EGERTON UNIVERSITY
TEGEMEO INSTITUTE OF AGRICULTURAL POLICY AND DEVELOPMENT

&

MICHIGAN STATE UNIVERSITY

TEGEMEO AGRICULTURAL POLICY RESEARCH ANALYSIS (TAPRA) PROJECT

HOUSEHOLD SURVEY 2007 DATA DOCUMENTATION

Support for this study was provided under the Tegemeo Agricultural Monitoring and Policy Analysis (TAMPA) project, supported by the United States Agency for International Development / Kenya. Supplementary support for this study is provided by the Office of Sustainable Development, Africa Bureau, and USAID/Washington.

2007

SAMPLING METHOD

The TAPRA sample was only composed of TAMPA households that were interviewed in 2004.

TAMPA Sample

The sampling method used was similar across all the sites and is described below:

1. Within the designated area of study (considering AEZs and other criteria), all the villages/sub-areas were listed with the help of the administration or chief.

   AEZ, population, and whether the district belonged to the "original" KAMPAP districts (districts where Tegemeo had conducted much research before and had some supplementary data and information on) were some of the key factors in this exercise.

   The first step was to identify the spatial distribution of AEZ in the district. The idea was to capture as much of the diverse conditions as possible in our sampling. From this step we were able to classify certain areas within AEZ with the help of the Ministry of Agriculture officers. Each district was in turn divided into divisions, locations and sub-locations and then villages/wards. From the district level we were able to pick representative divisions with the help of the district officers. I believe that we also took into account the populations and AEZ conditions within these areas to help us select these divisions. Because not all divisions could possibly be visited we picked a random sample of these divisions for further follow-up. These were selected with the idea of incorporating the diversities that were inherent in each district that we visited (a representative sample).

   At the division level, a similar exercise was carried out with the help of the Ministry officials. Then the locations were selected randomly. This was followed by sub-locations and then finally the villages/clusters below.

2. From this list (and considering the sample size required from the area) a number of villages were randomly selected by picking from the list above.

3. For the selected villages, and with the help of the administration and key informants, we listed all household units within the village by head of household.

4. In most cases the list above exceeded the sample size requirements for the area. Accordingly we used the 'universal' KAMPAP sampling technique to select households for interview.

   Universal KAMPAP sampling technique description: Most village elders/chiefs have a pretty comprehensive list of householders' names. Suppose we had a total list of 76 households for a village or cluster from the chief (numbered from 1 to 76). Assume too that all we needed was to interview 12 households from this village. The objective was to randomly select every sixth household to get the 12 we needed (approx 76/12=6). The question is, on a numerical list of 1 to 76 where do you start the selection (is it 1,2,3,4,5 or 6)? We wrote the numbers 1 to 6 on different pieces of paper of similar size, folded and mixed them up. Then we asked a villager or the chief to pick one of these papers and

reveal the number. Suppose the number picked is 3; then we proceeded to pick the households starting from the third on the list, i.e. 3,9,15,21,27 etc.

5. It happened that in some areas some of the selected households within a village had household heads that were related by marriage or some other kinship relationship (though the samples had been selected randomly in the first place). In such instances one could find cousins, brothers, uncles, etc who had bought farms in the same area and over the years subdivided their farms to their children, etc but all these were clearly separate households with different management styles and approached their household decisions separately. Relationships among households do not necessarily imply joint decision-making.

6. In conclusion the samples were as random as was possible and the data should be able to express this random nature despite some pockets here and there of 'relationships', if one may.

SUMMARY OF HOUSEHOLDS SURVEYED

Out of the 1397 households interviewed in 2004, there were 1342 households that were interviewed in 2007. Turkana and Garissa were not interviewed since these areas had been previously dropped from the sample in 2004. The argument was that the original sample was not typical of the area. Garissa for example, had households who were engaged in irrigation which gave an indication that the area was highly productive. Turkana district did not give the typical scenario of a nomadic pastoralist household. Moreover, in Turkana, it was difficult to generate panel data due to the nomadic nature of the household.

It is important to note that there was no replacement of households in the Tampa sample for this survey.

The data for page one of the survey instrument are contained in two files: allhhid07.sav and hhidfinal07.sav. The first file (allhhid07.sav) contains all the original households. The second file (hhidfinal07.sav) contains only those households that were interviewed for this 2007 survey (1342 hhids). This file should be used to merge the identifying characteristics to the other files as needed.

