Data Documentation for Suri (2011)

This document describes how the data extracts were created for Suri (2011). The full citation for the paper is "Selection and Comparative Advantage in Technology Adoption", *Econometrica*, 79 (1), 159–209.

The data comes from the Tegemeo Institute. The data used for the paper was last updated in August 2006.

The survey rounds used were: 1997 (that included retrospective data for 1996), 2000 and 2004. The 1998 and 2002 rounds were very short surveys that did not include detailed agricultural data. The data can be requested here: http://tegemeo.org/index.php/resources/data/230-request-for-data.html. However, note that the data used in the paper was last updated in 2006 and so will not account for any subsequent cleaning that was done.[1] Finally, it is worth noting that the 2007 and 2010 data have not been used in this paper – there have been some dramatic changes in the hybrid seed market since 2004, certainly a topic for future research.

The basic data extract used in the paper consists of a sample of 1202 households for 1997, 2000 and 2004. However, the 2000 data is largely not used as it did not include data on labor inputs (it is only used in Table IID in the paper to show switching behavior). In addition, the paper reports some results where I drop two districts where HIV had high prevalence. Below, I first describe how the sample of 1202 households was created and then I describe how each variable I created and used was constructed.

Also note that in each year of data, there is data for the main harvest season and for the short harvest season. This paper focuses on the data for the main season as it accounts for a large share of the harvest for a household in a given year and a large share of households do not grow anything in the short season. The short season is just treated very differently by farmers, given its very nature. Given the topic of the paper, it seemed important to restrict the farmer decisions being studied to the main season.

Finally, the 2004 survey included an extra sample of households that were not part of the original panel sample for a separate study. Make sure these households have been dropped (they were identified at the time as having a household ID greater than 5000).

A. Sample Restrictions

The methods developed in the paper require a balanced panel. The sample was therefore constructed as the set of households that grew maize in all the three main periods covered in the data: 1997, 2000 and 2004.

The sample restrictions can be implemented as follows:

1. Generate a variable that indicates both maize harvests and the hybrid dummy variable are not missing in each year (see below for how each of these variables are constructed).
2. Keep only the set of households where this dummy is always 1 for the years 1997, 2000 and 2004.

In some of the specifications I drop two districts that had a high incidence of HIV. These two districts can be found in the location file (key 8 in the data table below) and they are Kisumu and Siaya. In the 1997 location file (key 8), these are districts with the codes 42 and 43, respectively.

[1] We are happy to share the extracts of the data used if you can share a signed data release agreement with Tegemeo and get written permission for me to share the extracts.

B. Variable Construction

The specifications throughout the paper use the variables reported in the summary statistics in Tables IIA and IIB. Here, I describe the construction of each of these variables, variable by variable.

I also create a table with a file key that matches the data construction description to a particular file. This is shown immediately below. The table also identifies the names of the files for each of the two main years of data used (1997 and 2004). The files for 2000 are similarly named. Note that the file naming conventions may have changed since the data was shared with me by Tegemeo in 2006 but the table below should help identify the relevant file.

File Key

Key	File Description	File Name, 1997	File Name, 2004
1	Field data file *Unique observation at the level of household, year, season, field*	field97	field04
2	Crop level harvest file *Unique observation at the level of household, year, season, field, crop*	croplev97	croplev04
3	Labor file *Unique observation at the level of household, labor activity for 1997 and at the level of household, field, crop, labor activity for 2004*	labor97	labour04
4	Fertilizer field level data file *Unique observation at the level of household, year, season, field, fertilizer type, fertilizer unit*	NA	fert04
5	Fertilizer prices file *Unique observation at the level of fertilizer type, fertilizer unit, district for 2004 and at the level of district, division (the fertilizer type is wide) in 1997*	fertprc97	pricefert
6	Crop conversion file *Unique observation at the level of crop (this is a combination of old and new crop codes for 1997), harvest unit*	cropconv	cropconv
7	Household demographics and education file *Unique observation at the level of household, person*	demog97	demoga04, demoga_a04, demogc04
8	Location file *Unique observation at the level of household*	hhidfinl97	hhidfinl04
9	Rainfall data file *Unique observation at the level of village (years are wide)*	tampa_rain	tampa_rain
10	Remaining household level data file *Unique observation at the level of household*	hh97	hh04

Harvests

1. Use the crop level harvest file (key 2 in the table below). Keep the main season observations.
    a. It will be useful to use this file to create a field level file that identifies which fields have any maize on them (this will be used below).

