UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

COLEGIO DE CIENCIAS Y HUMANIDADES

MAESTRÍA EN CIENCIAS DE LA COMPLEJIDAD

GRUPO: DINÁMICA NO LINEAL Y SISTEMAS COMPLEJOS

PREDICCIÓN DE LA CONTRIBUCIÓN DE LA AGRICULTURA A LA SEGURIDAD ALIMENTARIA DE CAMPESINOS EN EL ALTIPLANO OCCIDENTAL DE GUATEMALA, APLICANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING ALGORITHMS)

TESIS

QUE PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS DE LA COMPLEJIDAD

PRESENTA

Biól. Luis Barba Escoto

Director: Dr. Santiago López Ridaura

Ciudad de México, Junio 2019

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

CA Coordinación Académica

SISTEMA BIBLIOTECARIO DE INFORMACIÓN Y DOCUMENTACIÓN

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO
COORDINACIÓN ACADÉMICA

RESTRICCIONES DE USO PARA LAS TESIS DIGITALES

DERECHOS RESERVADOS©

Table of Contents

Key words

Food security, Western Highlands Guatemala, Land availability, Machine Learning Algorithms

Abstract

Food security is a major challenge in Guatemala, one of the poorest countries in the world. Food insecurity is dominantly concentrated in the Western Highlands of Guatemala (WHG) where indigenous communities have been the main victims of social, political and economic marginalization. In this study we present an analysis to identify the main sources of food for different types of farm households and assess their food security status through a simple, yet robust, potential food availability (PFA) indicator. Based on a large and rich dataset of nearly 5,000 farm households, our results show the diversity of farming systems in the region dominated by maize and coffee production as well as the large differences in terms of their potential food availability. In our model, 52% of farm households in the WHG do not have the means to attain sufficient energy from their agricultural activities. We then developed a series of predictive models of potential food availability through the application of machine learning techniques of regression and classification. A neural network model performed best in both, predicting the continuous PFA (R^2= 0.83) as well as the binary status of food-secure/food insecure households (accuracy: 84.8%). The analysis of the interactions of the model predictors allows detecting land availability per person as the main constraint for food security attainment (a PFA of 2500 kcal/day /male adult equivalents) identifying a threshold of 0.065 ha per person as a pre-requisite to obtain enough kcal from agricultural production and commercialization, the median value of land availability per person is 0.085 ha, that could explain why 52% of the sample does not reach an enough food security from agriculture. On top of the low contribution of agriculture to food security and the fact that it reaches acceptable limits when commercial crops like coffee are grown, the sustainability of these systems should be quickly addressed as great risks are taken when policies and intervention programs prioritize only the productive element as the solution for malnutrition and undernourishment without considering its ecological and social consequences.

List of Figures

List of Tables

Acknowledgements

I would like to thank M.Sc. José Luis Gutiérrez Sánchez and Dr. Damián Hernández Herrán for their remarks that greatly improved this thesis, and specially, to Dr. Fernando Ramírez Alatriste and Dr. Santiago López Ridaura for their invaluable contributions that shaped and guided this research work.

Also, I would like to thank Universidad Autónoma de la Ciudad de México for supporting the printing of this thesis.

Chapter 1: Introduction

1.1 BACKGROUND

Uneven progress towards food security remains an intractable development challenge, and will remain so for decades to come. Sub-Saharan Africa has received most attention regarding this topic, but achieving food security remains a challenge in other parts of the world (Ericksen et al. 2009).

For the 20th century policymaker's food security was a key concept and a perspective about the future. This perspective centered on raising production as the solution to under-consumption and hunger. But another perspective, still emerging and less popular, more social and ecological, accepts the need to address a complex array of problems not just production. Production by itself in the current agenda, the production oriented one, requires a renewal and change from an agricultural-focused point of view to a food systems approach in order to be sustainable, incorporating the complex range of evidence from social, environmental and economic sources into an integrated policy response (Lang and Barling 2012). In this study we focus on the availability of food, as a first diagnostic of the problem of food security in the WHG, further studies should address also the sustainability of the food systems in the region.

In Guatemala, rural poverty and food insecurity are endemic. The World Food Program ranks Guatemala as the country with the highest level of children undernutrition in the western hemisphere and the fourth highest level in the world (WFP, 2018). The International Fund for Agricultural Development (IFAD, 2011) indicates that approximately 70% of the impoverished population lives in rural areas, where agricultural production is the main livelihood activity and source of food. Poverty and malnutrition are especially prevalent amongst indigenous communities, which comprise 38% of the total population. These communities are mainly concentrated in the Western Highlands of Guatemala (WHG), and historically have suffered from discrimination leading to structural exclusion, social inequality and violence (Steinberg and Taylor 2008). In 2010, the United States Agency for International Development (USAID) launched a strategy to address poverty and

chronic malnutrition in Guatemala. The strategy document specifically highlighted the concentration of poverty and malnutrition in the WHG (USAID 2013).

Food security in the WHG is, as in many rural areas of the world, a complex, multifaceted phenomena where farm households are, at the same time, producers and consumers of food, and their livelihoods relies on both agricultural and non-agricultural activities. This is especially the case in the WHG because of high population density and limited land availability: 62% of households in the region have less than 0.7 ha of arable land and 85% have less than 1.4 ha (MAGA 2011). Faced with limited land availability, farm households in the WHG have had to find alternative sources of food and income. These include daily agricultural labor on nearby farms, handcraft production (notably weaving), construction work, national and international migration (either seasonal or permanent) and, for the more fortunate, small businesses and waged jobs. Despite these multiple sources of income, agriculture is still the most important livelihood strategy for a majority of the population in the WHG and the backbone of local food security (IFAD, 2011).

Although the Western Highlands is recognized for being a diverse environment in which different historical processes have shaped distinct survival strategies among smallholders, the nature of that diversity is understudied (Smith 1989). Accounting for differences among smallholder farmers is key to addressing problems such as food insecurity. FAO (2016) recognizes that differences among regions and among social identities (like gender) affect development efforts to meet the challenges related to hunger listed in the Millennium Development Goals.

In the case of Guatemala food insecurity, only 19% of inhabitants are food secure (Instituto Nacional de Estadítica 2012) and landless small-scale farmers and subsistence farmers are recognized as among the most insecure (Taylor et al. 2006). Food insecurity in the country is caused by problems of food access that are rooted in structural problems such as inequality (Gobierno de Guatemala 2005; Guardiola et al. 2006). Inequality in Guatemala is characterized by a stratified society based on ethnicity, land ownership, health, education, gender, and age (Bruni et al. 2009). Small-scale farmers from the Western Highlands with indigenous origins, as well as women, are consistently identified as those who suffer most from the country's high levels of inequality. They are also the ones who have become the most vulnerable since the 1990s´ maize crisis (van Etten & Fuentes, 2004). Interventions aiming to

improve their food security need to take into account how these social differences interact with diverse maize production-consumption strategies.

1.2 CONTEXT

Even dough agriculture in this region is an inherent socio-ecological complex system, a call for, quick alternative methods of food security assessment, might be attractive for targeted policy interventions. And even more this complex system might follow a consistent pattern found in similar situations across the world (Frelat et al. 2016, Rietzema et al. 2017).

Process-based models are powerful tools for prediction, for example for yield estimation of crops at the field scale, as they simulate several interactions between the crops and the environment, nevertheless this models require intensive data collection and calibration of each of the process that is modelled. On the other hand, statistical modeling estimates direct relationships between predictor variables, without considering the underlying processes. Statistical models can provide simple but reasonable predictions, provided that sufficient and reliable data were used for model training and that the predictions are made within the boundaries of training data. Statistical models are less dependent on calibration data and may provide commonly used performance assessment measures useful for uncertainty analyses (Jeong et al. 2016).

Machine learning has progressed dramatically, from laboratory curiosity to a practical technology in widespread commercial use. Within artificial intelligence (AI), machine learning has emerged as the method of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications. Many developers of AI systems now recognize that, for many applications, it can be far easier to train a system by showing it examples of desired input-output behavior than to program it manually by anticipating the desired response for all possible inputs. The effect of machine learning has also been felt broadly across computer science and across a range of industries concerned with data-intensive issues, such as consumer services, the diagnosis of faults in complex systems, and the control of logistics chains. There has been a similarly broad range of effects across empirical sciences, from biology to cosmology to social science, as

machine-learning methods have been developed to analyze high-throughput experimental data in novel ways (Jordan and Mitchel, 2015).

Machine learning algorithms are sometimes referred to as black boxes as data goes in, decisions comes out but the processes between input and output are obscure, the inputs often suffer complex transformations (The, L. R. M., editorial 2018). An extensive literature attest the superiority of machine learning in minimizing predictive error, accuracy requires more complex precition methods, simple interpretable functions do not make the most accurate predictions. With machine learning although we gain in predictive power but loos in interpretability of the complex interactions (Goldstein et al. 2015).

1.3 HYPOTESIS

We hypothesize that based on a simply small set of predictor variables, the food security status of the smallholders' farm households of the wester highlands of Guatemala can be predicted constructing models with machine learning

1.4 OBJECTIVES

We attempt i) to estimate the contribution of agriculture to the food security measured in kcal/day/ male adult equivalents, of a large sample of smallholders farm households in the WHG based on a simple model of potential food availability which captures the dynamics of inputs and outputs of energy in the household through consumption or commercialization of farms products. ii) Develop, test and compare a series of models to predict the potential food availability status of WHG farm households based on a small set of predictors, and iii) to explore the variables interactions and thresholds that constrain or promote potential food availability.

Chapter 2: Methodology

2.1 METHODOLOGY

2.1.1 Study site and data

We use data from the baseline survey "*Encuesta de Monitoreo y Evaluación del Programa del Altiplano Occidental* (EMEPAO)", conducted by USAID and Guatemala's Ministry of Health and Social Assistance (MSPAS) between July and November 2013 as part of USAID's Feed the Future Program (Angeles et al. 2014). The survey targeted 6,301 households in 55 municipalities that belong to the departments of Totonicapán, San Marcos, Quiche, Quetzaltenango and Huehuetenango in the WHG (Fig.1).

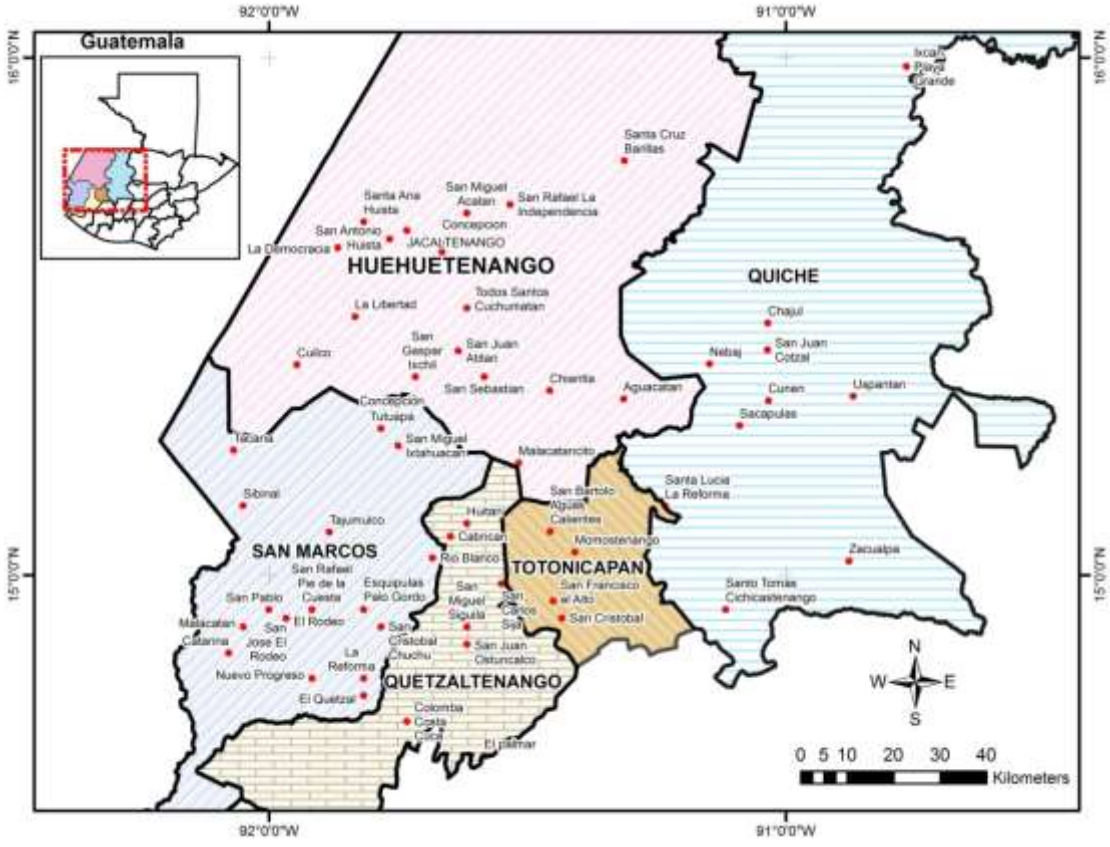

Figure 1. Municipalities in the WHG surveyed in the "Encuesta de Monitoreo y Evaluación del Programa del Altiplano Occidental (EMEPAO)"

The EMEPAO survey included information on household characteristics, their assets and living conditions, their participation in cooperatives or farmers associations, as well as the technical support received. The survey also included expenditure incurred

by the household for all kinds of goods and services as well as detailed information on cropping and livestock activities. The sample included participating and non-participating households in the USAID rural value chains program, and was randomly selected within the delimited national census sectors. See Angeles et al. (2014) for more details on the sampling process, survey tool and application. After removing households with no agricultural activity or with inconsistent data (e.g. more land on maize than total land available), data from 4,790 households were included in the analysis.