DATA FILE DESCRIPTIONS FOR RURAL HOUSEHOLD SURVEY

Directory Structure: - First level subdirectory off the root directory is called “docs”, the next level is called “Kenya” and the next level is called “Kenyahh2007”**.** There are eight subdirectories off this directory:

Under Kenyahh2007 the directories are: augdata, anal, docs, lookup, origdata, syntax, tmp.

C:\\TAPRA07

\anal- analysis files and syntax.
\augdata- final data files to be used for analysis

\docs- documentation of all files including the survey instruments and enumerator manual
\lookup- lookup data files and syntaxes.
\NewVars- files and syntaxes that have been computed and ready for analysis
    \demog – adults' equivalents and household size
    \income – income variables
\origdata- original files—not to be used for analysis – most people will not have this subdirectory.
\syntax- syntax files to be use for analysis and to clean data.
\tmp- used to store temporary files that the analyst does not plan to retain.

Variables to identify location:

Aez - agricultural ecological zones
Aezsmall - aez subdivided into more specific zones
Zone – habitat zones
Prov (province)
Dist (district),
div (division),
loc (location),
subloc (sub-location),
vil (village)

In addition to the identifying variables listed above GPS coordinates were collected and recorded for all the households that were interviewed.

DATA FILES

Directory: C:\tapra07\augdata

Type of data	File name	Key variables	Number of cases	Computed variables	Comments
Household	allhhid07.sav	hhid	1397		All households that were to be interviewed – **use only if want to know how many households were not interviewed**
Household assets	asset07.sav	hhid, asset	8837		Household agricultural assets
Business / informal labour	business07.sav	hhid, mem, activity	1625	Low, medium, high=# of low, medium and high income months	Business and informal labour activities
Fruits and vegetables purchased for home consumption	consumpt07.sav	hhid, fooditem	6363	totkgs = Total kgs purchased – 30 days kgsjan = total kgs purchased in January	Expenditure on food items over the past 30 days and in January
cow milk produced and sold	cowmilk07.sav	hhid, milk	1693		Details of cow milk produced, fresh and sour milk sold, prices & buyer
Cropping patterns	Croplev07.sav	hhid, harvest, field, crop	23066	Seedkg=Quantity of seed used in kgs Hvtkg=quantity of produce harvested in kgs kgsold=quantity of harvest sold in kgs	Crop level data - crops grown, seed information, harvest, sales & buyers, amount spoiled for fruits and vegetables
household members from previous surveys	demog07.sav	hhid, mem	9389		Household members listed in 2007. Data are: sex, age, relation to head, currently in school, years of schooling, months living at home, why left, engage in business/informal labor or salaried employment, if

					chronically ill for 3 or more months.
Additional members	demogA07.sav	hhid, mem	1327		household members not listed in 2004, same questions plus why joined household and if had other income before joining
willingness to pay an extension worker	Extension07.sav	hhid, serv	4036		Willingness to compensate an extension worker training individual or group of 20 members, for 3 hours on new technology that is urgently needed, gauged by the amount of money one will be willing to pay.
Fertilizer used	fert07.sav	hhid, harvest, field	8941	Fertotal – amount used was standardized to kgs	Field level file - types and amounts of fertilizer used
Cropping patterns	Field07.sav	hhid, harvest, field	9339		Field level data - acreage, tenure, land preparation types and costs
Household identification	Hhidfinal07.sav	hhid	1342		All households that **completed** the interview – **use this file to merge in location variables**
Household level questions	hh07.sav	hhid	1342		General household level questions
Inventory of crops	incrop07.sav	hhid, cropm, crops	17867		Crop inventory- field crops, fruit trees & vegetables
Land inheritance.	inheritance07.sav	hhid, mem	2689		Information about land ownership and inheritance from the original families of household head and spouse.
Inputs, including fertilizers purchased both in cash and on credit	input07.sav	hhid, inputype	5664		Fertilizer and other inputs purchased/hired.
Labour inputs	labour07.sav	hhid, activity	9199		Labour inputs for largest maize

					field
Livestock products	liveprod07.sav	hhid, liveprod	1286		Livestock production and sales
Livestock costs	livescost07.sav	hhid, animsp	1624		Livestock costs and livestock services costs
Livestock/ livestock inputs purchased on credit	livestinput07.sav	hhid, input	117		Livestock/ livestock inputs received on credit.
Livestock	livestock07.sav	hhid, livecode	4788		Livestock inventory and sales
Maize seed	MaizeSeed07.sav	hhid, season, field, sdtype	2879		Types of maize seed varieties used and their sources
Mortality since 2004	mortality07.sav	hhid, pdmem	194		Deaths since 2004, cause, sex, year born and died, relation to head, level of education
Non- Agricultural credit	nagcred07.sav	hhid, crduse	535		Credit used for non-agricultural purposes (in cash and in kind), source, value, and how it is to be repaid.
Home consumption purchases	purch07.sav	hhid, purch	3651	kgqty1, kgqty2, kgqty3, totkg, tkgs = total kgs purchased for the whole year	Purchases for home consumption by 4-month periods
Salaries and pensions	salwg07.sav	hhid, mem, activity	1048		Salaries / permanent employment-pensions and remittances
Savings account held by all household members	savings07.sav	Hhid, mem	1179		Savings accounts held by all household members, including ROSCAS and cooperatives.

