MINISTRY OF AGRICULTURE AND FISHERIES

Directorate of Economics

Research Paper Series

A Methodology for
Estimating Household Income in Rural Mozambique
Using Easy-to-Collect Proxy Variables

By
David Tschirley
Donald Rose
Higino Marrule

Research Report No. 38
February 2000

Republic of Mozambique

DIRECTORATE OF ECONOMICS

Research Paper Series

Through its Food Security Project, the Directorate of Economics of the Ministry of Agriculture and Rural Development maintains two publication series for results of research on food security issues. Publications under the *Flash* series are short (3-4 pages), carefully focused reports designed to provide timely research results on issues of great interest. Publications under the Research Paper series are designed to provide longer, more in-depth treatment of food security issues. The preparation of *Flash* reports and Research Reports, and their discussion with those who design and influence programs and policies in Mozambique, is an important step in the Directorates's overall analysis and planning mission.

Comments and suggestions from interested users on reports under each of these series help identify additional questions for consideration in later data analysis and report writing, and in the design of further research activities. Users of these reports are encouraged to submit comments and inform us of on-going information and analysis needs.

Sérgio Chitará
National Director
Directorate of Economics
Ministry of Agriculture and Fisheries

ACKNOWLEDGMENTS

The Directorate of Economics is undertaking collaborative research on food security with Michigan State University Department of Agricultural Economics.

We wish to acknowledge the financial and substantive support of the Ministry of Agriculture and Fisheries of Mozambique and the United States Agency for International Development (USAID) in Maputo to complete food security research in Mozambique. Research support from the Africa Bureau and the Bureau for Global Programs of AID/Washington have also made it possible for Michigan State University researchers to participate in this research, and to help conduct field activities in Mozambique.

The final views expressed here are those of the authors and do not necessarily reflect the official position of the Ministry of Agriculture and Fisheries, nor of USAID.

Duncan Boughton
Country Coordinator
Department of Agricultural Economics
Michigan State University

MAP/MSU RESEARCH TEAM MEMBERS

Sérgio Chitará, National Director, Directorate of Economics

Danilo Carimo Abdula, SIMA Coordinator

Rafael Achicala, SIMA Technician

Simão C. Nhane, SIMA Technician

Jaquelino Anselmo Massingue, MAP trainee Research and Agricultural Policy Analyst

Arlindo Rodrigues Miguel, MAP trainee Research and Agricultural Policy Analyst

Raúl Óscar R. Pitoro, MAP trainee Research and Agricultural Policy Analyst

Pedro Arlindo, Research Associate

Ana Paula Manuel Santos, Research Associate

Higino Francisco De Marrule, Research Associate

Paulo Mole, Research Associate

Maria da Conceição Almeida, Administrative Assistant

Francisco Morais, Assistant

Abel Custódio Frechaut, Assistant

Duncan Boughton, MSU Country Coordinator

Jan Low, MSU Analyst

Julie Howard, MSU Analyst

Donald Rose, MSU Analyst

David L. Tschirley, MSU Analyst

Michael T. Weber, MSU Analyst

Table of Contents

Foreword

Adapting INCPROX and INCPROX Lite to Other Data Sets

This report is a slightly modified version of a report originally prepared for use by USAID-funded NGOs in Mozambique in developing household income estimates for evaluation of their programs and reporting to USAID. Readers interested in the income proxy methodologies but not specifically in Mozambique might skip section II.A (Data Collection and Processing), as it contains primarily information very specific to Mozambique.

The methodologies reported on here represent a general approach applied to specific circumstances. The approach described in section II.B (INCPROX: A Structural Approach to Estimating Income) and II.C. (INCPROX Lite: A Simpler Alternative) could be applied in other countries or in other geographical areas of Mozambique, but would need to be adapted to those circumstances. Adapting INCPROX or INCPROX Lite to other areas would involve:

1. Collecting or gaining access to an existing household level data set that contains all the data needed to (a) directly calculate income for each household, and (b) develop income proxy variables for each household similar to those utilized in this report;

2. Utilizing regression techniques to develop INCPROX or INCPROX Lite models based upon this data set; and

3. Developing standard procedures for (a) collecting the proxy variables and (b) converting those proxy variables into estimates of household income and income components.

Income-expenditure surveys are done in many developing countries on a regular basis, for example every three- to four years. Thus, one wishing to develop and utilize these income proxy methodologies would typically *not* need to collect a data set specifically for that purpose; work could focus on developing the models and the standard procedures for utilizing the models to obtain income estimates. Once these models and procedures are developed, various organizations can collect a much reduced set of simple proxy variables on a regular basis (for example, yearly), and easily produce estimates of household income and income components. These organizations do not need sophisticated research capabilities, but do need access either in-house or through consultants to data collection and management skills typical of monitoring & evaluation operations.

Two key issues would benefit from further research. First, how well do the models perform over time? The value of these approaches as cost effective monitoring tools is predicated on the income estimates they generate being acceptably accurate over the course of several years (e.g., 2-4 years). If the models are robust over such a time period, then a rich set of monitoring information -- household income and its structure -- can be tracked regularly without the burdensome, complex, and costly work of collecting and processing income-expenditure data

sets.[1] In Mozambique, the lack of comparable data sets separated in time has not permitted testing the temporal durability of these models. A country with comparable income-expenditure data sets separated by 2-4 years would be an ideal candidate for such research.

Second, how can the models better deal with changing relative prices? Agriculture is a key component of income for most rural households in developing countries. Prices of agricultural commodities change every year, often in unexpected ways, and these price changes will affect income. Like the issue of temporal durability, developing an approach to deal effectively with changing relative prices requires comparable data sets separated in time (since relative prices will in all likelihood be different for each data set).

Section I of the paper provides a brief introduction. Section II reviews the work that was done to develop the models in Mozambique, and presents basic statistical results. Section III evaluates the performance of the models over space within the research area, and Section IV is a guide to NGOs on how to use the models - how to collect the proxy variables and develop the income estimates. In all these sections, much of the detail is in Annexes.

[1] These models are based on objective measures of the *intensity* of a household's involvement in each economic activity, and on the *productive resources* the household had available to dedicate to those activities. These simple proxy variables are complemented by quantitative measures of the production of two key crops - maize and cotton. Thus, this approach should, in theory, be reasonably sensitive to changes in weather (proxied by the production of maize and cotton), in a household's portfolio of economic activities (proxied by the intensity variables), and in the quantity of productive resources available to the household (proxied by *production function variables*). Factors not accounted for in these models which could affect income include changing relative prices, and pest or other production problems which affect a crop other than maize or cotton. Changes in the productivity of the household's productive assets will also affect income; these are partially accounted for by the quantitative estimates of maize and cotton production, holding constant the household's productive assets. The actual success of the approach in controlling for all these factors is, of course, an empirical issue requiring further analysis.

I. Introduction

This report outlines a method for estimating household income in rural areas of Mozambique using a proxy approach. It is based on collaborative work between Michigan State University and USAID-funded NGOs, and is meant for use by them in their areas of operation.

The development of such a methodology prompts two important questions. First, why focus on household income? Second, why use a proxy approach?

An important overall development goal for Mozambique is the reduction of poverty and improvement in the incomes and well-being of rural households. Thus, measurement of household income is a logical choice for monitoring the effects of policies and programs oriented towards accomplishing this goal. To be sure, there are other measures of household well-being. For example, some economists have argued that welfare levels are more appropriately determined by measuring household consumption expenditures, in part because of the extensive data collection activities needed to accurately assess household income. But, since so much of consumption in Mozambique is from own production, accurately measuring consumption in practice may be no easier than measuring income.

Income is difficult to measure in rural settings of developing countries, in part because there are so many different sources of income. Households in Mozambique earn income from the production and sale of seven different food staples, such as maize or manioc, seven different cash crops, like cotton or tobacco, and 20 different fruits and vegetables. In addition, income is obtained from the production and sale of livestock, from fishing, from wage labor, and from any of over three dozen different microenterprise activities, such as the weaving of baskets or the production and sale of alcoholic beverages. Thus, surveys attempting to measure household income need to ask questions on all of these activities and collect quantitative information on each.

In addition to the sheer number of sources of income, each of these sources presents different methodological challenges. For example, to get information on income from the production of maize, one needs to know how much maize was produced. This involves getting the farmer to remember how many bags or cans of which size were obtained from the harvest as well as the state of the maize, dried or fresh, on the cob or in grain. Conversion factors are needed for the size of the bag or can , and density factors are needed for the state of the maize. While all this is doable for one or two crops, it becomes very time-consuming and expensive when done for the vast array of crops that are grown in Mozambique. The expense in human and other resources is beyond the capacity of all but dedicated research projects.

An income-proxy methodology provides the possibility of obtaining regular (for example, yearly) information on household income without performing cumbersome quantitative surveys each time. This report outlines the development and use of such a methodology.

II. Development of the Proxy Methodology

Development of the proxy methodology involved data collection in collaboration with USAID-funded NGOs, followed by extensive data analysis. This section describes the design of data collection, the conceptual and statistical approaches utilized in developing the income proxy models, and presents selected statistical results and confidence intervals for the income estimates generated by the models. Two models are discussed. INCPROX utilizes 40 proxy variables to provide estimates of total household income and ten income components. INCPROX Lite uses 16 variables to estimate total household income, with no breakdown by component.

A. Data Collection and Processing

During June and November 1998, MSU collaborated with USAID-funded NGOs in two rounds of data collection that provided the basis for the development of these income proxy models. The purpose of the data collection was to obtain a high quality data base that had all data needed to calculate income, plus potential proxy variables. The data were cleaned and an income variable was calculated and used as the "gold standard" for which other easier-to-collect variables would proxy.

To improve data quality, two rounds of data collection were undertaken. The period of reference for the first round in June was from the beginning of the rains the previous year (October-November, depending on geographic location) until the time of the interview. The period of reference for the final round in November was from the previous (first) interview to the time of the final interview.

i. Sample Design

The NGO sample was stratified to ensure sufficient observations across all geographic areas in which the NGOs work. Districts in which NGOs work were grouped into seven zones (Table 1), based on available information about their agroecology and predominant economic activities. Within these zones, the universe for the sample was limited to villages in which NGOs had development activities; villages not directly served by NGOs were excluded. NGOs were asked to provide MSU with a list of all villages in which they worked, with information on their location and population. Ten villages were then randomly selected (using systematic sampling) within each of the seven zones, for a total of 70 villages. Within each village, 7 households were randomly selected using a spatial approach, giving a total sample size of 490 households. Households were selected regardless of whether they had received any direct assistance from a NGO.

Table 1. Stratification Zones for NGO Income Proxy Survey, 1998

Zone		Districts (NGO)
R1	Zambezi Valley	Marromeu (FHI), Caia (WV), Mutarara (WV), Chemba (WV)
R2	Central Zambêzia	Maganja da Costa (ADRA), Namacurra (WV), Nicoadala (WV), Morrumbala (WV), Milange (WV)
R3	Northern Zambêzia-South Nampula	Gurue (WV), Gilé (WV), Malema (CARE), Ribaué (CARE), Murrupula (WV,CARE), Nampula (CARE)
R4	Cotton Belt	Mogovolas (CARE), Meconta (CARE, WV), Nacaroa (WV), Erati (WV, CARE), Muecate (WV), Mecuburi (CARE)
R5	Coastal Nampula	Memba (SC-US), Nacala-a-Velha (SC-US)
R6	Central Sofala/Manica	Nhamatanda (FHI). Gorongoza (FHI), Gondola (Africare)
R7	Manica	Manica (Africare), Barue (Africare), Guro (Africare), Sussundenga (Africare)

WV = World Vision, FHI = Food for the Hungry International, ADRA = Adventist Development Relief Association, CARE = CARE, SC-US = Save the Children, US.

The spatial approach to selecting households was necessary because of the near impossibility of developing complete lists of all households in each of the villages. Dispersion of homes, population mobility, and lack of strong central authority at the village level combine to make the development of such lists exceptionally difficult. The approach was as follows:

1. After meeting with the village leaders, the enumerators and supervisor located the geographic center of the village.

2. Once in that geographic center, they spun a pencil or bottle and waited for it to stop.

3. Once stopped, the supervisor/enumerators asked the village leaders for how many minutes one would have to walk in that direction to reach the outer limits of the village.

4. This walking time was then divided by the number of interviews to be conducted along that route (3 or 4). This number was the *temporal section interval*; enumerators needed to walk for this amount of time in the randomly selected direction between each interview. For example, if the leaders indicated that it took about 45 minutes to reach the edge of the village in that direction, then 45/3 = 15 minutes. In this case, the enumerator

walked 15 minutes and then selected the first household encountered; the next interviewed household was 15 minutes from the first, and likewise for the third interview.

5. The second enumerator repeated steps 2-4, randomly selecting a new direction, determining the estimated walking time to reach the edge of the village, and dividing that time by 3 (if the previous enumerator is doing four interviews) or 4 (if the previous enumerator is doing three interviews).

6. If the enumerator reached the edge of the village and had not achieved his/her quota of interviews, the enumerator returned to the village center, informed the supervisor, and once again selected a direction in which to walk, dividing now the walking time by the number of additional interviews needed to be completed.

ii. Questionnaire Design

The questionnaires were carefully designed to elicit information on all of the in-kind and cash income earning activities in which households were involved. Sections in the questionnaires were:

I. Demographics
II. Remittances sent and received
III. Cultivated and fallow land
IV. Production of annual staple food and cash crops
V. Fresh production of food staples
VI. Agricultural sales
VII. Wage labor
VIII. Microenterprise activities
IX. Vegetable production
X. Fruit production
XI. Livestock holdings and production
XII. Cashew production (castanha and sub-products)
XIII. Fishing
XIV. Coconut production
XV. Expenditures (yes/no questions regarding the purchase of 17 items)
XVI. Construction of the home
XVII. Ownership of farm implements and household goods

Since the first round was conducted in June/July , the harvest of some crops for some households was not yet complete. In these cases, enumerators were instructed to record the fact that the household cultivated the crop, but had not finished the harvest. Total production and other information regarding that crop were then determined during the second round.

Selected information from the first round of interviews was entered by hand on the second round questionnaires prior to the second round field work, to be checked and also to serve as a guide in conducting the second round interview. Table cells that were filled-in this way during the second round are indicated on the questionnaire by a bold "**XX**".

B. INCPROX: A Structural Approach to Estimating Income

The conceptual approach used to estimate household income in INCPROX is "structural" in that it attempts to estimate different components of household income and, by summing these components, derives total income. Such an approach mirrors that used in most income surveys, which identify the different sources of income that a household may have, then ask the questions needed to quantify each of those income components. There are a number of advantages to such an approach:

1. For every household one knows unambiguously if it had zero or positive income from each of the components.

2. For each component, one can identify proxy variables which have a clear conceptual link to the level of income the household may have earned. For example, in estimating income from food crop production, variables such as the number of food crops cultivated, whether the household sold any food crops, the number of fields the household cultivated, the number of farm implements the household owns, and the number of adults available to work on the fields, should all be positively correlated with this component of income. For off-farm wage earnings, variables such as the number of household members engaged in such work, and whether the work is full- or part-time should both be correlated with the household's total wage earnings. These conceptual links between the proxy variables and the income components should improve the accuracy of a given model over time.

3. Estimating components of income, as opposed to total income only, provides a substantially richer set of insights into the evolution of household income strategies and of the rural economy in general. For example, knowing that an increasing (or decreasing) proportion of income is coming from off-farm activities, or from cash crops, is useful for policy formulation, program design, and related development planning activities in the agricultural sector.

Conceptually, income can be broken into a very large number of components; the specific components chosen should be a function of their relevance for understanding rural households and the rural economy, and the accuracy with which they can be predicted. For a given level of desired accuracy in the estimate of total income, estimating more income components will require the collection of more proxy variables. At some point, the number of variables collected becomes excessive given the fundamental objective of the proxy approach: reducing the cost of obtaining defensible estimates of household income. The analyst's challenge is to define a set of components which strikes a balance between accuracy, richness of information, and the amount of data collection and processing required.

The income components chosen for modeling in this analysis mirrored the sections of the survey instrument. They are income from:

1. **Food crop production**, defined as the value of production, when harvested in their mature state, of the basic staples: maize, all types of beans, manioc, rice, groundnuts, sorghum, and millet.

2. **"Non-food crop" production**, comprising the value of production of all other annual crops. The most important of these is cotton, but the group includes tobacco, sunflower, sesame, sugar cane, and seven other annual crops mentioned by interviewed households.

3. **Fresh production**, defined as the value of all annual crops that were harvested in a fresh state. Principal among these are fresh maize, beans, peanuts, and sweet potato. Though it is always harvested fresh, manioc is categorized in the food crop group due to its importance as a staple food crop.

4. **Vegetable production**, limited to the value of all production from the family's gardens (*hortas*). The most frequently produced vegetable crops were tomatoes, a dark leafy green known as "couve", pumpkin squash *(abóbora*) and onions. A total of 15 different vegetables were identified by respondents in the survey.

5. **Fruit production**, including production from all fruit trees. Key fruit crops were mangos, banana, papaya, and oranges. A total of 16 different fruits were identified in the data base.

6. **Fishing**, including the value of fresh fish (approximately 80% of all observations), dried fish, shrimp, and lobster (*lagosta*)

7. **Cashew production**, comprising raw cashew (50% of all observations), processed nut (*amendoa*), dried fruit (21% of observations), fresh fruit, and juice.

8. **Livestock production**, including cows, goats, pigs, chickens and other birds, rabbits, and other animals.

9. **Wage labor**, any off-farm activity where a household member is paid for his or her time, and does not have ownership of the activity. The most common types of wage labor were working on a neighboring smallholders' farm (55% of all observations) and working on the farm of a larger "privado" farmer (17%).

10. **Microenterprise activities**, defined as income from all sources other than wage labor or agricultural production and the sale of that production. The most commonly observed microenterprise activities were commerce, production of alcoholic beverages, craft activities such as carving, and weaving of baskets or mats. A total of 38 different microenterprise activities were identified in the survey.

All agricultural production was valued at mean sales prices by region. See Annex A for a list of the specific prices used.

In attempting to estimate each of these components, emphasis was placed on identifying proxy variables that would be straightforward to collect and process, and which had strong logical and empirical links to the level of income from the component. In general, three types of proxy variables were utilized: (1) measures of the *intensity* of the household's involvement in each area, (2) measures of the resources that the household could bring to bear on this productive activity (we will refer to these latter measures as *production function variables*), and (3) zone

variables which allowed the relationship between the proxy variables and component income to vary across space. Measures of intensity varied by component, but typically included the number of items within the category that the household produced (for example, the number of food crops that the household cultivated), and the number of items that it sold (or whether it sold any, or not). Production function variables were the same across all agricultural components: land proxied by the number of fields cultivated), labor (the number of non-elderly adults resident in the household), and capital (defined as the number of types of farm implements that the household owned). There were seven dichotomous zone variables, which indicated whether or not a household was situated in each of the different zones.

In addition to these intensity, production function, and zone variables, two quantitative production variables were included in the analysis: the quantity of maize grain produced and the quantity of seed cotton produced. These quantitative variables are more complex to collect and process than typical proxy variables, but are needed because production levels can fluctuate substantially from year-to-year based on rainfall and other factors. By quantifying the production of the most important food crop and cash crop, these quantities can themselves proxy for yield levels of other crops within their category. This should substantially improve the performance of the method over time.

Other variables were utilized in some estimations; see Annex B for the variables utilized in each component estimation.