2.1.2 Potential Food Availability Indicator

Food security is defined as everyone having continued access to a sufficient quantity and quality of food (FAO, 2003). Four dimensions of food security have been defined: availability, access, utilization and stability (FAO, 1996). Here we focus on the food availability dimension by applying the model of Potential Food Availability (PFA) at the household level from Frelat et al. (2016) (Figure 2) that quantifies the potential food availability as an index calculated on the basis of daily kilocalories per individual. Farm products are converted to kilocalories either if households consume these products directly or sell them. In the case of sold products, we convert the income into potential staple food that farmers can theoretically buy and, in turn, we convert this into kilocalories. The kilocalories potentially bought and consumed are summed up and, hence, correspond to total potential food availability per year. The household's energy requirements per year are calculated based on the composition of the household's members (Frelat et al. 2016) (Figure 2).

Detailed testing of the food availability indicator (e.g. Frelat et al., 2016; Hammond et al. 2017) showed that it is well related to other indicators of food security (e.g. the Household Food Insecurity Access Scale, number of months with hunger and the Household level Dietary Diversity Score) when agricultural production and off-farm income are constraining food security. In more intensive agricultural systems, the correlation is less robust and the indicator does not function well. Further validation of this simple indicator is presented here by correlating the

output of the PFA indicator with binary variables of the sample related to food, which are described in detail after the description of the model.

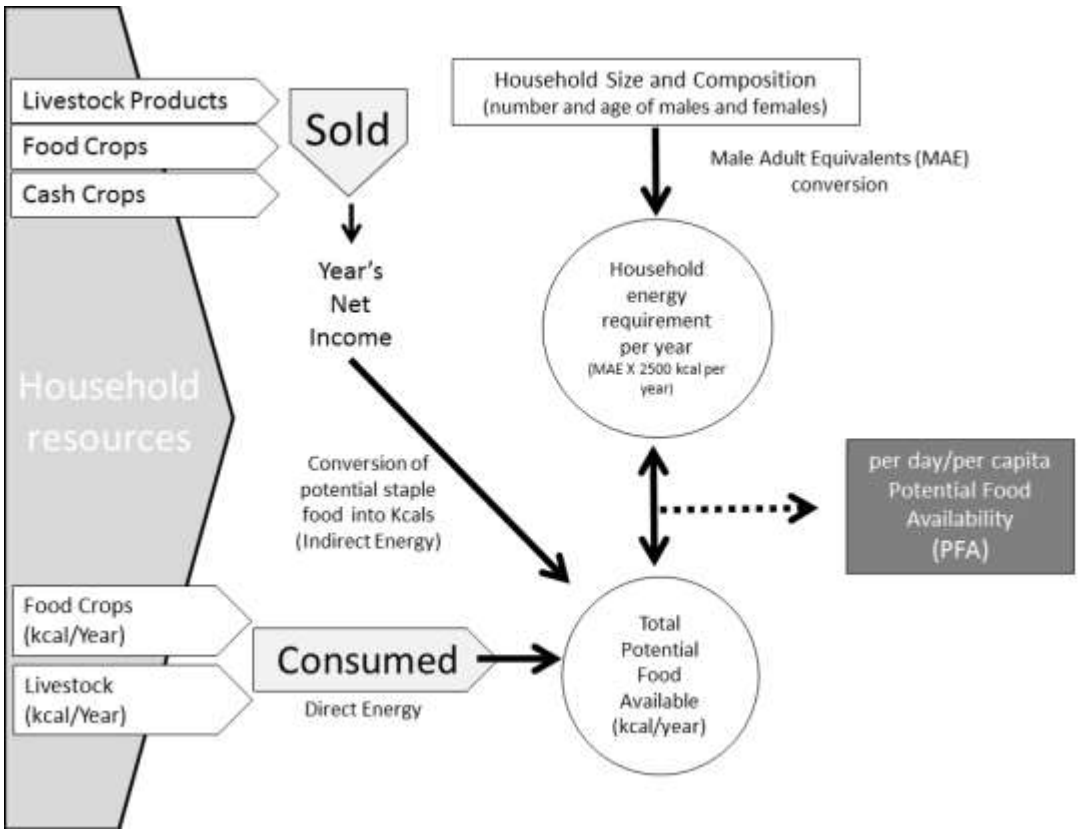

Figure 2. Model representation of household food availability expressed as kcal. Energy is derived from direct on-farm products that households consume and the transformation of the products' sale income into potential staple food purchase and its conversion into calories. The availability is based on male adult equivalents (MAE).The PFA, expresses that when it is ≥ 2500 kcal/day/MAE, the household has enough or more Potential Food Availability than needed per day per family member and is in consequence food secure. If the ratio is <2500 kcal/day/MAE, the household energy requirements is larger than potential food availability and thus the household is food insecure. Based on Frelat et al. (2016).

The quantification of the indicator is based on the household members' potential acquisition of kilocalories on a yearly basis by direct consumption of on-farm products and indirectly by the conversion of cash from products sales into staple food. The direct Potential Food Availability (PFA_{direct}) acquisition (i.e. consumed or produced) by households in kilocalories (per year) is determined by crop and livestock derived energy:

$$PFA_{direct} = \sum_{fC} Y_{fc} \times E_{fc} \times C_{fc} + \sum_{l} Y_{l} \times E_{l} \times C_{l}$$

Where: fc denotes a certain food crop, and l a certain livestock species, Y represent the production or number of animals or quantity of products (eggs, milk, meat, etc.). For fc or l, E denotes the energy (kcal kg^{-1}) content of each animal or food crop product, and C denotes the proportion of Y consumed by household. Energetic coefficients for crop and livestock products were determined from USDA (2015) values, while the production and consumption proportions are based on survey information.

Indirect energy acquisition by households ($PFA_{indirect}$) is a function of the potential quantity of staple food that can be bought by means of income derived from selling farm produce, expressed in calories. The Cash by Income in Quetzales per year, (CI_Q) the national currency, derived from farm produce total sales per year, is defined as:

$$CI_Q = \sum_{fC} Y_{fc} \times P_{fc} \times (1 - C_{fc}) + \sum_{l} Y_l \times P_l \times (1 - C_l)$$

Where P_{fc} and P_l denote the median market price per kilogram, reported by survey respondents, of a certain crop or animal product respectively, and where the term $(1 - C)$ denotes the proportion of farm product not consumed directly either from crops or livestock. Cash crops are defined as those where farmers sell more than 90% of the total annual production. With the available cash farmers buy staple food, which in this region is maize, at the median value of the market prices reported in the survey (S_{price}). Using the metabolic energy content of one kg of maize (E_{maize}) $PFA_{indirect}$ is then calculated as:

$$PFA_{indirect} = \frac{CI_Q}{S_{price}} \times E_{maize}$$

The total amount of food potentially available PFA_{total} for households is then defined as:

$$PFA_{total} = PFA_{direct} + PFA_{indirect}$$

On the energy needs side we use household composition, gender and age to calculate total household energy requirement (E_{hhr}). According to FAO (2001), a male adult has a daily need of 2500 kcal to sustain a fine nutrition with an average daily activity. Humans have a different energy requirement depending on age and gender. Hence, we took each household and assigned each member a male adult equivalent (MAE) according to the following ranges, new born: 0.29, children 1-3 years: 0.51, children 4-6 years: 0.71, children 7-10 years: 0.78, males 11-14 years: 0.98, males 15-18 years: 1.18, males 19-50 years: 1.14, males > 51 years: 0.9, females 11-50 years: 0.86, females > 51 years: 0.75. Daily E_{hhr} is calculated as:

$$E_{hhr} = 2500 \sum_{i}^{n} MAE_i$$

where *n* denotes the total number of individuals in household.

Considering the above calculations, we can define the Potential Food Availability (PFA) as:

$$PFA = \frac{PFA_{total}}{E_{hhr} \times 365\ _{days/year}}$$

which quantifies how much energy per capita per day is potentially available.

For validation of the results from the PFA indicator, we first made classes of the farm household in relation to their PFA. The classes were based on distribution of PFA in quartiles from the overall sample. Then, we used two questions in the survey to examine the validation of the PFA model. Question 1 - “In the last month, was there at some point no food because of lack of resources?” 29.2 % of those interviewed responded ‘yes’. The second question was “In the last month, have you gone to sleep without eating because of lack of food?” to which 10.5 % of the responded said ‘yes’. A chi-squared test was performed to see if the distribution of affirmative and negative responses was correlated with the PFA results.

2.2 POTENTIAL FOOD AVAILABILITY PREDICTION

Several methods were compared to find the best fit for PFA prediction, assuming that data may present non-linear relationships. Besides a multiple linear regression (MLR), we compare the machine learning classifiers Random Forest (RF) (Breiman, 2001) and Artificial Neural-Network (ANN) (Rosenblatt, 1958; Rumelhart et al., 1986), for predicting PFA.

Data selection for PFA prediction

Our target is to predict how agriculture contributes to food availability for smallholders in Guatemala, also, Maize cultivation is an essential part of the diet in the region but some other crops are cultivated like coffee. As many factors are related in food production in the WHG, the selected variables we chose for food availability prediction give a systems picture of the management and availability of land among smallholders, and also represent, a set of easy to capture or measure variables as very few time would take to ask for the data to a farmer. We selected as predictor variables of PFA, those related to land availability, crops diversity, crops yield and land allocation to crops. Specifically we initially had 4970 households with their predicted PFA and 14 variables: total land, land availability per person, total number of crops grown, land allocated to maize, land allocated to coffee, land allocated to potato, land allocated to other crops, tropical livestock units (TLU index) and the yield of each of the following crops: maize, coffee, bean, faba, pea and potato. Land, TLU and family size have been previously identified in other studies as important variables of household-level food security (Herrero et al. 2010; Ehui et al. 1998 and Frelat et al. 2016).

Exploratory data analysis of predictors vs target variable

All analysis and algorithms fitting were carried on R environment (R version 3.4.1, R core Team, 2017) with the IDE Rstudio (Version 1.1.423) using several packages as explained in the following table 1.

Table 1. R packages used for analysis and model construction

Package	Version	Used for:	Author
psych	1.8.4	exploratory descriptive analysis, scatter plots panels	Revelle 2018
corrplot	0.84	visualization of correlation plots	Taiyun and Simko 2017
randomForest	4.6-14	random forest algorithm for regression and classification	Liaw and Wiener,2002
caret	6.0-80	for training and parameter tunning of machine learning algorithms	Kuhn et al. 2018
pdp	0.7.0	for visualization of partial dependence plots of machine learning algorithms output, 3D visualizations and response surfaces praphs	Greenwell 2017
nnet	7.3-12	artificial neural networks fitting for regression and classification	Venables and Ripley 2002
NeuralNetTools	1.5.2	neural networks architechture visualization	Beck 2018

First data was screened for the detection of variables with zero or near-zero variance. Land on potato and land on other crops, faba, pea and potato yield proved to possess near-zero variance and therefore were excluded at this point.

After this, data was subjected to outliers detection for which a cut-off for registers elimination was selected of above quantile 99% and below quantile 1%, a total of 4499 FHH were left after outliers were detected.

We then followed a pairwise Pearson's product-moment correlation analysis to detect collinearity between the predictors, as well as their correlation with the target

variable. Predictors that were highly correlated ($r \geq 0.8$) in regression analysis were excluded from modelling. Total land and land availability per person (HaPerPerson) presented a correlation of 0.82 and therefore total land was excluded of the analysis (Fig. 3). We also estimated the descriptive statistics of the target variable (PFA) and predictors (Table 2) and present scatterplots for exploring the behavior of predictors especially with the target variable (Fig. 4).

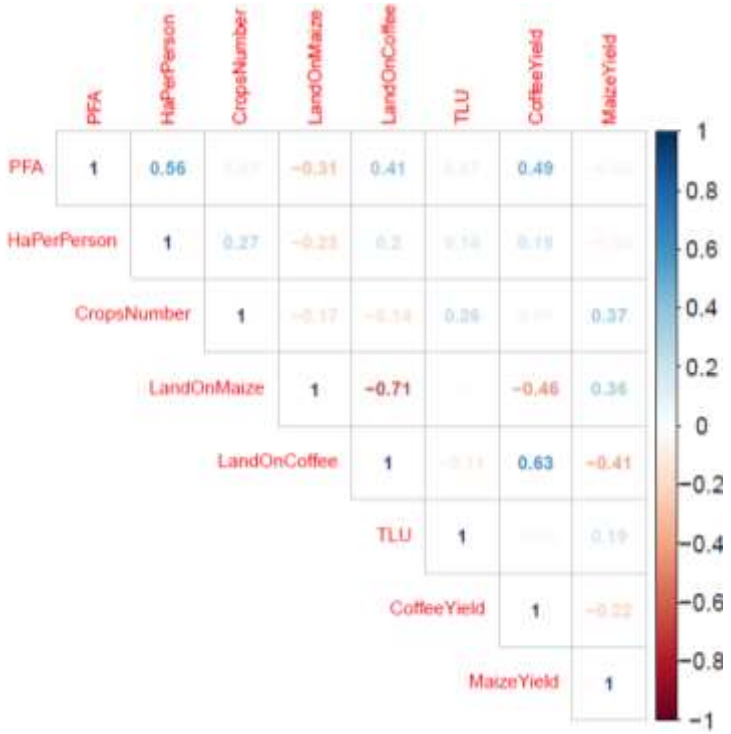

Figure 3. Correlation between variables included in the analysis

Table 2. Descriptive statistics of the target variable (PFA) and predictors

Variables	Achronym	Units	mean	sd	median	min	max	range	se
Potential Food Availability	PFA	kcal/person/day	4761	7368	2324	0	69135	69135	110
Land availability per person	HaPerPerson	ha/person	0.127	0.131	0.085	0.004	0.977	0.972	0.002
Land allocated to maize cultivation	LandOnMaize	percent	53	35	50	0	100	100	1
Total Number of Crops Cultivated	CropsNumber	number	2.5	1.4	2.0	1.0	6.0	5.0	0.0
Land allocated to coffee cultivation	LandOnCoffee	percent	23	35	0	0	100	100	1
Tropical Livestock Units	TLU	Index	0.47	0.66	0.20	0.00	3.79	3.79	0.01
Maize Yield	MaizeYield	ton/ha	2.25	2.01	1.98	0.00	11.69	11.69	0.03
Coffee yield	CoffeeYield	ton/ha	0.37	0.62	0.00	0.00	4.22	4.22	0.01

Data Preprocessing

Data was only preprocessed for ANN, as the activation function used was logistic; data was transformed to interval [0, 1] with the following calculation

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

where: $x_i = (x_1 \ldots . x_n)$, x, is the set of points of the variable to be transformed and z_i is the transformed variable.