Lookup tables

C:\tapra07\lookup

Type of data	File name	File to be used with	Key variables	Number of cases	Comments
Crop quantity conversion to kgs	Cropconv.sav	croplev07.sav	crop, unit	769	Used this file to convert all harvested/sold crop units to kgs.
Fruits and vegetables conversion to kgs	Consumptconv.sav	consumpt07.sav	Item, unitpur	63	Convert consumption units into kgs. This has already been done
Fertilizer quantity conversion to kgs	fertconv.sav	fert07.sav, Tfert07.sav	ferttype, fertunit	155	File used to convert fertilizer units into kgs
Purchase price	priceconsump.sav	consumpt07.sav	fooditem, dist	255	Created by PriceConsumpt.sps. District price conversion for consumpt04 file
Crop prices	pricecrop.sav	croplev07.sav	crop, dist	964	Created with PriceCrop.sps. DT (David Tschirley) developed this using standard approach: district median if >=10 observations, otherwise zonal median if >=10 observations, otherwise national median. **Issues-Mary Mathenge** ***Method*** *Computed pkg first using district price, then zone, then province (as long as respective cases >10) and finally national price.* *Note that only 7038 cases out of a total of 23066 were any sales made. We use these to compute prices for districts where no sales were made.* *After merging back the TL file to croplev.sav, we get 59 cases where pkg is missing (21 crops). These correspond to cases where no sales were made in the whole data. Tegemeo ascertained a national price which was added for these crops.*

Type of data	File name	File to be used with	Key variables	Number of cases	Comments
					N/B: No elimination of any irrelevant slunits or outlier values has been done in this case.
Milk prices	priceMP.sav	cowmilk07.sav	Milk, dist	32	Created by priceMP.sps. District price conversion for cow milk sold. (Both fresh & sour milk).
Fertilizer prices	pricefert.sav	fert07.sav, input07.sav	ferttype, fertunit, dist, inptype	371	Created with PriceFert.sps. Followed standard approach as in PriceCrop.sav. Note that we also used a fertilizer price lookup file in the 2004 data set.
Prices for livestock products	priceLP.sav	liveprod07.sav	liveprod, dist	70	Created by PriceLP.sps. District price conversion for livestock products
Livestock selling prices	priceLS.sav	livestock07.sav	livecode, dist	242	Created by PriceLS.sps. District price conversion for livestock sales
Prices of purchases	pricepurchase.sav	purch07.sav	purch, dist	94	Created by PricePurchase.sps. Price conversion for the prices of household purchases.
Livestock purchase price.	pricepurchLS.sav	livestock07.sav	livecode, dist	242	Created by PricePurchLS.sps. Conversion for the purchase prices of livestock
Prices of seed	Priceseed.sav	Croplev07.sav	Crop, sdtype, sunit, dist	1844	Convert prices of seed into district prices **Issues-Mary Mathenge** *Price of seed computed as in the other TL files.* *Since seed data was collected with the main crop file, I used cases where seed was purchased (scost>0) to create the table lookup file. Note that only about* 6389 *out of total of* 23,067 *cases have scost>0.* *Note that by merging the computed TL file with croplev.sav, we will be getting the value of seed used as opposed to the cost of seed purchased since we will be*

Type of data	File name	File to be used with	Key variables	Number of cases	Comments
					imputing a value for seed that may not have been purchased (similar approach as with Pricecrop.sav) unless we recode Pseed to zero for such cases during programming . *9947 cases in croplev.sav have missing scost:All N/A. 6732 cases had scost=0, and only* 6389 out of total of 23,068 cases have scost>0. *See summary in syntax.*
Purchases conversion to kgs	purchconv.sav	purch07.sav	purch, unit	30	Conversion of purchase units into kgs.

