DATA PREPARATION AND ANALYSIS[1]

Margaret Beaver and Rick Bernsten
Revised June 2009

INTRODUCTION

Survey research is carried out through a series of steps. You must define a survey population, identify a sampling frame (list) which includes most, if not all, of the population units, select a sample of elements (*i.e.*, villages, households, farmers) from the frame, identify a reporting unit to whom questions are asked, interview respondents, prepare the data for computer entry, create data files, enter the data and analyze the data using various SPSS commands.

The following discussion of data preparation uses terms associated with the microcomputer version of the SPSS Inc. analysis package, now called **P**redictive **A**nalysis **S**oft**w**are (PASW), Version 17, to illustrate general principles. For more complete, detailed instructions about these concepts, consult the manual accompanying your statistical package.

OVERVIEW

We will use two terms that are basic to understanding survey data collection and analysis when using PASW or any other statistical software: the *unit of observation* which refers to the main focus of the survey or the highest level of the type of data being collected. The unit of observation can be a household level survey where multiple questions refer to the household (the household head is usually the respondent), a village level survey where questions are asked about the village (the village headman or the major is usually the respondent) or some other category. In contrast, the *unit of analysis* is the unit to which the data you have collected refer. For example, the unit of analysis for household level questions is the household, the unit of analysis for member level questions is the household – member. You will see below how these two basic terms are used to organize data for storage and for analysis.

Before you can analyze your raw data, you must organize it into files which can be entered into a computer and stored for subsequent manipulation. PASW Statistics uses several different types of files during data processing. Three of the most important types of files are:

the data files which use an extension of ".sav",
the syntax files which contain SPSS commands and use the extension of ".sps" and
viewer (output) files that contain the results of commands and use the extension of ".spv".

The *data file* includes both the information that defines your data in the file (*i.e.*, variable names, format, variable labels, value labels, and missing value codes) and the data values associated with this file. A data file must be created before you can analyze your data. SPSS can read in several different types of files such as Excel, dBase, SAS, Stata and text files. The user can also enter the data directly into the Data Editor window.

The PASW Statistics program has several modules, one of which is SPSS DATA ENTRY. This program allows you to define variables and enter data, using either a customized "facsimile questionnaire" or a spreadsheet data entry form. This program also allows you to correct data already entered, define valid-entry specifications (ranges) and define cleaning rules that check relationships between variables — such as the age with respect to marital status or education. After all data are entered, the file is saved in the

[1] Information Technologist, Food Security Research Project; and Professor, Department of Agricultural, Food and Resource Economics, Michigan State University, East Lansing, MI.

SPSS data file format. The data entry program automatically associates the variables, variable labels, value labels and missing value codes with the data when the data file is created.

Several terms have SPSS-specific definitions. The *case* is the basic building block of the data file. The case is the *unit of analysis* on which you have collected the data. Depending on the nature of the survey, the case may be the village, the household, the crop, the plot, a sales transaction, or some other unit of observation. Each question for which data is collected is called a *variable*. Each variable holds a position in the data file. In a single data file, each case must include a single unit of analysis and a common set of variables. While each case may have different data values for each variable, each variable must be assigned to the same data field in the questionnaire.

The *active file* is the file that is currently "active" in the SPSS for Windows program. The later versions of SPSS permit multiple data files to be open at the same time. However, only one data file can be "active" at a time. The data file that is designated as "active" is the one used when a command is run. It will have a green plus in the title bar on the left. A temporary name is assigned to the datasets as they are opened. The name can be changed using the DATASET command. You create an active file by calling up an existing file using the GET, IMPORT, MATCH FILES, ADD FILES, or AGGREGATE command. The GET command is used to open a data file on disk that has created in a previous PASW Statistics session. The IMPORT command is used to read into PASW Statistics a portable ASCII data file (or text file) created using another program. The MATCH FILES command allows you to combine two or more previously created SPSS data files. The AGGREGATE command allows you to create a new active file from the current active file by aggregating groups of cases into single cases. The active file must be saved as a data file after being created or it will be lost when you shut down the program, or, if you are using earlier versions of SPSS, when another file is opened.