2. Merge on the field level maize identifier (created in 1a) and merge on the field data file (key 1).

3. Keep only the fields that grow maize and keep only the harvests of maize as identified by the crop variable in the file (keep both regular maize and green maize).

4. Create harvest in kg by converting the various units to kg by using the crop conversion data file (key 6).

5. Aggregate/collapse harvests to the household-year level.

6. Note that one complication when using the 1997 and 2004 files for more than maize (in this paper I focus on maize and so this is not an important issue) is that the crop codes changed (the crop conversion file (key 6) contains the conversion of the old crop codes to the new ones).

Acres

1. Use the field data file (key 1) and keep the main season observations.

2. Merge on the field level maize identifier (created above) and keep only the maize fields.

3. Aggregate/collapse acres to the household-year level.

Inputs of Seed and Hybrid Identifier

1. Use the crop level harvest file (key 2). Keep the main season observations and the observations for maize using the crop identifier in the file (again keep both regular maize and green maize).

2. Drop if the seed type is missing or has a code that is not identified in the codebook (I dropped seed types with codes 0 or 4900).

3. Standardize the seed units into kg.

4. Collapse this data to the household-year-seed type level. Reshape this wide by seed type.

5. Create the indicator for whether a household used hybrid which is a dummy for the seed quantity planted of purchased hybrid being greater than zero.

6. Create a variable that sums all the quantities of seed planted for a household in a given year.

Land Preparation Costs

1. Use the field data file (key 1) and keep the main season observations.

2. Merge on the field level maize identifier (created above) and keep only the maize fields.

3. Aggregate/collapse the land preparation cost variable to the household-year level.

Fertilizer, 1997

1. Use the use field data file (key 1) and keep the main season observations.

2. Merge on the field level maize identifier (created above) and keep only the maize fields.

3. Drop the observations where the fertilizer type is missing.

4. Reshape the file wide by household-field so that the quantities of the different fertilizer types are different columns (rather than different rows).

5. Note that there are two fertilizers by field so add these two quantities of fertilizer separately for each of the types of fertilizer to get the total kg of fertilizer of each type used in kg.

6. Aggregate/collapse the fertilizer quantity variables to the household-year level.

7. Merge on fertilizer prices from the fertilizer price file (key 5). Use the median district price by type of fertilizer to value the fertilizer. If that is missing, use the median sample price by fertilizer type to value the fertilizer.

8. Aggregate all the fertilizer expenditures on each type of fertilizer to get a total expenditure by the household on fertilizer in that year.

9. Recode missing values to zero as a number of households do not use any chemical fertilizers at all.

10. Create a dummy for whether the household had used any manure. To do this, there are two separate places in the 1997 survey that the household reports using manure. The first is in the remaining household level data file (key 10) – use this to create a dummy for whether the household reported using manure on maize. The second place is in field file (key 1). Here, restrict the observations to those for the main season and to just the maize fields. Then, create a dummy for the household using positive quantities of fertilizer types 13 or 18. For the overall dummy for manure use, take the maximum for the household across these two measures.

Fertilizer, 2004

In 2004, the fertilizer data is in a different file, a fertilizer specific field file (key 4), but the process is somewhat similar to the one described above for 1997.

1. Use the fertilizer field file (key 4) and keep the observations for the main harvest season and for the maize fields.

2. Merge on the location file (key 8).

3. Merge on fertilizer prices using the fertilizer prices file (key 5) and, as above, merge by district, year, fertilizer type and fertilizer unit.