"Stepwise" linear regression analysis was utilized to estimate the relationship between component income and the set of proxy variables. This approach tests a set of "candidate" proxy variables and selects those whose observed correlation to the dependent variable (component income) was strong enough that it was unlikely to be due to chance alone (i.e. statistically significant).[2] The results of this analysis yielded a regression model for each component of income. The regression models are simple algebraic relationships between the selected proxy variables and the dependent variables:

(1) $$Y_i = a_i + b_{i1} X_{i1} + b_{i2}X_{i2} + ... + b_{in}X_{in}$$

where,

Y_i	is income from component i,
a_i	is the constant (or intercept) calculated by the regression technique for each income component i,
$b_{i1} ... b_{in}$	are the coefficients (fixed numbers) calculated by the regression technique for each proxy variable in each income component i, and
$X_{i1} ... X_{in}$	are the selected proxy variables for income component i.

Utilizing this approach, a total of 39 different proxy variables across the ten income components were identified as having sufficient explanatory power to merit inclusion in the models. Including household size to calculate per capita income brings the total number of required

[2] More formally, the 95% confidence interval on the regression coefficient of the candidate variable had to *exclude* zero for that variable to enter the model.

proxy variables to 40. Table 2 lists these variables and their mean values across the NGO target areas. Each income component has its own algebraic relationship for generating predictions based on the proxy variables; these relationships are the foundation of INCPROX. Table 3 lists the coefficient estimates which describe the algebraic relationship of each proxy variable to each income component and provides an example of how one income component is calculated. See Annex B for more complete statistical output for each regression.

C. INCPROX Lite: A Simpler Alternative

Executing INCPROX requires the collection and processing of a relatively modest amount of data, and provides substantial insight into household income strategies and, over time, of the evolution of the rural economy. Nevertheless, to provide users with a more easily implemented alternative, the principles of INCPROX were used to develop a methodology requiring fewer variables to estimate total and per capita household income. This Total Income Proxy Methodology (INCPROX Lite) does not provide a breakdown of income by component, but the accuracy of its estimates are comparable to those of INCPROX.

To develop INCPROX Lite, a single stepwise linear regression was run utilizing total household income as the dependent variable, and all the candidate proxy variables previously tested in the INCPROX relationships as potential independent variables. Thus, any variable that could have entered into any of the ten INCPROX relationships was given the opportunity to enter into the INCPROX Lite relationship. In fact, only 15 candidate variables entered, meaning that users of INCPROX Lite need utilize only 16 (15 plus household size) variables to develop estimates of total and *per capita* household income.

Table 2. Proxy variables names, descriptions, and means over NGO sample (INCPROX)

Variable Number	Variable Description	Variable Name	Sample Mean
1	Number of types of farm implements owned	NINST	3.197
2	Number of cultivated fields	NMACH	3.196
3	Number of adults resident in the HH (age 10 to 65)	NADULT	3.164
4	Number of food crops cultivated	NCULT_AA	3.694
5	Number of food crops sold	NVEND_AA	0.788
6	Are BEANS the household's key food crop?	KEYFJ	0.006
7	Is MANIOC the household's key food crop?	KEYMD	0.592
8	Is RICE the household's key food crop?	KEYAZ	0.043
9	Is SORGHUM the household's key food crop?	KEYMP	0.069
10	kg MAIZE GRAIN produced	QPROD_MH	184.542
11	Number of other field crops cultivated	NCULT_CC	0.836
12	kg seed cotton produced	QPROD_AL	107.362
13	Number of fresh crops produced	NVERDE	2.726
14	Did the HH sell any fresh production? (0=no, 1=yes)	VEND_VR	0.040
15	Number of vegetables produced	NHORTA	0.533
16	Is ONION the HH's most important vegetable crop? (0=no, 1=yes)	KEY26	0.021
17	Did the HH produce vegetables? (0=no, 1=yes)	HT	0.270
18	Number of fruit trees of all types	NTREE_FT	19.059
19	Number of fish products sold	NVEND_PX	0.117
20	Did the HH produce fish? (0=no, 1=yes)	PX	0.237
21	Number of types of cashew products produced	NCAJU	0.915
22	Did the HH sell cashew? (0=no, 1=yes)	VEND_CJ	0.341
23	Did the HH produce cashew? (0=no, 1=yes)	CJ	0.378
24	Number of goats/sheep owned	NCABRA	1.249
25	Number of pigs owned	NSUINO	1.063
26	Number of chickens/ducks/other birds owned	NAVE	7.694
27	Number of other livestock owned	NOUTRO	0.864
28	Did the HH own any livestock? (0=no, 1=yes)	PEC	0.911
29	Number of formal sector jobs held	NFORMAL	0.055
30	Total number of people working off-farm, any activity	NTF	0.811
31	Did the HH have anyone work off the farm in any activity? (0=no, 1=yes)	TF	0.444
32	Did the HH own and operate a hammer mill? (0=no, 1=yes)	MOAG	0.005
33	Did the HH operate a trading business? (0=no, 1=yes)	COMERCIO	0.196

Variable Number	Variable Description	Variable Name	Sample Mean
34	Number of different MSEs the hh operated	NMSE	1.134
35	Is the HH in Zone 1? (0=no, 1=yes) (Marromeu, Caia, Mutarara, Chemba, Morrumbala, Milange))	ZONE1	0.104
36	Is the HH in Zone 3? (0=no, 1=yes) (Gurue, Gile, Malema, Ribaue, Morrupula, Nampula)	ZONE3	0.400
37	Is the HH in Zone 4? (0=no, 1=yes) (Mogovolas, Meconta, Nacaroa, Erati, Muecate, Mecuburi)	ZONE4	0.297
38	Is the HH in Zone 5? (0=no, 1=yes) (Memba, Nacala-a-Velha)	ZONE5	0.024
39	Is the HH in Zone 6? (0=no, 1=yes) (Nhamatanda, Gorongoza, Gondola)	ZONE6	0.052
40	Mean HH size (all resident members)	NMEM	5.250

D. Statistical Results and Confidence Intervals

INCPROX and INCPROX Lite deliver nearly identical accuracy in their estimates of total household income. INCPROX Lite gives an adjusted R^2 of 0.698, meaning that about 70% of all the variation of calculated income around its mean is explained by the single INCPROX Lite regression model. The standard error of the estimate for INCPROX Lite is 132.94. See Annex C for statistical output from the INCPROX Lite regression.

INCPROX is based on separate regressions for each of 10 different income components. Goodness of fit and standard errors of the regression are available for each of these individual components directly from the separate regression results. To obtain estimates of the goodness of fit of the overall INCPROX approach, and to calculate confidence intervals around the INCPROX estimate of total household income, a different approach was necessary. Essentially this approach consisted of estimating total household income by summing the estimated values of each component of income, then regressing this estimate of total income against calculated income. The adjusted R^2 from this regression is called the INCPROX Pseudo R^2. See Annex Cfor more detail, and statistical results.

The pseudo-R^2 from this approach was 0.698, with a standard error of the estimate of 132.88. Statistical output from the 10 component regressions can be found in Annex B results for the Pseudo-R^2 regression are in Annex C.

Table 3. Relationship between proxy variables and component income

Variable Name	Income Component									
	Food Crops	**Other Crops**	**Fresh production**	**Vegetables**	**Fruit**	**Cashew**	**Fishing**	**Livestock**	**Wage Labor**	**Micro-enterprise**
	----- Regression coefficients -----									
Constant	-45.913	-3.137	-2.236	-5.739	-6.411	-6.548	-4.107	0.000	-1.081	-1.028
NINST	6.339			2.980						
NMACH	4.646					2.144				-4.663
NADULT				-1.269	2.645		0.868			
NCULT_AA	7.181									
NVEND_AA	11.443									
KEYFJ	57.658									
KEYMD	23.092									
KEYAZ	49.344									
KEYMP	45.132									
QPROD_MH	0.138	0.013		-0.007						0.076
NCULT_CC		20.078								
QPROD_AL		0.110								
NVERDE			6.768							
VEND_VR			10.449							
NHORTA				17.264						
KEY26				64.118						
HT				-20.563						
NTREE_FT					0.834					
NVEND_PX							26.846			
PX							7.769			
NCAJU						9.779				
VEND_CJ						16.229				
CJ						-12.420				
NCABRA								8.130		
NSUINO								12.725		

Variable Name	Income Component									
	Food Crops	**Other Crops**	**Fresh production**	**Vegetables**	**Fruit**	**Cashew**	**Fishing**	**Livestock**	**Wage Labor**	**Micro-enterprise**
NAVE								2.048		
NOUTRO								18.376		
PEC								11.946		
NFORMAL									111.558	
NTF									8.502	
TF									38.405	
MOAG										260.119
COMERCIO										5.167
NMSE										21.795
ZONE1			24.374				19.013			
ZONE3			5.165	3.905						
ZONE4	17.612									
ZONE5						17.270				
ZONE6		19.225			30.198				41.190	

NOTES

1. Component income is equal to the sum of each coefficient (found in this table) multiplied by the sample mean (Table 3) for that variable. For example, mean income from wage labor (WLI) across the entire NGO area is:

WLI = -1.081 + 111.558(0.055) + 8.502(0.811) + 38.405(0.444) + 41.19(0.052) = $31.33

2. To calculate this number for a specific NGO, sample means for that NGO would be substituted for the sample means used here

3. Total household income is equal to the sum of income from each component

Confidence intervals can be calculated around the estimates of mean household income using the standard errors of the estimates (SEE) from the overall predicted INCPROX and INCPROX Lite regressions. These confidence intervals will include the true *sample mean* with 95% probability. In other words, these intervals will indicate the reduced precision of using INCPROX or INCPROX Lite as opposed to conducting a full income survey and calculating household income from that sample. The sampling error around calculated income is itself an important and additional source of error that is not treated in the calculations below.

SEE is equal to the standard deviation of the error terms from the regression; it indicates the accuracy with which the regression predicts income for an *individual household*. NGOs are interested in predicting mean income over a sample of households. The accuracy of this prediction depends on the standard error of the mean, which depends on the sample size used in the proxy survey. Specifically, the 95% confidence interval for INCPROX and INCPROX Lite estimates is:

$$\hat{Y} \; +/- \; \frac{1.96\sigma_y}{\sqrt{N}}$$

Where $\hat{Y}$ is the mean household income calculated from INCPROX or INCPROX Lite, N is sample size, and we substitute SEE for σ_y.

Thus, for INCPROX, the 95% confidence interval is given by:

(1) $$\hat{Y} \; +/- \; \frac{1.96(132.88)}{\sqrt{N}}$$

For INCPROX Lite, the 95% confidence interval is:

(2) $$\hat{Y} \; +/- \; \frac{1.96(132.94)}{\sqrt{N}}$$

For sample sizes above 100, these numbers are identical to two decimal places. Table 4 shows the 95% confidence interval resulting from different sample sizes; you can calculate your own interval using equation (1) or (2) and your actual sample size.

Table 4. 95% confidence interval on estimates of total household income from INCPROX and INCPROX Lite, by sample size

Sample Size	INCPROX/INCPROX Lite 95% confidence interval around *sample mean* is $\hat{Y}$ +/- .....[1]
200	18.4
300	15.0
400	13.0
500	11.6
600	10.6
700	9.8

[1] $\hat{Y}$ is estimated total household income derived from your application of INCPROX or INCPROX Lite. The interval includes the *sample mean* with 95% probability. The sampling error of that sample mean is in addition to the error defined in this table.

III. Performance of INCPROX and INCPROX Lite Across Zones

INCPROX and INCPROX Lite give identical estimates of total household income across all target zones, equal to the calculated income from the survey data (US$299.18). Table 5 examines how these two methods perform across zones. The table presents zonal means, and the ranking of those means across the seven zones, of household income, predicted income from INCPROX, and predicted income from INCPROX Lite. It also presents the percentage error of the INCPROX and INCPROX Lite estimates. Perfect performance across zones would mean that each approach exactly predicts calculated income in each zone and, as a result, gives the same correct income ranking of zones. Of course such perfect performance is not to be expected, but Table 5 shows that in general the two approaches do quite well distinguishing income levels by zone. Specifically, INCPROX Lite results in the same income ranking as calculated income (though specific estimates differ), while INCPROX switches zones 3 and 5 but otherwise ranks all zones correctly. Mean absolute error is slightly smaller for INCPROX - 6.2% compared to 6.6% for INCPROX Lite.

Tables 6 and 7 examine the performance of INCPROX from additional perspectives. Table 6 examines how well INCPROX predicts and ranks income components within zones. This is important to NGOs and donor agencies to know at a point in time the relatively importance of different economic activities, and over time as they track the evolution of the economy in an area. To produce the table, each income component was first ranked within each zone, then 1) the number of incorrect rankings, 2) the mean number of incorrect places in the rankings, and 3) the number of times a component is ranked incorrectly by more than one place, are summarized in the table. An example of an incorrect ranking of one place is if food crop income, for example, were actually the third most important income source in a given zone, but was ranked

Table 5. Zone-by-zone comparison of INCPROX and INCPROX Lite in level and ranking of predicted income

	Calculated Income		INCPROX Estimate			INCPROX Lite Estimate		
Zone	Income (US$/hh)	Rank	Income (US$/hh)	Rank	% Error[2]	Income (US$/hh)	Rank	% Error[3]
7	536.35	1	483.03	1	-9.9%	509.98	1	-4.9%
6	482.92	2	464.09	2	-3.9%	425.79	2	-11.8%
1	419.33	3	390.11	3	-7.0%	379.47	3	-9.5%
4	309.61	4	316.16	4	2.1%	306.50	4	-1.0%
2	281.93	5	282.37	5	0.2%	289.88	5	2.8%
3	218.42	6	227.68	7	4.2%	239.20	6	9.5%
5	200.66	7	233.36	6	16.3%	214.00	7	6.6%
All Zones[1]	299.18		299.18			299.18		

[1] Mean is weighted by zone level sample weights
[2] Mean absolute error = 6.23%
[3] Mean absolute error = 6.59%

by INCPROX as second or fourth. This table shows that, while on average each zone has 2.8 income components incorrectly ranked, these errors are generally of only one place. In other words, ranking errors typically involve the switching of adjacent income components. Most and least important components are nearly always correctly identified.

Table 7 examines how well INCPROX ranks income components across zones. For example, which zones have most and least production of non-staple crops, or of cashew, or depend most or least on off-farm earnings? This type of information is important for USAID to know with

Table 6. Summary performance of INCPROX ranking income components within zones

Zone	# of incorrect rankings of income components (out of 10)	Mean # of incorrect places in ranking	# of times an income component is ranked incorrectly by more than one place
1	0	0.0	0
2	2	0.2	0
3	5	0.8	2
4	4	0.4	0
5	9	1.6	4
6	3	0.4	1
7	5	0.7	2
mean	2.8	0.59	1.29

Table 7. Summary performance of INCPROX ranking zones by income component

Income Component	# of incorrect rankings of a zone (out of 7)	Mean # of incorrect places in ranking	# of times a zone is ranked incorrectly by more than one place
Food crops	0	0.00	0
Other crops	2	0.29	0
Fresh production	0	0.00	0
Vegetables	4	0.57	0
Fruit	4	0.86	1
Cashew	2	0.57	1
Fishing	3	0.29	0
Livestock	0	0.00	0
Wage labor	2	0.29	0
Microenterprise	2	0.29	0
Mean	1.9	0.315	0.20

what confidence it can compare NGO estimates from one zone with those from another. To produce this table, zones were first ranked by income component. For example, within the food crop component, zones were ranked according to their mean value of food crop income. The table summarizes how accurately INCPROX and INCPROX Lite predict these rankings by presenting the same indicators as in Table 6: number of incorrect rankings, mean number of incorrect places in ranking, and number of times a zone is ranked incorrectly by more than one place. In general, ranking of zones by income component is quite good; the mean number of incorrect places in the ranking is less than one-third of a place, and in only two cases is a zone ranked incorrectly by more than one place. See Annex D for the complete results used to generate Tables 6 and 7.

IV. Using INCPROX and INCPROX Lite

Using INCPROX or INCPROX Lite to generate estimates of total household income (and ten components in the case of INCPROX) entails three broad steps:

1. Conducting the proxy survey,

2. Processing the data to develop the proxy variables,

3. Using the proxy variables to generate estimates of household income and income component.

A. Conducting the Proxy Survey

Potential users of INPROX or INCPROX Lite typically have a great deal of survey experience, so details of conducting a survey will not be covered in this report. This section will briefly discuss sampling issues, referring the reader to other reports for more detail; it will also briefly review the questionnaires that have been developed for each of the approaches, and discuss when during the year the survey should be done.

Sampling: To report results with greater accuracy and reliability across the different areas where NGOs operate, and to increase the comparability of reporting across NGOs, it would be appropriate that all organizations followed some basic steps in the design of their samples. The suggested steps are:

- In addition to the usual target group, include a comparison group
- Draw samples of similar size in the comparison and target groups;
- Design samples that are probability proportional to size (PPS) in both target and control groups;
- Present results separately for target and control groups

See Benfica and Tschirley (1999) , included here as Annex E , for more detail on how to implement each of these steps. Note that INCPROX and INCPROX Lite can be utilized to generate estimates of household income regardless of the sampling approach used to obtain the data. However, the validity of the estimates will be in part a function of the rigor of the sampling technique applied.

Questionnaires: Michigan State University has developed separate questionnaires for INCPROX and INCPROX Lite. Each is designed to collect the required data as efficiently as possible. See Annex F for copies of each questionnaire. It is strongly recommended that users of INCPROX and INCPROX Lite utilize the respective questionnaire in its entirety. Spreading the required questions through other questionnaires that the NGO is implementing for other purposes will require greater care on the part of the user to avoid errors in extracting only the relevant variables for the proxy estimates. Using a question whose wording is "similar" to one in the proxy questionnaire to substitute for that "similar" question can cause even greater problems, as the question may be understood differently and thus generate different data.

Timing of the survey: The results of any survey are influenced by the timing of that survey. This influence comes primarily through:

1. The ability of respondents to recall information, depending on when in the year it is asked. For example, farmers asked in January to recall production from the previous May will have more difficulty doing so than if they had been asked the same questions in June or July; and

2. The influence of the timing of the survey on the effective period of reference for certain questions. This effect is most often seen in questions about what the farmer has done with the most recent harvest of annual crops. For example, if farmers are asked in June whether they have sold a crop from the harvest in May, the number of positive answers will be fewer than if the same question were asked in November.

The original survey to develop INCPROX and INCPROX Lite was conducted in two rounds, during June/July and November, 1998. Thus, this survey had the advantage of short recall on recent production (during the first round) and more time to get more complete information on crop sales (second round). NGOs will conduct the proxy survey in only one round, and so need to achieve a balance between the two sources of error in deciding on the timing of their own income proxy surveys. A rule of thumb is to attempt to schedule the survey during September - the midpoint between June/July and November. Farmers at this point should still have reasonably accurate recall of maize and cotton production quantities (the only two quantities that enter into INCPROX and INPCROX Lite), and will have had more time to engage in marketing activities than if the survey is conducted in June. Only under extenuating circumstances should the survey be done prior to June 1, as some farmers may not yet have concluded the harvest of maize or cotton.