FILE ORGANIZATION DECISION CRITERION

Data collected during a single survey may be organized into a single file or several different files, depending on the level of the data (unit of analysis). The key point that you must keep in mind in determining which variables should be placed in a file is the meaning of the concept, "the case". Each case should have a variable (or multiple variables) that identify the case as being unique.

Type of Analysis

To analyze the relationship between several variables, all the variables must exist in the same data file. Data may be put into a single file by (1) initially entering all of the data in the same file when it is created; by (2) adding data (variables) in one file to another file; and (3) by creating a new file which joins data (variables) from two or more files into a third file. For convenience you may want to enter all of the data into a single file. Data in a single file may be less time consuming and cumbersome to manipulate than if you have to add variables from other files or join file to complete your analysis. However, depending on the types of data collected, it could be much more difficult to manipulate data and compute new variables if all the data are entered into just one file, rather than following the rules for the different levels of data (unit of analysis).

Size of the Data Set

The number of variables in a single file for the later versions of PASW Statistics is no longer a limitation. With earlier versions, there is a limit of 500 variables per file. For these earlier versions, as you proceed with your analysis you will probably want to create new variables using various transformation procedures. Therefore, as a rule of thumb, it is advisable to initially create files with no more than 75 percent the total number of variables that your statistical program can analyze in a single file.

The larger the data file, the more time will be required for the computer to execute a specific analytical

procedure. Keep in mind that the size of the data file is a function of both the number of cases and the number of variables.

Unit of Analysis

A single file may include all the information collected from a questionnaire or multiple files may be required to record the data. The key criterion for determining if you can process the questionnaire as a single file, or as several separate files, is whether the unit of observation and the unit of analysis are the same for all of the data in the questionnaire.

Often, the unit of observation is the same as the unit of analysis. This is true if all the data solicited through a questionnaire are obtained from a single observer and there is only one response to each question. If these conditions are met, all the data may be included in a single data file.

On the other hand, sometimes information is solicited from one observer about other individuals or phenomena. For example: in a household survey, questions would be asked about the members as well as about what the household does as a household. The unit of analysis for the member questions would be household member whereas the unit of analysis for the household questions would be the household. In this instance, the unit of observation and the unit of analysis are different — each household may have a different number of members, resulting in a different number of cases for each household. Consequently, you must construct a file for the information for the household, where the household id identifies each case as being unique, and another file for the information about the members where the household id and member number are required to identify a case as being unique.

It may be easier to understand this concept using the term “levels of data”. All data that represent a single level may be included in a single file. When the data represents more than one level, a separate file is required for each. In practice, a single survey in which data are collected from one unit of observation often involves several units of analysis (different levels). In such instances, the data referring to each different unit of analysis (level) must be entered into a separate file. These concepts are explained in greater detail below.

QUESTIONNAIRE TYPES: SINGLE AND MULTIPLE UNITS OF ANALYSIS

While surveys may differ greatly in terms of their subject focus and specific information sought, all data collected in a single questionnaire may be classified in terms of two general types — those in which the unit of observation and the unit of analysis are the same and those in which they are different, or as previously noted, single level and multiple level data. The following examples illustrate these two types of questionnaires, using questions that are commonly asked in agricultural survey research.

Single Unit Of Observation Per Unit Of Analysis

In agricultural surveys, a sample of households from different villages is selected and the head of the household is designated as the unit of observation. Each household head is asked the same questions and is permitted to give only one "response" to each question (*e.g.*, each interview generates the same number of cases). Questions that illustrate this type of data solicit information from the household head about the number of household members, age of head, education of head, hectares of land cultivated and total maize produced. Each of these items represents a variable in the data file and each case (household) gives only one answer for each question. In this instance, the unit of observation is the same as the unit of analysis — all the data are collected at the same level. Consequently, all the data may be included in a single data file which we will call HOUSEHOLD. The “key variable” that identifies each case is HHID.