As several models were built to compare its performance, data was split into training and testing sets in a proportion of 0.75 (n= 3375 FHH) and 0.25 (n=1124 FHH) respectively. Both data sets were used for training and testing all the models to make them comparable.

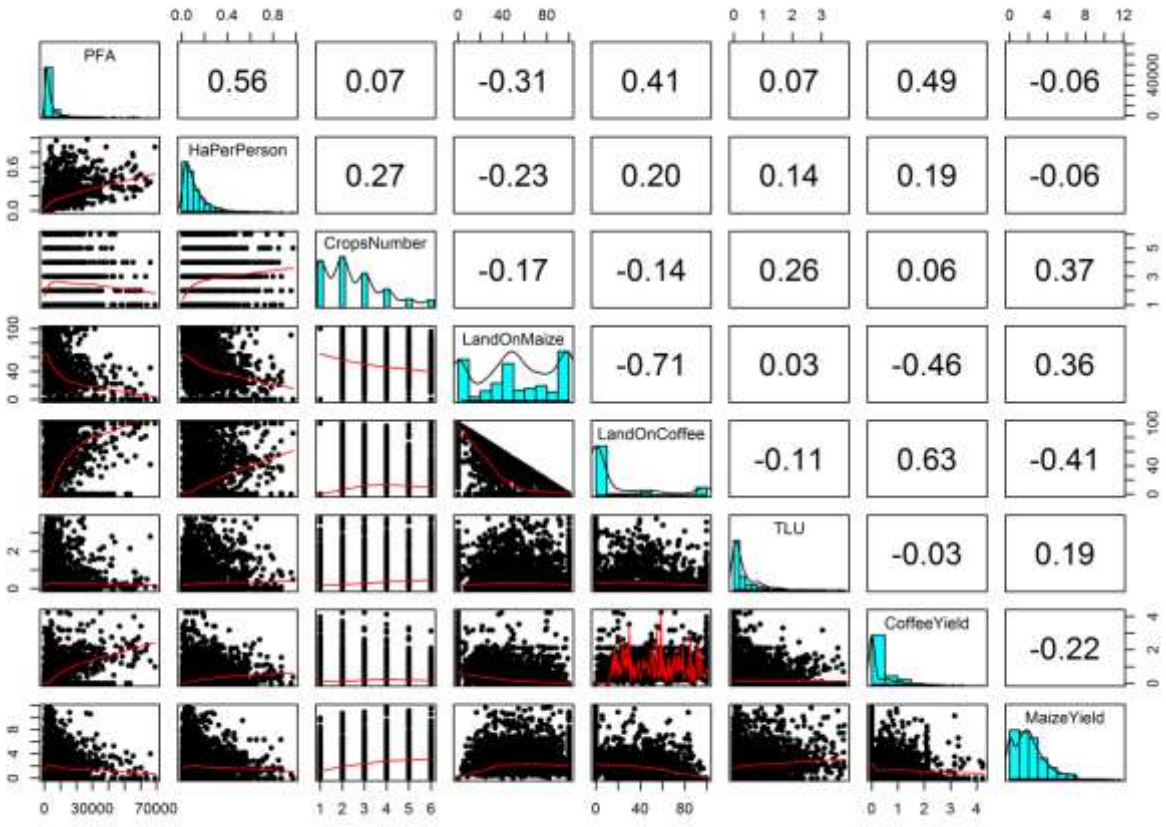

Figure 4. Scatterplots, histograms and correlation between target variable (PFA) and predictors

2.3 PREDICTION OF POTENTIAL FOOD AVAILABILITY AS A REGRESSION OR CLASSIFICATION PROBLEM (CONTINUOUS VS CATEGORICAL DATA).

We divided the prediction exercise into: i) continuous data which are the actual values of PFA, considered as a regression problem; and ii) A classification problem, a binary prediction of food secure/insecure FHH based on the threshold of 2500 kcal/day/MAE.

Models testing and comparison

The testing data was used to validate the MLR, RF, and ANN models, and derive statistical measures to compare their performance. The root mean squared error (RMSE) and R^2 of observed vs predicted values were computed from the differences between the predicted PFA values and those calculated with the food security model to determine the precision and bias of the predictions, respectively. RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}}$$

where n is the number of observation in the data set, and $X_{obs,i}$ are the values of the target variable from the testing or training data set and the $X_{model,I}$ are the values predicted by the model. Also scatterplots of predicted vs observed values were constructed to compare the models accuracy in the dispersion of points, linear model regression lineas and 1:1 lines were added to each plot for ease of interpretation.

.

2.4 THE TESTED ALGORITHMS, PREDICTION OF PFA AS CONTINUOUS VARIABLE

2.4.1 Linear model

Technical description

Adjusting a line to predict a relationship between two variables is called linear regression. As a predictive analysis, multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. Historically, linear regression in its earliest form as least squared method published by Legendre in 1805 and Gauss in 1809 (Yan and Su 2009), was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend

linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine. Linear regression models are often fitted using the least squares approach. Several drawbacks of using linear regression for model fitting are that it assumes linearity, *i.e.*, the response variable is a linear combination of the parameters and their predictor variables, homoscedasticity, multivariate normality, no multicollinearity, and no autocorrelation. We applied a generalized linear model to fit the PFA as response variable to the seven other predictors.

2.4.2 Random Forest

Technical description

We trained and applied RF, a binary tree based machine-learning method, to predict PFA. RF can be used for both classification and regression purposes, and the scope of our study is to use it as a regression tool. Briefly, to train RF models, many classification and regression trees (CARTs) are grown with a 'random' subset of predictors without pruning, and the 'forest' of CARTs is averaged. Source data for model training are bootstrapped to make various subsets to generate a large numbers of trees randomly. Predictor variables are evaluated by how much they decreased node impurity when they are selected for the splits or how often they make successful predictions in the forest of CARTs. Node impurity is defined as mean square error (MSE) of the node in RF regression. Briefly the random forest algorithm grows many classification trees, to classify an object from an input vector, the vector is put down each of the trees in the forest, each tree gives a classification (each tree votes for a specific class), then the forest chooses the classification having the moost votes overall the trees in the forest. Each tree is grown by first, sample N cases in the training set at random, but with replacement, from the original data. This sample will be the training set for growing the tree. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing. Each tree is grown to the largest extent possible. There is no pruning. Increasing the correlation between the trees increases forest error, single strong classifier trees increase the strength of the forest. See Breiman

(2001) for more details on the RF algorithm. For analysis we used the package *randomForest* in R .

Random forest algorithm in *randomForest* package allows parameter tuning for the number of trees and the number of randomly tried variables on each split. In RF modelling, the training parameters that we selected to tune were: i) the number of trees to grow in the forest (*ntree*), ii) the number of randomly selected predictor variables at each node (*mtry*).

A variable analysis tool is available from the same package and is used for analysis to extract variable importance. The mean decrease accuracy ("%IncMSE") is used as a measure of variable importance for regression models. The %IncMSE plot shows the mean increase of MSE in nodes that use a predictor in the model, when values of the predictor are randomly permuted. The most important variables are ranked on how much %IncMSE is increased if the variable is removed from the analysis. For classification models the % mean decrease accuracy (%MDA) is used as the importance measure.

2.4.3 Artificial Neural Networks

Technical description

Artificial neural network algorithm simulates human learning processes through establishment and reinforcement of linkages between the input and output data. The linkages then connect input and output data in the absence of training data (Campbell

2002). Numerous ANN algorithms have been proposed, such as Radial Basis Function (Vojislav 2001), Elman recurrent (Rakkiyappan and Balasubramaniam, 2008), and Hopfield neural networks (Nguyen et al., 2006); however, Multi-layer perceptron neural networks (MLP Neural Nets) with back-propagation algorithm may be the most popular, and was selected for this study (Haykin, 1994).

Because ANN models allow an illustration of complex and non-linear relationships without rigorous assumptions regarding the distribution of samples (Bishop 1995; Breiman et al. 1984), the method is gaining popularity for research areas where there is little or incomplete understanding of the problem to be solved, but where training data are available. The ANN structure is based on the human brain's biological neural processes. Interrelationships of correlated variables that symbolically represent the interconnected processing neurons or nodes of the human brain are used to develop models (see figure 5). ANN models find relationships by observing a large number of input and output examples to develop a formula that can be used for predictions (Pachepsky et al. 1996). Nonlinear relationships overlooked by other methods can be determined with little a priori knowledge of the functional relationship (Elizondo et al. 1994). A minimum of three layers is required in an ANN model: the input, hidden and output layers (Lee and Evangelista 2006; Conforti et al. 2014).

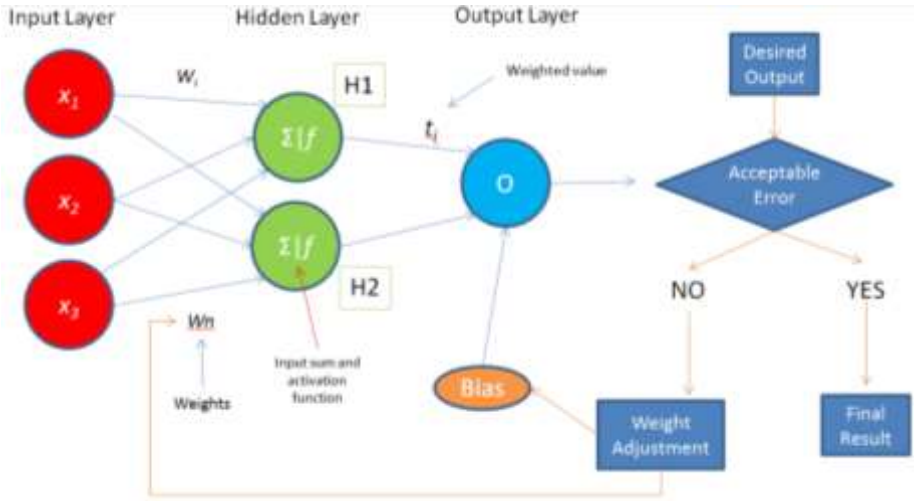

Figure 5. Feed forward, multilayer artificial neural network and back-propagation algorithm.

The input and output layers contain nodes that correspond to input and output variables, respectively. In our case, the output layer is the PFA values and the Input layer is the set of *n* predictor variables Data move between layers across weighted connections. A node accepts data from the previous layer and calculates a weighted sum of all its inputs:

$$t_i = \sum_{j=1}^{n} w_{ij} x_j$$

where n is the number of inputs, w is the weight of the connection between node i and j, and x is the input from node j. A transfer function is then applied to the weighted value, t, to calculate the node output,

$$o_i = f(t_i)$$

The most commonly used transfer function is a sigmoidal function for the hidden and output layers and a linear transfer function is commonly used for the input layer. The number of hidden nodes determines the number of connections between inputs and outputs and may vary depending on the specific problem under study. If too many nodes are used then the ANN may become over-trained, causing it to memorize the training data and resulting in poor predictions (Lawrence 1994). The learning rate determines the amount the weights change during a series of iterations to bring the predicted value within an acceptable range of the observed value. The training tolerance refers to the maximum error rate at which the network must converge during training. Once the network converges, an approximate function is developed and utilized for future predictions (Schmueli 1998). The trained network is then tested with a separate data set with its output information omitted.

We used the package *nnet* combined with caret, which allows training ANNs and tune some parameters to find the most accurate model.

2.5 PREDICTION OF PFA AS A CATEGORICAL VARIABLE

We constructed models for predicting the categorical variables, either for the PFA categories or the binary food security of the FHH. We ran some tests with random forests and ANNs algorithms which also can be used for classification tasks. The models performance was compared by the accuracy of the predictions with the training as well as with the test data. Also confusion matrices were built to help visualizing the accuracy of predictions, with thee associated error by category.

Cohen's kappa coefficient (κ) is a statistic which measures inter-rater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as κ takes into account the possibility of the agreement occurring by chance (Cohen 1960). The values of k were used also for selecting the most accurate model.

2.6 SENSITIVITY ANALYSIS: PARTIAL DEPENDENCY PLOTS (PDPs) AND RESPONSE SURFACES WITH CONTOUR LINES

Complex non-parametric models like neural networks, random forests, and support vector machines are more common than ever in predictive analytics, especially when dealing with large observational databases that don't adhere to the strict assumptions imposed by traditional statistical techniques (e.g., multiple linear regressions which assume linearity, homoscedasticity, and normality). Unfortunately, it can be challenging to understand the results of such models and explain them to management. Partial dependence plots (PDPs) offer a simple solution. Partial dependence plots are low-dimensional graphical renderings of the prediction function so that the relationship between the outcome and predictors of interest can be more easily understood. These plots are especially useful in explaining the output from black-box models (Greenwell 2017). Contour lines in bi-dimensional plots allow the representation of limits in which a third variable has a particular value and is allowed to change in response to the other two variables. Traditionally, for classification problems, partial dependence functions are on a scale similar to the logit,

PDPs for single predictor variables vs target variable (PFA) were constructed to understand the individual variables effect on overall PFA when all the rest of the variables are left constant at its median value.

Three-dimensional response surface curves were plotted to understand the interaction of the variables for the continuous prediction and to determine the optimum level of variables combination for maximum response.

