on management from focal plots alone (Giller et al., 2006: 14-15). Furthermore, input resource intensity and wealth are not enough to classify productive and non-productive plots, or even to predict soil fertility (Tittonell et al., 2005). The average household income of TN9 was *124,000TSh* ($54) as compared to *789,000TSh* ($343) for TN8 yet both achieved similar yields. This introduces another dynamic in terms of the interaction of wealth-flows and their effects on soil nutrients in the long term. The constraints in terms of land:labour ratios are especially relevant, for example TN7 farmers had an average income of 700,000TSh (compared to the mean of *462,000TSh*) yet limited labour (12.9 days per year compared to the mean of 21.4). Hence, even in resource intensive households with access to fertiliser, soil degradation can occur in the long run due to the lack of labour for applying manure, highlighting a substitution effect between inorganic-organic inputs (Omamo et al., 2002).

The issue of capturing the nuanced aspects of these interactions is shown in figure 15 which highlights the increasing variation with more nodes. This is potentially from bias in variable selection as well as omitted structural factors. Interestingly however, the low-resource soil degraded nodes are more homogeneous, possibly due to the increasing limitations soil nutrients have on yield. A regression tree could be constructed by including other plots as well as farm level inputs, while cross-validation, either by splitting the data or through new observations, could improve reliability.

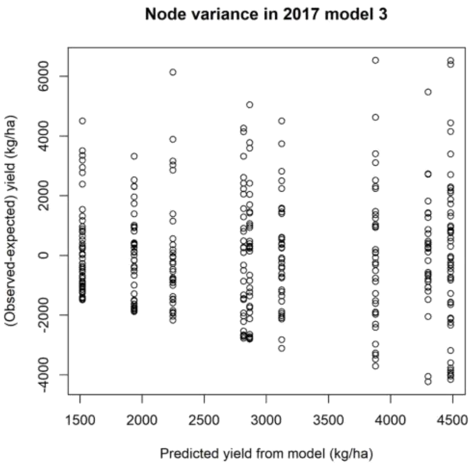


Figure 15: Variance of actual values for predicted classes.

Conclusion

Soil fertility heterogeneity has large impacts on crop yield variation across most of Sub-Saharan Africa (Kihara et al., 2016). Although intensive resource usage can mitigate nutrient limitations alongside good management techniques, the sustainability of this process is limited unless combined with techniques that counter soil degradation (Giller et al., 2006; Vanlauwe et al., 2006). Furthermore, even though perceived fertility did not seem to directly affect resource allocation, closer fields were prioritised with higher levels of inputs such as manure, which is possibly intensified by the high level of land fragmentation in Tanzania. Analysis of potential yield gaps across all terminal nodes of the final model suggest that increasing yield potential should not be the goal in Sub-Saharan Africa, but rather improving soil property and management technique interaction to attempt to reach locally obtainable yield (Fermont et al., 2009; van Vugt and Franke, 2018; Tittonell and Giller, 2013). This is especially crucial with the large variation in terms of response to fertiliser application, where out of three terminal nodes with degraded organic carbon content, only one had other nutrient limitations. Hence, improved allocation and application of resources could lead to better resource efficiency across all plots.

Though the models demonstrated the importance of planting delay as a second-highest level factor in explaining yield production, this has to be reconciled with local knowledge and prioritisation. Increasing variability of rainfall patterns and discontinuous access to fertiliser both highlight the difficulty in planning long-term rehabilitation and management improvement strategies (Tittonell and Giller, 2013; Mahoo et al., 1999). Management techniques must be optimised against local perceptions and goals regarding production, where food security and uncertainty in access to markets and transport may suggest maximising yield is offset against growing length, storage ability and response to limited rainfall (Giller et al., 2006). Differing wealth-flows and land:labour interaction mean that soil fertility and hence yield are not homogeneous across wealth classes, and affect management decisions across all levels (Omamo et al., 2002). However, CART does offer an avenue of exploring the complex relationships between soil properties and management within a socio-economic context. Transitive nodes with the lowest opportunity cost can be targeted, such as classifying fields which are non-fertile but responsive as opposed to non-responsive degraded. Thus, increasing food security and yield in some instances must be considered in the infrastructural and socio-economic context of Africa through a balance between resource use intensity, long-term sustainability as well as local knowledge and expectations.

References

Achard, F. and Banoin, M. (2003), ‘Fallows, forage production and nutrient transfers by livestock in niger’, *Nutrient Cycling in Agroecosystems* **65**(2), 183–189.

Achieng, J., Ouma, G., Odhiambo, G., Muyekho, F. et al. (2010), ‘Effect of farmyard manure and inorganic fertilizers on maize production on alfisols and ultisols in kakamega, western kenya.’, *Agriculture and Biology Journal of North America* **1**(4), 430–439.