Multiple Units Of Analysis Per Unit Of Observation

Alternatively, the household head (unit of observation) may provide detailed information about

phenomena such as individuals in the household, crops grown, parcels farmed or activities pursued by household members — for which the numbers of responses per household varies from household to household. Since, each of these "levels of data" represents a different record type, has a different meaning for the "case" and may have a different number of cases per household, each type of data must be entered into a separate file. Several such record types frequently included in a single survey are illustrated below. In each example, note that "unit of observation does not equal unit of analysis":

1) *Household Member Level Data.* Suppose you plan to collect detailed data from each household about characteristics of each member such as age, education, sex, months living at home, income from off-farm employment, etc.. Because all households will not have the same number of members, these data fail to meet the criterion noted above. Therefore, this information about each household member must be placed in a data file separate from the household head, which we will call the HOUSEHOLD MEMBER level data file.

In this file, the case is the individual household member. Each case must include — at a minimum — a code to identify the household to which the member belongs. In addition, you should include a sequence number to uniquely identify each member (*i.e*. to indicate the first, second, third etc. household member). The remaining variables will store the responses to the questions regarding the specific member. You must enter this ID information for each household member (case), followed by the coded answer to each question (variable) about the household member's characteristics. The "key variables" that identify a case are HHID, MEM.

2) *Parcel Level Data*. Suppose you plan to collect information about each parcel (a piece of land with a single set of management practices) owned and/or managed by each household. Information may be asked about tenure status, area, distance from homestead, if the parcel grows one crop only or multiple crops, etc. Here, the unit of observation is the household head, but the unit of analysis is the parcel. Therefore, the information about each parcel must be placed in a separate data file we will call, the HOUSEHOLD PARCEL level data file.

In this file, the case is the individual parcel. Each record must include a code to identify the household owning and/or managing the parcel. In addition, you must include a parcel ID code and any additional information which you plan to use to sort and/or analyze this file. You must enter this ID information for each parcel (case), followed by the coded answer to each question (variable) about the parcel's characteristics. The "key variables" are HHID, PARCEL_ID.

3) *Household Parcel Crop Level Data*. Suppose you plan to collect information about the characteristics of each crop grown by the household, such as: crop type (*e.g.*, maize, sorghum, wheat), variety name, planting date, harvest date, area planted, production, kilograms of fertilizer applied, etc. You also want to keep track of the parcel where each crop is grown. Here, the unit of observation is the household, but the unit of analysis is the parcel - crop. Therefore, the crop information must be placed in a separate data file which we will call, the HOUSEHOLD PARCEL_ID CROP level data file.

In this file, the case is the individual crop within each parcel. Each case must include a code to identify the household plus a unique code to identify the parcel as well as the code to identify the crop grown within the parcel. Additional information about each crop within the parcel is entered into the variables that below with this parcel-crop. You must enter the ID information for each crop (case), followed by the coded answer to each question (variable) about the crop's characteristics. The "key variables" are HHID, PARCEL_ID, CROP.

4) *Transactions Characteristics*. In some instances, you may want to collect information about various types of transactions engaged in by each household, such as: labor used in farming, off-farm employment, consumption, product purchases and/or product sales. In each instance, the unit of analysis is the specific

type of transaction. Different data are collected for each transaction; therefore, each transaction type must be placed in a separate data file we will call, FARM LABOR, OFF-FARM LABOR, CONSUMPTION, PRODUCT PURCHASES and PRODUCT SALES level data files, respectively. However, if the same questions are asked about each transaction type, these data may be included in a single file. However, the type of transaction (crop sale, remittance, purchase, etc.) must be included as a variable with codes to reflect the type of transaction, so you can analyze the different transaction types separately.