Also PDPs were constructed to understand the effect of single variables on the probability of belonging to one of the discrete categories of PFA or to the binary food-secure/food-insecure status, also two-dimensional response surfaces were plotted to find the interactions defining the probability of membership to each category.

Chapter 3: Results

3.1 FOOD SECURITY INDICATOR

The contribution of agriculture to the Potential Food availability (PFA) for the farm households in the WHG varies from almost no contribution to more than ten times the kcal needs for the family (Figure 6-A). For more than half of the households (52 %), agricultural production does not meet the kcal needs of the family and therefore farmers need other sources of food/income (e.g. off-farm income, remittances from family members working in the United States). The contribution of energy by consumption of farm produced food crops is relatively low, but is of significant importance in households with low PFA. The absolute contribution from consumption of farm produced food crops increases as PFA increases, but only up to a certain point and for households with higher levels of PFA, it decreases again (Figure 6-A).

Full food self-sufficiency is never reached from own produced crops, with market orientation taking off when farm households are able to produce between 50 and 70% of their food needs. This section of maximum PFA is characterized by households in which sales of cash crops contribute most to PFA, while also sales food crops in general increasing with increasing PFA. Livestock do contribute to PFA of households with low PFA scores, although the contribution is not major, and thereby play a complementary role to the consumption of food crops. As the PFA value of households gets larger, energy coming from livestock consumption or livestock sales becomes larger as well. However, for households with the higher PFA values cash crops are more important and the role of livestock gets quite small (Figure 6-A and 6-B).

We defined four food security classes by partitioning the 4790 households in quartiles: i) the NEFA (Not Enough Food Available) encompasses those households with PFA below 1090 kcal/capita/day; ii) the REFA class (Roughly Enough Food Available) with those that PFA value falls between 1090-2390 kcal/capita/day; iii) the SFA class (Sufficient Food Available) between 23905240 kcal/capita/day and finally iv) the MEFA class (More than Enough Food Available) >5240

kcal/capita/day. For both the NEFA and REFA classes, the average PFA scores do not meet the daily kcal requirements of the household with agricultural activities. For the first three quartiles, the most important contributor to PFA indicator are the consumed own grown crops, with increasing importance of cash crop and decreasing importance of livestock when moving up in the quartiles (Figure 6-B).

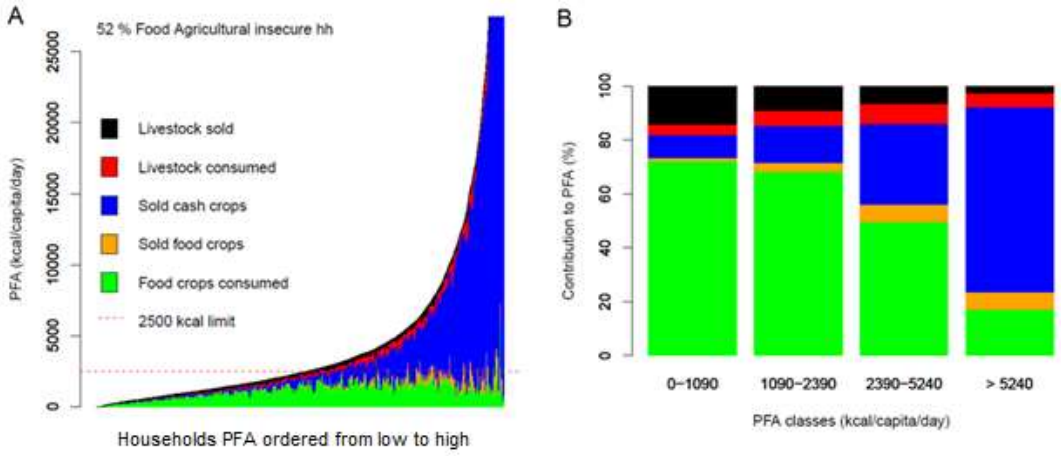

Figure 6. A) Potential food availability among 4790 households from the EMEPAO-WHIP survey. Farm households are ordered from low to high PFA values. The red dashed line indicates the 2500 kcal//day/MAE (PFA=1). Barstacked are the sources contributing to PFA. Colors indicate potential sources of energy for households. A moving average was applied with a window length of five households for ease of interpretation. B) Contribution of different sources of energy for the different quartiles of the PFA indicator

To test the PFA indicator we compared the PFA with the results obtained from two binary questions asked in the survey related to lack of food available. The chi-square test showed a significant positive correlation between the questions related to food limitations with the lowest quartile class of the PFA indicator, as well as negative correlation with the highest quartile (Figure 7).

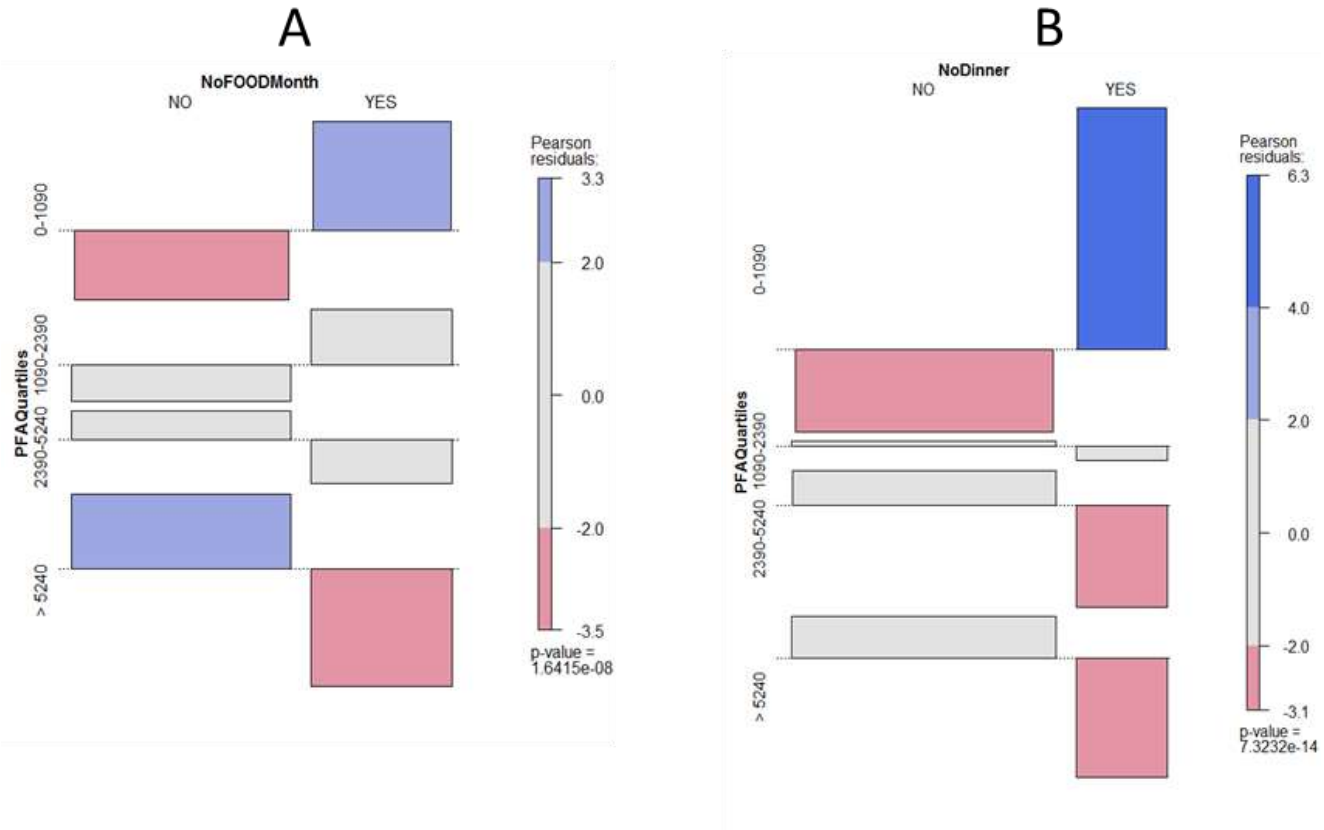

Figure 7*. Pearson residuals visualization after chi-squared test of PFA classes vs food insecurity survey binary questions. A) Perception of lack of food in the last month, B) Household members going to sleep without eating in the last month. Blue rectangles show positive significant associations, and red rectangles show negative significant associations, gray rectangles show non-significant associations.*

Maize is the most important food crop in the WHG. Farm households in the first two PFA classes, given their large food deficiency, will certainly need to use other livelihood activities and income sources that were not captured in the survey, to buy maize. For these two classes, the average annual household deficit to reach the 2500 kcal/day/MAE threshold is 898 kg maize per year (which is equivalent to 355 USD/year) for the NEFA class, 367 kg maize per year (equivalent to 145 USD/year) for the REFA class (1 quetzal=0.13 USD)(See supplemental material 1).

3.2 POTENTIAL FOOD AVAILABILITY PREDICTION MODELS

Tuning of models

Randomforest

The first step in random forest tuning was to assess the best *mtry*, this value is searched based on the out of bag error (OOB) of each *mtry* measured as the mean of the squared residuals (MSE). The OBB is a measure of the error of the model performance based on the data not considered in training on each subsample of the bootstrap conducted on each random forest run. The results of this step are shown in figure 8 and table 3, the best value for *mtry* =4.

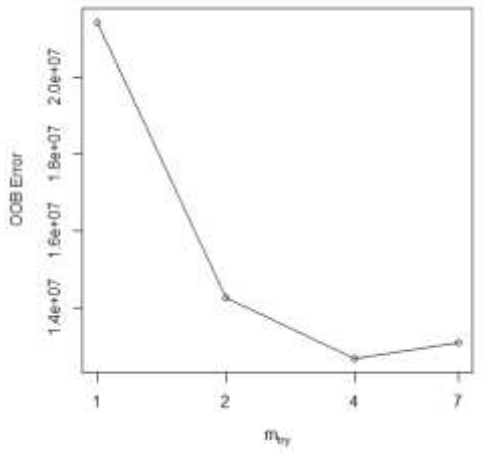

Figure 8. Results for the estimation of the mtry parameter for randomforest tuning.

Table 3. The number of variables to test at each node split in randomforest algorithm (mtry) and their accuracy in kcal

mtry	OOBError (MSE)	sqrtOOBError (RMSE)
1	21445014	4630
2	14275190	3778
4	12678489	3560
7	13102484	3619

After finding the best *mtry*, different numbers of *ntree* parameter were tested. The best model was selected based on the highest R^2 and lowest RMSE resulting in 1500 trees (See table 4).

Table 4. MSE, RMSE and R2 for different ntree parameter values in the RF

ntree	MSE	RMSE	R^2
3000	**12631486**	**3554.08**	**0.765**
1500	**12584440**	**3547.45**	**0.766**
1000	**12611397**	**3551.25**	**0.766**

Artificial Neural Network

The *caret* and *nnet* packages combined allow tuning the weight decay during training and the number of nodes in the hidden layer. After trying several maximum numbers of iterations by trial and error, 1500 iterations allowed to convergence of all the training rounds and lasted not too much time on computing. With this parameter fixed we compared several combinations of decay and number of nodes, we selected the ANN configuration with the lowest RMSE and best R^2.

Table 5. Accuracy of the ANN after training with different combinations of weight decay and node size, model selection was based on the lowest RMSE which is presented in scaled form.

decay	node size	RMSE	Rsquared
0.001	2	0.056	0.720
0.01	2	0.057	0.713
0.1	2	0.067	0.612
0.5	2	0.082	0.470
0.001	3	0.053	0.746
0.01	3	0.053	0.750
0.1	3	0.066	0.615
0.5	3	0.082	0.465
0.001	4	0.053	0.748
0.01	4	0.052	0.759
0.1	4	0.066	0.614
0.5	4	0.082	0.465
0.001	5	0.052	0.759
0.01	5	0.051	0.768
0.1	5	0.066	0.614
0.5	5	0.082	0.467
0.001	6	0.052	0.759
0.01	6	0.051	0.770
0.1	6	0.066	0.613
0.5	6	0.081	0.473
0.001	7	0.052	0.761
0.01	7	0.051	0.771
0.1	7	0.066	0.611
0.5	7	0.080	0.478
0.001	8	0.052	0.763
0.01	8	0.051	0.771
0.1	8	0.067	0.610
0.5	8	0.080	0.479

The best ANN model selected after the tuning process is a combination of the following parameters, a decay of 0.01, 8 hidden nodes which resulted with a RMSE of 0.51 (scaled) and a R2 of 0.771 see Table 5 .The final ANN architecture (in terms of nodes and layers: (7-8-1) is s a 7 input variables with 8 hidden nodes, one bias node for the hidden layer and one bias for the output node (Fig. 9).

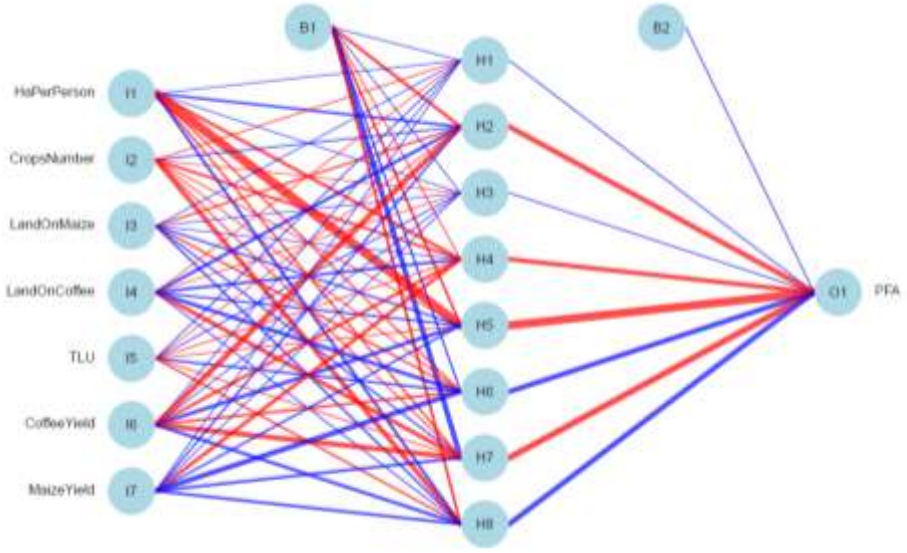

Figure 9. Artificial neural network architecture trained for predicting continuous PFA, H denotes hidden layer nodes, B1 and B2 the bias for each layer.