New Computed Variables

C:\Tapra07\NewVars

\demog

\income

Type of data	File name	Key variables	Number of cases	Variables	Syntax File
Subdirectory "demog"					
Adult equivalents and size of household	ae_hhsize_07.sav	Hhid	1342	aehh07 – adult equivalents hhsize07 – household size	ae_hhsize_07.sps – see note at end of documentation regarding method used to compute adult equivalents
Subdirectory "income"					
All income variables in one file	income07.sav	hhid	1342	Main variables are: Income07 (sum of crpinc07, offarminc07, vnetlv07)	merge_income.SPS
Crop income computation	cropinc07.sav	Hhid	1334	crpinc07, totcost (vprod, vsold, vret, lpcost, fertcost, seedcost)	cropinc07.sps
Off farm income	offarminc07.sav	Hhid	1222	Vsalrem, vinform, offrino7	offfarm income 07.SPS
Live animal valuation	vlivestock_net07.sav	Hhid	1342	Vcost_lv (Vetserv, sallvstk animfeed) – costs vnet_ls, vprod_lp, vsold_ls, vpur_ls	livestock_income07.SPS

Documentation files

C:\Tapra07\docs

File name	Contents
2007_Original_Questionnaire.pdf	Questionnaire used in the field
2007_Synthetic_Questionnaire.pdf	Field questionnaire restructured to reflect the data file structure
2007_SurveyDocumentation.pdf	Documentation of data files, sampling methods, specific issues with the data set
2007_Enumerator_Manual.pdf	Instructions to enumerators

Data files pertaining to TAMPA surveys conducted in 1997, 2000, and 2004.

Purpose	File name	Number of Cases	Comments
C:\docs\Kenya\KenyaGen\data			
Consumer Price Index	CPI_allyears.sav	4	The consumer price index is based on the year 2003/2004, using raw CPI data from the Ministry of Finance, Government of Kenya. The period is from June xxxx to May xxxx (xxxx refers to the various years). To reflate all years to 2003/04, divide by these CPIs for their respective years. The years are: 1995/96, 1996/97, 1999/2000 and 2003/2004. 2006/07.
Rain information for the villages covered in the TAMPA surveys	tampa_rain.sav tampa_rain.dta	107	File contains data at the prov, dist, div. village level. Altitude, latitude, longitude, rainfall for the long and short harvests as well as fraction of 20 day periods with <40mm rain for each season
C:\docs\Kenya\KenyaGen\docs			
Documentation of rainfall data	Kenya Rainfall Data.pdf		
Main and short season rain periods defined	Rainfall Periods for Tegemeo Sample Villages.pdf		

Miscellaneous Notes on the Rural Household Survey 2007
Egerton University - Tegemeo Institute / MSU
Updated – Nov 2007

Household Numbers

Original household numbers (hhid) from the 1997, 2000 and 2004 survey range from 1 to 1838 for a total of 1347 households interviewed.

There were gaps in numbering in the Tapra sample because of the households that have been missed since 1997.

Brief Documentation for all files

All the files except field07 and fert07 contain a variable **'comment'**. This variable consist any issues that were noted during cleaning that are specific to the particular case or set of cases.

1. **allhhid07:** It is preferred that analyst use the hhidfinal07 file which is contains only the households that were interviewed. This file contains all the households that were supposed to be interviewed. No major issues were noted in this file

2. **hhidfinal07:** This is a generated file. It contains all the households that were interviewed. It is at household level and contains the identifying variables for the household. The total number of cases is 1342.

3. **hh07:** This file contains the household level questions. Question 21.1 and 21.2 had quite a number of households with recall problems hence many 'don't know' response. There was a perception problem with the same questions in cases where a son or daughter has inherited a household that was previously headed by the parents. It was not clear whether the questions should have been addressed to the previous or the current head.

4. **incrop07:** This file is at crop level. It contains a Yes entry for the annual crops planted the perennial crops produced or planted. During cleaning, more emphasis was directed to the crop file. Notice that commercial trees and sisal were transferred to the informal income section.

5. **field07:** This file is generated from the original crop file. It contains field level information. Some acreages were noted to be very small especially when related to the yields and seed quantities. Confirmation was got from the questionnaires. No major issues noted.

6. **croplev07:** This file is generated from the original crop file. The file contains details of the cropping pattern for the main and the short season. The file is at HHID, harvest, field, crop level. 7 households did not have any cropping activity in the year. The file also contains information on amounts harvested and amounts sold from the harvests. The seed cost for maize is repeated on the maizeseed file but sometime with some minor discrepancies. Analyst should work with the details on the maize seed file where applicable. Note that commercial trees and sisal were transferred to the

informal income section. In the 2000 survey the "harvest" variable was called "season".

