Chapter XV
A guide for data management of household surveys

Juan Muñoz
Sistemas Integrales
Santiago, Chile

Abstract

The present chapter describes the role of data management in the design and implementation of national household surveys. It starts by discussing the relationship between data management and questionnaire design, and then explores the past, present and future options for survey data entry and data editing, and their implications for survey management in general. The following sections provide guidelines for the definition of quality control criteria and the development of data entry programs for complex national household surveys, up to and including the dissemination of the survey data sets. The final section discusses the role of data management as a support for the implementation of the survey sample design.

Key terms: consistency check, data cleaning, data editing, data management, household survey, quality control criteria

A. Introduction

1. Although the importance of data management in household surveys has often been emphasized, data management is still generally seen as a set of tasks related to the tabulation phase of the survey, in other words, activities that are conducted towards the end of the survey project, that use computers in clean offices at survey headquarters, and that are generally under the control of data analysts and computer programmers.

2. This restrictive vision of survey data management is changing. Experience from the past two decades shows that data management can and should play a critical role beginning with the very earliest stages of the survey effort. It is also becoming clear that data management does not terminate with the publication of the first statistical reports.

3. The clearest demonstration of effective data management efforts prior to the analytical phases has been given by the World Bank's Living Standards Measurement Study and other surveys that have successfully integrated computer-based quality controls with survey field operations. Even when data entry is not implemented as a part of fieldwork, data managers should participate in the design of questionnaires to ensure that the statistical units observed by the survey are properly recognized and identified, that skip instructions for the interviewers are explicit and correct, and that deliberate redundancies are eventually incorporated into the questionnaires that can be later used to implement effective consistency controls.

4. At the other extreme of the survey project timeline, the notion that the end product expected from the survey is a printed publication, with a collection of statistical tables, has been replaced by the concept of a database that not only can be used by the statistical agency to prepare the initial tables, but will also be accessible to researchers, policymakers and the public in general. The descriptive summary report of survey results is no longer seen as the final step, but rather as the starting point of a variety of analytical endeavours that may last for many years after the project is officially closed and the survey team is disbanded.

5. The present chapter begins with a discussion of the relationships between survey data management and questionnaire design, followed by an exploration of the past, present and future options for survey data entry and data editing, and their implications for survey management in general. The subsequent sections provide guidelines for the definition of quality control criteria and the development of data entry programs for complex national household surveys, up to and including the dissemination of the survey data sets. The final section discusses the role of data management as a support for the implementation of the survey sample design.

B. Data management and questionnaire design

6. Survey data management begins concurrently with questionnaire design and may to a large extent influence the latter. The data manager should be consulted on each major draft of the questionnaire, since he or she will have an especially sharp eye for flaws in the definition of units of observation, skip patterns, etc. The present section explores some of the formal aspects of the questionnaire that deserve attention at this point.

7. *Nature and identification of the statistical units observed*. Every household survey collects information about a major statistical unit - the household - as well as about a variety of subordinate units within the household - persons, budget items, plots, crops, etc. The questionnaire should be clear and explicit about just what these units are, and it should also ensure that each individual unit observed is properly tagged with a unique identifier.

8. The identification of the household itself generally appears on the cover page of the questionnaire. It sometimes consists of a lengthy series of numbers and letters that represent the geographical location and the sampling procedures used to select the household. Although it may seem self-evident, the use of all these codes as household identifiers should be critically assessed, because it is cumbersome, error-prone and expensive (often 20 digits or more may be needed to identify just the few hundred households in the sample); sometimes it does not even ensure unique identification of the unit as, for instance, when geographical codes on the cover page identify the dwelling but do not consider the case of multiple households in a dwelling. An easier and safer alternative is to identify the households by means of a simple serial number that can be handwritten or stamped on the cover page of the questionnaire, or even pre-printed by the print shop. Geographical location, urban/rural status, sampling codes and the rest of the data on the cover page then become important attributes of the household, which as such must be included in the survey data sets, but not necessarily for identification purposes. A good compromise between these two extremes (the list of all detailed sampling codes and a simple household serial number) is to give a three- or four-digit serial number to the primary sampling units (PSUs) used in the survey, and then a two-digit serial number to the households within each PSU.

9. The nature of the subordinate statistical units is often obvious (for instance, the members of the households are individual persons), but ambiguities may present themselves when what seems like an individual unit is in fact a multiplicity of units of a different kind. This may occur, for instance, when a man who has been asked to report on the main activity of his job conducts multiple, equally important activities at the same time or has more than one job in a given reference period. Similarly, ambiguity is possible when a woman who has been asked about the gender or weight of her last child gave birth to boy-girl twins with different weights. However, although such situations should of course be averted through good questionnaire design and piloting, they often arise in subtle ways, and this is where the critical vision of an experienced data manager can offer invaluable assistance to the subject matter specialists in spotting them.

10. Whatever its nature, subordinate units within the household should always be uniquely identified. This can be done by means of numerical codes assigned by the interviewer, but it is generally better to have these identifiers pre-printed in the questionnaire whenever possible.

11. *Built-in redundancies*. The design of the questionnaire may consider the inclusion of deliberate redundancies, intended to detect mistakes of the interviewer or data entry errors. The most common examples are:

- Adding a bottom line for “totals” under the columns that contain monetary amounts. Generating these totals may often be the interviewer’s task, but even when this is not the case, their inclusion is convenient because they are a very

effective way (often the only way) of detecting data entry errors or omissions. In fact, totals may be added for quality control purposes at the bottom of any numerical column, even when the sum of the numbers does not represent a meaningful measure of magnitude (for instance, a total may be added at the bottom of a column containing the quantities (not the monetary amounts) of various food items purchased, even if that means adding heterogeneous numbers, such as kilos of bread and kilos of potatoes (or even litres of milk). This point is further elaborated in the discussion of typographic checks below.