Comparison of Models Performance for the regression problem

The models performance showed that the ANN presents a RMSE of 3176 kcal/day/MAE. The performance of the RF algorithm is less satisfactory as the RMSE of the model is 3547 kcal/day/MAE, while the MLR model presented even more deviations with an RMSE of 9150 kcal/day/MAE. Thus the ANN is the best model. The comparison of performance based on the predictions of testing data of each one of the models is presented in the next table 6.

Table 6. Multiple linear regression, Random Forest ,and Artificial Neural Network model performance evaluation statistics

	MLR	RF	ANN
RMSE	9150	3547	3176
R2	0.512	0.767	0.836

The comparison between the Observed vs. Predicted values indicates that the ANN predictions correspond highly with the observations (R2=0.836), the RF is less accurate (R2=0.767) but still predicts a large variation of observed PFA, while the MLR shows the lowest accuracy in predicting the observed values (R2=0.512) Table 6 and figure 10.

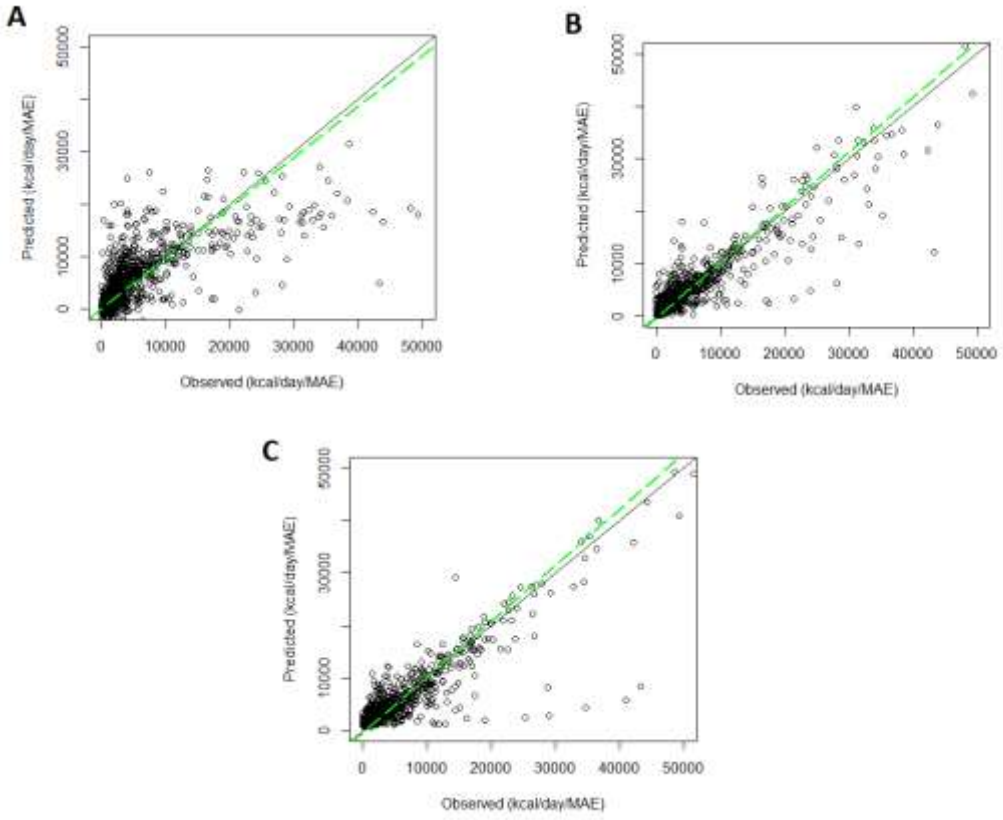

Figure 10. Multiple Linear Regression, Random Forest and Artificial Neural Network models performance for test datasets. Observed vs. Predicted plots are shown for each case. Black line represents the 1:1 relationship, while the green-dashed line represents the linear regression of Observed vs Predicted

Variables importance in PFA prediction.

After random forest which shows a high power to predict PFA, the most important variables are HaPerPerson (%IncMSE~250), CoffeeYield (%IncMSE~190) and LandOnCoffee (%IncMSE~100), less important variables are LandOnMaize, MaizeYield, CropsNumber and TLU (Figure 11). Considering removing those less important variables decreased the accuracy of the models to an R2 of the best ANN to 0.53

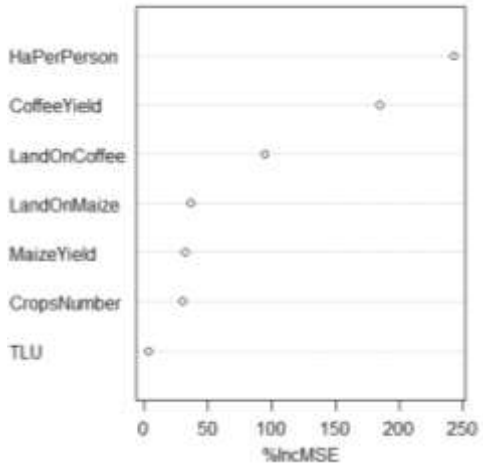

Figure 11. Variable importance for prediction of continuous PFA from random forest algorithm

3.3 POTENTIAL FOOD AVAILABILITY, THRESHOLDS AND CONSTRAINTS, THE REGRESSION APPROACH.

The seven variables selected for prediction of the continuous values of PFA with the trained ANN, ha per person, land allocated to coffee, land allocated to maize, yields of coffee and maize, TLU and number of crops, explain a substantial variation of PFA (R2=0.84) (Table 6) .

3.3.1 PDPs for regression

The response curves identified with the partial PDP plots for individual variables show some non-linear relationships (Fig. 12). Between PFA and ha per person (Fig 12-A), the relationships is somehow logistic with a tendency of saturation after 0.5 ha per person. This suggests that land productivity decreases in households with more land. Coffee and Maize Yield (Fig. 12-B and 12-E) also show a similar curve as ha per person but without a saturation curve (but with a more linear tendency), which is to be expected as, higher yields may allow to higher productivity of the land and in consequence higher PFA. The same happens with TLU (Fig. 12-F), but it shows a definitive linear relationship, increasing TLU increases PFA monotonically. In terms of the land allocated to coffee (Fig. 12-C), the relationship is somehow linear, and shows a slight change in the rate of increase of PFA at around a value between 25 and 50 % of land on coffee with a slight tendency to exponential growth, while the effect of the proportion of land allocated to maize on PFA (Fig. 12-G), shows an inverse relationship. The relationship of PFA with the number of crops (Fig. 12-D) shows a non-linear decay as the number of crops increases. As it can be seen from the scale of the PDPs, the effects of each variable are different and may cause larger or smaller effects on the PFA. The variable ha per person shows the largest effect on the change of PFA, as it may affect PFA from values of 0 to 20,000 kcal/day/MAE, followed by coffee yield, with 2500-12500 kcal/day/MAE.

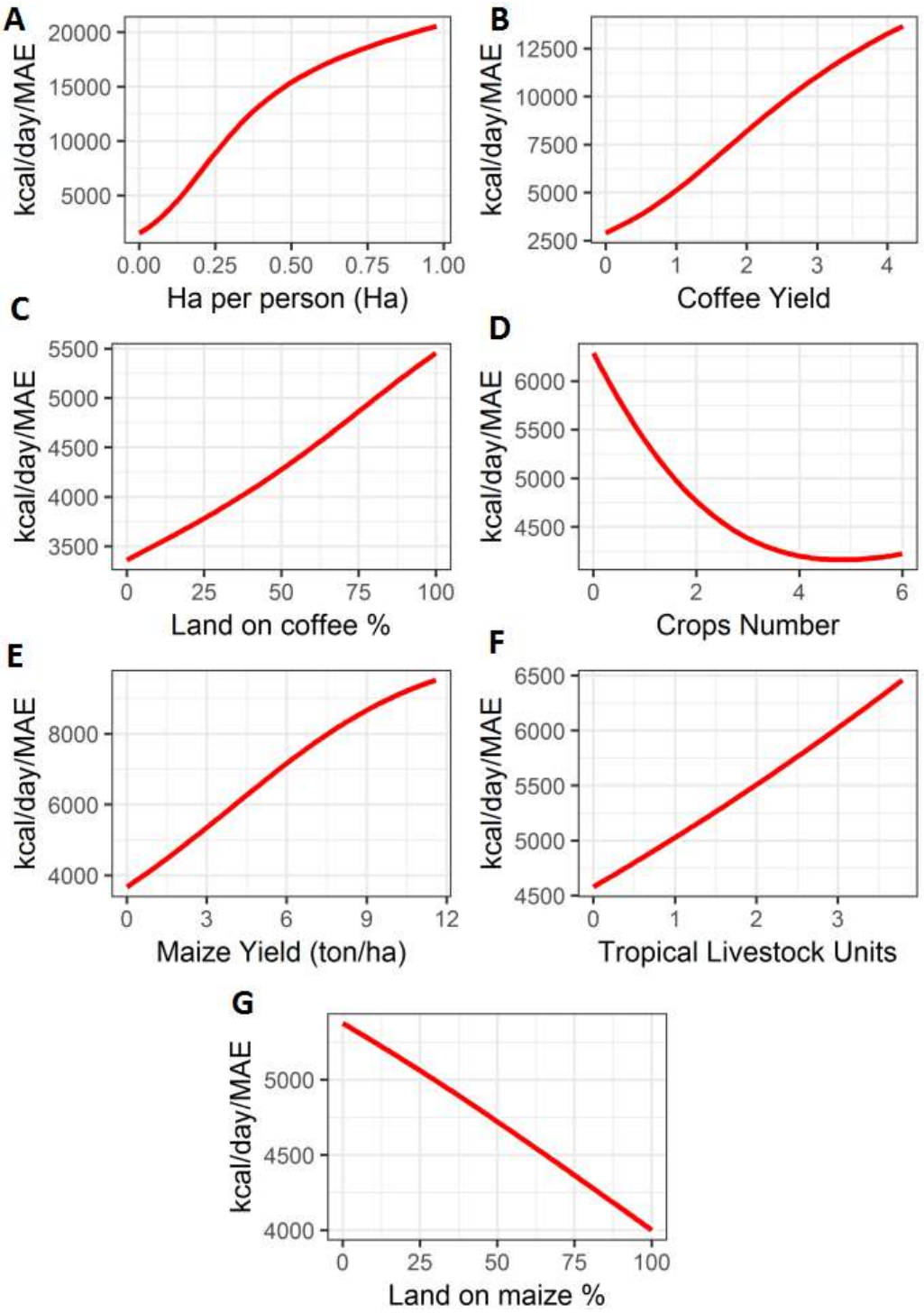
A
kcal/day/MAE
20000
15000
10000
5000
0.00
0.25
0.50
0.75
1.00
Ha per person (Ha)
B
kcal/day/MAE
12500
10000
7500
5000
2500
0
1
2
3
4
Coffee Yield
C
kcal/day/MAE
5500
5000
4500
4000
3500
0
25
50
75
100
Land on coffee %
D
kcal/day/MAE
6000
5500
5000
4500
0
2
4
6
Crops Number
E
kcal/day/MAE
8000
6000
4000
0
3
6
9
12
Maize Yield (ton/ha)
F
kcal/day/MAE
6500
6000
5500
5000
4500
0
1
2
3
Tropical Livestock Units
G
kcal/day/MAE
5000
4500
4000
0
25
50
75
100
Land on maize %

Figure 12. Partial dependence plots for the predictor variables and PFA, resulting from ANN

3.3.2 Partial dependency plots for variables interaction effects on continuous PFA

As ha per person has the strongest effect over PFA when inspecting the interaction of predictors, the thresholds and constraints found affecting PFA are discussed around ha per person.

With the aid of the response surfaces and two-dimensional PDPs of pairwise variables (Figure 13), the interaction of HaPerPerson and LandOnCoffee, show that no matter if land on coffee is large, if land availability is short, the PFA values are very low, only when lots of land is allocated to coffee and land availability is large the PFA reaches higher values (Fig. 13-B) .

When it is observed the interaction of HaPerPerson with LandOnMaize (Fig. 13-C), a similar effect is seen as high land on maize but low land availability produces low PFA. As land availability increases, no matter if there are low and high values of land on maize PFA increases, and it reaches slightly highest values when land on maize is above 40.

The interaction of crops number with ha per person, show that multiple crops and large land sizes produce low values of PFA, only having fewer crops with large land sizes may allow higher values of PFA, but the largest effect is seen with large land availability and monocrops (Figure 13-D). An interesting effect on PFA when looking at the interaction of maize yield with the number of crops, only when high yield are attained and there are fewer crops the PFA start to augment, no matter if there is a high yield of maize having lots of crops maintain PFA in its lowest values.