There will be a downward bias in estimated income from conducting the survey earlier than November (the timing of the final round in the original survey), but this bias is not likely to exceed 1%. This downward bias comes from households having less time to have engaged in marketing activities. INCPROX use four sales variables in its estimates: number of food crops sold (NVEND_AA), did the household sell any fresh crops (VEND_VR), number of fish products sold (NVEND_PX), and did the household sell any cashew products (VEND_CJ). Of these, only NVEND_AA is likely to be affected by the timing of the survey. Any survey done after 1 June will catch virtually all fresh sales, the period of reference for fish sales is 12 months regardless of the timing of the survey, and questions about cashew refer to the last harvest and require only a simple yes/no answer, not a continuous number. Thus, if there had only been one round of the survey and it had been fielded in June, estimated household income would have been only US$3.43, or 1.1 percent, lower than the value we obtained.[3] The closer to November that the survey is conducted, the smaller this error would be.

INCPROX Lite does not use NVEND_AA in its estimates, and thus should not suffer from even this small downward bias as a result of the survey being conducted prior to November.

B. Developing the Proxy Estimate of Household Income

Estimates of household income using INCPROX or INCPROX Lite can be developed with one of two packages developed by MSU: the spreadsheet package with accompanying manual for each, and the SPSS/Windows package. Use of the spreadsheet package is covered in detail in

[3] This number is derived by comparing the value of NVEND_AA using only first round data (0.49) to the value based on both rounds (0.79), and combining this with the value of the estimated regression parameter on NVEND_AA in the food crops regression (11.443): (0.79-0.49)*11.443 = 3.43. On estimated total household income of US$299.18, this comes to 1.1%.

their respective manuals: “Manual for Calculating Total Household Income and Income Components Using the Income Components Proxy Methodology (INCPROX)”, and “Manual for Calculating Total Household Income Using the Total Income Proxy Methodology (INCPROX Lite)”. See Annex G for copies of these manuals.

Access to SPSS for Windows will substantially reduce the amount of data processing work needed to develop the estimates. We recommend that any NGO with access to SPSS/Windows and a data anlayst well-versed in its use utilize the SPSS/Windows package instead of the spreadsheet package. See Annex J for the procedures needed to implement this approach.

Annex A

Prices Used in Valuing Agricultural Production

Crop	Region	mts/kg
maize	Nampula	1,345
maize	Zambezia	1,143
maize	Tete, Sofala, Manica	1,316
beans	Nampula	2,394
beans	Zambezia	2,742
beans	Tete, Sofala, Manica	3,898
manioc	Nampula	1,168
manioc	Zambezia	846
manioc	Tete, Sofala, Manica	688
rice	Nampula	1,481
rice	Zambezia	1,358
rice	Tete, Sofala, Manica	1,295
groundnut	Nampula	2,917
groundnut	Zambezia	1,469
groundnut	Tete, Sofala, Manica	2,144
sweet potato	Nampula	2,908
sweet potato	Zambezia	2,908
sweet potato	Tete, Sofala, Manica	2,908
sorghum	Nampula	1,744
sorghum	Zambezia	1,744
sorghum	Tete, Sofala, Manica	1,850
tobacco	Nampula	8,436
tobacco	Zambezia	8,436
tobacco	Tete, Sofala, Manica	8,436
sunflower	Nampula	1,574
sunflower	Zambezia	1,551
sunflower	Tete, Sofala, Manica	2,143
sesame	Nampula	2,441
sesame	Zambezia	3,679
sesame	Tete, Sofala, Manica	3,514
sugar cane[1]	Nampula	20,833
sugar cane[1]	Zambezia	20,833
sugar cane[1]	Tete, Sofala, Manica	20,833
onion	Nampula	1,744
onion	Zambezia	1,744
onion	Tete, Sofala, Manica	1,850
Pineapple[2]	Nampula	1,000
Pineapple[2]	Zambezia	1,000

Crop	Region	mts/kg
Pineapple[2]	Tete, Sofala, Manica	1,000

[1] Price is per “molho”, a bundle of cane stalks
[2] Price is per pineapple

Annex B

Results of INCPROX Component Regressions

General Note

In most cases we present the results of the full stepwise procedure. Both the Model Summary and Coefficients output include results from every model, including those sub-optimal models prior to the final, optimal model. It is the results of the final model that were used in the development of INCPROX and INCPROX Lite

In the Coefficients output, the column labeled "B" contains the coefficients used in INCPROX and INCPROX Lite. These are identical to those found in Table 3 in the body of the text.

Food Crops Regression

As in all other regressions, a stepwise linear regression approach was utilized in the food crops regression. This regression went through 10 iterations (models) before arriving at the final model. To economize on space, we present below the results of a simple linear regression (SPSS subcommand ENTER) which included all the independent variables which entered in the stepwise approach. Results are identical between the two.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.780[a]	.609	.600	40.8427

a. Predictors: (Constant), ZONE4, NINST, KEYFJ, KEYMP, KEYAZ, NVEND_AA, NMACH, QPROD_MH, NCULT_AA, KEYMD

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-45.913	8.626		-5.322	.000
	QPROD_MH	.138	.007	.721	18.848	.000
	NCULT_AA	7.181	1.948	.133	3.687	.000
	NVEND_AA	11.443	2.300	.157	4.975	.000
	KEYFJ	57.658	25.813	.067	2.234	.026
	KEYMD	23.092	5.597	.176	4.126	.000
	KEYAZ	49.344	10.590	.156	4.659	.000
	KEYMP	45.132	8.679	.177	5.200	.000
	NMACH	4.646	1.574	.107	2.952	.003
	NINST	6.339	1.629	.120	3.890	.000
	ZONE4	17.612	4.488	.125	3.924	.000

a. Dependent Variable: VPROD_AA valor da producao dos alimentos basicos

Other Crops Regression

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.733[a]	.537	.536	41.3659
2	.784[b]	.615	.614	37.7673
3	.789[c]	.622	.620	37.4541
4	.792[d]	.627	.624	37.2603

a. Predictors: (Constant), QPROD_AL

b. Predictors: (Constant), QPROD_AL, NCULT_CC

c. Predictors: (Constant), QPROD_AL, NCULT_CC, QPROD_MH

d. Predictors: (Constant), QPROD_AL, NCULT_CC, QPROD_MH, ZONE6

e. Dependent Variable: VPROD_CC valor da producao de culturas de rendimento

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	15.386	2.023		7.605	.000
	QPROD_AL	.126	.005	.733	22.984	.000
2	(Constant)	.752	2.397		.314	.754
	QPROD_AL	.109	.005	.633	20.485	.000
	NCULT_CC	19.693	2.056	.296	9.581	.000
3	(Constant)	-2.081	2.565		-.811	.418
	QPROD_AL	.110	.005	.642	20.844	.000
	NCULT_CC	19.508	2.039	.293	9.565	.000
	QPROD_MH	1.531E-02	.005	.085	2.936	.003
4	(Constant)	-3.137	2.590		-1.211	.227
	QPROD_AL	.110	.005	.641	20.924	.000
	NCULT_CC	20.078	2.043	.302	9.828	.000
	QPROD_MH	1.312E-02	.005	.073	2.492	.013
	ZONE6	19.225	8.036	.070	2.392	.017

a. Dependent Variable: VPROD_CC valor da producao de culturas de rendimento

Fresh Production Regression

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.342[a]	.117	.115	24.9061
2	.429[b]	.184	.180	23.9709
3	.437[c]	.191	.186	23.8907
4	.444[d]	.197	.190	23.8288

a. Predictors: (Constant), NVERDE

b. Predictors: (Constant), NVERDE, ZONE1

c. Predictors: (Constant), NVERDE, ZONE1, ZONE4

d. Predictors: (Constant), NVERDE, ZONE1, ZONE4, VEND_VR

e. Dependent Variable: VPROD_VR valor da producao em verde

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	1.211	2.767		.438	.662
	NVERDE	7.151	.920	.342	7.769	.000
2	(Constant)	-1.118	2.690		-.416	.678
	NVERDE	7.149	.886	.342	8.070	.000
	ZONE1	22.376	3.670	.259	6.097	.000
3	(Constant)	-1.816	2.703		-.672	.502
	NVERDE	6.778	.902	.324	7.515	.000
	ZONE1	24.084	3.755	.278	6.414	.000
	ZONE4	5.161	2.564	.089	2.013	.045
4	(Constant)	-2.236	2.706		-.826	.409
	NVERDE	6.768	.900	.324	7.523	.000
	ZONE1	24.374	3.748	.282	6.502	.000
	ZONE4	5.165	2.557	.089	2.020	.044
	VEND_VR	10.449	5.704	.077	1.832	.068

a. Dependent Variable: VPROD_VR valor da producao em verde

Vegetable Production Regression

Model Summary[h]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.334	.332	21.3302
2	.656[b]	.431	.428	19.7359
3	.676[c]	.458	.454	19.2870
4	.683[d]	.467	.462	19.1404
5	.691[e]	.477	.472	18.9726
6	.695[f]	.483	.477	18.8846
7	.699[g]	.489	.481	18.8108

a. Predictors: (Constant), KEY26

b. Predictors: (Constant), KEY26, NHORTA

c. Predictors: (Constant), KEY26, NHORTA, HT

d. Predictors: (Constant), KEY26, NHORTA, HT, NINST

e. Predictors: (Constant), KEY26, NHORTA, HT, NINST, QPROD_MH

f. Predictors: (Constant), KEY26, NHORTA, HT, NINST, QPROD_MH, NADULT

g. Predictors: (Constant), KEY26, NHORTA, HT, NINST, QPROD_MH, NADULT, ZONE3

h. Dependent Variable: VPROD_HT valor da producao de hortas

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.910	1.009		2.884	.004
	KEY26	103.919	6.887	.578	15.088	.000
2	(Constant)	-1.250	1.046		-1.195	.233
	KEY26	75.094	7.165	.417	10.480	.000
	NHORTA	8.965	1.019	.350	8.800	.000
3	(Constant)	1.831E-14	1.056		.000	1.000
	KEY26	63.527	7.417	.353	8.565	.000
	NHORTA	17.194	2.005	.672	8.577	.000
	HT	-19.962	4.221	-.340	-4.730	.000
4	(Constant)	-6.545	2.546		-2.571	.010
	KEY26	63.854	7.362	.355	8.674	.000
	NHORTA	17.062	1.990	.667	8.574	.000
	HT	-20.095	4.189	-.342	-4.797	.000
	NINST	2.078	.737	.097	2.821	.005
5	(Constant)	-6.600	2.524		-2.615	.009
	KEY26	66.348	7.344	.369	9.034	.000
	NHORTA	17.049	1.973	.667	8.643	.000
	HT	-20.006	4.152	-.341	-4.818	.000
	NINST	2.544	.746	.119	3.408	.001
	QPROD_MH	-8.15E-03	.003	-.106	-3.005	.003
6	(Constant)	-4.050	2.749		-1.473	.141
	KEY26	64.517	7.354	.359	8.773	.000
	NHORTA	17.298	1.966	.676	8.797	.000
	HT	-20.115	4.133	-.342	-4.867	.000
	NINST	2.950	.764	.138	3.861	.000
	QPROD_MH	-7.54E-03	.003	-.098	-2.778	.006
	NADULT	-1.272	.558	-.081	-2.282	.023
7	(Constant)	-5.739	2.851		-2.013	.045
	KEY26	64.118	7.328	.356	8.750	.000
	NHORTA	17.264	1.959	.675	8.814	.000
	HT	-20.563	4.122	-.350	-4.988	.000
	NINST	2.980	.761	.139	3.915	.000
	QPROD_MH	-6.64E-03	.003	-.086	-2.427	.016
	NADULT	-1.269	.555	-.081	-2.284	.023
	ZONE3	3.905	1.834	.073	2.130	.034

a. Dependent Variable: VPROD_HT valor da producao de hortas

Fruit Production Regression

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.702[a]	.493	.492	41.8509
2	.711[b]	.506	.503	41.3690
3	.715[c]	.511	.508	41.1858

a. Predictors: (Constant), NTREE_FT

b. Predictors: (Constant), NTREE_FT, ZONE6

c. Predictors: (Constant), NTREE_FT, ZONE6, NADULT

d. Dependent Variable: VPROD_FT valor da producao de frutas

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.799	2.112		1.325	.186
	NTREE_FT	.872	.041	.702	21.024	.000
2	(Constant)	1.662	2.114		.786	.432
	NTREE_FT	.849	.042	.684	20.457	.000
	ZONE6	30.172	8.838	.114	3.414	.001
3	(Constant)	-6.411	4.165		-1.539	.124
	NTREE_FT	.834	.042	.671	19.889	.000
	ZONE6	30.198	8.799	.114	3.432	.001
	NADULT	2.645	1.178	.075	2.246	.025

a. Dependent Variable: VPROD_FT valor da producao de frutas

Fish Production Regression

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.621[a]	.385	.384	17.7722
2	.670[b]	.449	.447	16.8428
3	.681[c]	.464	.461	16.6305
4	.684[d]	.468	.464	16.5860

a. Predictors: (Constant), NVEND_PX

b. Predictors: (Constant), NVEND_PX, ZONE1

c. Predictors: (Constant), NVEND_PX, ZONE1, PX

d. Predictors: (Constant), NVEND_PX, ZONE1, PX, NADULT

e. Dependent Variable: VPROD_PX valor da producao de peixe

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1.424	.867		1.641	.101
	NVEND_PX	35.767	2.118	.621	16.887	.000
2	(Constant)	-7.78E-02	.848		-.092	.927
	NVEND_PX	31.075	2.109	.539	14.734	.000
	ZONE1	19.643	2.709	.265	7.250	.000
3	(Constant)	-1.354	.911		-1.487	.138
	NVEND_PX	26.734	2.413	.464	11.077	.000
	ZONE1	19.227	2.678	.260	7.180	.000
	PX	7.697	2.163	.145	3.558	.000
4	(Constant)	-4.107	1.741		-2.358	.019
	NVEND_PX	26.846	2.408	.466	11.150	.000
	ZONE1	19.013	2.673	.257	7.113	.000
	PX	7.769	2.158	.146	3.601	.000
	NADULT	.868	.468	.064	1.853	.065

a. Dependent Variable: VPROD_PX valor da producao de peixe

Cashew Regression

Model Summary[f]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.676[a]	.456	.455	16.6452
2	.689[b]	.474	.472	16.3878
3	.700[c]	.489	.486	16.1656
4	.705[d]	.497	.493	16.0566
5	.711[e]	.506	.500	15.9377

a. Predictors: (Constant), NCAJU

b. Predictors: (Constant), NCAJU, NMACH Quantas machambas a sua familia cultivou a campanha passada?

c. Predictors: (Constant), NCAJU, NMACH Quantas machambas a sua familia cultivou a campanha passada?, ZONE5

d. Predictors: (Constant), NCAJU, NMACH Quantas machambas a sua familia cultivou a campanha passada?, ZONE5, VEND_CJ Vendeu alguma quantidade?

e. Predictors: (Constant), NCAJU, NMACH Quantas machambas a sua familia cultivou a campanha passada?, ZONE5, VEND_CJ Vendeu alguma quantidade?, CJ

f. Dependent Variable: VPROD_CJ valor da producao de caju

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.270	.939		.288	.774
	NCAJU	11.195	.573	.676	19.537	.000
2	(Constant)	-5.951	1.835		-3.243	.001
	NCAJU	10.869	.570	.656	19.061	.000
	NMACH Quantas machambas a sua familia cultivou a campanha passada?	2.040	.520	.135	3.924	.000
3	(Constant)	-7.085	1.836		-3.859	.000
	NCAJU	10.637	.566	.642	18.792	.000
	NMACH Quantas machambas a sua familia cultivou a campanha passada?	2.321	.518	.154	4.478	.000
	ZONE5	18.343	4.982	.126	3.682	.000
4	(Constant)	-6.952	1.825		-3.810	.000
	NCAJU	8.403	1.006	.507	8.353	.000
	NMACH Quantas machambas a sua familia cultivou a campanha passada?	2.094	.522	.139	4.013	.000
	ZONE5	16.740	4.985	.115	3.358	.001
	VEND_CJ Vendeu alguma quantidade?	7.854	2.933	.165	2.678	.008
5	(Constant)	-6.548	1.817		-3.604	.000
	NCAJU	9.779	1.114	.590	8.779	.000
	NMACH Quantas machambas a sua familia cultivou a campanha passada?	2.144	.518	.142	4.137	.000
	ZONE5	17.270	4.951	.118	3.488	.001
	VEND_CJ Vendeu alguma quantidade?	16.229	4.184	.341	3.879	.000
	CJ	-12.420	4.456	-.267	-2.787	.006

a. Dependent Variable: VPROD_CJ valor da producao de caju

Off-farm Labor Regression

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.465[a]	.217	.215	71.6726
2	.573[b]	.328	.325	66.4486
3	.583[c]	.340	.336	65.9185
4	.587[d]	.345	.339	65.7440

a. Predictors: (Constant), NFORMAL

b. Predictors: (Constant), NFORMAL, TF

c. Predictors: (Constant), NFORMAL, TF, ZONE6

d. Predictors: (Constant), NFORMAL, TF, ZONE6, NTF

e. Dependent Variable: VTF valor do trabalho fora da mach

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	23.460	3.423		6.853	.000
	NFORMAL	139.438	12.437	.465	11.211	.000
2	(Constant)	5.579E-14	4.169		.000	1.000
	NFORMAL	115.892	11.846	.387	9.784	.000
	TF	55.793	6.429	.343	8.678	.000
3	(Constant)	-1.078	4.152		-.260	.795
	NFORMAL	109.952	11.930	.367	9.216	.000
	TF	54.156	6.403	.333	8.458	.000
	ZONE6	41.078	14.235	.113	2.886	.004
4	(Constant)	-1.081	4.142		-.261	.794
	NFORMAL	111.558	11.930	.372	9.351	.000
	TF	38.405	10.659	.236	3.603	.000
	ZONE6	41.190	14.198	.113	2.901	.004
	NTF	8.502	4.607	.119	1.846	.066

a. Dependent Variable: VTF valor do trabalho fora da mach

MSE Regression

Model Summary[f]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.406[a]	.165	.163	93.3358
2	.483[b]	.233	.229	89.5702
3	.515[c]	.265	.260	87.7617
4	.545[d]	.297	.291	85.9173
5	.549[e]	.302	.294	85.7284

a. Predictors: (Constant), NMSE

b. Predictors: (Constant), NMSE, QPROD_MH

c. Predictors: (Constant), NMSE, QPROD_MH, COMERCIO

d. Predictors: (Constant), NMSE, QPROD_MH, COMERCIO, MOAG

e. Predictors: (Constant), NMSE, QPROD_MH, COMERCIO, MOAG, NMACH Quantas machambas a sua familia cultivou a campanha passada?

f. Dependent Variable: VMSE valor da renda da micro empresa

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-3.750	5.980		-.627	.531
	NMSE	34.174	3.603	.406	9.485	.000
2	(Constant)	-14.472	5.984		-2.418	.016
	NMSE	30.689	3.501	.365	8.765	.000
	QPROD_MH	7.952E-02	.013	.263	6.328	.000
3	(Constant)	-14.961	5.864		-2.551	.011
	NMSE	22.953	3.844	.273	5.971	.000
	QPROD_MH	7.407E-02	.012	.245	5.986	.000
	COMERCIO	52.365	11.741	.204	4.460	.000
4	(Constant)	-15.640	5.743		-2.723	.007
	NMSE	21.619	3.775	.257	5.727	.000
	QPROD_MH	7.531E-02	.012	.250	6.216	.000
	COMERCIO	55.714	11.518	.217	4.837	.000
	MOAG	258.768	56.950	.180	4.544	.000
5	(Constant)	-1.028	10.206		-.101	.920
	NMSE	21.795	3.768	.259	5.785	.000
	QPROD_MH	7.635E-02	.012	.253	6.308	.000
	COMERCIO	55.167	11.497	.215	4.799	.000
	MOAG	260.119	56.830	.181	4.577	.000
	NMACH Quantas machambas a sua familia cultivou a campanha passada?	-4.663	2.695	-.068	-1.730	.084

a. Dependent Variable: VMSE valor da renda da micro empresa

Livestock Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.662[a]	.439	.437	70.6953
2	.861[b]	.741	.740	48.0633
3	.942[c]	.887	.887	31.7187
4	.970[d]	.941	.941	22.9776
5	.971[e]	.942	.942	22.7659

a. Predictors: (Constant), NOUTRO

b. Predictors: (Constant), NOUTRO, NCABRA

c. Predictors: (Constant), NOUTRO, NCABRA, NSUINO

d. Predictors: (Constant), NOUTRO, NCABRA, NSUINO, NAVE

e. Predictors: (Constant), NOUTRO, NCABRA, NSUINO, NAVE, PEC

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	50.184	3.415		14.694	.000
	NOUTRO	18.535	.983	.662	18.847	.000
2	(Constant)	37.321	2.388		15.627	.000
	NOUTRO	18.804	.669	.672	28.119	.000
	NCABRA	10.115	.439	.550	23.023	.000
3	(Constant)	25.740	1.647		15.631	.000
	NOUTRO	18.112	.442	.647	40.957	.000
	NCABRA	8.515	.297	.463	28.637	.000
	NSUINO	13.338	.550	.393	24.271	.000
4	(Constant)	10.092	1.421		7.103	.000
	NOUTRO	18.465	.321	.660	57.552	.000
	NCABRA	8.154	.216	.443	37.724	.000
	NSUINO	12.834	.399	.378	32.176	.000
	NAVE	2.122	.105	.233	20.273	.000
5	(Constant)	-7.77E-15	3.574		.000	1.000
	NOUTRO	18.376	.319	.657	57.576	.000
	NCABRA	8.130	.214	.442	37.939	.000
	NSUINO	12.725	.397	.375	32.072	.000
	NAVE	2.048	.107	.225	19.231	.000
	PEC	11.946	3.888	.036	3.073	.002

a. Dependent Variable: VPEC valor da producao pecuaria

Annex C

Goodness of Fit and Standard Errors of the Estimate for INCPROX and INCPROX Lite

INCPROX Pseudo-R Squared Regression

INCPROX is based on separate regressions for 10 different income components. Goodness of fit and standard errors of the regression (and thus confidence intervals) are available for each of these individual components directly from the separate regression results. To obtain estimates of the goodness of fit of the overall INCPROX approach, and to calculate confidence intervals around the INCPROX estimate of total household income, the following procedures were utilized:

1. The predicted value of component income for each household from the final model of each of the 10 component regressions was saved.