4. Use the variable pfert to value the fertilizer quantities in KShs.

5. Aggregate all the fertilizer expenditures on each type of fertilizer to get a total expenditure by the household on fertilizer in that year.

6. Recode missing values to zero as a number of households do not use any chemical fertilizers at all.

7. Create a dummy for manure use, using the fertilizer field file (key 4). Create a dummy for any of the fields having positive quantities of fertilizer types 13 or 18. Collapse this by taking the maximum of this dummy by household-year.

Labor, 1997

1. Use the labor file (key 3). Note that for 1997, the data is only for the main season maize crop (another reason it is good that the analysis in the paper is for main season maize only).

2. Generate the total hired labor hours worked by adding up the number of men and women hired and multiplying by the hours worked. If the hours worked are missing, replace it with the reported average hours or, if that is missing, then use the average for hours worked for the sample by that activity (if that is missing replace with just average hours worked in the sample).

3. Generate the total family labor hours worked by adding up the number of men, women and children in the family that worked and multiplying by the hours worked. If the hours worked are missing, replace it with the reported average hours or, if that is missing, then use the average for hours worked for the sample by that activity (if that is missing replace with just average hours worked in the sample).

4. Generate the total labor hours for hired and family labor by multiplying the two created above by the days.

5. Collapse the total family hours and total hired hours to the household-year level.

Labor, 2004

The labor data was collected differently in 2004. So, there are two potential ways to compute the relevant labor, though the first one seems the correct way (which is what is used in the paper). I use the maize

specific labor in the paper (not the maize field specific labor). However, I did a robustness check to using maize field specific labor the and results do not change. Here I describe how to compute both versions.

Maize specific labor

1. Use the labor file (key 3) and keep the observations for the main season and if the crop is maize (regular or green).

2. Recode all the -888/-889/-899 to missing.

3. Multiply the three "time" labor variables (lb01, lb02, lb03) to get the hired labor hours. If this is missing replace it with the contract labor variable, lb04. This is the measure of hired labor used in hours.

4. To generate family labor hours, add together all the hours variables for the family members, lb06, lb08 and lb10.

5. Aggregate/collapse the hired and family labor hours variables to the household-year level.

Maize field specific labor

1. Use the labor file (key 3) and keep if the observations for the main season. Merge to the maize field identifier (created above) and keep only the fields that have maize on them. Drop any activities that do not specifically apply to maize.

2. Recode all the -888/-889/-899 to missing.

3. Multiply the three "time" labor variables (lb01, lb02, lb03) to get the hired labor hours. If this is missing replace it with the contract labor variable, lb04. This is the measure of hired labor used in hours.

4. To generate family labor hours, add together the hours variables for the family members, lb06, lb08 and lb10.

5. Aggregate/collapse the hired and family labor hours variables to the household-year level.

Demographics

1. Use the household demographics and education file (key 7).

2. Drop the people whose age or sex is missing.

3. Generate age-sex categories for (i) males ages 0-6, (ii) females ages 0-6, (iii) males ages 6-16, (iv) females aged 6-16, (v) males aged 17-39, (vi) females aged 17-39, (vii) males aged over 40, and (viii) females aged over 40.

4. Create a measure of household size and create fractions of the household that are in each of these age-sex categories.

5. Note that since the demographic files for 2004 come as three separate files (adults who were tracked from the previous round, additional adults and children) all these files have to be appended first before creating these measures of age-sex composition of the household.

Education

1. Use the household demographics and education file (key 7). Keep the adults (age is greater than equal to 17 years) and the household residents (they are resident for more than at least two months).

2. Drop the observations where education is missing.

3. Create dummy variables for the adult member having the following levels of education: (i) none, (ii) some primary, (iii) completed primary, (iv) some secondary, (v) completed Form 4, (vi) completed Form 6, and (vii) any university.