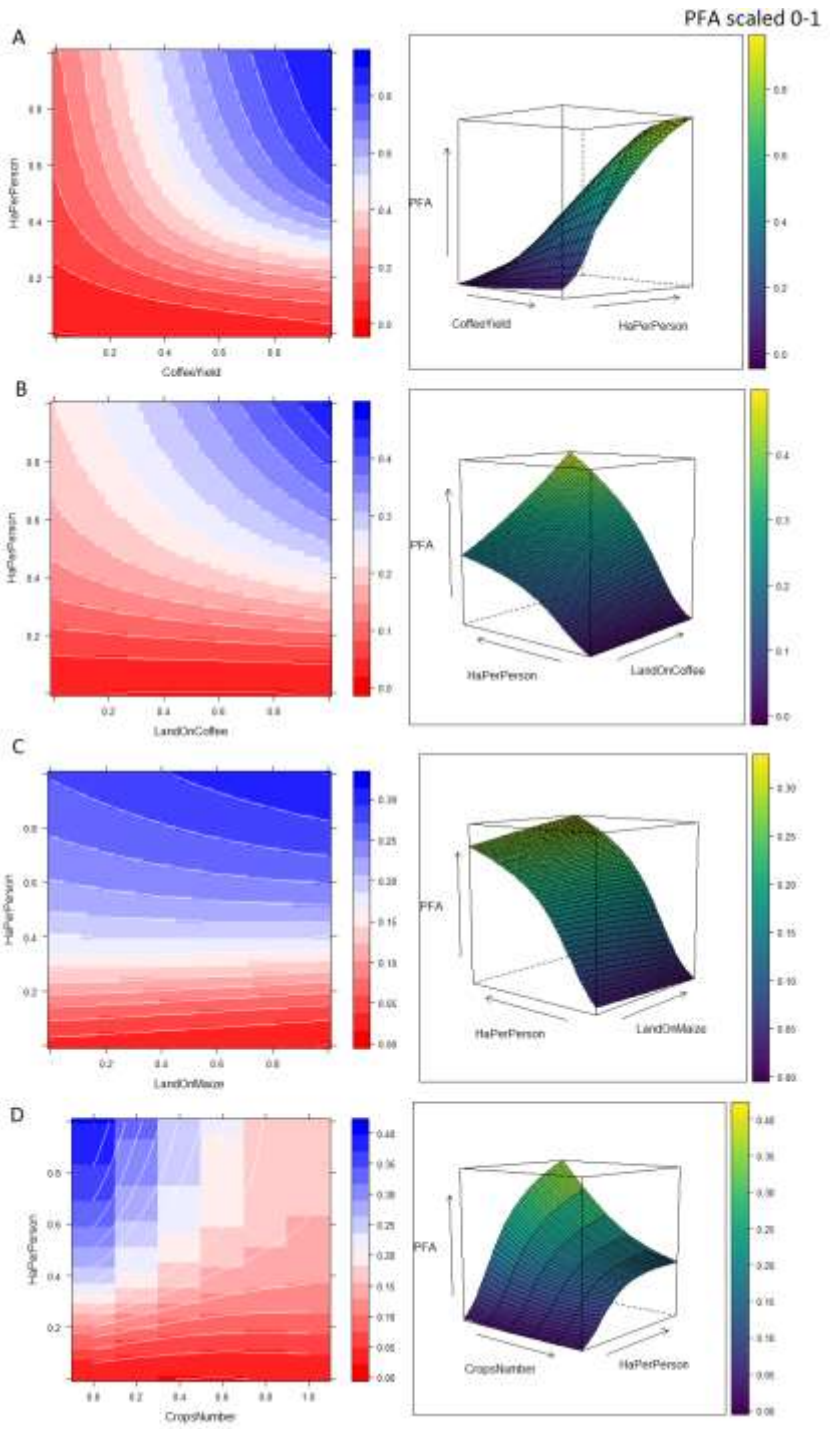

Figure 13. Tri-dimensional partial dependency plots and PFA, values are scaled 0-1. Response surfaces are also shown as 3-D representations. A) HaPerPerson vs CoffeeYield, B) HaPePerson vs LandOnCoffee, C) HaPerPerson vs LandOnMaize and D) HaPe Person vs CropsNumber

3.4 POTENTIAL FOOD AVAILABILITY THRESHOLDS AND CONSTRAINTS, THE CATEGORICAL APPROACH.

When we attempt to predict the food secure-food insecure status (can a household, yes or no, produce, and/or purchase enough food to feed the family?) based on the threshold of 2500 kcal/day/MAE, and with the exact same variables as in the prediction of continuous PFA, the ANN model predicts correctly 84.88 % (k=0.696) of the houeholds, performing best than the randomForest algorithm (84.07% predicted correctly, k=0.679) (see Table 7.) Both models struggle more when predicting food-secure households as they present more misclassified households (RF: 24.9 %, ANN: 18.4 %) than when predicting food-insecure households (RF: 12.3 %, ANN: 12.1 %). This might be due to the highly skewed nature of the PFA variable. We chose the ANN model as it predicts slightly better but also as the PDPs are less expensive in computation resources.

Table 7. Confusion matrix for results of RF and ANN in predicting a FHH is food secure (>2500=Yes)

n= 1124		Observed			
		RF		ANN	
		Food Insecure	Food Secure	Food Insecure	Food Secure
Predictied	Food Insecure	520	106	521	98
	Food Secure	73	425	72	433
Statistics	Accuracy %	84.07		84.88	
	Kappa (k)	0.679		0.696	

The best ANN model after training has 7 hidden nodes and a weight decay of 0.1 (see Table 8 and figure 14).

Table 8. Statistics of ANN training for predicting food security membership swapping different parameters

decay	size	Accuracy	Kappa	AccuracySD	KappaSD
0.01	5	0.856129	0.710569	0.009	0.0182
0.1	5	0.856681	0.711629	0.007	0.014
0.01	6	0.858214	0.714796	0.008	0.017
0.1	6	0.858415	0.715119	0.007	0.015
0.01	7	0.857995	0.714301	0.010	0.021
0.1	7	0.858902	0.716107	0.007	0.014
0.01	8	0.85812	0.714527	0.011	0.022
0.1	8	0.858415	0.71511	0.007	0.015

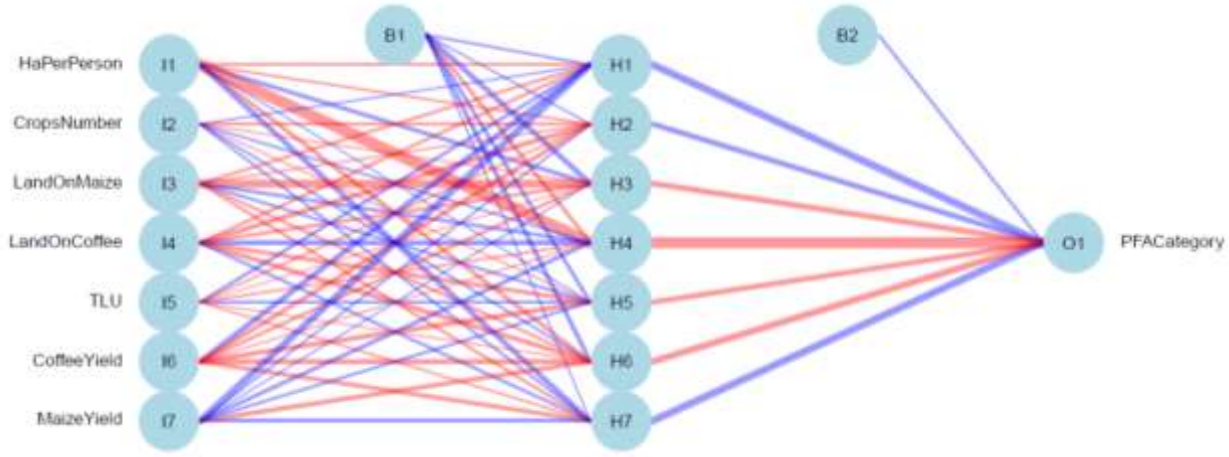

Figure 14. The final ANN model used for predicting FHH yes/no food security

Variables importance on predicting food-secure/food insecure status of households. The most important variable for classification into food-secure or food-insecure FHH is HaPerPerson (%MDA~400), followed by CoffeeYield (%MDA ~180), MaizeYield (%MDA~170) and LandOnCoffee (%MDA~160), less important variables are LandOnMaize, CropsNumber and TLU (Fig. 15.)

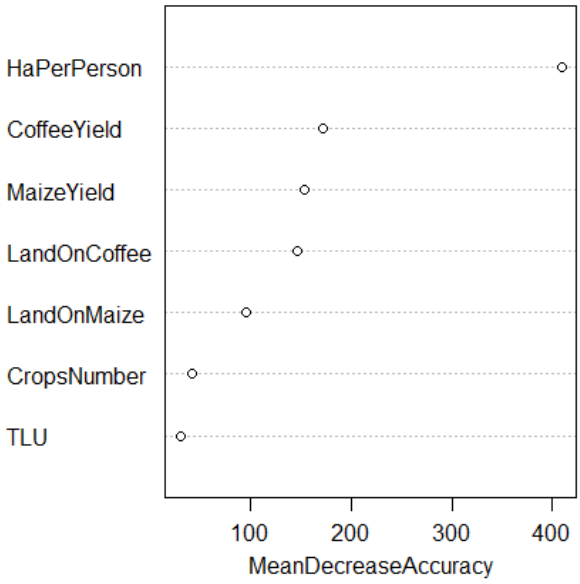

Figure 15. Variables importance after random forest algorithm applied for predicting the binary yes/no food-secure output

3.4.1 PDPs Response curves of probability of being food-secure

The response curves found through PDPs are presented based on the probability of being food-secure. For the case of the effect of single variables on the probability of being a food-secure household several non-linear relationships are found (Fig. 16). For the variable HaPerPerson the relationship with the probability of being food-secure is a high non-linear saturation curve, the increase in probability rises quickly with relative small amounts of land availability per person, a threshold is found around 0.25 ha per person in which almost the probability is 0.87, this confirms the pattern found with the continuous approach, land productivity decreases but also food security is achieved with the appropriate amount of land (Fig. 16-A). For land on coffee the relationship is also non-linear, the increase in probability augments gradually but with a first shift in the rate of change of the probability by increasing land on coffee is found around the value of 12.5% of LandOfCoffee, from this point onwards a logistic increase of the probability can be observed, then, at around between 50 and 75% land on coffee a threshold is also found in which the rate of change of the probability flattens and start to be negative close to 70% (Fig. 16-B).

For maize yield also can be observed a high non-linear relationship with probability, but less steep in comparison with ha per person and land on coffee, a threshold is found around the value of 3 ton/ha, in which the rate of change of probability start to decrease, saturates around 6 ton/ha and start to be negative after 9 ton/ha (Fig. 16-C). For land on maize a somehow hyperbolic function seems to represent the relationship with the probability of being food-secure. For low values of land on maize a high probability of being food-secure is achieved, but as land on maize increases the probability start to decrease (a negative rate of change of probability), a threshold around 30 % of land on maize an inflexion point is found, the rate of change is zero, and changes to positive for values around 50 and 75% of land on maize, but again at around 75 % of land on maize, another inflexion point is found and the probability starts to decrease again, reaching the lowest values when land on maize is close to 100% (Fig. 16-D). For coffee yield, the relationship with the probability of being food-secure is close to linear and direct (Fig. 16-E)., for tropical livestock units is quite similar but slightly exponential, an increase in TLU

increases the probability of being food-secure (Fig. 16-F), while for crops number the relationships in linear but inverse (Fig. 16-G).

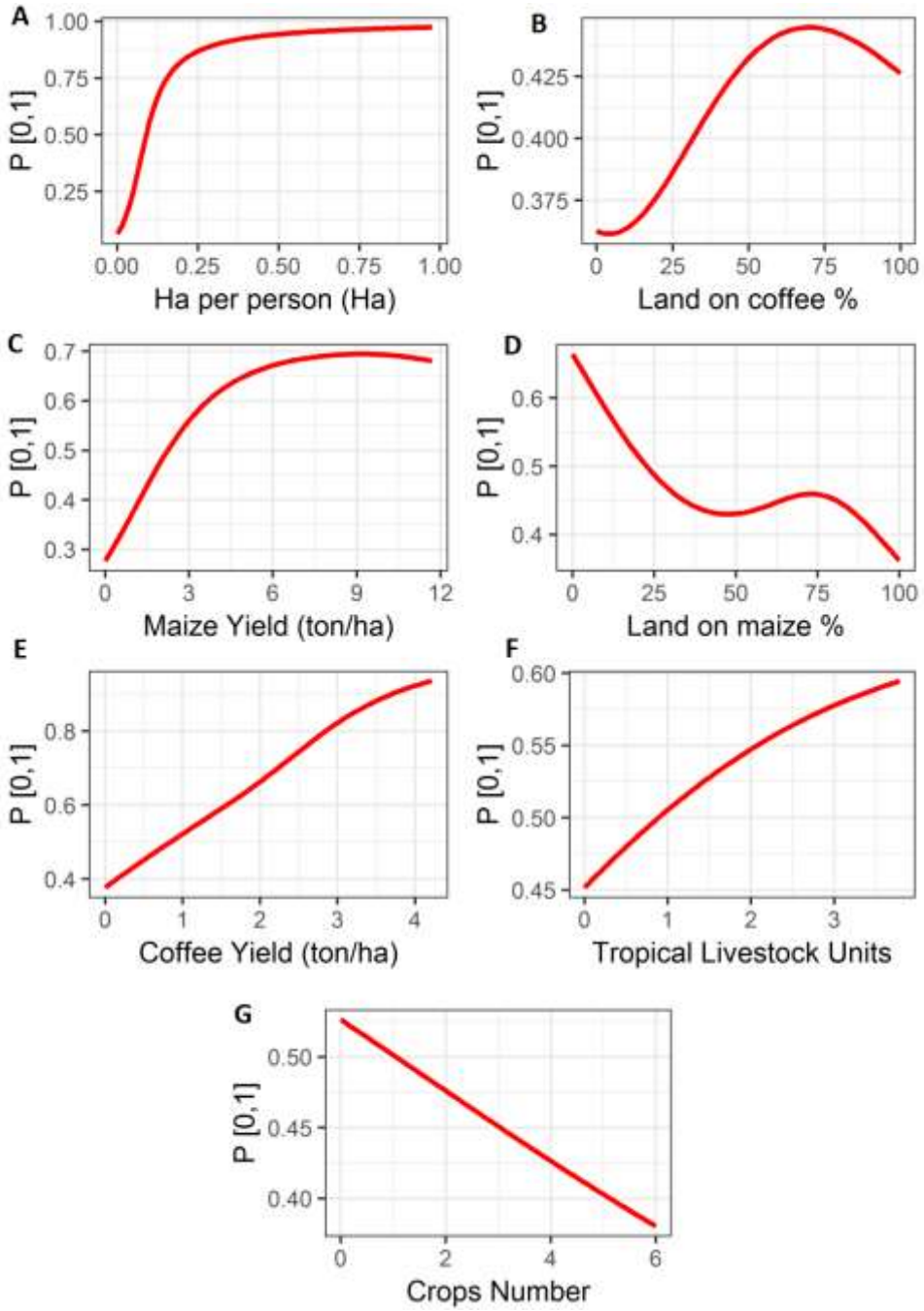

Figure. 16. Response curves for the effect of single variables on PFA.