Bivand, R. S., Pebesma, E. and Gomez-Rubio, V. (2013), *Applied spatial data analysis with R, Second edition*, Springer, NY.
URL: *http://www.asdar-book.org/*

Braun, A. R., Smaling, E. M., Muchugu, E. I., Shepherd, K., Corbett, J. D. et al. (1997), ‘Maintenance and improvement of soil productivity in the highlands of ethiopia, kenya, madagascar and uganda: an inventory of spatial and non-spatial survey and research data on natural re-sources and land productivity’, *AHI technical report series/International Centre for Research in Agroforestry. African Highlands Initiative; no. 6*.

Chamberlin, J., Masuki, K., Karwani, G., Nord, A., Simbogoso, V., Craufurd, P., Snapp, S., Jayne, T. S., Mushongi, A. and Magoye, L. (2018), ‘Tanzania agronomy panel survey (aps) 2017 - taking maize agronomy to scale in africa (tamasa) project’.
URL: *http://hdl.handle.net/11529/10548038*

Collier, P. and Gunning, J. W. (1999), ‘Explaining african economic performance’, *Journal of eco-nomic literature* **37**(1), 64–111.

Cooksey, B. (2011), ‘Marketing reform? the rise and fall of agricultural liberalisation in tanzania’, *Development Policy Review* **29**, s57–s81.

Cornia, G. A. (1985), ‘Farm size, land yields and the agricultural production function: An analysis for fifteen developing countries’, *World development* **13**(4), 513–534.

Croon, I., Deutsch, J. and Temu, A. (1984), ‘Maize production in tanzania’s southern highlands: current status and recommendations for the future’.

De’ath, G. and Fabricius, K. E. (2000), ‘Classification and regression trees: a powerful yet simple technique for ecological data analysis’, *Ecology* **81**(11), 3178–3192.

Ekbom, A., Sterner, T. et al. (2008), Production function analysis of soil properties and soil con-servation investments in tropical agriculture, Technical report.

Ellis, F. (1982), ‘Agricultural price policy in tanzania’, *World Development* **10**(4), 263–283.

Fermont, A. M., Van Asten, P. J., Tittonell, P., Van Wijk, M. T. and Giller, K. E. (2009), ‘Closing the cassava yield gap: an analysis from smallholder farms in east africa’, *Field Crops Research* **112**(1), 24–36.

Fick, S. E. and Hijmans, R. J. (2017), ‘Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas’, *International Journal of Climatology* **37**(12), 4302–4315.

Food and Agricultural Organisation of the United Nations (2012), ‘Soils map of tanzania’. [Online; accessed September 9, 2018].
URL: *http://data.fao.org/ref/3e7aa06e-2736-4b91-a70b-9babd6e14ff8.html?version=1.0*

Food and Agricultural Organisation of the United Nations (2018*a*), ‘FAO Crop Calendar’. [Online; accessed August 26, 2018].

URL: *http://www.fao.org/agriculture/seed/cropcalendar/welcome.do*

Food and Agricultural Organisation of the United Nations (2018*b*), 'FAOSTAT Database'. [Online; accessed September 13, 2018].
URL: *http://www.fao.org/faostat/en/data/QC*

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A. et al. (2015), 'The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes', *Scientific data* **2**, 150066.

Gibbon, P., Havnevik, K. J. and Hermele, K. (1993), *A blighted harvest: the World Bank & African agriculture in the 1980s*, Africa World Press.

Giller, K. E., Rowe, E. C., de Ridder, N. and van Keulen, H. (2006), 'Resource use dynamics and interactions in the tropics: Scaling up in space and time', *Agricultural Systems* **88**(1), 8–27.

Global Administrative Areas (2012). [Online; accessed August 26, 2018].
URL: *http://www.gadm.org/home*

Hammer, R. D., Philpot, J. W. and Maatta, J. M. (1990), 'Applying principal component analysis to soil-landscape research-quantifying the subjective'.

Harrington, L. W. and Grace, P. (1998), 'Research on soil fertility in southern africa: Ten awkward questions'.

Heathfield, D. F. and Wibe, S. (1987), *An introduction to cost and production functions*, Humanities Pr.

Hijmans, R. J. (2018), *raster: Geographic Data Analysis and Modeling*. R package version 2.7-8/r3387.
URL: *https://R-Forge.R-project.org/projects/raster/*

Ionescu, C., Klein, R. J., Hinkel, J., Kumar, K. K. and Klein, R. (2009), 'Towards a formal framework of vulnerability to climate change', *Environmental Modeling & Assessment* **14**(1), 1–16.

Isham, J. (2002), 'The effect of social capital on fertiliser adoption: Evidence from rural tanzania', *Journal of African Economies* **11**(1), 39–60.