In each file, the case is the individual transaction type. Each case must include a code to identify the household involved in the transaction. In addition, you may want to include other pieces of information which you plan to use to sort and/or analyze each file (*i.e.*, for farm labor data, you may want to code the parcel on which the work was carried out). You must enter the ID information for each transaction, followed by the coded answer to each question (variable) about the characteristics of the transaction. The number of "key variables" required to identify a case as being unique can vary with these types of data.

5) *Multiple Visit Surveys*. In some research designs the same respondents are interviewed several times during a season or year, using the same questionnaire. This strategy is employed to minimize recall error which arises when there is a long interval between when an event occurs and when the enumerator solicits information about the event. Most commonly, multiple visit surveys are used to collect transactions type data described above.

Each type of multiple visit transaction data must be entered into a separate file which you should create after the first round of data collection. Each subsequent round of data is initially entered into a separate file. In each file (for each round), the case is the individual transaction. Each case must include, at a minimum, a unique ID code to identify the household involved in the transaction and a coded value to indicate the date the transaction occurred or the visit number or the date when the data was collected. In addition, you may want to include other information which you plan to use to sort and/or analyze each file. You must enter the ID information for each transaction, followed by the coded answer to questions (variable) about the characteristics of each transaction. An example of what the primary "key variables" might be are: HHID, ROUND where the variable "ROUND" would contains the value to represent the visit – 1 = first visit, 2 = second visit, 3 = third visit. Other variables may be required depending on the level of the data.

To analyze the combined data for each transaction type across two or more data collection rounds, you must restructure the files to create an inclusive new file which will contain the data for all visits. You do this by merging individual transactions files of the same type (*i.e.*, the file created after each round) using the ADD FILES command. This command allows you to add cases (*i.e.*, individual transactions) across files with the same unit of analysis to create a single large file which includes all the transactions data of a single type. Individual transactions files may be joined at the completion of each round of data collection or after several such files have been constructed. In the example below, five separate data files, each representing a round of labor data, are joined together. The variables in each of the files have identical variable names.

```
ADD FILES FILE='c:\survey\labor_r1.sav'
  /FILE='c:\survey\labor_r2.sav'
  /FILE='c:\survey\labor_r3.sav'
  /FILE='c:\survey\labor_r4.sav'
  /FILE='c:\survey\labor_r5.sav'
EXECUTE.
SAVE OUTFILE='c:\survey\all_labor.sav'
  /COMPRESSED.
```

where: ADD FILES identifies the command to be executed

labour_r1.sav ... labour_r5.sav are the file names of the previously created labor transactions data files, for rounds one through five, and

SAVE OUTFILE tells SPSS to save the combined data to a new file named all_labor.sav

Once this new file, which includes transaction data from each round, has been created, it may be handled in the same way you would manipulate a transaction file where all of the data were collected during a single round.

Multiple Units Of Analysis Per Unit Of Observation – Constructed as a "Flat File"

As with all rules, there are exceptions or special cases. In some instances, you may have questions for which, strictly speaking, the unit of observation does not equal the unit of analysis. Yet, you do not have to create a separate file for these data if you know that for each case, the phenomena occur only a few times. For example, assume the household head is the unit of observation. Each household head is asked to provide information about each tractor owned. You know that most likely no household owns more than three tractors. In such instances – where the range in the number of possible responses is small – you do not have to prepare a separate TRACTOR file. Instead, you may code the information about each tractor as a variable in the HOUSEHOLD file. Each of the 3 variables is given a unique variable name. The case is still a household level case. The term for this type of file is "flat file".

For example, suppose for each unit, you collect data on the tractor's age, horsepower, hectares plowed and purchase price. You could code these data for the three tractors using variable names like:

TRAGE1, TRHP1, TRHA1, TRPRICE1,
TRAGE2, TRHP2, TRHA2, TRPRICE2
TRAGE3, TRHP3, TRHA3, TRPRICE3.