3.4.2 The frontiers and thresholds found for the interaction of variables.

The relationship of land on maize and ha per person shows that for being food secure with a probability of one, it can be reached when the land availability per person, is around 0.7 and 1 ha, even when land on maize is 100 %. A still good chance of being food secure P ~ (0.9, 1), can be reached with a land availability per person between 0.2 and 0.4 ha with any % of land on maize, but if land availability per person is less than 0.1 ha, having either large or low amounts of land on maize greatly decreases the probability of being food secure (Fig. 17-A).

For the case of land on coffee and ha per person a similar situation presents alike with land on maize- ha per person, but achieving a probability of 1 of being food secure is achieved with almost full on land on coffee but with less ha per person than with maize, around 0.4 and 0.7 of land availability per person. Less probability for being food-secure occurs even when 100% of land is cultivated with coffee but land availability is less 0.2 ha (Fig. 17-C).

The relationship of land on coffee and land on maize shows the effect of having an intermediate amount of land on maize, but 0 on land on coffee, in this case the probability of being food secure is very low. While having certain amount of land on coffee of even less than 20%, but less land on maize than 10 %, increases the probability of being food secure, also having land on coffee above 60 % and below 80% and no land on maize, highly increase the probability of being food-secure (Fig. 17-B). A high yield on maize increases the probability of being food-secure when either land on maize is high between 0.7 and 0.9 or when it is less than 0.2. A low yield on maize induces a lower probability of being food-secure when land on maize is either high or low (Fig. 17-D).

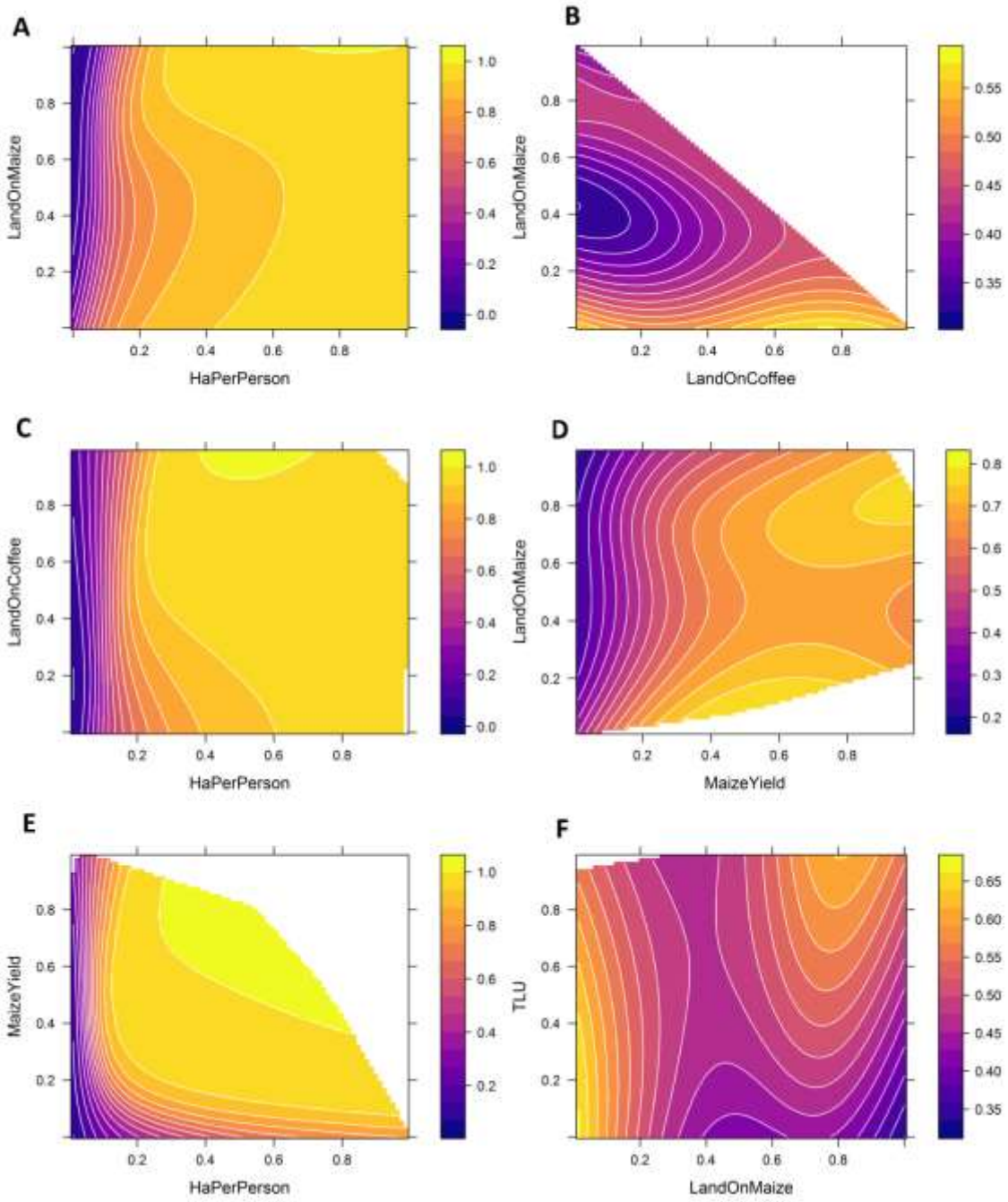

Figure 17. Bi-dimensional partial dependence plots effect of interacting variables, for the probability of being food secure

The relationship of maize yield and ha per person highly increases the probability of being food-secure when a high amount of ha per person, between 0.3 and 0.8, is combined with a high yield of maize a between 0.4 and 0.9, good chances (but still less) are also possible for lower amounts of land per person of above 0.2 and maize yields above around 0.3. When there is no sufficient land availability, below 0.2 ha,

not even a high yield on maize provides a good chance for being food-secure (Fig. 17-E).

The effect of land on the interaction of maize and TLU on the probability of being food-secure (Fig. 17-F). is that when low land on maize is cultivated and there exist low values of TLU high chances of being food secure is achieved, a small amount of both combined probably with the cultivation of a cash crop may triggers that high probability of being food secure. Large amounts of land on maize above 40 % with low values of TLU trigger a less probability of being food-secure, while having large amounts of land on maize and also large values on TLU increase the probability of being food-secure.

Chapter 4: Discussion

Despite the potential biases resulting from, variability or uncertainty on data from agricultural production (subjective farmers own assessment of production), market fluctuation on crops prices, farmer's individual attitudes towards risk, variability on land and assets management, familiar households dynamics, biophysical constraints and overall, farms intrinsic diversity, but even more lacking information on off-farm income, we have been able to capture that complexity and provide an accurate assessment of the food security status of farm households in the WHG, as it is shown by the model validation in which we found a relationship between the lack of everyday meals and low PFA values.

Also, through the application of machine learning algorithms, we have been able to construct models with a high level of precision for predicting PFA, with only a small subset of variables. But even more, through the partial dependencies and surface plots, we now possess an analytical tool to assess the effect of the interaction of these variables, their thresholds and frontiers, affecting the amount of PFA or the food-secure/food-insecure status of households.

Our results show the large diversity of farming systems in the region is dominated by maize and coffee production, as well as the large differences in terms of their potential food availability (PFA), the indicator we used to approximate food security. In our PFA calculation, 52% of farm households in the WHG do not have the means to attain sufficient energy from their agricultural activities. Overall, our results are consistent with other studies on the severity of food insecurity and malnutrition in the WHG. For example, USAID (2018) has estimated that approximately 50% of Guatemalan children under five years of age are stunted due to chronic food insecurity and, within indigenous areas such as the WHG, nearly 70 percent of the population is chronically malnourished.

The consumption of food crops (notably maize) plays a key role to food security of a large part of the farm households in the WHG (Fig. 6). However, as illustrated by Figure 6-A, the contribution of grown and consumed crop has a bell shape in relation to PFA in which, for very low and very high PFA values, the

contribution of consumed food crops is relatively low compared to the middle section of the PFA value range. Similar results have been documented in other studies (Frelat et al. 2016, Ritzema et al. 2017, Lopez-Ridaura et al. 2018) showing a common pattern among small scale farming systems in completely different regions of the world.

Maize is critical for food security in the WHG and it represents a high percentage of the PFA for most farm households. Maize yields in the region range from 1 to 2 tons per ha (Hellin et al. 2017) and increasing the yield and yield stability of this crop might or might be not critical, if we take into account that maybe land could be a critical point defining food security of small holders in the WHG.

Small-scale farmers in Guatemala are key actors in maize production-consumption. Farmers owning less than seven ha of land represent 53% of the farmers growing maize. Over two-thirds of the maize they produce is for self-consumption, covering approximately 8 months of their needs (USAID/Guatemala, 2009). However, these subsistence farmers have experienced accelerated land fragmentation (Isakson, 2014). Currently 55% of the national maize production is produced by farmers with less than 3.5 ha (Fuentes-López, et al. 2005). This land fragmentation has been more dramatic in the Western Highlands. Hellin et al. (2017), found that, in 2016, farmers from this region owned an average of 4.3 ha in the 1980s, but currently they have only 0.4 ha. These findings coincide with Bellowet al. (2008), and Sigüenza Ramírez et al. (2010) who find that the average land size for a small scale farmer was 0.34 ha. Hellin et al. (2017) also find a large variability in maize farmers´ average landholdings, depending on their location within the five departments that comprise the Western Highlands.

The results of the present predictive analysis coupled with the partial dependence plots and variables interaction give important information on the drivers for household's food security attainment. Our simple response model allows to explain nearly 83.6 % of variation for the continuous prediction of PFA, and to correctly predict the food-secure status of 84.8 % of the households. The results pertaining variables interaction show that the main constraints for food security in the WHG might be land availability and not necessarily crops productivity, although high probabilities of being food secure augment if crops yield increases.

When land availability per person is large, higher values of PFA and higher chances of being food secure are possible, independently of the crop produced or the surface cultivated. Nevertheless increasing crops diversity is a main concern for the system as it is shown that high crops diversity greatly affects PFA as apparently, not even when several crops are cultivated on large plots of enough size to provide enough land availability per person, food security might be attained, there exist a threshold in the number of crops that can be cultivated in small plots to attain large PFA values. The other way around, monocrop cultivation on large plots allow attaining high values of PFA.

Nevertheless PFA reaches a saturation point when land availability reaches a point at around 0.25 ha per person that agrees with other studies like Frelat et al. (2016), where they explain this saturation could be explained by a decline in productivity per unit land (kcal/ha), when land increases. This pattern of inverse land size-productivity is found in many studies of smallholders farmers (Ali et al. 2014; Otsuka et al. 2014), showing also that medium sized farmers are more efficient per unit area (Muyanga and Jayne, 2014).

On respect how land is split into staple crop –cash crop, (land on maize vs land on coffee) highlights that it is possible that food security might be attained easier if land is allocated completely to coffee or maize (provided that a large land is available), but when both crops are sown by the same household, not having a significative piece of the land allocated to the cash crop if maize is also cultivated affects the attainment of food security, as maybe, the cash crops supplies the kcal needed for rising the PFA of the households. Nevertheless, small amounts of both crops seem to give good chances of attaining food security, probably because maize contribute with kcal directly to the household, cash crops by means of commercialization, and the rest either because of land allocated to livestock, or other food, or cash crops, that would mean a diversified household might have good chances of being food secure, but again provided enough land per person is available as we have seen it is a constraint to all the rest of the variables in augmenting PFA.

The main concern here would be that land availability of the sample is on average 0.127 ha and more, a median of 0.085 ha, that would mean that nearly half of the households population does reach the minimum amount of ha per person to achieve 2500 kcal/day/MAE, which is around 0.0625 ha, but the other half might be

not reaching not even this value. This point (HaPer Person=0.0625, 2500 kcal/ha/MAE) is represented in figures 12-A and 16-A. for the case of the continuous PFA, and could be an inflection point on which the PFA starts to increase more quickly, that means a change to a steepest rate of change. As well, at around this point of HaPerPerson, in the case of the classification approach, also the rate of increase in probability changes dramatically.

Given this information, the main concern here is that increasing crops yield might be not the proper intervention for these households as attaining high yield but not having enough land sizes PFA doesn't reaches high values.

Another important interaction found, is that PFA relationship with crops number is inverse. The subsistence smallholders in Guatemala traditionally grow polyculture systems incorporating maize, beans, potato, squash and many other crops (Milpa). As polyculture cropping systems have proved to be more sustainable and provide more ecological services, interventions targeting food insecurity would need to take into account that increasing diversity of crops would increase food diversity and even food quality, but also that food security from agriculture in the region is attained mostly by coffee monocrops or by growing maize but in rare large areas which are scarce.

An important analysis that could be performed is to identify the effect of market access on increased PFA. As Frelat et al (2016) find, access or means of commercialization is crucial to ensure or improve the livelihoods of smallholders, and might impact more than crop productivity interventions.

