Combining principal component analysis and logistic regression models to assess household level food security among smallholder cash crop producers in Kenya

Elizabeth Kabura Nyaga and **Werner Doppler**
University of Hohenheim, Stuttgart, Germany

Abstract

Concerns about rising food insecurity in developing countries have prompted a lot of research. This paper deals with household survey of cash crop smallholders in Murang'a District, Kenya. The paper focuses on how factors related to cash crops production affect food security. Principal component analysis was applied in developing a relative food security index that was used as an outcome variable in logistic regression to predict factors affecting food security in this region.

Results indicate that the percentage of food secure households is lower in the tea zone. This is despite relatively higher income from tea production. Income from tea is used in food purchase because of insignificant own production. Coffee farmers concentrate more in food than coffee production. They may therefore produce adequate for their households and sell any surplus. Further results indicate that increasing land under cash crop would increase the likelihood of a household being food insecure.

Keywords: cash crops, food security, Kenya
JEL: Q12, D13, R20

1. Introduction and background

Kenya is a major producer and exporter of tea and coffee. It ranks second after Sri Lanka in tea exports and is the fourth major black tea producer in the world. The tea industry employs about 2 million people. Small-scale tea production accounts for 65% of the total area under tea production in Kenya and about 62% of production (NYANGITO, 2000a). Coffee was the leading export crop in Kenya since independence to 1988. Between 1975 and 1986, it contributed over 40% of the total Kenyan export value but by 1992 it only contributed less than 9% of the country's export value (REPUBLIC OF KENYA, 2001). Between 1976 and 1986, the average price of Kenyan coffee was 141.66 US cents per lb, but in 1992, it was only 28.90 US cents per lb.

Between 2000 and 2004, the average price was 64.31 US cents per lb (INTERNATIONAL COFFEE ORGANIZATION, 2007).

Tea and coffee production have therefore played crucial roles in creating employment, foreign exchange earnings and have thus contributed to income of many households in Kenya. During the period following the introduction of the two crops in Kenya and the late 1980's or early 1990's, farmers in coffee and tea growing regions earned handsomely from the crops. The two cash crops contributed to relatively higher income and food security among farmers in this region as compared to farmers without the cash crops in the same and neighbouring districts. Currently, these contributions have declined. Access to food also remains an issue of great concern in this region due to poor harvests caused by unreliable rainfall and small farm sizes. Small farm sizes limits expansion of either food or cash crop production. The high costs of farm inputs have adversely affected coffee and tea production leading to high costs of production. Coffee farmers have also experienced reduction in purchasing power due to decline in coffee production and low coffee payments. Coffee bushes are unpruned, unweeded and diseased (NYANGITO, 2000b). Despite these problems, coffee farmers cannot uproot their coffee to give room to other crops and activities due to the fact that they need approval by the Coffee Board of Kenya. The Coffee Board of Kenya in consultation with the minister responsible for agriculture formulates policies and makes rules to regulate the coffee industry (REPUBLIC OF KENYA, 2002). The Board is therefore responsible in supervising the rules that regulate establishment of a coffee farm and also uprooting of coffee bushes (REPUBLIC OF KENYA, 2001) as spelled in the Coffee Act. The laws governing coffee are spelled out in the Coffee Act Chapter 333, which controls production, marketing and exportation of coffee (NYANGITO, 2000b) among other activities. The specific laws governing the tea industry are stated in the Tea Act (Chapter 343) which provides for control and regulation of production, manufacture and export of tea (NYANGITO, 2000a) among other activities. The act allows the Tea Board of Kenya to control and regulate the tea industry. Therefore, rules for uprooting of tea bushes are monitored by the Board (REPUBLIC OF KENYA, 2000) as spelled in the Tea Act. The act also empowers the Minister for agriculture, in consultation with the Tea Board of Kenya to make regulations for protecting and promoting the tea industry and carrying out of the provisions of the act (NYANGITO, 2000a). This implies that the Minister of Agriculture together with the Tea Board of Kenya can regulate expansion of tea bushes by smallholder tea producers. However, in this area, most farmers do not have anymore land remaining to allow them to expand tea production

According to MASLOW (1943) hierarchy of needs theory, as humans meet 'basic needs', they seek to satisfy successively 'higher needs' that occupy a set of hierarchy. Studies have shown that, in most developing countries, Kenya inclusive, food is still a 'basic need'. This need is yet to be satisfied for most farmers who are therefore unable to move to higher levels in the hierarchy of needs. In order to move to another level in the hierarchy, households should have sufficient income (or other entitlements) to purchase food (LORENZO et al., 2005) when own production is not adequate. Situation characterised by a combination of poor harvests of food crops and low returns from cash crops is a common phenomenon in the research area. The area therefore suffers from what AGBOLA et al. (2004) termed as transitory food insecurity.