2. Predicted total household income for each household was calculated as the sum of the predicted values for each of the 10 components.

3. Predicted income from (2) was regressed as the independent variable against the actual household income computed from the survey data.

4. The Adjusted R^2 from this regression is called the INCPROX Pseudo- R^2.

5. The Standard Error of the Estimate from this regression is used to calculate a confidence interval around the INCPROX estimate of total household income.

Results of the pseudo-R^2 regression are presented below.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.836[a]	.699	.698	132.8831

a. Predictors: (Constant), PRE_INC

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-17.430	11.556		-1.508	.132
	PRE_INC	1.058	.033	.836	32.504	.000

a. Dependent Variable: INCOME

INCPROX Lite Regression

INCPROX Lite was estimated using a stepwise linear regression approach, as in INCPROX. The actual stepwise regression went through 15 iterations before arriving at a final solution. To economize on space, below we present output from a simple linear regression (SPSS subcommand ENTER) using all the variables which entered in the stepwise approach. Results are identical to the stepwise approach.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.841[a]	.708	.698	132.9377

a. Predictors: (Constant), NOUTRO, NTREE_FT, NCABRA, MOAG, NFORMAL, NVEND_PX, NINST, NCAJU, COMERCIO, NCULT_CC, NAVE, NSUINO, QPROD_AL, QPROD_MH, NMSE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	17.531	19.097		.918	.359
	NINST	14.457	5.463	.073	2.646	.008
	QPROD_MH	.228	.021	.318	10.681	.000
	NCULT_CC	24.441	7.550	.092	3.237	.001
	QPROD_AL	.105	.020	.153	5.163	.000
	NVEND_PX	57.248	16.492	.093	3.471	.001
	NTREE_FT	.837	.136	.163	6.161	.000
	NFORMAL	84.210	23.628	.094	3.564	.000
	NCAJU	14.242	4.894	.080	2.910	.004
	NMSE	26.519	6.167	.133	4.300	.000
	COMERCIO	43.663	18.538	.072	2.355	.019
	MOAG	531.946	90.024	.156	5.909	.000
	NCABRA	8.106	1.338	.172	6.056	.000
	NSUINO	19.097	2.394	.219	7.977	.000
	NAVE	4.064	.652	.174	6.230	.000
	NOUTRO	21.347	1.991	.297	10.720	.000

a. Dependent Variable: INCOME

Annex D

Complete INCPROX Ranking Performance Results

INCPROX Performance
NGO Data

XVIII. RANKING OF TOTAL INCOME BY ZONE

	Calculated Income		INCPROX Estimate			INCPROX Lite Estimate		
Zone	Income (US$/hh)	Rank	Income (US$/hh)	Rank	% Error[2]	Income (US$/hh)	Rank	% Error[3]
7	536.35	1	483.03	1	-9.9%	509.98	1	-4.9%
6	482.92	2	464.09	2	-3.9%	425.79	2	-11.8%
1	419.33	3	390.11	3	-7.0%	379.47	3	-9.5%
4	309.61	4	316.16	4	2.1%	306.50	4	-1.0%
2	281.93	5	282.37	5	0.2%	289.88	5	2.8%
3	218.42	6	227.68	7	4.2%	239.20	6	9.5%
5	200.66	7	233.36	6	16.3%	214.00	7	6.6%
All Zones[1]	299.18		299.18		0.0%	299.18		0.0%

[1] Mean is weighted by zone level sample weights
[2] Mean absolute error = 6.23%
[3] Mean absolute error = 6.59%

II. RANKING OF COMPONENTS WITHIN ZONES (INCPROX)

Zone 1

Income Component	Calculated Value		Estimated Value		Incorrect Ranking?	# of Incorrect Places
	Value	Rank	Value	Rank		
Livestock	92.30	1	96.64	1		
Food crop	85.83	2	78.88	2		
Microenterprise	72.29	3	57.99	3		
Wage earnings	50.75	4	46.95	4		
Fresh	40.75	5	40.75	5		
Fishing	34.16	6	34.16	6		
Non-food crop	22.61	7	17.74	7		
Fruit	15.52	8	15.78	8		
Vegetables	5.12	9	4.14	9		
Cashew	0.00	10	0.00	10		

Zone 2

Income Component	Calculated Value		Estimated Value		Incorrect Ranking?	# of Incorrect Places
	Value	Rank	Value	Rank		
Food crops	56.51	1	64.92	1		
Livestock	53.56	2	54.74	2		
Microenterprise	47.04	3	43.41	3		
Wage earnings	36.26	4	40.00	4		
Fruit	35.55	5	27.91	5		
Other crops	22.69	6	19.22	7	x	1
Fresh	18.23	7	22.46	6	x	1
Fishing	8.59	8	6.51	8		
Cashew	2.40	9	4.40	9		
Vegetables	0.84	10	0.00	10		

Zone 3

Income Component	Calculated Value		Estimated Value		Incorrect Ranking?	# of Incorrect Places
	Value	Rank	Value	Rank		
Food crops	65.55	1	65.57	1		
Livestock	44.54	2	45.50	2		
Microenterprise	23.66	3	26.66	3		
Fruit	16.98	4	16.13	6	x	2
Other crops	16.63	5	16.43	5		
Fresh	16.27	6	15.92	7	x	1
Wage earnings	12.41	7	18.39	4	x	3
Vegetables	9.25	8	9.25	9	x	1
Cashew	8.92	9	10.23	8	x	1
Fishing	1.77	10	2.62	10		

Zone 4

Income Component	Calculated Value		Estimated Value		Incorrect Ranking?	# of Incorrect Places
	Value	Rank	Value	Rank		
Food crops	77.60	1	77.60	1		
Livestock	77.17	2	77.53	2		
Other crops	51.94	3	55.71	3		
Wage earnings	25.12	4	22.54	5	x	1
Fresh	24.45	5	24.45	4	x	1
Cashew	19.99	6	18.50	7	x	1
Microenterprise	15.29	7	18.85	6	x	1
Fruit	11.60	8	15.35	8		
Vegetables	1.92	9	2.96	9		
Fishing	1.40	10	1.07	10		

Zone 5

Income Component	Calculated Value		Estimated Value		Incorrect Ranking?	# of Incorrect Places
	Value	Rank	Value	Rank		
Food crops	55.26	1	46.76	2	x	1
Livestock	55.14	2	56.05	1	x	1
Cashew	33.75	3	33.75	3		
Fresh	16.96	4	21.99	5	x	1
Wage earnings	16.68	5	32.50	4	x	1
Other crops	5.79	6	3.55	9	x	3
Fruit	5.35	7	7.16	8	x	1
Fishing	4.85	8	3.40	10	x	2
Microenterprise	3.69	9	19.69	6	x	3
Vegetables	2.73	10	7.47	7	x	3

Zone 6

Income Component	Calculated Value		Estimated Value		Incorrect Ranking?	# of Incorrect Places
	Value	Rank	Value	Rank		
Wage earnings	109.85	1	109.85	1		
Food crops	90.84	2	86.00	2		
Microenterprise	80.89	3	71.30	5	x	2
Livestock	77.21	4	82.40	3	x	1
Fruit	75.23	5	75.23	4	x	1
Other crops	33.61	6	33.61	6		
Fresh	10.23	7	6.14	7		
Vegetables	2.70	8	3.36	8		
Fishing	2.17	9	1.44	9		
Cashew	0.19	10	0.00	10		

Zone 7

Income Component	Calculated Value		Estimated Value		Incorrect Ranking?	# of Incorrect Places
	Value	Rank	Value	Rank		
Livestock	139.35	1	110.52	3	x	2
Food crops	131.28	2	142.78	1	x	1
Microenterprise	120.99	3	112.78	2	x	1
Wage earnings	99.11	4	57.56	4		
Other crops	18.91	5	14.38	5		
Fruit	17.37	6	11.81	6		
Fresh	7.54	7	5.50	7		
Vegetables	1.40	8	0.00	9/10	x	1.5
Fishing	0.40	9	0.19	8	x	1
Cashew	0.00	10	0.00	9/10	x	0.5

III. RANKING OF ZONES BY INCOME COMPONENT (INCPROX)

Income Component	Zone	Rank by Calculated Value	Rank by Estimated Value	Incorrect Ranking?	# of Incorrect Places
Food Crops	7	1 (highest)	1		
	6	2	2		
	1	3	3		
	4	4	4		
	3	5	5		
	2	6	6		
	5	7 (lowest)	7		
Other Crops	4	1 (highest)	1		
	6	2	2		
	2	3	3		
	1	4	4		
	7	5	6	x	1
	3	6	5	x	1
	5	7 (lowest)	7		
Fresh	1	1 (highest)	1		
	4	2	2		
	2	3	3		
	5	4	4		
	3	5	5		
	6	6	6		
	7	7 (lowest)	7		
Vegetables	3	1 (highest)	1		
	1	2	3	x	1
	5	3	2	x	1
	6	4	4		
	4	5	5		
	7	6	7	x	1
	2	7 (lowest)	6	x	1

Income Component	Zone	Rank by Calculated Value	Rank by Estimated Value	Incorrect Ranking?	# of Incorrect Places
Fruit	6	1 (highest)	1		
	2	2	2		
	7	3	6	x	3
	3	4	3	x	1
	1	5	4	x	1
	4	6	5	x	1
	5	7 (lowest)	7		
Fishing	1	1 (highest)	1		
	2	2	2		
	5	3	3		
	6	4	5	x	1
	3	5	4	x	1
	4	6	6		
	7	7 (lowest)	7		
Cashew	5	1 (highest)	1		
	4	2	2		
	3	3	3		
	2	4	4		
	6	5	7	x	2
	1	6	5	x	1
	7	7 (lowest)	6	x	1
Livestock	7	1 (highest)	1		
	1	2	2		
	6	3	3		
	4	4	4		
	5	5	5		
	2	6	6		
	3	7 (lowest)	7		

Income Component	Zone	Rank by Calculated Value	Rank by Estimated Value	Incorrect Ranking?	# of Incorrect Places
Wage Earnings	6	1 (highest)	1		
	7	2	2		
	1	3	3		
	2	4	4		
	4	5	6	x	1
	5	6	5	x	1
	3	7 (lowest)	7		
Microenterprise	7	1 (highest)	1		
	6	2	2		
	1	3	3		
	2	4	4		
	3	5	5		
	4	6	7	x	1
	5	7 (lowest)	6	x	1

Annex E

Sampling Guidelines for Income Proxy Surveys

Income Proxy Surveys:
Guidelines for PVO Sampling

By Rui Benfica

and

David L. Tschirley

June 1999
Maputo, Mozambique

1. Introduction

To report results with greater accuracy and reliability across the different areas where PVOs operate, and to increase the comparability of reporting across PVOs, it would be appropriate that all organizations followed, to the extent possible, some basic steps in the design of their samples. The guidelines presented here are aimed at providing PVOs with some key principles to be applied and steps to be followed, in order to improve the quality of their data and reporting, given constraints on time, personnel, and money. These guidelines do not represent USAID "policy", but rather technical suggestions to be applied whenever possible. The closer these guidelines are followed the better the USAID Mission will be able to track performance and impact across the board. Some PVOs are already implementing their surveys using the approach suggested here or a version that is close to it.

This paper is in no way meant to be a comprehensive guide to survey sampling. Consult surey sampling texts for questions which may emerge from reading this paper. A helpful and relatively accessible guide to survey sampling is Graham Kalton, "Introduction to Survey Sampling", Quantitative Applications in the Social Sciences Paper No. 35, Sage Publications. 1985.

2. Basic Principles of the Sampling Approach

The basic principles suggested are:

- Besides the usual target group, include a control group in the sample;
- Draw samples of similar size in the control and target groups;
- Design samples that are probability proportional to size (PPS) in both target and control groups;
- Present results separately for target and control groups

Background and, where relevant, specific steps to follow in applying these principals are presented in the following sections.

2.1. Control and Target Groups

To compare households assisted and not assisted by PVO programs, the sample should include both a target and a control group. The question then is how to develop a definition of these two groups that is *workable* in terms of available time and resources, and *meaningful* in a reporting context. Given the various types of programs in place and the likely indirect impact over undefined areas, there is seldom a straightforward, "correct" definition of the two. Therefore, each PVO needs to develop a definition they consider workable and meaningful, according to their specific circumstances.

In doing so, be clear about the *level* at which you make the definition:

- Defining the two groups at the ***household level*** implies that you can have both target and control households in a single village. This may be most meaningful for interventions which are easily targeted to specific households and which have little spillover or demonstration effect on other households. However, if the intervention does have significant spillover or demonstration effects, then a household level definition may not

be the most meaningful. In any case, a household level definition will require lists of all households stratified (classified) as target and control. Developing such lists may require substantial additional work prior to fielding the survey. Thus, in general, a household level definition will typically require more time and resources - will be less workable - than a village level definition.

- Defining the two groups at the ***village level*** assumes that entire villages are affected by the interventions of the PVO, or not. Such a definition is most meaningful when an intervention has significant spillover or demonstration effects. Preparing the sample using a village level definition may require significantly less time and effort than using a household level definition, so in general the village level approach is the most workable.

Since many PVO interventions have spillover and demonstration effects, defining target and control groups using a village level approach will typically provide the best combination of workability and meaning for PVO impact surveys. If a PVO already has lists of target and control (participant and non-participant) households for its villages, and if it is confident that its interventions have few spillover or demonstration effects, then it might consider using a household level approach. The discussion in this paper is oriented towards a village level approach.

2.2. Sample Size

The size of the sample must be decided at three levels:

6. The total sample size in each group - target and control. We will refer to this number as n.
7. The distribution of that sample over villages i.e., the number of villages in each group (v).
8. The number of households to interview in each village (h).

Total sample size in each group: The primary purpose of defining control and target groups is to compare the means of selected variables across those groups. For example, you may want to know whether the maize yield in the target group is significantly higher than in the control group. This comparison of means across groups is most statistically efficient when the samples in the two groups are of equal size. Allowing the sample size in the groups to differ, for example by allowing each sample to be proportional to the size of its group, reduces the efficiency of the comparisons to be made. Thus, your design should call for total samples of equal size in the target and control groups. Given the practical problems of fielding surveys, actual sample sizes might differ slightly, but these differences should be minimized.

But what size should the sample be? There is no easy answer to this question for various reasons. First, a theoretically recommended sample size is a function of the desired level of accuracy, which in turn depends on the variance in the variable to be estimated. In this case, we have many variables to be estimated, each with different and unknown variances. Second, the sample size is a function of available time and resources, particularly human and financial.

However, as a rule of thumb, having a sample size of at least 200 households, preferably more, in each group is desirable.[4]

Number of villages and number of households in each village: The determination of number of villages and number of households per village can proceed in two ways:

- If you first decide how many villages to work in, then the number of households to be interviewed in each village is determined by n/v, where n is the total sample size and v is the number of villages you have decided to visit. For example, if desired sample size in each group is 250 and you decide that you have the resources to work in 20 villages in each group, then the number of households to be interviewed in each village is 250/20 = 12.5. You would interview 13 households per village and achieve a sample size of n = 260.

- Alternatively, you can first decide how many households to interview in each village. In this case, the number of villages is determined by n/h, where h is the number of households you wish to interview in each village. If your desired sample size is again 250 and you decide to interview 15 households per village, you will need to work in 250/15 = 16.67 villages. Rounding, you would work in 17 villages, achieving a sample size of n = 255.

A common approach would be to decide that you want to spend one day conducting interviews in each selected village. You would then estimate how many interviews you can conduct in one day: that number becomes h. You then calculate v (number of villages in each group) as n/h.

It should be clear from this discussion that the determination of v and h is based primarily on pragmatic considerations. However, a statistical principle to keep in mind is that, for a given n (total sample size), the efficiency of your estimates will generally be greater if you have ***more villages and fewer households per village***.[5] Thus, subject to your constraints ot time, money, and personnel, you should spread your sample over as many villages as possible.

---

[4] As an example of the results you can expect from a sample of 200, if you are estimating maize yield with a *simple random sample* of 200, and your sample mean is 1,200 kg/ha, with a sample standard deviation of 500 kg/ha (variance of 250,000; these would not be atypical numbers), then a 95% confidence interval for that mean is 1,200 +/- 1.96 * sqrt(250,000/200) = 1,200 +/-35. In other words, you have 95% confidence that the true mean is between 1,165 kg/ha and 1,235 kg/ha. Note again that this calculation is based on a simple random sample. The approach suggested here (called *cluster sampling*) results in wider confidence intervals for a given sample size (its use is nevertheless often justified because it is a much more workable design than a simple random sample). The increase in the confidence interval with cluster sampling depends principally on the number of households interviewed per village (for a given total sample size n, fewer households per village - and more villages - gives a better estimate) and the degree of homogeneity within villages. It would not be unusual for the confidence interval in a cluster sample design to be 2-3 times larger than the interval from a simple random sample. This means that if the same data were obtained from the procedures recommended here (same sample size, mean, and standard deviation), the 95% confidence interval on maize yield could be as large as 1,200 +/- 105 kg. Note also that this example ignores issues of non-normal distribution of yield data, a treatment of which is beyond the scope of this paper.