4. Collapse this down to the fraction of adults in the household having each of these levels of education.

5. Separately compute and merge on the education of the household head (use the relationship to the head variable in this file).

6. Note that since the demographic files for 2004 come as three separate files (adults who were tracked from the previous round, additional adults and children) all these files have to be appended first before creating these education measures. In addition, 2004 records years of education and so these have to be converted into the categories above. The same applies to the categories of education for the household head.
    a. Converted years of education to categories as follows: (i) none = 0 years of education, (ii) some primary = 1-7 (inclusive) years of education, (iii) completed primary = 8 years of education, (iv) some secondary = 9-11 (inclusive) years of education, (v) completed Form 4 = 12 years of education, (vi) Form 6 = 13 or 14 years of education, and (vii) any university = more than 14 years of education.

Maize Prices

1. Use the crop level data (key 2) and keep the main season observations and the observations for the maize crop (again keep both regular and green).

2. Drop if the unit for the price variable is missing.

3. Standardize the price variable to be price per kg by using the crop conversion units in the crop conversion file (key 6).

4. Merge on the location file (key 8).

5. Count how many observations there are per district for prices. If there more than 20, then compute the mean and median price by district. Else (i.e. if there are less than 20 observations on prices per district) use the overall sample mean and median. In computing the overall sample mean and median, remove the outliers, in particular the prices that are in the top and bottom two percentiles of the sample.

Wage Rates

1. Use the remaining household data file (key 10) which has the data on wage rates at the household level.

2. Merge on the location data file (key 8).

3. Drop the outliers in wages. For 1997, drop observations with wages >100 and wages <20. For 2004, drop observations with wages >250.

4. Count how many observations there are per district for wages. If there more than 5, then compute the mean and median wage by district. Else (i.e. if there are less than 5 observations on wages per district) use the overall sample mean and median.

Other Data

The last table in the paper also uses data from both years on the following: (i) distance to the closest fertilizer seller (this is also used as the excluded instrument in Table IV), (ii) distance to a motorable road, (iii) distance to a matatu stop, (iv) distance to extension services, (v) whether the household tried to get credit, (vi) whether the household tried to get credit but did not get it, (vii) whether the household received credit.

These variables come from the remaining household level data file (key 10). This file just requires some basic cleaning as follows:

1. Convert all the credit variables into dummy variables.

2. Replace values in the credit variables of 3 to missing (these are mistakes in the data).

3. Replace the distance variables as missing if they are -888/-889/-899.

Rainfall Data

I use rainfall data from the NOAA, which is available at the 0.1 degree by 0.1 degree of latitude and longitude. It is online at http://www.cpc.ncep.noaa.gov/products/international/africa/africa.shtml.

The rainfall data from NOAA is already provided as merged onto the Tegemeo household data in a file called tampa_rain (key 9). This was created by me and passed along to Tegemeo to provide alongside their data. Accompanying this file are data documents that describe how it was created. Please refer to those for more detail. I use only the relevant year's main season rainfall in the analysis throughout the paper.

Final Dataset

1. Merge all the files/variables created in the steps above to each other within year.

2. Append the different years to create a panel dataset.

3. Make sure to merge on the rainfall data (key 9), the maize prices (as computed above), the wages (as computed above), and the location data (key 8) to be able to control for province level dummies.

4. Compute all the relevant per acre variables.

5. Yields are measured in logs of harvest (in kg per acre) throughout the tables and fertilizer use in KShs of total expenditure on fertilizer per acre.

6. For the labor variables, I used family labor hours per acre and hired labor in KShs per acre. To compute the hired labor value in KShs, multiply the hours by the mean wage and divide by 6 as the mean wage is per day and I assume 6 hours of work per day. Replace the missing hired and family labor variables with zeros as those that did not use any family or hired labor did not fill in those variables in the survey.

7. I only use the maize prices to value the harvest and when I use either generalized revenue or profits as the dependent variable as a robustness check (see footnotes 22 and 51 in the paper). To compute profits, I valued family labor at the same wage rate as hired labor.