The key variable for this file would be HHID. For households owning three tractors, you would have data for each of these four variables for all three units of analysis (*i.e.*, the tractor). For households (cases) that do not own a tractor all of these variables would be blank. If they have only one tractor variables trage2 through trprice3 would be blank. The problem with deciding to collect the tractor data as a "flat" file where there is only one case per household is that should a household have more than 3 tractors, the information for the remaining tractors cannot be asked because variables have not been defined to accept information for more than 3 tractors.

Analysis can be more difficult with a file that is structured as a "flat file". If you want to know the average age of the tractors owned by the household, you would have to create a new variable to store the information, e.g.

```
COMPUTE tract_mean_age = MEAN(trage1, trage2, trage3).
VAR LAB tract_mean_age 'Average age of tractors owned'.
EXECUTE.
```

If a file had been created with a case for each tractor within the household, the average age for the tractors could have been obtained by using the DESCRIPTIVES command, e.g. DESCRIPTIVES trage.

You may use the above method for defining variables whenever you know there are a limited number of possible responses to a multiple response question (*i.e.*, you want to ask one question where the respondent may give several answers to that question). An example would be – Which magazines do you

read most frequently? Three responses are provided. All three variables contain the same choices. The respondent can fill in all three variables, or only one or two, if they don't read 3 magazines.

Additional situations in which you may want to use this procedure include instances where you collect data about multiple credit sources, multiple marketing points, and multiple extension contacts. As a rule of thumb, you should avoid this approach if there are more than three or four possible responses for a given line of questioning.

WITHIN FILE ANALYSIS

After you create each data file, defined according to the rules specified by your statistical package, you may proceed to analyze the data in each file. Two general types of within file analysis are discussed below: whole file analysis and whole file analysis of lower level data.

Whole File Processing

In whole file analysis, you process your whole data file as a unit, to produce statistics that are computed across all cases in the file, or selected subsets of the data using conditional selection procedures. This type of analysis is appropriate when you want to analyze all cases (or subsets) in the file (*i.e.*, households, crops, parcels, and transactions) as a single group.

Lower-Level Data Processing

For files which include varying numbers of cases per household, you will also want to analyze lower-level data for each household. For example, from the HOUSEHOLD MEMBER file you may want to count the number of adult male, adult female and child members in each household. From the PARCEL data file, you may want to calculate the total hectares for each type of tenure status for each household. From the FARM LABOR transactions file, you may want to calculate the total workdays of labor used for each operation by each household. From the PRODUCT SALES transactions file, you may want to estimate the total value of sales made each month by each household. The following example illustrates this type of analysis, using PASW Statistics commands.

You can generate household level analysis of lower-level data using the PASW Statistics MEANS command. It is assumed that the household will raise a crop on more than one parcel. For example, to calculate the average yield per hectare for each crop harvested by the household, using the PARCEL CHARACTERISTICS file, you would first use the COMPUTE command to calculate a new variable, yield (*i.e.*, production divided by area for each parcel). Then, use the MEANS command to calculate the average yield for each household by crop, as shown below:

```
MEANS  TABLES= yield BY hhid BY crop
  /CELLS MEAN COUNT STDDEV.
```

where: MEANS identifies the command to be executed,

The key word "TABLES=" is optional.

"yield" is the name of the dependent variable, yield.

"hhid" is the name of the first independent variable, household number,

"crop" is the name of the second independent variable, crop type, for which each crop has been assigned a different value.

/CELLS MEANS COUNT STDDEV identifies the statistics to be generated.

This combination of commands will produce results for each household which show the average yield/ha for each crop grown.

When initially creating each data file, you must include all the information needed to specify the subcategories you plan to use to carry out the desired analysis. For instance, in the above example, to calculate household level results from lower-level data, you had to include a unique identification number for each household and each crop type.