[5] This statement assumes that households are more similar to their neighbors in the same village than they are to households in other villages. This assumption is generally appropriate in rural Africa.

2.3. Selection of Villages and Households

Once you have determined n, v, and h, you need to choose the actual villages in which to work, and the households to interview.

Selection of Villages: The sampling method recommended in this case is the selection of villages with **Probability Proportional to Size (PPS)**. This means that the probability of a village being selected is proportional to the size of that village. Thus, for example, a village with 400 households would have twice the probability of being selected of a village with only 200 households. Why use PPS and not another sampling method? First, PPS eliminates the need for weighting the results in the analysis by ensuring that each household has the same probability of being selected. Second, PPS allows one to draw equally sized samples in each village, regardless of its size. Having the same number of households to be surveyed in each village makes it easier to program the fieldwork – assuming that interviews take approximately the same time in each village.

With n, v, and h defined, the next step consists of classifying and listing by target and control group, all villages which could potentially be included in the survey. You must then obtain data on the population (or number of households) of each village. The selection of villages is done separately in the target and the control group, using the same procedures. PPS sampling is straightforward and described in the hypothetical example below.

The first step in this method is to list the villages and their total population. If population numbers are not available, you can use the total number of households in each village. You must then construct the cumulative ranges (cr) and probabilities (p) for each group. The example here is for the target area group of villages and assumes that the number of villages to be selected is 4. For the control group of villages, the same method is followed.

Table 1: Organization of village data for PPS sampling

Villages	# of HHs (*)	Cumulative Range (cr)	Probability (p)
Josina Machel	100	1-100	100/1500
1 de Maio	120	101-220	120/1500
3 de Fevereiro	220	221-440	220/1500
Agostinho Neto	80	441-520	80/1500
Lipilichi	160	521-680	160/1500
Napipine	240	681-920	240/1500
25 de Junho	90	921-1010	90/1500
Spartan	100	1011-1110	100/1500
Ujamaa	80	1111-1190	80/1500
Buckeye	310	1191-1500	310/1500

(*) Can also be in terms of total population.

There are 1500 households in the population to be sampled. The cumulative range (cr) keeps track of the interval of numbered households in each village. The order in which the villages appear in the list is not important. In this list, Josina Machel Village has the first 100 households, 1 de Maio has households 101-220, and so on. The probability (p) for each village is simply the number of households in that village divided by the total number of households in the survey area. The villages with greater numbers of households have larger probabilities of selection.

You may choose a sample of 4 villages in two ways: using a random number table, or using *systematic sampling*. Using a random number table, you select 4 random numbers between 1 and 1500 from the table. This can also be done using a computer application – simple spreadsheets have a statistical function for these purposes. Suppose that the numbers selected in this random selection are 20, 530, 1099 and 1420. These numbers should be located in the cr column and the villages corresponding to those cumulative range intervals will constitute the sample: Josina Machel, Lipilichi, Spartan and Buckeye. These villages have been selected with probabilities proportional to their numbers of households.

An alternative approach is to use systematic sampling. This consists in dividing the total number of households (1500) by the number of villages to be sampled (4) to get the sampling interval (375). A random number between 0 and 375 is chosen randomly from the random number table to determine the first village selection. If the random number selected is 150, then 1 de Maio is the first village. Then 375 is added to the random number to give 525, making Lipilichi the second selection, adding 375 again gives 900, making Napipine the third selection. Finally, adding another 375 gives 1,275 and makes Buckeye the last village selected.

Selection of Households: Once villages have been selected, for each of them the entire list of households is necessary – no detailed data on the household are needed, except for the name of the household head that identifies it. The actual selection of households is done using **Systematic Sampling (SS)**. First, number all households in the village from 1 to n. The total number of households in each village j is THHj. Then, the actual selection process is made using lists for each village with the following steps for each village:

Definition of Sampling Intervals (SI). SI for Village j (SIj) is given by SIj= THHj/h. Note that, while h is the same across all villages sampled, SIj between villages varies because of the differences in their sizes. If h is 10 in each village, and THH for a given village j is 120, then the SIj is 120/10 = 12.

For each Village, the first household to be selected in its list is obtained by choosing a random number between 0 and its SIj (a simple scientific calculator or spreadsheet can be used to select random numbers). The corresponding household in the list of numbered households is picked. For example, with a selection interval of 12, the first random number between 0 and 12 might be 4: the fourth household on your list is selected.

Then the process continues by systematically picking up every "+ SIj" household in the list until the desired number of households for the Village is reached. This process allows for a selection of households uniformly distributed along the Village list. In our example, you would select households 4, 16, 28, 40, 52, 64, 76, 88, 100, and 112, for a total of the desired 10 households.

2.4 Summary of Sampling Procedures

In summary, we are suggesting that you engage in the following steps to design and execute your sample:

1. Define target and control groups. You should probably do this at the village level, rather than the household level. There is no single correct way to define these groups, so think through the issues and present your reasoning in the results report.

2. Define the total sample size in each group. Try to do at least 200 in each group, more if your resources permit. Design the sample to deliver equal sample sizes in each group, recognizing that final numbers may differ slightly.

3. Determine the number of villages (v) and the number of households per village (h) that you will interview. The final decision is based on pragmatic considerations (time, personnel, money), but remember that, for any given n, your statistical estimates will be more accurate if you spread your sample over more villages, implying fewer household interviews in each village; 200 interviews spread over 10 villages (20/village) are better than 200 spread over 5 villages (40/village). ***Conduct the survey in as many villages as your resources of time, personnel, and money will permit.***

4. Select v villages with ***probability proportional to size*** (PPS). See the discussion above on how to do this.

5. Select h households in each village using systematic sampling. See above.

2.4. Reporting of Results

In reporting your results, follow these principles:

1. Present clearly your definition of target and control groups. Recognize the limitations of your definition (none is ever perfect), but highlight the strengths and explain why you made the decision you did.

2. Present a clear but concise description of your sampling strategy in each group.

3. Whenever relevant, present results broken down by control and target groups.

4. In your breakdowns, indicate the number of observations that contributed to any given mean. This will assist the reader in assessing the numbers you present. For example, if you have a sample size of 200 in your target group but have a table reporting results for target households in one specific area, the number of observations for that table will be less than 200. Include this number in each of the cells of your tables.

5. Remember that most statistical packages assume simple random sampling when conducting statistical tests (e.g., for a difference in means). We have seen that the cluster sampling approach advocated here results in wider confidence intervals than does simple random sampling. As a result, for a given n it will be more difficult to conclude that there are statistically significant differences in means or proportions. Put another way, if you present the results of unadjusted statistical tests, you will sometimes be concluding that there are statistically significant differences when, in fact, there are not.

   If you want to present statistical tests, you need to adjust them to take into account the *sample design effect*. Consult a sampling text such as Kalton for how to do this.

Annex F

INCPROX and INCPROX Lite Questionnaires

Inquérito sobre Indicadores de Rendimento Familiar

Income Components Proxy Methodology (INCPROX)

AVISO

O Sr(a). tem direito a não participar nesta entrevista. A sua participação é inteiramente voluntária. No entanto vale a pena indicar que, caso do Sr(a). participar na entrevista, toda a informação recolhida será completamente confidencial - em nenhuma circunstancia o seu nome será associado a nenhuma resposta.

Provincia	PROV
Distrito	DIST
Aldeia	ALD
Número do AF	AF
Nome do Chefe do AF	
Nome da pessoa entrevistada	
Nome do inquiridor	INQ
Nome do supervisor	SUP

Prov____ Dist _____ Ald _____ Af _____

I. MEMBROS RESIDENTES

- Gostariamos perguntar algumas coisas sobre cada pessoa que costumava comer aqui nesta casa durante os últimos 12 meses

Tabela 1. Pessoas que regularmente tomavam as refeições nesta casa durante os últimos 12 meses

Nome	**No.**	**Relação ao Chefe** 1 chefe 2 esposa/o 3 filha/o 4 pai/mãe 5 outra fam. 6 outro (esp)	**Sex**o 1 m 2 f	**Idade**	Durante os últimos 12 meses, esta pessoa fez trabalho a **CONTA PROPRIA**? 0 Não 1 Sim	Durante os últimos 12 meses, esta pessoa fez **TRABALHO FORA DA MACHAMBA**? 0 Não 1 Sim
NOME	**MEM**	**I1**	**I2**	**I3**	**I4**	**I5**
	1					
	2					
	3					
	4					
	5					
	6					
	7					
	8					
	9					
	10					

Prov____ Dist _____ Ald _____ Af _____

II. PRODUÇÃO AGRICOLA

AF1 _____ Incluindo todas as culturas, quantas machambas cultivou este agregado durante a última campanha?

- Quais das seguintes culturas produziu/vendeu o seu agregado durante os últimos 12 meses? (**Só produção da última campanha**)
-

Tabela 2. Culturas alimentares, outras culturas, e produção em verde

Culturas Alimentares			**Outras Culturas**		**Produção em Verde**		
Cultura	O seu agregado **PRODUZIU** esta cultura alimentar durante os últimos 12 meses? 0 Não 1 Sim	O seu agregado **VENDEU** esta cultura alimentar durante os últimos 12 meses? 0 Não 1 Sim	Outra Cultura	O seu agregado **PRODUZIU** esta outra cultura durante a última campanha? 0 Não 1 Sim	Cultura em Verde	O seu agregado **PRODUZIU** esta cultura **em verde** durante os últimos 12 meses? 0 Não 1 Sim	O seu agregado **VENDEU** esta cultura **em verde** durante os últimos 12 meses? 0 Não 1 Sim
CULTALIM	**II1**	**II2**	**CULTOUTR**	**III1**	**CULTVERD**	**IV1**	**IV2**
1 Milho			1 Algodão		1 Maçaroca		
2 Feijoes			2 Batata doce		2 Feijão verde		
3 Mandioca seca			3 Tabaco		3 Mandioca fresca		
4 Arroz			4 Girassol		4 Folhas de mand.		
5 Amendoim			5 Gergelim		5 Amend. em verde		
6 Mapira			6 Cana Doce		6 Batata doce		
7 Mexoeira			7 Ananás				
			Outro (esp.)				

Prov____ Dist _____ Ald _____ Af _____

AF2 Se produziu **milho**, quanto produziu?

AF2a _____ quantidade

AF2b _____ Unidade

3	kilo	50	saco de 50 kilos
5	lata de 5 litros	90	saco de 90 kilos
10	lata de 10 litros	100	saco de 100 kilos
20	lata de 20 litros	999	outro (especificar)_____

AF3 _____ Esta quantidade, estava em grão ou em espiga?

1 milho em grão
2 milho em espiga

AF4 _____ Qual cultura alimentar lhe deu **MAIOR PRODUÇÃO** durante a última campanha?

1	milho	5	amendoim
2	feijoes	6	mapira
3	mandioca	7	mexoeira
4	arroz		

AF5 Se produziu **algodão**, quanto produziu? **(Algodão carroço)**

AF5a _____ quantidade

AF5b _____ Unidade

3	kilo	
50	saco de 50 kilos	
90	saco de 90 kilos	
999	outro (especificar)	_____

AF6 _____ O seu agregado produziu alguma HORTICOLA durante os últimos 12 meses?

0 Não
1 Sim

Prov____ Dist _____ Ald _____ Af _____

AF7 _____ O seu agregado produziu alguma FRUTA durante os últimos 12 meses?

0 Não
1 Sim

AF8 _____ O seu agregado produziu CAJU durante os últimos 12 meses?

0 Não
1 Sim

Tabela 3. Hortícolas, frutas, e cajú

Hortícolas			**Frutas**		**Cajú**		
Hortícola	O seu agregado **PRODUZIU** esta hortícola durante os últimos 12 meses? 0 Não 1 Sim	O seu agregado **VENDEU** este hortícola durante os últimos 12 meses? 0 Não 1 Sim	Fruta	Quantos **ARVORES** deste tipo possui o seu agregado?	Cajú	O seu agregado **PRODUZIU** este produto de cajú durante os últimos 12 meses? 0 Não 1 Sim	O seu agregado **VENDEU** este produto de cajú durante os últimos 12 meses? 0 Não 1 Sim
HORTIC	**V1**	**V2**	**FRUTA**	**VI1**	**CAJU**	**VII1**	**VII2**
1 Feijões (só folhas)			1 Banana		1 Castanha		
2 Tomates			2 Manga		2 Amendoa		
3 Alface			3 Laranja		3 Fruta seca		
4 Abóbora			4 Papaia		4 Fruta fresca		
5 Piri-piri			5 Limão		5 Sumo de cajú		
6 Alho			6 Abacate		6 Aguardente de		
7 Cebola			7 Goiaba				
8 Repolho			8 Tangerina				

Prov____ Dist _____ Ald _____ Af _____

Hortícolas			**Frutas**		**Cajú**		
Hortícola	O seu agregado **PRODUZIU** esta hortícola durante os últimos 12 meses? 0 Não 1 Sim	O seu agregado **VENDEU** este hortícola durante os últimos 12 meses? 0 Não 1 Sim	Fruta	Quantos **ARVORES** deste tipo possui o seu agregado?	Cajú	O seu agregado **PRODUZIU** este produto de cajú durante os últimos 12 meses? 0 Não 1 Sim	O seu agregado **VENDEU** este produto de cajú durante os últimos 12 meses? 0 Não 1 Sim
HORTIC	**V1**	**V2**	**FRUTA**	**VI1**	**CAJU**	**VII1**	**VII2**
9 Pimentão			9 Maçanica				
10 Pepino			Outro				
11 Couve							

AF9 _____ Qual hortícola lhe deu maior produção durante os últimos 12 meses?

1 Feijões (só folhas) 4 Abóbora 7 Cebola10 Pepino
2 Tomates 5 Piri-piri 8 Repolho11 Couve
3 Alface 6 Alho 9 PimentãoOutro (especificar)_____

AF10 ____ Alguma pessoa no seu agregado dedicou-se ao **PESCADO** durante os últimos 12 meses?

0 Não
1 Sim

AF11 ____ O seu agregado tem **ANIMAIS**?

0 Não
1 Sim

AF12 ____ O seu agregado tem **INSTRUMENTOS DE PRODUCAO**?

0 Não
1 Sim

Prov____ Dist _____ Ald _____ Af _____

Tabela 4. Pescado, pecuaria e instrumentos de produção

Pescado			**Pecuaria**		**Instrumentos de Produção**	
Peixe	O seu agregado **PESCOU/ PRODUZIU** este tipo de peixe durante os últimos 12 meses? 0 Não 1 Sim	O seu agregado **VENDEU** este tipo de peixe durante os últimos 12 meses? 0 Não 1 Sim	**Tipo de animal**	**Quantos tem agora?**	**INSTRUMENTO**	O seu agregado possui **pelo menos um** deste instrumento? 0 Não 1 Sim
PEIXE	**VIII1**	**VIII2**	**PEC**	**IX1**	**INST**	**X1**
1 Peixe fresco			1 cabrito/ovelha		1 Enxadas	
2 Peixe seco			2 porcos		2 Catanas	
3 Camarão			3 galinhas/patos/ outras aves		3 Machados	
4 Carangueijo			4 Outros (especificar)		4 Pás	
5 Lagosta					5 Ancinhos	
6 Outro (esp.)					6 Foices	
					7 Limas	
					8 Charruas de Tracção	
					9 Carroça	
					10 Motobomba	

Prov____ Dist _____ Ald _____ Af _____

III. TRABALHO FORA DA MACHAMBA E A CONTA PROPRIA

AF13 _____ Alguma pessoa do seu agregado trabalhou fora da machamba (recebendo em dinheiro ou em espécie) durante os últimos 12 meses?
0 Não
1 Sim

AF14 _____ Alguma pessoa membro do seu agregado trabalhou a conta própria durante os últimos 12 meses?
0 Não
1 Sim

Tabela 5. Trabalho fora da machamba e actividades a conta própria

Trabalho fora da machamba		**Actividades a conta própria**	
Tipo de trabalho fora	Número de membros residentes que participaram na actividade durante os últimos 12 meses	Tipo de actividade a conta própria	Algum membro deste agregado fez este tipo de trabalho a conta própria durante os últimos 12 meses? 0 Não 1 Sim
TRABFORA	**XI1**	**CONTPROP**	**XII1**
Trabalho a tempo inteiro		1 Ser dono **e operar** uma MOAGEM	
1 Machamba da companhia		2 Compra/venda de qualquer producto	
2 Fábrica da companhia		3 Artesanato	
3 Função pública		4 Venda de bebida	
4 Professor		5 Carpintaria	
5 Outro trabalho ***a tempo inteiro*** (especificar)		6 Curandeiro	
Trabalho NAO a tempo inteiro		7 Alfaiate	
6 Machamba de um vizinho		8 Reparador de bicicletas	
7 Machamba de um privado		9 Fabrico de cestos/esteiras	
8 Outro (especificar)		10 Pedreiro	
		11 Lenhador/carvoeiro	
		12 Oleiro	

Prov____ Dist _____ Ald _____ Af _____

Inquérito sobre Indicadores de Rendimento Familiar

Total Income Proxy Methodology (INCPROX Lite)

AVISO

O Sr(a). tem direito a não participar nesta entrevista. A sua participação é inteiramente voluntária. No entanto vale a pena indicar que, caso do Sr(a). participar na entrevista, toda a informação recolhida será completamente confidencial - em nenhuma circunstancia o seu nome será associado a nenhuma resposta.

Provincia	PROV
Distrito	DIST
Aldeia	ALD
Número do AF	AF
Nome do Chefe do AF	
Nome da pessoa entrevistada	
Nome do inquiridor	INQ
Nome do supervisor	SUP

Prov____ Dist _____ Ald _____ Af _____

I. MEMBROS RESIDENTES

- Gostariamos perguntar algumas coisas sobre cada pessoa que costumava comer aqui nesta casa durante os últimos 12 meses

Tabela 1. Pessoas que regularmente tomavam as refeições nesta casa durante os últimos 12 meses

Nome	No.	Relação ao Chefe 1 chefe 2 esposa/o 3 filha/o 4 pai/mãe 5 outra fam. 6 outro (esp)	Sexo 1 m 2 f	Idade
NOME	**MEM**	**I1**	**I2**	**I3**
	1			
	2			
	3			
	4			
	5			
	6			
	7			
	8			
	9			
	10			

II. PRODUÇÃO AGRICOLA

AF1 Produziu milho durante a última campanha agrícola?

0 Não
1 Sim

AF2 Se produziu milho, quanto produziu?

AF2a _____ quantidade

AF2b _____ Unidade

3	kilo	50	saco de 50 kilos
5	lata de 5 litros	90	saco de 90 kilos
10	lata de 10 litros	100	saco de 100 kilos
20	lata de 20 litros	999	outro (especificar) _____

AF3 _____ Esta quantidade, estava em grão ou em espiga?

1 milho em grão
2 milho em espiga

- Quais das seguintes culturas não alimentares produziu o seu agregado durante os últimos 12 meses? (**Só produção da última campanha**)

Tabela 2. Culturas não alimentares

Culturas nao alimentares	
Cultura	O seu agregado **PRODUZIU** esta outra cultura durante a última campanha? 0 Não 1 Sim
CULTOUTR	**II1**
1 Algodão	
2 Batata doce	
3 Tabaco	
4 Girassol	
5 Gergelim	
6 Cana Doce	
7 Ananás	
Outro (esp.)	