In this paper, descriptive and econometric analysis is applied to household data in order to understand food security status of households as well as factors contributing to this status. The paper is structured as follows: section two deals with the research area and methodology used in the paper. Section three presents results of principal component analysis while section four deals with specification of the logistic model used in the econometric analysis. Section five discusses the results of the regression analysis. The paper closes with some concluding remarks.

2. Research area and methodology

The analyses presented in this paper relate to data collected in two cash crop growing zones of Murarandia location, Murang'a District in the year 2005. Murarandia location was purposively selected based on three reasons. First, the area has tea and coffee in defined zones. Secondly, each of the above cash crops are grown in respective zones within the location with each of the dominant crop in each zone playing a crucial role in household income presently and/or in the past. Thirdly, the regional development of the two cash crop growing zones in terms of infrastructures and regional resources is almost the same. Murarandia location is divided into several agro-ecological zones (JAETZOLD and SCHMIDT, 1983), but those that grow tea and coffee and considered in the paper are UM2 (coffee zone) and LHI (tea zone). UM1 which is a transitional zone separating the tea and coffee zones and that has both crops was left out because the farmers earn from both cash crops and it would therefore be difficult to attribute the changes in variables under observation to either tea or coffee. To select families in the coffee zone, a sampling frame was obtained from the two main coffee factories, namely Gatuya and Kianjogu. The names of all farmers with accounts in the two coffee factories were listed alphabetically and systematic random sampling was applied to select 60 families. The factory officials provided information on villages where the 60 families lived. To select families in the tea zone, a list of tea buying centres was obtained from Githambo tea processing factory. Eight buying centres were randomly selected. The names of all families selling tea in the eight buying centres were listed alphabetically and systematic random sampling was applied to select 60 families. The two sampling processes therefore yielded a sample of 120 households.

Growers in the research area depend mostly on income from these cash crops as a source of income for smoothing consumption in cases of insufficient own production. This paper therefore seeks to determine the role of tea and coffee production in meeting food security needs of the population in this region. For this purpose, Principal Component Analysis (PCA) was used to construct a relative household food security index based on a range of food security related indicators. The method creates an index by isolating the food security component embedded in the various food-related indicators. It allows for the aggregation of a large number of interrelated variables with as little loss of information as possible (DOPPLER, 2004). The index represents a particular household's food security status in relation to all other households in the sample. The indicators used to develop the index differ in their measurement scale. PCA, however, converts all indicators into standardized variables with a mean of zero and standard deviation of one (ZELLER et al., 2005). Hence the resulting index is also a normally distributed variable (BASILEVSKY, 1994). Upon construction, the index was converted into a qualitative variable and logistic regression models were then used to estimate the effect of smallholder tea and coffee production on household food security. In order to identify variables to be used in construction of the index, this research adopted the definition of OBAMIRO (2004) that defines food security in terms of food availability, food accessibility and nutritional factors. This definition is extended to include food shortage coping strategies (MAXWELL, 1995) which captures element of 'sufficiency' and 'security' in the judgement of the people responsible in provision of food within the household. OBAMIRO (2004) defined food availability as issues related to farm resources and input markets, production and processing, home consumption and family size; food accessibility as issues related to markets, cash to family, sex, age, family income, family size and nutritional factors as issues related to work load, age and sex.

The PCA method used in this analysis creates several components or factors which are linear combinations of the original indicators of food security. The first step in conducting PCA is to run a correlation analysis between the benchmark indicator and several variables used in the survey that are related to the underlying issue of concern. The benchmark indicator is one that appears to be closely identifying the issue of concern. In this paper, the issue to be analysed is food security and several food security related indicators were targeted in the survey. The amount of maize from the farm that is stored for home consumption was used as the benchmark indicator. This was correlated with all variables relating to food availability, food accessibility and food coping strategies. This variable was chosen as the benchmark indicator because maize is considered a staple food in the region and families would consider themselves as food insecure if they have no maize in the house. Based on the strength of correlation, weaker variables are filtered to ensure that only variables that have a strong correlation with the benchmark indicator are included in the PCA model

(HENRY et al., 2003). While this helps to reduce the complexity of calculated components and yields stronger food security component, care should be taken to ensure that the final index reflects various dimensions of food security. Table 1 shows correlation results.