ACROSS FILE ANALYSIS: COMBINING FILES

In addition to within file analysis, you may want to analyze data from several different data files (*i.e.*, use variables in two or more data files in a given procedure). The steps required to execute this analysis depends on whether the unit of observation and unit of analysis are the same in each file. In the latest versions of PASW Statistics, there is no limit to the number of files that can be merged together. In older versions, a maximum of five files could be combined in a single ADD or MATCH operation, but you could perform as many JOIN operations as you needed.

Same Unit of Analysis in Both Files

Before you can analyze variables across these data files, you must first combine the files by linking them together using the MATCH command. This command merges all of the variables specified in selected input files (*i.e.*, the files which will be merged) into a single data file. In this example, two household surveys were conducted asking different questions in each of the surveys. In both instances, the household head was both the unit of observation and the unit of analysis.

This type of MATCH requires that both files be sorted in order of the "BY" variable which is the household ID to identify the household. If the files are not already sorted in this order, you must use the SORT CASES command before performing the MATCH.

The commands required to merge these files (assuming the second file has already been sorted and saved) are:

```
SORT CASES BY hhid.

MATCH FILES /FILE='c:\survey\hh1.sav'
   /FILE='c:\survey\hh2.sav'
   /BY hhid.
SAVE OUTFILE='c:\survey\hh_all.sav'
   /COMPRESSED.
```

where: hh1.sav is the first set of questions and hh2.sav is a different set of questions.

BY hhid - the cases from the two files should be matched by the variable "hhid", the variable required to identify the household - both files are at the same level - the household level.

SAVE OUTFILE tells SPSS to save the combined variables to a new file called "hh_all.sav"

The new data file includes all the variables from each of the original files. If a variable has the same name in each original data file, SPSS for Windows includes only the variable from the first data file listed in the MATCH command and drops the variable with the same name from the second file. If you wish,

you can specify that only a subset of variables from each data file are to be included in the combined data file, using the DROP or KEEP subcommand. If a case is missing from one of the files, that case will have missing values for all of the variables that would have come from that file.

Different Unit of Analysis in Each File

If two or more files have different units of analysis and you want to make them the same unit of analysis before you merge them together, you must first restructure these data files to the same unit of analysis. This process builds on the concept described in the discussion on household level analysis of lower-level data. Assume you want to jointly analyze data from the HOUSEHOLD level data file and the PARCEL level data file. In the HOUSEHOLD file, the unit of analysis is the household head (case), but in the PARCEL file it is the parcel (case). You must use the AGGREGATE command to summarize (aggregate) data in the PARCEL data file to produce a new dataset containing one case per household. In other words, this command creates a new dataset by aggregating groups of cases into single cases. Once this is accomplished, this new dataset can be matched to the HOUSEHOLD file to create a new inclusive data file. To use AGGREGATE, you should specify the name of the new aggregated dataset to be created, the variables that define break groups (*i.e*., variable required to uniquely identify the new level of data) and the functions to be used to create the aggregated variables.

You want to compare the yields of maize, groundnuts, and sorghum across all parcels planted by the household with the some information collected at the household level.

The required PASW Statistics commands are illustrated below:

```
GET
  FILE='c:\survey\parcel.sav'.
DATASET  NAME parcel WINDOW=FRONT.

*compute yields.
IF (CP=1) maz_yld=PROD/AREA.
VARIABLE LABEL maz_yld 'Maize yield'.
IF (CP=2) gn_yld=PROD/AREA.
VARIABLE LABEL gn_yld 'Groundnut yield'.
 IF (CP=3) sor_yld=PROD/AREA.
VARIABLE LABEL sor_yld 'Sorghum yield'.
EXECUTE.

*aggregate to household level.
DATASET DECLARE parcel_hh.
AGGREGATE
   OUTFILE= 'parcel_hh'
   /BREAK= hhid
   /maz_yld gn_yld sor_yld. = MEAN(maz_yld gn_yld sor_yld ).
DATASET ACTIVATE parcel_hh.
*merge the household level file into the newly created dataset.

MATCH FILES /FILE=*
   /FILE='c:\survey\hh_all.sav'
   /BY hhid.
```

where: GET FILE='c:\survey\parcel.sav' tells SPSS to open the data file named parcel.sav, gives the dataset the name of "parcel" and makes it the active file.