AF4 Se produziu algodão, quanto produziu? **(Algodão carroço)**

AF4a _____ quantidade

AF4b _____ Unidade 3 kilo
50 saco de 50 kilos
90 saco de 90 kilos
999 outro (especificar) _____

- Quantas árvores de fruta a familia possui?

Tabela 3. Arvores de fruta

. Fruta	Quantas **ARVORES** deste tipo possui o seu agregado?
FRUTA	**III1**
1 Banana	
2 Manga	
3 Laranja	
4 Papaia	
5 Limão	
6 Abacate	
7 Goiaba	
8 Tangerina	
9 Maçanica	
Outro (especificar)	

Prov____ Dist _____ Ald _____ Af _____

- Quais dos seguintes tipos de **PEIXE** e **CAJU** produziu/vendeu o seu agregado durante os últimos 12 meses?

Tabela 4. Peixe e cajú

	Peixe		**Cajú**	
Peixe	O seu agregado **PESCOU/ PRODUZIU** este tipo de peixe durante os últimos 12 meses? 0 Não 1 Sim	O seu agregado **VENDEU** este tipo de peixe durante os últimos 12 meses? 0 Não 1 Sim	Cajú	O seu agregado **PRODUZIU** este produto de cajú durante os últimos 12 meses? 0 Não 1 Sim
PEIXE	**IV1**	**IV2**	**CAJU**	**V1**
1 Peixe fresco			1 Castanha	
2 Peixe seco			2 Amendoa	
3 Camarão			3 Fruta seca	
4 Carangueijo			4 Fruta fresca	
5 Lagosta			5 Sumo de cajú	
6 Outro (esp.)			6 Aguardente de cajú	

Tabela 5. Pecuaria e instrumentos de produção

Pecuaria		**Instrumentos de Produção**	
Tipo de animal	**Quantos tem agora?**	**INSTRUMENTO**	O seu agregado possui **pelo menos um** deste instrumento? 0 Não 1 Sim
PEC	**VI1**	**INST**	**VII1**
1 boi/vaca		1 Enxadas	
2 cabrito/ovelha		2 Catanas	
3 porcos		3 Machados	
4 galinhas/patos/ outras aves		4 Pás	
5 Outros (especificar)		5 Ancinhos	
		6 Foices	
		7 Limas	
		8 Charruas de Tracção	
		9 Carroça	
		10 Motobomba	

Prov____ Dist _____ Ald _____ Af _____

III. TRABALHO FORA DA MACHAMBA

AF5 _____ Alguma pessoa do seu agregado trabalhou fora da machamba (recebendo em dinheiro ou em espécie) durante os últimos 12 meses?

0 Não
1 Sim

AF6 _____ Alguma pessoa membro do seu agregado trabalhou a conta própria durante os últimos 12 meses?

0 Não
1 Sim

Tabela 6. Trabalho fora da machamba e actividades a conta própria

Trabalho fora da machamba		**Actividades a conta própria**	
Tipo de trabalho fora	Número de membros residentes que participaram na actividade durante os últimos 12 meses	Tipo de actividade a conta própria	Algum membro deste agregado fez este tipo de trabalho a conta própria durante os últimos 12 meses? 0 Não 1 Sim
TRABFORA	**VIII1**	**CONTPROP**	**IX1**
Trabalho a tempo inteiro		1 Ser dono **e operar** uma MOAGEM	
1 Machamba da companhia		2 Compra/venda de qualquer producto	
2 Fábrica da companhia		3 Artesanato	
3 Função pública		4 Venda de bebida	
4 Professor		5 Carpintaria	
5 Outro trabalho ***a tempo inteiro*** (esp.)		6 Curandeiro	
		7 Alfaiate	
		8 Reparador de bicicletas	
		9 Fabrico de cestos/esteiras	
		10 Pedreiro	
		11 Lenhador/carvoeiro	
		12 Oleiro	
		Outro (especificar)	

Annex G

INCPROX and INCPROX Lite Manuals
(Spreadsheet Version)

Manual for Calculating Total Household Income and Income Components Using the Income Components Proxy Methodology (INCPROX)

Michigan State University Food Security Project

June 1999

Introduction

The Michigan State University Food Security Project has substantially improved the income proxy methodology over what it was in 1997/98. NGOs are now in a position to use the new Income Components Proxy Methodology (INCPROX) to estimate not just total income, but 10 different components of income, and to do so with greater accuracy than in the past. Thus, compared to the approach used in 1997/98, INCPROX provides a substantially richer set of results, much greater insight into the evolution of household income strategies and of the rural economy in general, and greater confidence in the results.

Executing INCPROX requires the collection of somewhat more data than did the previous methodology. INCPROX utilizes 44 variables, while the previous approach required 23. The basic data approach is the same, meaning that both methodologies rely predominantly on yes/no questions which are easy to ask and easy to process. We believe that the modest increase in time of collection and processing that INCPROX requires is more than offset by 1) the increased accuracy of the results, and 2) the fact that INCPROX provides estimates of 10 different components of income in addition to total income. Nevertheless, to provide NGOs with a more easily implemented alternative, we have used principles of the INCPROX approach to develop a methodology that uses only 17 variables to estimate total and per capita household income. This Total Income Proxy Methodology (INCPROX Lite) does not provide any breakdown of income by component, and may be somewhat less accurate than INCPROX. However, we believe that it too is a substantial improvement over the method used in 1997, and provides NGOs with a statistically defensible, low-cost alternative to INCPROX. Implementing INCPROX Lite is documented in "Manual for Calculating Total Household Income Using the *Total Income Proxy Methodology* (INCPROX Lite)", accompanied by the QuattroPro spreadsheet file INCP Lite-CALC.WB3.

This present manual accompanies 1) the INCPROX questionnaire and 2) the QuattroPro file INCP-CALC.WB3 (this file can also be utilized in Microsoft Excel). Together, these three documents provide the details you will need to implement this new Income Components Proxy Methodology.

The Questionnaire

After the cover page with identifier variables, the questionnaire for the income components proxy methodology begins with a simple demographic table to identify all resident members' age, sex, and relationship to the head of household. To assist in obtaining later information about wage and microenterprise earnings, this table also asks which members participated in these activities. Following this demographic table, the questionnaire consists primarily of a series of tables, one for each of the 10 income components. In nearly all cases, these tables ask two yes/no questions about a series of items - "did you produce this item?", and "did you sell this item?". For example, the Food Crop table asks these two yes/no questions about seven crops that we have defined as the "food crop" basket. These questions will be easy to ask, easy to record (0=no, 1=yes), and easy to clean.

The principal exceptions to this general pattern of yes/no questions are:

1. **Quantity produced of maize (questions AF2a, AF2b, AF3) and cotton (AF5a, AF5b)**: Agricultural production is a large proportion of total income for most

households, and this production can vary substantially from year-to-year with weather and pest conditions. Thus, to obtain acceptably accurate estimates of household income from year-to-year with a proxy approach, it is necessary to include quantity variables which can themselves serve as proxies for production of the whole range of crops that a household may cultivate. We have chosen maize and cotton to fulfill these roles, based on their importance in most households' "portfolio" of crops, and the relative ease of collecting data on quantities produced.

For both these sets of quantity questions, we provide detailed instructions in Annex A (Developing the Proxy Variables from the Proxy Questionnaire) about how to convert the answers into kilograms of each crop.

2. **Most important food (AF4) and vegetable (AF9) crops:** Econometric analysis found that these variables were helpful in predicting, respectively, the food crop and vegetable crop components of income. These questions are straightforward, asking the interviewee to indicate which crop from a list of crops gave the household the most production.

3. **Number of each type of livestock:** Analysis indicated that knowing the *number* of each type of livestock was substantially more useful than knowing simply if the household owned or did not own each type. The livestock table asks for present ownership numbers of five types of livestock.

4. **Number of members involved in different types of wage labor activities**: As in the livestock analysis, knowing the *number* was substantially more useful than knowing only *whether* anyone was involved in each activity.

After collecting your data, you must follow a three-step process to generate estimates of total and per capita household income and its 10 components:

1. **Enter and clean the data** you have collected in the software package of your choice. We will refer to the data you actually collect as the *questionnaire variables.*

2. **Perform selected transformations** on the questionnaire variables to develop the *proxy variables*; these proxy variables are the variables actually used in the calculation of income and its components.

3. **Develop a household level electronic file** containing these proxy variables. The file will consist of one row for each HH in your sample, one column for each of the 44 proxy variables, and additional columns as needed for the identifier variables you use to uniquely identify each household.

4. **Calculate the mean over your sample of each of these 44 proxy variables**, and

5. **Enter these mean values in the "Data" page of the QuattroPro spreadsheet INCP-CALC.WB3.**

The next sections provide details on steps 2-5.

Transforming the Questionnaire Variables, Developing the Household Level Electronic File, and Calculating Sample Means

This file must contain one row for every household in your sample, and one column for each of the 44 proxy variables that are used in calculating the income components. You will also want each row (each household) to have identifier variables such as province, district, village, and household number. These identifier variables may be different for different NGOs. If you have four identifier variables for each household, you will need 44+4=48 total variables (columns) in your file.

The data in this household level file are *derived from* the data you collect, but they are not *identical* to that data; you must perform certain transformations on the questionnaire variables to generate the proxy variables which are actually used in the calculation of household income and its components. In making the transformations on the questionnaire variables to create the proxy variables, you must refer to the tables in Annex I: Developing the Proxy Variables from the Proxy Questionnaire. These tables link the proxy variables to the questionnaire variables, give needed detail on how to use the questionnaire variables to calculate the proxy variables, and provide information on the acceptable range for individual values of proxy variables (the values in the data file you are developing) and the probable range for the sample means that you will calculate. Take some time now to look at some of these tables to familiarize yourself with the type of information they provide.

Most of the transformations are quite straightforward. For example, the value of proxy variable NINST (# of types of farm implements owned) for a given household is obtained by summing the values in the principal column (VII1) of the Farm Implements table. Some of the proxy variables are identical to questionnaire variables: for example, proxy variable NMACH (# of cultivated fields) is equal to questionnaire variable AF1.

The development of proxy variables QPROD_MH (kg of maize grain produced) and QPROD_AL (kg of seed cotton produced) involves a somewhat higher level of complexity than the others, because rural households often report production in non-standard units, while the income calculations require data in kilograms. These conversions are not, however, especially difficult, and Annex I provides the detail and examples needed to make them.

In calculating the sample means, it is imperative that every cell in the data file have a value. Specifically, cells where a value of zero defines the situation of that household must have the value zero entered, and not be left blank. For example, a household that did not produce maize (or cotton) must have zero as the value for QPROD_MH (or QPROD_AL); these cells must not be left blank. Likewise, a household that reported no fruit production must have values of zero entered for each of the fruit component proxy variables (NFRUTA, NTREE_FT, FT). ***Do not leave any cells blank!***

Once you have ensured that all cells have values, calculating the mean of each variable over all values is straightforward, though the specific commands will vary with different software packages. After calculating these means, you are ready to enter them in the spreadsheet file INCP-CALC.WB3, and obtain your estimates for the 10 household income components and total household and per capita income.

Obtaining the Income Estimates

The file INCP-CALC.WB3 contains 12 pages: one Summary page, one Data page, and one page for each of the 10 income components. For your purposes, however, you need only deal with 2 pages: Summary and Data.

The Data page: This is the only page where you will enter data. All other pages (and all sections of this page not requiring data entry) are protected so that you cannot change them. Please do not remove this protection, as doing so may result in alterations to the parameter and calculation sections of the spreadsheet that could invalidate your income estimates.

This page contains four columns: Variable Number, Variable Description, Variable Name, Sample Means. You must enter the sample means that you calculated in the previous steps in the shaded cells of this latter column. Once you have entered and checked these values, and saved the file, your work is done - estimated income and its 10 components will be automatically calculated in the Summary page.

The Summary page: This page lists the 10 income components, reports the 1998 US$ value of income and the percentage income share from each, and computes total and estimated per capita household income.

Appendix I: Developing the Proxy Variables from the Proxy Questionnaire

Variables Used in Several Calculations

There are three variables which are used in the calculation of several income components:

Proxy Variable Number (1st column of "Data" page)	**Proxy Variable Description** (2nd column of "Data" page)	**Proxy Variable Name** (3rd column of "Data" page)	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual household level values***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
1	# of types of farm implements owned	NINST	Sum all the values in column X1, Instrumentos de Produção section of Tabela 4 (pescado/pecuaria/instrumentos)	X1	<=10	Provincial means in 1998 were 2.5-3.5 with no clear geographical pattern. Could rise over time.
2	# of cultivated fields	NMACH	This Proxy Variable is equal to questionnaire variable AF1	AF1	1 - high	1998 zone means were 2.9-5.1. Values lowest in Manica, Sofala, Zambezi Valley; highest from Central Zambêzia north.
3	# of resident adults in HH	NADULT	The number of entries in Tabela 1 between 10 and 65 years of age, inclusive	I3, Tabela 1 (Membros Residentes)	0 - high	1998 zone means were highest in Manica (4.0), lowest in Coastal Nampula (2.3). Others fluctuated around 3.0.

Variables Used Only in the Food Crop Calculation

"Food crops" for this analysis are maize, beans, manioc, rice, groundnuts, sorghum, and millet. All other crops are allocated to the "other field crop" category. To obtain the estimated total value of production of these food crops for all households in your sample, you have to enter the sample mean of 10 additional Proxy Variables:

Proxy Variable Number (1st column of "Data" page)	**Proxy Variable Description**	**Proxy Variable Name**	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual proxy variables***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
4	# of food crops cultivated	NCULT_AA	Sum of all values in column II1, Culturas Alimentares section of Tabela 2	II1	0-7	1998 zone means were 2.8 (Manica/Sofala) to 4.2 (N. Zambêzia, Nampula)
5	# of food crops sold	NVEND_AA	Sum of all values in column II2, Culturas Alimentares section of Tabela 2	II2	0-7	**Zones 1,2,5,6,7**: 0.7-1.1 **Zones 3,4**: 1.4-1.6
6	Are beans the household's key food crop?	KEYFJ	Equal to 1 if questionnaire variable AF4 is equal to 2 (feijão). Otherwise, KEYFJ is equal to zero.	AF4	0 or 1	***Must*** lie between 0 and 1. Very unlikely to exceed 0.10 in any zone. Typically will lie between 0 and 0.02.
7	Is manioc the household's key food crop?	KEYMD	Equal to 1 if questionnaire variable AF4 is equal to 3 (mandioca). Otherwise, KEYMD is equal to zero.	AF4	0 or 1	***Must*** lie between 0 and 1. **Manica**: Could be below 0.1 **Sofala/Zambezi Valley**: 0.1 - 0.3 **Other zones**: As high as 0.8
8	Is rice the household's key food crop?	KEYAZ	Equal to 1 if questionnaire variable AF4 is equal to 4 (arroz). Otherwise, KEYAZ is equal to zero.	AF4	0 or 1	***Must*** lie between 0 and 1. As high as 0.10 in some zones; normally around .02 to .03
9	Is groundnut the household's key food crop?	KEYAM	Equal to 1 if questionnaire variable AF4 is equal to 5 (amendoim). Otherwise, KEYAM is equal to zero.	AF4	0 or 1	***Must*** lie between 0 and 1. Typically around 0.02. Could be higher in Coastal Nampula (Zone 5)

Proxy Variable Number (1st column of "Data" page)	Proxy Variable Description	Proxy Variable Name	Procedures to Calculate this variable at the *household level*	Questionnaire Variables Utilized	Acceptable range for *individual proxy variables*	Probable range for *proxy variable sample means*
10	Is sorghum the household's key food crop?	KEYMP	Equal to 1 if questionnaire variable AF4 is equal to 6 (mapira). Otherwise, KEYMP is equal to zero.	AF4	0 or 1	***Must*** lie between 0 and 1. **N. Zambêzia/ Nampula**: Up to 0.1 **Elsewhere**: Typically below 0.05.
11	Is millet the household's key food crop?	KEYMX	Equal to 1 if questionnaire varable AF4 is equal to 7 (mexoeira). Otherwise, KEYMX is equal to zero.	AF4	0 or 1	***Must*** lie between 0 and 1. 0 - 0.01 everywhere
12	kg maize grain produced	QPROD_MH	QPROD_MH = Quantity*weight*processing factor **Quantity** = AF2a **Weight:** *If quantity reported "em grão" (AF3=1)*: Kilo = 1.0 lata 5 litros = 4.375 lata 10 litros = 8.75 lata 20 litros = 17.5 saco de 50 kg = 50 saco de 90 kg = 90 saco de 100 kg = 100 *If quantity reported "em espiga" (AF3=2)*: Kilo = 1.0 lata 5 litros = 1.5 lata 10 litros = 3.0 lata 20 litros = 6.0 saco de 50 kg = 17.15 saco de 90 kg = 30.87 saco de 100 kg = 34.30 **Processing factor**: *If quantity reported "em grão" (AF3=1)*: 1.0 *If quantity reported "em espiga" (AF3=2)*: 0.75	AF2a, AF2b, AF3	0 - high. Will vary by zone, weather, expected prices. Highest in 1998 sample was 10 tons.	Will vary by zone, weather, expected prices.

Proxy Variable Number (1st column of "Data" page)	Proxy Variable Description	Proxy Variable Name	Procedures to Calculate this variable at the *household level*	Questionnaire Variables Utilized	Acceptable range for *individual proxy variables*	Probable range for *proxy variable sample means*
			Examples *30 latas de 10 litros em espiga*: AF2a=30, AF2b=10, AF3=2. Quantity = AF2a = 30 Weight = 3.0 kg (lata de 10 litros em espiga) Processing factor = 0.75 (espiga) QPROD_MH = (30)(3)(.75) = 67.5 kg *30 latas de 10 litros em grão*: AF2a=30, AF2b=10, AF3=1. Quantity = AF2a = 30 Weight = 8.75 (lata de 10 litros em grão) Processing factor = 1.0 (grão) QPROD_MH = (30)(8.75)(1.0) = 262.5 kg *10 sacos de 50 kg em espiga*: AF2a=10, AF2b=50, AF3=2. Quantity = AF2a = 10 Weight = 17.15 (saco de 50 kg em espiga) Processing factor = 0.75 (espiga) QPROD_MH = (10)(17.15)(.75) = 128.6 kg. *10 sacos de 50 em grão*: AF2a=10, AF2b=50, AF3=2. Quantity = AF2a = 10 Weight = 50 (saco de 50 kg em grão) Processing factor = 1.0 (grão) QPROD_MH = (10)(50)(1.0) = 500 kg.			

Other Field Crops

This category includes all other field crops that the household produced. The data are collected in the Outras Culturas section of Tabela 2 in the questionnaire. To get an estimate of household income from this source, you must enter two additional variables:

Proxy Variable Number (1st column of "Data" page)	**Proxy Variable Description**	**Proxy Variable Name**	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual proxy variables***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
13	# of other crops cultivated	NCULT_CC	Sum all the values in column III1, Outras Culturas section of Tabela 2	III1	<=7, because only 7 crops are asked about in the questionnaire	
14	kg of cotton produced	QPROD_AL	QPROD_AL = quantity*weight **Quantity** = AF5a **Weight:** Empirically we find that farmers always report this production in kilos, saco de 50, or saco de 90, with saco de 90 being the most common. *If kilo (AF5b=3)*: 1.0 kg *If saco de 50 kg (AF5b=50)*: 36 kg *If saco de 90 kg (AF5b=90)*: 64.8 kg **Example** 3 sacos de 90 kg = (3)(64.8) = 194.4 kg	AF5a, AF5b	0 - high. Will vary by zone, weather, expected prices. Highest in 1998 sample was 4.2 tons.	Will vary by zone, weather, expected prices.