Anyway, these interventions might be profitable for those accessing with at least enough land sizes but for those that do not, increasing yield, not even accessing to market might alleviate food insecurity, and different interventions outside the production and commercialization of crops, could be explored.

The sustainability of these systems must be established soon, since even those in which agriculture provides sufficient food, the management of monocultures could reduce the resilience of the ecosystems where they are practiced, in addition to reducing the biological diversity that is traditionally practiced. In the polycultures of the region and also, we should take into account the social dynamics that have led to the extreme decrease in the availability of land and its fragmentation as well as the

contribution of activities outside agriculture that could remedy to some extent the malnutrition in the area

Chapter 5: Conclusions

In this study the potential food availability reflecting the food security of smallholders' farm households was estimated through the application of a simple food indicator to 4790 households in 55 municipalities of five departments in the western highlands of Guatemala. Our results show that the large diversity of farming systems in the region is dominated by maize and coffee production and reveal the large differences in terms of their PFA. From our indicator we can estimate that 52% of farm households in the WHG do not have the means to attain sufficient energy from their agricultural activities. This suggests the importance of off/non-farm sources of income/food and indicates that interventions for improving agricultural productivity for improving household's food security and livelihoods should take into account such diversity.

Also, through the application of machine learning algorithms, we constructed a tool which allows predicting the farm households' potential food availability with a high degree of precision. This model is based on a simple short set of variables which could be obtained with very few questions to the farmers. Although information on yield could be not accurate, the variables which matter most for the model's accuracy are land availability and % of land allocation to crops, which are quite structural variables with a lower error range. The relationship of some of the predictor variables with PFA, extracted with the machine learning algorithm, reveal some non-linear patterns and allow us to extract some conclusion referring to the main constrains that prevent households from achieving food security. Large land availability and high yields of crops seem to ensure achieving food security, but around half of the population does not possess enough land.

Although this is a socio-ecological complex system, with many layers interacting in many ways, with bottom-up and top-down effects, it is remarkable that certain patterns can be extracted which could explain the dynamics of the food security status among the smallholder of Guatemala.

The design of agricultural interventions should take into account that farm households in the western highlands of Guatemala are land constrained, and that improving land productivity might be not a bullet proof recipe.

Bibliography

Ali DA, Deininger K (2014) Is there a farm-size productivity relationship in African agriculture? Evidence from Rwanda. World Bank Policy Research Paper 6770 (World Bank, Washington, DC).

Angeles, G., Hidalgo, E., Molina-Cruz, R., Taylor, T., Urquieta-Salomón, J., Calderón C, Fernández, J.C., Hidalgo, M., Brugh, K. & Romero, M. (2014). Encuesta de Monitoreo y Evaluación del Programa del Altiplano Occidental, Línea de Base 2013. 149 pp. USAID. Resource document. https://www.measureevaluation.org/resources/publications/tr-14-100-es. Accessed December 14 2017.

Beck MW (2018). "NeuralNetTools: Visualization and Analysis Tools for NeuralNetworks." Journal of Statistical Software, *85*(11), pp. 1-20. doi:10.18637/jss.v085.i11 (URL:http://doi.org/10.18637/jss.v085.i11).

Beck MW (2018). "NeuralNetTools: Visualization and Analysis Tools for NeuralNetworks." Journal of Statistical Software, *85*(11), pp. 1-20. doi:10.18637/jss.v085.i11 URL:http://doi.org/10.18637/jss.v085.i11)

Bellow, J. G., Hudson, R. F., & Nair, P. K. R. (2008). Adoption potential of fruit-tree-based agroforestry on small farms in the subtropical highlands. Agroforestry Systems, 73(1), 23–36. https://doi.org/10.1007/s10457-008-9105-x

Bishop, C. M. (1995). Neural Networks for Pattern Recognition. Oxford: Oxford University Press

Breiman, L., 2001. Random Forests, Machine Learning 45, pp. 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees, 1984: Belmont. CA: Wadsworth International Group.

Bruni, L., Fuentes, A., & Rosada, T. (2009). Dynamics of Inequality in Guatemala. UNDP project "Markets, the State, and the Dynamics of Inequality: How to Advance Inclusive Growth," coordinated by Luis Felipe López-Calva and Nora Lustig (http://undp. economiccluster-lac. org/).

Campbell, J.B., 2002. Introduction to Remote Sensing. Taylor & Francis, London

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". Educational and Psychological Measurement. 20 (1): 37–46. doi:10.1177/001316446002000104.

Conforti, M., Pascale, S., Robustelli, G., Sdao, F., 2014. Evaluation of prediction capability ofthe artificial neural networks for mapping landslide susceptibility in the Turbolo River catchment (northern Calabria, Italy). Catena 113, 236–250.

Cortes, C. & Vapnik, V., 1995. Support-Vector Networks, Machine Learning 20(3), pp. 273-297.

Ehui S, Li-Pun H, Mares V, Shapiro B (1998) The role of livestock in food security and environmental protection. Outlook Agric 27(2):81–87.

Elizondo, D. A., McClendon, R. W., & Hoogenboom, G. (1994). Neural network models for predicting flowering and physiological maturity of soybean. Transactions of the ASAE, 37(3), 981-988.

Ericksen PJ, Ingram JSI, Liverman DM (2009) Food security and global environmental change: Emerging challenges. Environ Sci Policy 12(4):373–377

FAO (1996). Food, agriculture and food security: developments since the World Food Conference and prospects for the future. World Food Summit technical background document No. 1. Rome.

FAO (2003). Trade reforms and Food Security, Conceptualizing the linkages. Food and Agriculture organization of the United Nations. Italy, Rome

FAO, I. (2016). WFP (2015), The State of Food Insecurity in the World 2015. Meeting the 2015 international hunger targets: taking stock of uneven progress. Food and Agriculture Organization Publications, Rome.

Frelat R., Lopez-Ridaura, S., Giller, K.E., Herrero. M., Douxchamps, S., Djurfeldt, A,A, Erenstein, O., Henderson, B., Kassie, M., Paul, B.K., Rigolot, C., Ritzema, R.S, Rodriguez, D., van Asten, P.J., & van Wijk, M.T. (2016). Drivers of household food availability in sub-Saharan Africa based on big data from small farms. Proceedings of the National Academy of Sciences, 113(2), 458-463.

Fuentes-López, M. R., Van Etten, J., Vivero Pol, J. L., & Ortega Aparicio, A. (2005). Maíz para Guatemala: Propuesta para la Reactivación de la Cadena Agroalimentaria del Maíz Blanco y Amarillo, SERIE "PESA Investigación", n 1, FAO Guatemala, Guatemala. Guatemala City: FAO.

Gobierno de Guatemala. (2012). Programa de Agricultura Familiar para el Fortalecimiento de la Economía Campesina (PAFFEC 2012-2015). elaborado por el Ministerio de Agricultura, Ganadería y Alimentación, MAGA, con el apoyo de la Organización de las Naciones Unidas para la Agricultura y la Alimentación, FAO.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics, 24(1), 44-65.

Greenwell B. M. (2017). pdp: An R Package for Constructing Partial Dependence Plots. The R Journal, 9(1), 421--436. URL https://journal.r-project.org/archive/2017/RJ-2017-016/index.html.

Guardiola, J., Cano, V. G., & Pol, J. L. V. (2006). La seguridad alimentaria: estimación de índices de vulnerabilidad en Guatemala. VIII Reunión de Economía Mundial, 20–22.

Hammond, J., Fraval, S., van Etten, J., Suchini, J.G., Mercado, L., Pagella, T., Frelat, R., Lannerstad, M., Douxchamps, S., Teufel, N., Valbuena, D. & van Wijk M.T. (2017). The Rural Household Multi-Indicator Survey (RHoMIS) for rapid characterisation of households to inform climate smart agriculture interventions: Description and applications in East Africa and Central America. Agricultural Systems, 151, 225-233.

Haykin, S. (1994). Neural networks: a comprehensive foundation. Prentice Hall PTR.

Hellin, J., Cox, R. & Lopez-Ridaura, S. (2017). Maize diversity, market access, and poverty reduction in the Western Highlands of Guatemala. Mountain Research and Development 37 (2): 188-197.

Herrero M, et al. (2010) Smart investments in sustainable food production: Revisiting mixed crop-livestock systems. Science 327(5967):822–825.

IFAD (International Fund for Agricultural Development). (2011). Enabling poor rural people to overcome poverty in Guatemala. Rome Italy. Resource document. https://www.ifad.org/documents/10180/16e68b93-2e7f-4804-8385-b8d53d784130 Accessed: January 30 2018

Instituto Nacional de Estadítica. (2012). Caracterización República Guatemala.

Isakson, S. R. (2014). Maize diversity and the political economy of agrarian restructuring in Guatemala. Journal of Agrarian Change. https://doi.org/10.1111/joac.12023

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., ... & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. PLoS One, 11(6), e0156571.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Kuhn Max. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2018). caret: Classification and Regression Training. R package version 6.0-80. https://CRAN.R-project.org/package=caret

Lang, T., & Barling, D. (2012). Food security and food sustainability: reformulating the debate. The Geographical Journal, 178(4), 313-326.

Lawrence, J. (1994). Introduction to neural networks: design, theory, and applications (p. 423). Nevada City, CA: California Scientific Software.

Lee, S., Evangelista, D.G., 2006. Earthquake-induced landslide susceptibilitymapping using an artificial neural network. Nat. Hazards Earth Syst. Sci. 6, 687–695.

Liaw A. and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Lopez-Ridaura, S., Frelat, R., Van Wijk, M. T., Valbuena, D., Krupnik, T. J., & Jat, M. L. (2018). Climate smart agriculture, farm household typologies and food

security: An ex-ante assessment from Eastern India. Agricultural systems, 159, 57-68.

MAGA (Ministerio de Agricultura, Ganadería y Alimentación de la República de Guatemala) (2011). Diagnóstico de la región de occidente de Guatemala. Guatemala City. 106 p.

Muyanga M, Jayne TS (2014) Is small still beautiful? The farm size-productivity relationship Revisited. Paper prepared for presentation at the 2014 conference on land policy in Africa African Union Conference Center, Addis Ababa, Ethiopia November 11- 14, www.afdb. org/en/aec-2015/papers/paper/is-small-still-beautiful-the-farm-size-productivity-relationshiprevisited-4736/.

Nguyen, M.Q., Atkinson, P.M., Lewis, H.G., 2006. Super-resolution mapping using Hopfield neural network with fused images. IEEE Trans. Geosci. Remote Sens. 44 (3), 736–749.

Otsuka K, Matsumoto T, Kilic T (2013) Should African rural development strategies depend on smallholder farms? An exploration of the inverse productivity hypothesis. Agric Econ 45(3):1–13. 27.

Pachepsky, Y. A., Timlin, D., & Varallyay, G. Y. (1996). Artificial neural networks to estimate soil water retention from easily measurable data. Soil Science Society of America Journal, 60(3), 727-733.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rakkiyappan, R., Balasubramaniam, P., 2008. Delay-dependent asymptotic stability for stochastic delayed recurrent neural networks with time varying delays. Appl. Math. Comput. 198 (2), 526–533.

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.4.

Ritzema, R. S., Frelat, R., Douxchamps, S., Silvestri, S., Rufino, M. C., Herrero, M., Giller, K., López-Ridaura, S., Teufel, N., Paul, B. & Van Wijk, M. T. (2017). Is production intensification likely to make farm households food-adequate?

A simple food availability analysis across smallholder farming systems from East and West Africa. Food Security, 9(1), 115-131.

Rosenblatt, F., 1958. The Perceptron: A probabilistic model for information storage and organization in the brain, Psychological Review 65, pp. 386-408

Rumelhart, D. E.; Hinton, G. E. & Williams, R. J., 1986. Learning internal representations by error propagation, MIT Press, Cambridge, MA, USA, pp. 318-362.

Shmueli, D. (1998). Applications of neural networks in transportation planning. Progress in Planning, 50(3), 141-204.

Sigüenza Ramírez, P., Winkler, K., Monzón, R., Gauster, S., Dürr, J., & Ozaeta, J. P. (2010). Nuestro maíz, nuestro futuro: Estudios para la reactivación de la producción nacional de maíz en Guatemala. Guatemala City.

Smith, C. A. (1989). Survival strategies among petty commodity producers in Guatemala. Int'l Lab. Rev., 128, 791.

Steinberg, M., & Taylor, M. (2008). Guatemala's altos de Chiantla: Changes on the high frontier. Mountain Research and Development, 28(3), 255-262.

The, L. R. M. (2018). Opening the black box of machine learning. The Lancet. Respiratory medicine, 6(11), 801.

Taiyun Wei and Viliam Simko (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

Taylor, J. E., Yuñez-Naude, A., Jesurun-Clements, N., Huard, A., Sanchez, M. A., Alvarez, V. M., & Baumesiter, E. (2006). Los posibles efectos de la liberalización comercial en los hogares rurales Centroamericanos a partir de un modelo desagregado para la economía rural.

USAID (2013). Integration of USAID in the Western Highlands. Resource document. http://pdf.usaid.gov/pdf_docs/pdacx493.pdf. Accessed September 27 2017.

USAID (2018). Food Assistance Fact Sheet Guatemala. Updated January, 2018. Resource document.

USAID/Guatemala. (2009). Evaluación rápida del sector agrícola guatemalteco y su estado para abordar los retos de seguridad alimentaria del país.

van Etten, J., & Fuentes, M. R. (2004). La crisis del maíz en Guatemala: las importaciones de maíz y la agricultura familiar. Anuario de Estudios Centroamericanos, 30(1/2), 51–66. Retrieved from http://www.jstor.org/stable/25661376

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Vojislav, K., 2001. Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models (Complex Adaptive Systems). The IMT Press

WFP (World Food Programme). (2018). Guatemala. Country profile. https://www.wfp.org/node/3475/4323/639382. Accessed May 2018

Yan, X., & Su, X. (2009). Linear regression analysis: theory and computing. World Scientific. , Singapore.