IF (CP=1) maz_yld=PROD/AREA creates a new variable maz_yld (maize yield) by dividing PROD (production) by AREA,

IF (CP=2) gn_yld=PROD/AREA creates a new variable gn_yld (groundnut yield) by dividing PROD (production) by AREA,

IF (CP=3) sor_yld=PROD/AREA creates a new variable sor_yld (sorghum yield) by dividing PROD (production) by AREA,

VARIABLE LABEL assigns a variable label to each of the newly created variables.

DATASET DECLARE 'parcel_hh' assigns a dataset name to a new data set that will be created.

.

AGGREGATE OUTFILE='parcel_hh' tells SPSS to place the aggregated data in the dataset named "parcel_hh",

/BREAK= hhid - identifies the break group as the variable HHID (*i.e.*, every time the HHID variable value change, write the summarized data to a new case in the dataset called "parcel_hh"), and

/maz_yld gn_yld sor_yld = MEAN(maz_yld gn_yld sor_yld) - defines the new variable names and states the type of function that should be used to compute the new variables (MEAN).

DATASET ACTIVATE parcel_hh - makes the new dataset that is created, the active dataset.

MATCH FILES merges the file containing household level questions on disk to the current active dataset (which is now at the household level) matching by the variable HHID.

The above commands will create a new active dataset which contains the new variables maz_yld, gn_yld sor_yld which are the average yields, across all parcels containing the crop, for each household. This active file may be saved as a new data file by including a SAVE command. For each household (case) the AGGREGATE command will create one case for the household with a new variable for each of the 3 crops. If a household does not grow one or more of the crops specified on the crop code designation, the yield for that crop will be designated as missing (*i.e.*, no observation). As this newly created file has one observation per household, the new variables (*i.e.*, average yield of each crop) may be subsequently merged with another HOUSEHOLD data file using the MATCH command described earlier.

As the need arises, you can create summary variables, such as yields, from several different data files that contain data on yields; and then create a new data file which includes all of these new variables--plus variables in other existing files--using the MATCH procedure. This sequence of file manipulations will enable you to join summary variables from several files, so you can perform analysis across data collected in all the various surveys completed. The important point to keep in mind, is that the BY variable must be specified with a MATCH command to assure that each household is matched to its own data between files.

CONCLUSIONS

Key concepts in data preparation and analysis are the "unit of observation", the "unit of analysis", "key variables", and the "case." The unit of observation refers to the person from whom the data were collected. The unit of analysis refers to the level of the data and will determine the level of the data file

that is to be created. The “key variables” are the variables required to identify the unit of analysis (case) as being unique.

Survey questionnaires often collect data that can be analyzed at a single level of analysis (i.e., all cases have only one response for each variable) or at multiple levels of analysis (i.e., the number of case varies across respondents; e.g., from each respondent, the enumerator would collect data about all family members, parcels, crops, transaction).

Analysis can be done using only the variable within a single file (within file analysis) or by combining variables from different files into one file (across file analysis). If files have different levels of data for the unit of observation, the files can be transformed to new units of analysis to be able to combined variables into one file for analysis.

Proper file organization is critical for data analysis. Before creating a data file, the researcher must carefully review the questionnaire and identify the level of the data, i.e., which variables can be entered into a single file (household level where the household is the primary level) (single level analysis) and which data must be entered into separate files where there are multiple cases per the primary (household) level (e.g. crop level or member level) (multiple level analysis).