Fresh Production

This category includes all production in a fresh state of the crops covered in the previous two sections - "food crops" and "other crops" refer to dry production of those crops. To estimate mean household income from fresh production for households in your sample, you need to enter two additional variables:

Proxy Variable Number (1st column of "Data" page)	**Proxy Variable Description**	**Proxy Variable Name**	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual proxy variables***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
15	# of fresh crops cultivated	NVERDE	Sum all the values in column IV1, Produção em Verde section of Tabela 2	IV1	<=5	Range across zones in 1998 was 1.0 - 4.5
16	did the household sell any fresh production?	VEND_VR	Equal to 0 if the HH did not sell any fresh production (all values in column IV2 = 0). Equal to 1 if the HH sold *any* fresh production (at least one value = 1 in column IV2)	IV2	0 or 1	Range across zones in 1998 was 0.015 - 0.095 (i.e., between 1.5% and 9.5% of HHs sold fresh production, on average, across zones)

Vegetable Production

To get an estimate of the mean value of vegetable production for households in your sample, you need to enter 5 additional variables:

Proxy Variable Number (1st column of "Data" page)	**Proxy Variable Description**	**Proxy Variable Name**	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual proxy variables***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
17	# of vegetables produced	NHORTA	Sum all the values in column V1, hortícolas section of Tabela 3	V1	<=11	1998 zone means were 0.25 to 0.79.,
18	Is onion the HH's key vegetable crop?	KEY26	Equal to 1 if AF9 = 7. Otherwise = 0.	AF7	0 or 1	1998 zone means were 0 to 0.034

Proxy Variable Number (1st column of "Data" page)	Proxy Variable Description	Proxy Variable Name	Procedures to Calculate this variable at the *household level*	Questionnaire Variables Utilized	Acceptable range for *individual proxy variables*	Probable range for *proxy variable sample means*
19	Did the HH produce any vegetables?	HT	This Proxy Variable is equal to the value of questionnaire variable AF6. **Note** that if AF6=1, then Tabela 3 must indicate positive numbers of at least one type of vegetable in column V1.	V1	0 or 1	1998 zone means were 0.17 to 0.36 (i.e., on average between 17% and 36% of HHs produced vegetables)

Fruit Production

To obtain the estimated mean value of fruit production for households in your sample, you must enter three additional variables:

Proxy Variable Number (1st column of "Data" page)	Proxy Variable Description	Proxy Variable Name	Procedures to Calculate this variable at the *household level*	Questionnaire Variables Utilized	Acceptable range for *individual proxy variables*	Probable range for *proxy variable sample means*
A	**B**	**C**	**D**	**E**	**F**	**G**
20	total # of fruit trees	NTREE_FT	Sum all the values in column VI1, Fruta section of Tabela 3	VI1	0 to high	1998 zone means were 0.03 to 0.17.

Fish

Proxy Variable Number (1st column of "Data" page)	Proxy Variable Description	Proxy Variable Name	Procedures to Calculate this variable at the *household level*	Questionnaire Variables Utilized	Acceptable range for *individual proxy variables*	Probable range for *proxy variable sample means*
A	B	C	D	E	F	G
21	# types of fish products sold	NVEND_PX	Sum all the values in column VIII2, Peixe section of Tabela 4	VIII2	0 - 6	1998 zone means ranged from 0.015 (Manica) to 0.54 (Zambezi Valley)
22	did the HH engage in fishing?	PX	This Proxy Variable is equal to the value of questionnaire variable AF10. **Note** that if AF10=1, then Tabela 4 must indicate positive numbers of at least one type of fish in column VIII1	AF10	0 or 1	1998 zone means ranged from 0.13 to 0.51. Highest values were in Zones 1 (Zambezi Valley) and 2 (Central Zambezia), each around 0.50. Manica was lowest at 0.06

Cashew Products

Proxy Variable Number (1st column of "Data" page)	Proxy Variable Description	Proxy Variable Name	Procedures to Calculate this variable at the *household level*	Questionnaire Variables Utilized	Acceptable range for *individual proxy variables*	Probable range for *proxy variable sample means*
A	B	C	D	E	F	G
23	# of types of cashew products produced	NCAJU	Sum all the values in column VII1, Cajú section of Tabela 3.	VII1	0 - 6	**1998 zone means:** *Zones 2-5 (all areas from Central Zambezia north)*: 1.3 to 2.7 *Zones 1, 6, 7*: 0.015 to 0.13
24	did the HH sell any cashew product?	VEND_CJ	Equal to 0 if the HH did not sell any cashew product (all values in column VII2 = 0). Equal to 1 if the HH sold *any* cashew product (at least one value = 1 in column VII2)	VII2	0 - 6	**1998 zone means:** *Zones 3-5*: 0.33 to 0.72 *Zone 3*: 0.095 (reflecting poor state of cashew orchard in Central Zambezia) *Zones 1, 6, 7*: Zero in each

Proxy Variable Number (1st column of "Data" page)	Proxy Variable Description	Proxy Variable Name	Procedures to Calculate this variable at the *household level*	Questionnaire Variables Utilized	Acceptable range for *individual proxy variables*	Probable range for *proxy variable sample means*
25	did the HH produce any cashew product?	CJ	This Proxy Variable is equal to the value of questionnaire variable AF8. **Note** that if AF8=1, then Tabela 3 must indicate positive numbers of at least cashew product in column VII1.	AF8	0 or 1	**1998 zone means:** *Zones 2-5 (all areas from Central Zambezia north)*: 0.43 to 0.83 (43% to 83% of HHs produced cashew) *Zones 1, 6, 7*: 0.015 to 0.078

Livestock

You must add 5 additional variables to obtain the estimate of mean household income from livestock:

Proxy Variable Number (1st column of "Data" page)	Proxy Variable Description	Proxy Variable Name	Procedures to Calculate this variable at the *household level*	Questionnaire Variables Utilized	Acceptable range for *individual proxy variables*	Probable range for *proxy variable sample means*
A	**B**	**C**	**D**	**E**	**F**	**G**
26	# of goats/sheep owned at time of interview	NCABRA	Equal to the number in the SECOND row (PEC=2) of column IX1, Pecuaria section of Tabela 4	IX1	0 - high. Highest value in 1998 was 130. Any value >10 should be checked.	**1998 zone means:** *Zones 1 and 7 (Zambezi Valley & Manica)*: Nearly 7 ***Other zones:*** 0.23 to 1.8
27	# of pigs owned at time of interview	NSUINO	Equal to the number in the THIRD row (PEC=3) of column IX1, Pecuaria section of Tabela 4	IX1	0 to high. Highest value in 1998 was 30. Any value >10 should be checked.	**1998 zone means:** *Zone 5*: Zero *All others:* 0.54 to 1.9

Proxy Variable Number (1st column of "Data" page)	**Proxy Variable Description**	**Proxy Variable Name**	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual proxy variables***	**Probable range for *proxy variable sample means***
28	# of chickens/ducks/ other birds owned at time of interview	NAVE	Equal to the number in the FOURTH row (PEC=4) of column IX1, Pecuaria section of Tabela 4	IX1	0 to high. Highest value in 1998 was 80. Any value >20 should be checked.	**1998 zone means:** *Zones 1, 6, 7*: 12 - 16 *All others:* 3.6 to 9.5
29	# of other livestock owned at time of interview	NOUTRO	Equal to the number in the SIXTH row (PEC=all other values) of column IX1, Pecuaria section of Tabela 4	IX1	0 - high. Highest value in 1998 was 112. Any value >20 should be checked.	**1998 zone means:** *Zones 4-7*: 3.0 to 4.6 *All others:* < 1.4
30	did the HH own any livestock of any kind?	PEC	This proxy variable is equal to the value of questionnaire variable AF11. **Note** that if AF11=1, then Tabela 4 must indicate positive numbers of at least one animal in column IX1.	AF11	0 or 1	**1998 zone means:** Above 0.80 everywhere.

Wage Labor Earnings

Estimated mean household income from wage labor is obtained by enter three additional variables:

Proxy Variable Number (1st column of "Data" page)	**Proxy Variable Description**	**Proxy Variable Name**	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual proxy variables***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
31	# of "formal sector" jobs held by resident members	NFORMAL	This is the total number of ***formal sector jobs*** held in the family. In Tabela 5 (Trabalho fora e actividades a conta própria) sum all values of XI1 for which TRABFORA is <= 5.	XI1	0 - high. Cannot be greater than number of resident members.	**1998 zone means:** *Zones 2, 6, 7*: 0.24 - 0.36 *Zones 1, 3, 4, 5*: 0.014 - 0.086
32	total # of resident members working off-farm, in any activity	NTF	This is the total number of jobs ***of any kind*** held in the family. In Tabela 5, sum all values in column of XI1, regardless of the value of TRABFORA	XI1	0 - high. Cannot be greater than number of resident members.	**1998 zone means:** *Zones 1, 6, 7*: 1.19 - 2.03 *Zones 2-5*: 0.67 - 0.87
33	did the HH have anyone working off the farm in any type of activity?	TF	Equal to the value of questionnaire variable AF13. **Note** that if AF13=1, then Tabela 5 must indicate positive numbers of at least one type of wage work in column XI1.	AF13	0 or 1	**1998 zone means:** *Zones 1, 6, 7*: 0.61 - 0.73 (61% to 73% of HHs had off-farm wage income). *Zones 2-5*: 0.33 - 0.49 (33% to 49%)

Microenterprise Earnings

To obtain estimated mean household income from microenterprises, enter three additional variables:

Proxy Variable Number (1st column of "Data" page)	**Proxy Variable Description**	**Proxy Variable Name**	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual proxy variables***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
34	Did the HH own and operate a hammer mill?	MOAG	Equal to 1 if the FIRST row of column XIII, Tabela 5 ("ser dono e operar uma MOAGEM") is equal to 1 Equal to zero otherwise	XII1	0 or 1	**1998 zone means:** Near zero in all zones, because so few HHs own and operate a hammer mill. Any sample mean above 0.02 should be checked.
35	Did the HH operate a trading business? (Buying and selling of any good for profit)	COMERCIO	Equal to 1 if the SECOND row of column XIII, Tabela 5 ("compra/venda de qualquer produto (agrícola ou não agrícola)") is equal to 1 Equal to zero otherwise	XII1	0 or 1	**1998 zone means:** Ranged from 0.10 to 0.45
36	# of different MSEs the HH operated	NMSE	This proxy variable is equal to the value of questionnaire variable AF14. **Note** that if AF14=1, then Tabela 5 must indicate positive numbers of at least one type of microenterprise activity in column XII1.	AF14	0 or 1	**1998 zone means:** Ranged from 0.36 to 0.79

Zone Variables

Zone variables are used in some of the calculations to adjust values in zones where that component of income was especially high or low, beyond that explained by the other variables. Zone variables are:

Proxy Variable Number (1st column of "Data" page)	**Proxy Variable Description**	**Proxy Variable Name**	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual proxy variables***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
37	Is the household in Zone 1?	ZONE1	Equal to 1 if the household is in the districts of Marromeu, Caia, Mutarara, or Chemba (**and others in this area if NGOs expand to them**) Equal to zero otherwise	DIST	0 or 1	
38	Is the household in Zone 2?	ZONE2	Equal to 1 if the household is in the districts of Mag. da Costa, Namacurra, Nicoadala, Morrumbala, Milange (**and others in this area if NGOs expand to them**) Equal to zero otherwise	DIST	0 or 1	
39	Is the household in Zone 3?	ZONE3	Equal to 1 if the household is in the districts of Gurue, Gile, Malema, Ribaue, Morrupula, Nampula (**and others in this area if NGOs expand to them**) Equal to zero otherwise	DIST	0 or 1	

Proxy Variable Number (1st column of "Data" page)	**Proxy Variable Description**	**Proxy Variable Name**	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual proxy variables***	**Probable range for *proxy variable sample means***
40	Is the household in Zone 4?	ZONE4	Equal to 1 if the household is in the districts of Mogovolas, Meconta, Nacaroa, Erati, Muecate, Mecuburi (**and others in this area if NGOs expand to them**) Equal to zero otherwise	DIST	0 or 1	
41	Is the household in Zone 5?	ZONE5	Equal to 1 if the household is in the districts of Memba, Nacala-a-Velha (**and others in this area if NGOs expand to them**) Equal to zero otherwise	DIST	0 or 1	
42	Is the household in Zone 6?	ZONE6	Equal to 1 if the household is in the districts of Nhamatanda, Gorongoza, Gondola (**and others in this area if NGOs expand to them**) Equal to zero otherwise	DIST	0 or 1	
43	Is the household in Zone 7?	ZONE7	Equal to 1 if the household is in the districts of Manica, Barue, Guro, Sussundenga (**and others in this area if NGOs expand to them**) Equal to zero otherwise	DIST	0 or 1	

Household Size

Household size is used to convert estimated total household income into estimated per capita household income.

Proxy Variable Number (1st column of "Data" page)	**Proxy Variable Description**	**Proxy Variable Name**	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual proxy variables***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
44	How many people are in this household?	NMEM	Equal to the number of rows filled-out in Tabela 1 (Resident members of the household)	MEM	1 to high	**1998 zone means:** Ranged from 4.3 to 5.8 for Zones 1-6. 6.99 in Zone 7 (Manica).

Manual for Calculating Total Household Income Using the Total Income Proxy Methodology (INCPROX Lite)

Michigan State University Food Security Project

April 1999

Introduction

The Michigan State University Food Security Project has developed two new approaches to estimating rural household income through a proxy methodology. Both represent substantial improvements over the methodology utilized in 1997. This manual explains the operation of the Total Income Proxy Methodology (INCPROX Lite). Using this approach, NGOs can estimate total and per capita household income in their project areas utilizing only 17 proxy variables.

INCPROX Lite is a simplified version of the Income Components Proxy Methodology (INCPROX), which utilizes 44 proxy variables to estimate not only total and per capita household incomes, but ten different income components. Thus, compared to INCPROX Lite, INCPROX requires somewhat more data collection and processing, but provides a substantially richer set of results, and much greater insight into the evolution of household income strategies and of the rural economy in general. Both INCPROX Lite and INCPROX require vastly less data collection and processing than would direct estimates of income, and we expect that each will also provide more accurate estimates than will the proxy methodology utilized in 1997.

This manual should be used in conjunction with 1) the INCPROX Lite questionnaire and 2) the QuattroPro file INCP Lite-CALC.WB3 (this file can also be utilized in Microsoft Excel). Together, these three documents provide the details you will need to implement this new Total Income Proxy Methodology.

The Questionnaire

After the cover page with identifier variables, the questionnaire for the income components proxy methodology begins with a simple demographic table to identify all resident members' age, sex, and relationship to the head of household. Following this demographic table, the questionnaire consists primarily of a series of tables for those aspects of the household's economic behavior found to be most useful in predicting total household income. In most cases, these tables ask one yes/no question about a series of items - often "did you produce this item?", sometimes complemented by "did you sell this item?". For example, the "Culturas Alimentares" table asks the production question about seven crops that we have defined as the "non-food crop" basket. The Peixe section of the "Peixe e Cajú" table asks both the production and the sales questions. These questions will be easy to ask, easy to record (0=no, 1=yes), and easy to clean.

The three exceptions to this general pattern of yes/no questions are:

1. **Quantity produced of maize (questions AF2a, AF2b, AF3) and cotton (AF4a, AF4b)**: Agricultural production is a large proportion of total income for most households, and this production can vary substantially from year-to-year with weather and pest conditions. Thus, to obtain acceptably accurate estimates of household income from year-to-year with a proxy approach, it is necessary to include quantity variables which can themselves serve as proxies for production of the whole range of crops that a household may cultivate. We have chosen maize and cotton to fulfill

these roles, based on their importance in most households' "portfolio" of crops, and the relative ease of collecting data on quantities produced.
For both these sets of quantity questions, we provide detailed instructions in Annex A (Developing the Proxy Variables from the Proxy Questionnaire) about how to convert the answers into kilograms of each crop.

2. **Number of members involved in different formal sector wage labor activities**: Analysis showed that knowing the *number* was substantially more useful than knowing only *whether* anyone was involved in each of these activities.

3. **Number of different types of livestock owned**: As in labor activities, knowing the *number* of each type owned was more useful than knowing only *whether* each type was owned.

After collecting your data, you must follow a three-step process to generate estimates of total and per capita household income:

1. **Enter and clean the data** you have collected in the software package of your choice. We will refer to the data you actually collect as the *questionnaire variables.*

2. **Perform selected transformations** on the questionnaire variables to develop the *proxy variables*; these proxy variables are the variables actually used in the calculation of total and per capita household income.

3. **Develop a household level electronic file** containing these proxy variables. The file will consist of one row for each HH in your sample, one column for each of the 17 proxy variables, and additional columns as needed for the identifier variables you use to uniquely identify each household.

4. **Calculate the mean over your sample of each of these 17 proxy variables**, and

5. **Enter these mean values in the QuattroPro spreadsheet INCP Lite-CALC.WB3.**

The next sections provide details on steps 2-5.

Transforming the Questionnaire Variables, Developing the Household Level Electronic File, and Calculating Sample Means

This file must contain one row for every household in your sample, and one column for each of the 17 proxy variables that are used in calculating the income components. You will also want each row (each household) to have identifier variables such as province, district, village, and household number. These identifier variables may be different for different NGOs. If you have four identifier variables for each household, you will need 17+4=21 total variables (columns) in your file.

The data in this household level file are *derived from* the data you collect, they are not *identical* to that data; you must perform certain transformations on the questionnaire

variables to generate the proxy variables which are actually used in the calculation of household income and its components. In making the transformations on the questionnaire variables to create the proxy variables, you must refer to the tables in Annex I: Developing the Proxy Variables from the Proxy Questionnaire. These tables link the proxy variables to the questionnaire variables, give needed detail on how to use the questionnaire variables to calculate the proxy variables, and provide information on the acceptable range for individual values of proxy variables (the values in the data file you are developing) and the probable range for the sample means that you will calculate. Take some time now to look at these tables to familiarize yourself with the type of information they provide.

Most of the transformations are quite straightforward. For example, the value of proxy variable NCULT_CC (# of types of non-food crops produced) for a given household is obtained by summing the values in the production column (II1) of the Culturas Não Alimentares table. NTREE_FT is obtained by summing the values in column III1 of the Arvores de Fruta table.

The development of proxy variables QPROD_MH (kg of maize grain produced) and QPROD_AL (kg of seed cotton produced) involves a somewhat higher level of complexity than the others, because rural households often report production in non-standard units, while the income calculations require data in kilograms. These conversions are not, however, especially difficult, and Annex I provides the detail and examples needed to make them.

In calculating the sample means, it is imperative that every cell in the data file have a value. Specifically, cells where a value of zero defines the situation of that household must have the value zero entered, and not be left blank. For example, a household that did not produce maize (or cotton) must have zero as the value for QPROD_MH (or QPROD_AL); these cells must not be left blank. Likewise, a household that reported owning no fruit trees must have values of zero entered for NTREE_FT. ***Do not leave any cells blank!***

Once you have ensured that all cells have values, calculating the mean of each variable over all values is straightforward, though the specific commands will vary with different software packages. After calculating these means, you are ready to enter them in the spreadsheet file INCP Lite-CALC.WB3, and obtain your estimates for total and per capita household income.