Table 1. Correlation between benchmark variable and other indicator variables

Indicator variable (N=120)	Pearson correlation
Amount of maize from farm stored for home consumption	1.000
Eggs production per year	0.439 (0.000)
Total maize output during the year	0.608 (0.000)
Total potatoes output during the year	0.227 (0.006)
Total size of land in hectares	0.351 (0.000)
Number of equipments for cultivation	0.389 (0.000)
Number of spraying equipments	0.235 (0.005)
Duration of storing potatoes in months	0.214 (0.009)
Total cost of labour in food production during the year	0.534 (0.000)
Cash from sale of milk during the year	0.157 (0.044)
Distance of source of drinking water from the house	-0.172 (0.030)

Source: 2004/05 family survey, Murarandia Location, Murang´a District, Kenya

Once stronger variable indicators have been selected, they are subjected together with the benchmark indicator to the PCA process which extracts various components that capture overall variance in all food security indicators. Following similar studies by ZELLER et al. (2005), the ultimate goal in our case is to produce a new variable called "food security component" that accounts for maximum variance in the original indicators as follows:

(1) $Y_i = q_1 x_1 + q_2 x_2 + ... + q_n x_n + \varepsilon i$

Applied to our analysis, Y_i is an index of food security, q_1, $q_2 \ldots q_n$ are vectors of weights of respective food security variables identified as significant in the PCA model and X_1, $X_2 \ldots X_n$ are vectors of food security related variables collected in the survey. A number of components are extracted through this process and each component reflects the variance exhibited by the original indicators. The first component registers the maximum variance in the selected food security indicators. Each of the components is unique in terms of the indicators included but subsequent components show less variance in the selected indicators. The components therefore represent different dimensions of food security, but the first component is a more representative indicator of food security since it shows maximum variance in the selected indicators of food security. Every respondent or household therefore has a score for the food security index and in a situation where the measurement scales of all indicator variables are similar, the factor score can be expressed mathematically as in equation (1).

3. Results of principal component analysis

The results of principal component analysis were generated using SPSS programme. In generating factors or components in PCA, one must select a method of calculating the factor scores. SPSS offers three methods of calculating factor scores: regression method, the Barlett method and the Anderson-Rubin method (see FIELD, 2005). Regression method was applied in this paper**.** This method is preferred because the influence of different scales of measurement used in different variables in the analysis is removed (FIELD, 2005). The component loadings are also adjusted to take account of the initial correlation between variables. There are also several methods of rotating the factors. By rotating the factors, loadings of each variable are maximized on one extracted factor but minimized on all the other factors. Varimax method was applied to rotate the factors. This method does not allow the factors to be correlated and loads a smaller number of variables highly onto each factor resulting in more interpretable clusters of factors (see FIELD, 2005).

Since the purpose of carrying PCA was to generate an index, not all the results have been presented. The most important outputs are the component matrix, component score coefficient matrix and the KMO-Barlett test and they are mainly used to assess the quality of the first component (ZELLER et al., 2005). PCA is therefore run several times until the best results according to these outputs are achieved. The component matrix shows the component loading coefficients which signifies the amount of correlation between the component variable or factor and the indicator variables (FIELD, 2005, and ZELLER et al., 2005). The component scores on the other hand do not only consider the relationship between the component variable and the indicator variables but also consider the original relationship between pairs of indicator variables (FIELD, 2005). PCA identified three principal components with an eigenvalue greater than one. ZELLER et al. (2005) points out that an eigenvalue greater than one is recommended as cut-off criteria for the component to be considered important in developing an index. The first component explains 34% of the variance, whereas the second component explains 11.99% and the third component explains only 9.90%. In table 2, we present the component loadings of the three components. The component loadings coefficients of the first component are high and the signs correspond to reality. The coefficients of all variables except distance of source of drinking water from the house are positive. This indicates that increasing the values of these variables