Isinika, A. C., Msuya, E. E., Djurfeldt, G. and Aryeetey, E. (2011), 'Addressing food selfsufficiency in tanzania: a balancing act of policy coordination', *African Smallholders: Food Crops, Markets and Policy, G. Djurfeldt et al* .

Ismail, I., Blevins, R. and Frye, W. (1994), 'Long-term no-tillage effects on soil properties and continuous corn yields', *Soil Science Society of America Journal* **58**(1), 193–198.

Kaliba, A. R., Verkuijl, H. and Mwangi, W. (2000), 'Factors affecting adoption of improved maize seeds and use of inorganic fertilizer for maize production in the intermediate and lowland zones of tanzania', *Journal of Agricultural and Applied Economics* **32**(1), 35–47.

Kangalawe, R. Y., Christiansson, C. and Östberg, W. (2008), 'Changing land-use patterns and farming strategies in the degraded environment of the irangi hills, central tanzania', *Agriculture, ecosystems & environment* **125**(1-4), 33–47.

Kassambara, A. and Mundt, F. (2017), *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.5.

URL: *https://CRAN.R-project.org/package=factoextra*

Katinila, N., Verkuijl, H., Mwangi, W., Anandajayasekeram, P. and Moshi, A. J. (1998), *Adoption of maize production technologies in Southern Tanzania*, CIMMYT.

Kelly, P. M. and Adger, W. N. (2000), 'Theory and practice in assessing vulnerability to climate change andfacilitating adaptation', *Climatic change* **47**(4), 325–352.

Kihara, J., Nziguheba, G., Zingore, S., Coulibaly, A., Esilaba, A., Kabambe, V., Njoroge, S., Palm, C. and Huising, J. (2016), 'Understanding variability in crop response to fertilizer and amend-ments in sub-saharan africa', *Agriculture, ecosystems & environment* **229**, 1–12.

Lal, R. (2006), 'Enhancing crop yields in the developing countries through restoration of the soil organic carbon pool in agricultural lands', *Land Degradation & Development* **17**(2), 197–209.

Lema, M. and Majule, A. E. (2009), 'Impacts of climate change, variability and adaptation strategies on agriculture in semi arid areas of tanzania: The case of manyoni district in singida region, tanzania', *African Journal of Environmental Science and Technology* **3**(8), 206–218.

Lopes, A. (1980), 'Micronutrients in soils of the tropics as constraints to food production', *Priorities for alleviating soil-related constraints to food production in the tropic* pp. 277–298.

Lungu, O., Temba, J., Chirwa, B. and Lungu, C. (1993), 'Effects of lime and farmyard manure on soil acidity and maize growth on an acid alfisol from zambia'.

Mahoo, H., Young, M. and Mzirai, O. (1999), 'Rainfall variability and its implications for the transfer-ability of experimental results in the semi arid areas of tanzania', *Tanzania Journal of Agricultural Sciences* **2**(2).

Mongi, H., Majule, A. E. and Lyimo, J. G. (2010), 'Vulnerability and adaptation of rain fed agriculture to climate change and variability in semi-arid tanzania', *African Journal of Environmental Science and Technology* **4**(6).

Morris, M. (n.d.), Understanding household coping strategies in semi-arid tanzania. department for international development strategy for research on renewable natural resources, *in* 'Institute, University of Greenwich. UK Deptarment for International Development', Citeseer.

Morton, J. F. (2007), 'The impact of climate change on smallholder and subsistence agriculture', *Proceedings of the national academy of sciences* **104**(50), 19680–19685.

Nkonya, E., Schroeder, T. and Norman, D. (1997), 'Factors affecting adoption of improved maize seed and fertiliser in northern tanzania', *Journal of Agricultural Economics* **48**(1-3), 1–12.

Omamo, S. W., Williams, J. C., Obare, G. and Ndiwa, N. (2002), 'Soil fertility management on small farms in africa: evidence from nakuru district, kenya', *Food policy* **27**(2), 159–170.

Pavelescu, F.-M. et al. (2011), 'Some aspects of the translog production function estimation', *Ro-manian Journal of Economics* **32**(1), 41.

Putterman, L. (1995), 'Economic reform and smallholder agriculture in tanzania: A discussion of recent market liberalization, road rehabilitation, and technology dissemination efforts', *World Development* **23**(2), 311–326

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Ray, S. C. (1982), 'A translog cost function analysis of us agriculture, 1939–77', *American Journal of Agricultural Economics* **64**(3), 490–498.

Reeves, D. (1997), 'The role of soil organic matter in maintaining soil quality in continuous cropping systems', *Soil and Tillage Research* **43**(1-2), 131–167.