Obtaining the Income Estimates

The file INCP Lite-CALC.WB3 contains only one page. Most of this page is protected, not allowing you to make any kind of change. Please do not remove this protection, as doing so may result in alterations to the parameter and calculation sections of the spreadsheet that could invalidate your income estimates. The only areas of the spreadsheet which are not protected are the shaded cells of the *Sample Mean* column. It is in these shaded cells of the Sample Mean column that you will enter your data - sample means for the 17 proxy variables required by INCPROX Lite. Take a few minutes now to look at this spreadsheet file and see where you will enter your data. Once you have entered these 17 values, your work is done - values for estimated total household income and estimated per capita HH income will be automatically calculated and shown in the top portion of the window.

Appendix I: Developing the Proxy Variables from the Proxy Questionnaire

Table A1. Procedures for developing proxy variables from questionnaire variables

Proxy Variable Number (Column A, INCP Lite-CALC.WB3)	**Proxy Variable Description** (Column B, INCP Lite-CALC.WB3)	**Proxy Variable Name** (Column C, INCP Lite-CALC.WB3)	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual household level values***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
1	# of types of farm implements owned	NINST	The sum of all values entered in column VII1 of the Instrumentos section of Tabela 5 (Pecuaria e instrumentos de produção)	VII1	1 - 10	1998 zone means ranged from 2.7 (zone 2) to 4.2 (zone 7)
2	kg maize grain produced	QPROD_MH	QPROD_MH = Quantity*weight*processing factor **Quantity** = AF2a **Weight:** *If quantity reported "em grão" (AF3=1)*: Kilo = 1.0 lata 5 litros = 4.375 lata 10 litros = 8.75 lata 20 litros = 17.5 saco de 50 kg = 50 saco de 90 kg = 90 saco de 100 kg = 100 *If quantity reported "em espiga" (AF3=2)*: Kilo = 1.0 lata 5 litros = 1.5 lata 10 litros = 3.0 lata 20 litros = 6.0 saco de 50 kg = 17.15 saco de 90 kg = 30.87	AF2a, AF2b, AF3	0 - high.	Will vary by zone, weather, expected prices. Highest in 1998 sample was 10 tons.

Proxy Variable Number (Column A, INCP Lite-CALC.WB3)	Proxy Variable Description (Column B, INCP Lite-CALC.WB3)	Proxy Variable Name (Column C, INCP Lite-CALC.WB3)	Procedures to Calculate this variable at the *household level*	Questionnaire Variables Utilized	Acceptable range for *individual household level values*	Probable range for *proxy variable sample means*
A	**B**	**C**	**D**	**E**	**F**	**G**
			saco de 100 kg = 34.30 **Processing factor**: *If quantity reported "em grão" (AF3=1)*: 1.0 *If quantity reported "em espiga" (AF3=2)*: 0.75 **Examples** *30 latas de 10 litros em espiga*: AF2a=30, AF2b=10, AF3=2. Quantity = AF2a = 30 Weight = 3.0 kg (lata de 10 litros em espiga) Processing factor = 0.75 (espiga) QPROD_MH = (30)(3)(.75) = 67.5 kg *30 latas de 10 litros em grão*: AF2a=30, AF2b=10, AF3=1. Quantity = AF2a = 30 Weight = 8.75 (lata de 10 litros em grão) Processing factor = 1.0 (grão) QPROD_MH = (30)(8.75)(1.0) = 262.5 kg *10 sacos de 50 kg em espiga*: AF2a=10, AF2b=50, AF3=2. Quantity = AF2a = 10 Weight = 17.15 (saco de 50 kg em espiga)			

Proxy Variable Number (Column A, INCP Lite-CALC.WB3)	**Proxy Variable Description** (Column B, INCP Lite-CALC.WB3)	**Proxy Variable Name** (Column C, INCP Lite-CALC.WB3)	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual household level values***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
			Processing factor = 0.75 (espiga) QPROD_MH = (10)(17.15)(.75) = 128.6 kg. *10 sacos de 50 em grão*: AF2a=10, AF2b=50, AF3=2. Quantity = AF2a = 10 Weight = 50 (saco de 50 kg em grão) Processing factor = 1.0 (grão) QPROD_MH = (10)(50)(1.0) = 500 kg.			

Proxy Variable Number (Column A, INCP Lite-CALC.WB3)	**Proxy Variable Description** (Column B, INCP Lite-CALC.WB3)	**Proxy Variable Name** (Column C, INCP Lite-CALC.WB3)	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual household level values***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
3	kg of cotton produced	QPROD_AL	QPROD_AL = quantity*weight **Quantity** = AF4a **Weight:** Empirically we find that farmers always report this production in kilos, saco de 50, or saco de 90, with saco de 90 being the most common. *If kilo (AF4b=1)*: 1.0 kg *If saco de 50 kg (AF4b=50)*: 36 kg *If saco de 90 kg (AF4b=90)*: 64.8 kg There is no **processing factor**, as cotton is always reported by farmers in terms of seed cotton, which is the form needed for income calculations. **Example** 3 sacos de 90 kg = (3)(64.8) = 194.4 kg	AF4a, AF4b	0 - high. Will vary by zone, weather, expected prices. Highest in 1998 sample was 4.2 tons.	Will vary by zone, weather, expected prices.
4	# of non-food crops cultivated	NCULT_CC	Sum all the values in column II1, Tabela 2 (Culturas não alimentares)	II1	0 - 7	**1998 zone means:** 0.28-0.37 in zones 5,6,7 0.77-1.07 in zones 1-4. Highest in zone 4 (1.07)

Proxy Variable Number (Column A, INCP Lite-CALC.WB3)	**Proxy Variable Description** (Column B, INCP Lite-CALC.WB3)	**Proxy Variable Name** (Column C, INCP Lite-CALC.WB3)	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual household level values***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
5	total # of fruit trees	NTREE_FT	Sum all the values in column III1, Tabela 3	III1	0 to high	1998 zone means were 0.03 to 0.17.
6	# types of fish products sold	NVEND_PX	Sum of all values in column IV1, Peixe section of Tabela 4	IV1	0 - 6	1998 zone means ranged from 0.015 (Manica) to 0.47 (Zambezi Valley)
7	# of cashew products produced	NCAJU	Sum of all values in column Vi, Cajú section of Tabela 4	V1	0-6	**1998 zone means:** 0 in zones 1, 7; 0.047 zone 6 0.71-0.86 zones 2, 3 1.6-1.7 zones 4, 5
8	# of "formal sector" jobs held by resident members	NFORMAL	This is the total number of formal sector jobs held in the family. In the Trabalho fora da Machamba section of Tabela 6, sum all values of VIII1	VIII1	0 - high. Cannot be greater than number of resident members.	**1998 zone means:** *Zones 2, 6, 7*: 0.24 - 0.36 *Zones 1, 3, 4, 5*: 0.014 - 0.086
9	Did the HH own and operate a hammer mill?	MOAG	Equal to 1 if the FIRST row of Tabela 6, section Actividades a Conta Própria ("ser dono e operar uma MOAGEM") is equal to 1 Equal to zero otherwise	IX1	0 or 1	**1998 zone means:** Near zero in all zones, because so few HHs own/operate a hammer mill. Check any sample mean above 0.02.

Proxy Variable Number (Column A, INCP Lite-CALC.WB3)	**Proxy Variable Description** (Column B, INCP Lite-CALC.WB3)	**Proxy Variable Name** (Column C, INCP Lite-CALC.WB3)	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual household level values***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
10	Did the HH operate a trading business? (Buying and selling of any good)	COMERCIO	Equal to 1 if the SECOND row of Tabela 6, section Actividades a Conta Própria ("compra/venda de qualquer produto (agrícola ou não agrícola)") is equal to 1 Equal to zero otherwise	IX1	0 or 1	**1998 zone means:** Ranged from 0.10 to 0.45
11	# of different microenterprises operated by the HH	NMSE	Equal to the sum of all 0/1 values in column IX1, Actividades a Conta Própria section of Tabela 6	IX1	0 - high	**1998 zone means:** *Zones 1, 2, 7:* 1.55 - 1.63 *Zones 3, 4, 6*: 0.94 - 1.39 *Zone 5*: 0.52
12	# of cows owned	NBOI	Equal to the number in the FIRST line of column VI1, Tabela 5, Pecuária section (boi/vaca)	VI1	0-high	**1998 zone means:** 0.0 in zones 4,5,6. 0.014 - 0.076 in zones 1,2,3 0.83 in zone 7
13	# of goats/sheep owned	NCABRA	Equal to the number in the SECOND line of column VI1, Tabela 5, Pecuária section (cabrito/ovelha)	VI1	0-high	**1998 zone means:** 0.23-0.46 in zones 3,4,5 1.7-1.8 in zones 2,6 5.5-6.1 in zones 1,7
14	# of pigs owned	NSUINO	Equal to the number in the THIRD line of column VI1, Tabela 5, Pecuária section (porcos)	VI1	0-high	**1998 zone means:** 0.0 in zone 5 0.5 - 1.5 in all other zones

Proxy Variable Number (Column A, INCP Lite-CALC.WB3)	**Proxy Variable Description** (Column B, INCP Lite-CALC.WB3)	**Proxy Variable Name** (Column C, INCP Lite-CALC.WB3)	**Procedures to Calculate this variable at the *household level***	**Questionnaire Variables Utilized**	**Acceptable range for *individual household level values***	**Probable range for *proxy variable sample means***
A	**B**	**C**	**D**	**E**	**F**	**G**
15	# of chickens/ducks/ other birds owned	NAVE	Equal to the number in the FOURTH line of column VI1, Tabela 5, Pecuária section (galinhas/patos/outras aves)	VI1	0-high	**1998 zone means:** 4.1-7.7 in zones 2-5 13-17 in zones 1,6,7
16	# of other livestock owned	NOUTRO	Equal to the number in the FIFTH line of column VI1, Tabela 5, Pecuária section (outros)	VI1	0-high	**1998 zone means:** <2 in all zones
17	How many people are in this household?	NMEM	Equal to the number of rows filled-out in Tabela 1 (Membros Residentes)	MEM	1 to high	**1998 zone means:** Ranged from 4.3 to 5.8 for Zones 1-6. 6.99 in Zone 7 (Manica).

Annex H

Procedures for Using SPSS/Windows to Generate INCPROX Estimates of Income and Income Components

To generate estimates of income and income components using SPSS/Windows, the data will first have to be cleaned just as with the spreadsheet approach. After cleaning, all conversion of questionnaire variables to proxy variables will be done by an SPSS syntax file developed by MSU. The steps for using the SPSS for Windows package are as follows:

1. Enter the questionnaire data in the following files:

File #	File Name	Relation to Questionnaire	Variables
1	AF.SAV	Household level file - all "AF_" questions	key variables AF1 ... AF14
2	MEM.SAV	Member level file - all data from Tabela 1	key variables mem I1 ... I5
3	CULTALIM.SAV	Food crop file - Culturas Alimentares section of Tabela 2	key variables cultalim II1, II2
4	CULTOUTR.SAV	Other annual crops file - Outras Culturas section of Tabela 2	key variables cultoutr III1
5	CULTVER.SAV	Fresh production file - Produção em Verde section of Tabela 2	key variables cultverd IV1, IV2
6	HORTIC.SAV	Vegetable production file - Hortícolas section of Tabela 3	key variables hortic V1, V2
7	FRUTA.SAV	Fruit production file - Frutas section of Tabela 3	key variables fruta VI1
8	CAJU.SAV	Cashew production file - Cajú section of Tabela 3	key variables caju VII1, VII2
9	PEIXE.SAV	Fish production - Pescado section of Tabela 4	key variables peixe VIII1, VIII2
10	PEC.SAV	Livestock holdings - Pecuaria section of Tabela 4	key variables pec IX1
11	INST.SAV	Farm implement ownership - Instrumentos de Produção section of Tabela 4	key variables inst X1

File #	File Name	Relation to Questionnaire	Variables
12	TRABFORA.SAV	Wage work - Trabalho fora da machamba section of Tabela 5	key variables trabfora XI1
13	CONTPROP.SAV	Microenterprise - Actividades a Conta Própria section of Tabela 5	key variables contprop XII1

2. Save these uncleaned files in a folder of your choice. This will be your copy of the original, uncleaned data, which should not be changed.

3. Create the folder c:\proxy99\incprox\data and copy all 13 uncleaned files to it.

4. Clean the files in c:\proxy99\incprox\data using procedures your NGO has developed with other surveys, and save the files to the same names. You will now have uncleaned, original data in a folder of your choice, and cleaned data in c:\proxy99\incprox\data.

5. Create the folder c:\proxy99\incprox\syntax and copy the file IncproxVarsNGOs.sps to it. You can obtain a floppy diskette with this file from the MSU Food Security Project.

6. Run IncproxVarsNGOs.sps. This will create the data file IncproxVarsNGOs.sav, which contains all proxy variables needed for INCPROX.

7. Copy the file IncproxEstimate.sps into c:\proxy99\incprox\syntax. You can obtain a floppy diskette with this file from the MSU Food Security Project.

8. Add ZONE1 through ZONE7 variables to IncproxVarsNGOs.sav, as instructed in IncproxEstimate.sps, and save to the same name.

9. Run IncproxEstimate.sps. It will deliver your results in the Output Navigator.

NDAE Working Papers

1. Informing the Process of Agricultural Market Reform in Mozambique: A Progress Report, October 1990

2. A Pilot Agricultural Market Information and Analysis System in Mozambique: Concepts and Methods.

3. Inquérito ao Sector Familiar da Província de Nampula: Observações Metodológicas, 9 de Novembro de 1991

3E. A Socio-Economic Survey of the Smallholder Sector in The Province of Nampula: Research Methods (**translated from Portuguese**), January 1992

4. Inquérito ao Sector Familiar da Província de Nampula: Comercialização Agrícola, 30 de Janeiro de 1992

4E. A Socio-Economic Survey in The Province of Nampula: Agricultural Marketing in the Smallholder Sector (**translated from Portuguese**), January 1992

5. Inquérito ao Sector Familiar da Província de Nampula: O Algodão na Economia Camponesa, 9 de Novembro de 1991

5E. A Socio-Economic Survey in The Province of Nampula: Cotton in the Smallholder Economy (**translated from Portuguese**), January 1992

6. The Determinants of Household Income and Consumption in Rural Nampula Province: Implications for Food Security and Agricultural Policy Reform, August 1992

6P. Determinantes do Rendimento e Consumo Familiar nas Zonas Rurais da Província de Nampula: Implicações para a Segurança Alimentar e as Reformas de Política Agrária (**Traduzido do Inglés**), 24 de Fevereiro de 1993

8. Dengo, Maria Nita, "Household Expenditure Behavior and Consumptio Growth Linkages in Rural Nampula Province, Mozambique", M.Sc. Thesis, Dept. of Agricultural Economics, Michigan State University (**Reprint**), December 18 1992

9. The Maputo Market Study: Research Methods, March 8 1993

9P. O Estudo do Mercado de Maputo: Observações Metodólogicas, 1 de Junho de 1993

10. The Organization, Behavior, and Performance of the Informal Food Marketing System, May 28 1993

12. The Pricing and Distribution of Yellow Maize Food Aid n Mozambique: An Analysis of Alternatives, October 18 1993

14. Liedholm, Carl and Donald Mead, "Small-scale Enterprises: a Profile", in Economic Impact: A Quarterly Review of World Economics, no. 63 (**Reprint**)

14P. Liedholm, Carl and Donald Mead, "Pequenas Empresas: Um Perfil", em Economic Impact: A Quarterly Review of World Economics, no. 63 (**Reprint, translated from English)**

15. Mini-SIMA e Análises Específicas: Um Ensaio Aplicado aos Mercados de Maputo, 15 de Julho de 1993

16. The Evolution of the Rural Economy in Post-War Mozambique: Insights from a Rapid Appraisal in Monapo District of Nampula Province

17. Padrões de Distribuição de Terras no Sector Familiar em Moçambique: A Similaridade entre duas Pesquisas Distintas e as Implicações para a Definição de Políticas, May 1994

18. Who Eats Yellow Maize? Some Preliminary Results from a Survey of Consumer Maize Preferences in Maputo, October 1994

18P. Quem Come Milho Amarelo? Alguns Resultados Preliminares de um Inquérito sobre as Prefencias dos Consumidores de Milho na Cidade de Maputo (**Traduzido do Inglés**), 10 de Novembro de 1994

19. Diagnóstico da Estrutura, Comportamento, e Desempenho dos Mercados Alimentares de Moçambique, 4 de Julho de 1995

20. Inquérito ao Sector Moageiro de Pequena Escala em Moçambique: Observações Metodológicas, 30 de Janeiro de 1995

21. O Sector da Castanha de Caju - Lucros Potenciais Perdidos por Africa? (**Reimpressão)**, Novembro de 1995

22. Smallholder Cash Cropping, Food Cropping and Food Security in Northern Mozambique: Research Methods, March 1996

22P. Culturas de Rendimento, Culturas Alimentares e a Segurança Alimentar do Sector Familiar no Norte de Moçambique: Métodos do Estudo, Novembro de 1996

23. Plan of Activities for Food Security Research Project, September 1995 through August 1997, 1996

24. Strasberg, Paul, “Smallholder Cash-Cropping, Food-Cropping and Food Security in Northern Mozambique”, Ph.D.Dissertation, Dept. of Agricultural Economics, Michigan State University (**Reprint**), May 1997

25. Smallholder Cash-Cropping, Food-Cropping and Food Security in Northern Mozambique: Summary, Conclusions, and Policy Recommendations, June 1997

26. Agricultural Market Information for Family Farms in Mozambique, June 1997

26P. Informação de Mercado Agricola para o Sector Familiar em Moçambique, Junho 1997

27. Micro and Small Enterprises in Central and Northern Mozambique: Results of a 1996 Survey, September, 1997.

27P. Micro e Pequenas Empresas no Centro e Norte de Moçambique: Resultados do Inquerito Realizado em 1996, Setembro de 1997.

28. Desafios Para Garantir a Concorrência e Reduzir os Custos no Sistema Alimentar de Moçambique, 12 de Maio de 1998.

29. Planning for Drought in Mozambique: Balancing the Roles of Food Aid and Food Markets, May 14, 1998

30. Séries Históricas dos Preços de Grão de Milho Branco e suas Tendências Reais em Alguns Mercados do País, 18 de Maio de 1998.

31. What Makes Agricultural Intensification Profitable for Mozambican Smallholders? An Appraisal of the Inputs Subsector and the 1996/97 DNER/SG2000 Program, Volume I: Summary, October, 1998.

32. What Makes Agricultural Intensification Profitable for Mozambican Smallholders? An Appraisal of the Inputs Subsector and the 1996/97 DNER/SG2000 Program, Volume II: Main Report, October, 1998.

33. Household Food Consumption in Mozambique: A Case Study in Three Northern Districts, February, 1999.

34. The Effects of Maize Trade with Malawi on Price Levels in Mozambique: Implications for Trade and Development Policy, November,1999.

35. Séries Históricas dos Preços de Grão de Milho Branco e Suas Tendências Reais em Alguns Mercados do País no Periódo Compreendido Entre Abril 1993 e Setembro 1999, November, 1999.

36. A Simplified Method for Assessing Dietary Adequacy in Mozambique. January, 2000.

37. Implementing A Simplified Method for Assessing Dietary Adequacy in Mozambique: A User's Manual. January, 2000.