7. **fert07:** This file is generated from the original crop file and contains information on types and amounts of fertilizer used on every field. No major issues were noted. In the 2000 survey the "harvest" variable was called "season".

8. **labour07:** The labour file contains details on labour for the largest monocrop maize field or the largest intercrop maize field if maize was not grown on its own. The interest was to collect all the labour inputs on maize in case of intercrop maize fields.

   The 2007 questionnaire had variables LB11, LB12 and LB13 to capture details of work done by salaried workers who are paid on a monthly basis. This was not previously captured in 2004. A total of 23 households did not have any labour activity of which 16 did not produce maize in the main season while the rest did not have any cropping activity.

9. **Maizeseed07:** The file contains details on maize seed type, purchase and prices. Note that seed information is also collected at the crop level in "croplev07.sav". However, the question asked in that table referred to the total quantity and did not ask for detailed information. The maizeseed07.sav file asked for expanded detail allowing the respondent to indicate the different seed types used in the same field. In many instances the information in this file will be the same as in the croplev07.sav file. It is recommended that researchers use the information in this file for analysis of seed types used for maize.

10. **Input07:** This file contains inputs that were bought both **on cash** and **on credit** basis by the households. The inputs are mainly fertilizers, agrochemicals and other farming inputs. Land preparation cost was only included in this table if procured on credit, else the details can be obtained from field07.sav. No table lookup to standardize prices was created except in the case of fertilizers. The actual price quoted for other inputs should be used.

11. **Nagred07:** this file contains details of credit obtained and used for non-agricultural purposes by the households. There are no issues to note concerning the file.

12. **Purch07.** The file contains details of purchases on key items (dry food stuffs) in 4 month groupings within the year and if the respondent could not answer in 4 month grouping, the response was given for the whole year. There are notes in the comment field for cases that may need some explanation.

13. **Consumpt07:** File gives details of purchase of selected horticultural crops over the last 30 days and also for January. If the respondent did not purchase the product in the last 30 days, the enumerator was to record the unit of measure for the January purchases in the section for the last 30 days. No major issues noted.

14. **Livestock07:** Gives livestock inventory details. Purchases and sales were collected for cattle, sheep, goats, camels, pigs and chicken unlike in 2004 where such details were only collected for cattle. No major issues noted.

15. **Cowmilk07:** Gives details of the quantities of cow milk produced, sold as fresh and in sour form on month by month basis.

16. **Liveprod07:** Gives details of production and sale of livestock products. Under the "other (specify)" some people sold broilers in large numbers. We choose to let it remain here instead of transferring it to the informal income because we did not have all the cost involved in the production process.

17. **demog07:** This file contains details of the demographic characteristics of the household. Adult household members listed in 1997, 2000 , 2002 and 2004 are in this file. Most of the household heads are in this file. However, some heads of household are in the additional adults file.

18. **demogA07:** This file contains details on all additional members (both adult and children) of the household not listed in 2004. There were 8 households where the head of the household is in this file. The variable "mem" starts at 71.

19. **business07:** This file contains details on informal business household income. For hhid 367 the high average cost for livestock selling was higher than the high cost. The activity was accruing losses in the low and average months. In hhid 1488, the matatu business was making losses.

20. **Salwg07:** Gives details on salaried income for the household. Remittance data were collected in this file.

21. **Asset07:** Gives details on the number and value of selected assets for the household. No major issues noted so far.

Adult equivalence

The table shows the recommended conversion of different age categories and gender into adult equivalence. This table may be used together with the 3 demography tables for various computations.

The file called **ae_hhsize_07.sav** in the "c:\tapra07\NewVars\demog" subdirectory has already computed the adult equivalents using the breakdown outlined in the table below.

Gender	Age	AE
Both	<1 year	0.33
Both	1-2 years	0.46
Both	2-3 years	0.54
Both	3-5 years	0.62
Male	5-7 years	0.74
Male	7-10 years	0.84
Male	10-12 years	0.88
Male	12-14 years	0.96
Male	14 -16 years	1.06

Male	16 -18 years	1.14
Male	18-30 years	1.04
Male	30-60 years	1.00
Male	>60 years	0.84
Female	5-7 years	0.70
Female	7-10 years	0.72
Female	10-12 years	0.78
Female	12-14 years	0.84
Female	14 -16 years	0.86
Female	16 -18 years	0.86
Female	18-30 years	0.80
Female	30-60 years	0.82
Female	>60 years	0.74

Document name: C:\tapra07\docs\2007_SurveyDocumentation.doc