Rowhani, P., Lobell, D. B., Linderman, M. and Ramankutty, N. (2011), 'Climate variability and crop production in tanzania', *Agricultural and Forest Meteorology* **151**(4), 449–460.

Seyoum, E., Battese, G. E. and Fleming, E. (1998), 'Technical efficiency and productivity of maize producers in eastern ethiopia: a study of farmers within and outside the sasakawa-global 2000 project', *Agricultural economics* **19**(3), 341–348.

Skarstein, R. (2005), 'Economic liberalization and smallholder productivity in tanzania. from promised success to real failure, 1985–1998', *Journal of Agrarian Change* **5**(3), 334–362.

Smaling, E. M. and Braun, A. (1996), 'Soil fertility research in sub-saharan africa: New dimensions, new challenges', *Communications in Soil Science and Plant Analysis* **27**(3-4), 365–386.

Tabu, I., Obura, R., Bationo, A. and Nakhone, L. (2005), 'Effect of farmers' management practices on soil properties and maize yield', *J. Agron* **4**, 1–7.

The World Bank (2017), 'Agricultural inputs'. data retrieved from World Development Indicators:
http://data.worldbank.org/indicator/SP.DYN.LE00.FE.IN.

Therneau, T. and Atkinson, B. (2018), *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13.
URL: *https://CRAN.R-project.org/package=rpart*

Therneau, T. M., Atkinson, E. J. et al. (1997), 'An introduction to recursive partitioning using the rpart routines'.

Tittonell, P. and Giller, K. E. (2013), 'When yield gaps are poverty traps: The paradigm of ecological intensification in african smallholder agriculture', *Field Crops Research* **143**, 76–90.

Tittonell, P., Shepherd, K. D., Vanlauwe, B. and Giller, K. E. (2008), 'Unravelling the effects of soil and crop management on maize productivity in smallholder agricultural systems of western kenya—an application of classification and regression tree analysis', *Agriculture, ecosystems & environment* **123**(1-3), 137–150.

Tittonell, P., Vanlauwe, B., De Ridder, N. and Giller, K. E. (2007), 'Heterogeneity of crop productivity and resource use efficiency within smallholder kenyan farms: Soil fertility gradients or management intensity gradients?', *Agricultural systems* **94**(2), 376–390.

Tittonell, P., Vanlauwe, B., Leffelaar, P., Rowe, E. C. and Giller, K. E. (2005), 'Exploring diversity in soil fertility management of smallholder farms in western kenya: I. heterogeneity at region and farm scale', *Agriculture, ecosystems & environment* **110**(3-4), 149–165.

Tsien, C. L., Fraser, H., Long, W. J. and Kennedy, R. L. (1998), 'Using classification tree and logistic regression methods to diagnose myocardial infarction', *Medinfo* **98**.

Umar, H., Girei, A. and Yakubu, D. (2017), 'Comparison of cobb-douglas and translog frontier mod-els in the analysis of technical efficiency in dry-season tomato production', *Agrosearch* **17**(2), 67– 77.

van Vugt, D. and Franke, A. (2018), 'Exploring the yield gap of orange-fleshed sweet potato vari-eties on smallholder farmers' fields in malawi', *Field Crops Research* **221**, 245–256.

Vanlauwe, B., Tittonell, P. and Mukalama, J. (2006), 'Within-farm soil fertility gradients affect re-sponse of maize to fertiliser application in western kenya', *Nutrient Cycling in Agroecosystems* **76**(2-3), 171–182.

Vu, V. Q. (2011), *ggbiplot: A ggplot2 based biplot*. R package version 0.55.
URL: *http://github.com/vqv/ggbiplot*

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
URL: *http://ggplot2.org*

Wickham, H., François, R., Henry, L. and Müller, K. (2018), *dplyr: A Grammar of Data Manipulation*.
R package version 0.7.6.
URL: *https:CRAN.R-project.org/package=dplyr*

Wiig, H., Aune, J. B., Glomsrød, S. and Iversen, V. (2001), 'Structural adjustment and soil degrada-tion in tanzania a cge model approach with endogenous soil productivity', *Agricultural Economics* **24**(3), 263–287.

Yonah, I., Oteng'i, S. and Lukorito, C. (2006), 'Assessment of the growing season over the unimodal rainfall regime region of tanzania', *Tanzania Journal of Agricultural Sciences* **7**(1).

Zeileis, A., Hothorn, T. and Hornik, K. (2008), 'Model-based recursive partitioning', *Journal of Com-putational and Graphical Statistics* **17**(2), 492–514.

Zheng, H., Chen, L., Han, X., Zhao, X. and Ma, Y. (2009), 'Classification and regression tree (cart) for analysis of soybean yield variability among fields in northeast china: The importance of phosphorus application rates under drought conditions', *Agriculture, Ecosystems & Environment* **132**(1-2), 98–105.