- Adding a check digit to the codes of some important variables (such as the occupation or activity of a person, or the nature of the consumption item). A check digit is a number or letter that can be deducted from the rest of the digits in the code by means of arithmetic operations performed at data entry time. A common check digit algorithm is the following: multiply the last digit in the code by 2, the second from last by 3, etc. (if the code is longer than six digits, repeat the sequence of multipliers 2, 3, 4, 5, 6, 7), and add the results. The check digit is the difference between this sum and the nearest higher multiple of 11 (the number 10 is represented by the letter K). Check digit algorithms are constructed so that the more common coding mistakes, such as transposing or omitting digits, will produce the wrong check digit.

C. Operational strategies for data entry and data editing

12. Many household surveys still consider data entry and editing as activities to be conducted in central locations, after the survey is fielded, whereas other surveys are already implementing the concept of integrating data entry into field operations. In the near future, the idea may evolve towards the application of computer-assisted interviewing. The present section discusses the organizational implications of the various strategies and the common and specific features of the data entry and data editing software developed under each alternative.

13. *Centralized data entry.* Centralized data entry was the only known option before the emergence of microcomputers, and it is still used today in many surveys. It considers data entry an industrial process, to be conducted in centralized data entry workshops after the end of the interviews. The objective of the operation is to convert the raw material (the information on the paper questionnaires) into an intermediate product (machine-readable files) that needs to be further refined (by means of editing programs and clerical processes) in order that a so-called clean database may be obtained as a final product.

14. During the initial data entry phase, the priorities are speed and ensuring that the information on the files perfectly reflects the information gathered in the questionnaires. Data entry operators are indeed not expected to "think" about what they are doing, but rather to just faithfully copy the data given to them. Sometimes, the questionnaires are submitted to double-blind data entry, in order to ascertain that this is done correctly.

15. Until the mid-1970s, data entry was carried out with specialized machines having very limited capabilities. Although, at present, the process is almost always carried out with microcomputers that can be programmed with quality control checks, this capability is seldom used in practice. The prevalent belief is that few quality control checks should be included in the data entry process, since the operators are not trained to make decisions as to what to do if an error is found. Besides, the detection of errors and their solutions slow down the data entry process. This school of thought considers that quality control checks should be solely reserved for the editing process.

16. *Data entry in the field*. Starting in the mid-1980s, the integration of computer-based quality controls into field operations has been identified as one of the keys to improving the quality and timeliness of household surveys. These ideas were initially developed by the World Bank's Living Standards Measurement Study (LSMS) surveys, and have been applied later to various other complex household surveys. Under this strategy, data entry and consistency controls are applied on a household-by-household basis as a part of field operations, so that errors and inconsistencies are solved by means of eventual revisits to the households.

17. The most important and direct benefit of integration is that it significantly improves the quality of the information, because it permits the correcting of errors and inconsistencies while the interviewers are still in the field rather than by office "cleansing" later. Besides being lengthy and time-consuming, office cleansing processes at best produce databases that are internally consistent but do not necessarily reflect the realities observed in the field. The uncertainty stems from the myriad of decisions - generally undocumented - that need to be made far from where the data are collected, and long after the data collection.

18. The integration of computer-based quality controls can also generate databases that are ready for tabulation and analysis in a timely fashion, generally just a few weeks after the end of field operations. In fact, databases may be prepared even as the survey is conducted, thus giving the survey managers the ability to effectively monitor field operations.

19. Another indirect advantage of integration is that it fosters the application of uniform criteria by all the interviewers and throughout the whole period of data collection, which is hard to achieve in practice with pre-integration methods. The computer indeed becomes an incorruptible and tireless assistant of the survey supervisors.

20. The integration of computer-based quality controls to field operations also has various implications for the organization of the survey, the most important being that it requires the field staff to be organized into teams. A field team is usually headed by a supervisor and includes a data entry operator in addition to two to four interviewers.

21. The organization of field operations depends on the technological options available. The two most used set-ups involve desktop and notebook computers and entail the following steps:

- Have the data entry operator work with a desktop computer in a fixed location (generally a regional office of the statistical agency,) and organize fieldwork so that the rest of the team visits each survey location (generally a primary sampling

unit) at least twice, so as to give the operator time to enter and verify the consistency of the data in between visits. During the second and subsequent visits, the interviewers will re-ask questions where errors, omissions or inconsistencies are detected by the data entry program.

- Have the data entry operator work with a notebook computer and join the rest of the team in its visits to the survey locations. The whole team stays in the location until all the data are entered and certified as complete and correct by the data entry program.

22. Both options have external requirements that need to be carefully considered by the survey planners and managers. One of them entails ensuring a permanent power supply for the computers, which may be an issue in poorly electrified countries. If desktops in fixed locations are used, this may require installing generators and ensuring that fuel for the generators is always available. If mobile notebooks are used instead, this may require the use of portable solar panels.

23. An obvious but important difference between the two strategies is that if computer-based quality controls are to be integrated into fieldwork, the data entry and editing program needs to be developed and debugged before the survey starts. With centralized data entry, this is also convenient (so that data entry can proceed in parallel with field operations,) but not absolutely necessary.

24. *Paperless interviews*. The use of hand-held computers to get rid of the paper questionnaires altogether is very appealing because of the advantages of automating certain parts of the interviews, such as skip instructions. However, although the technology has been available for almost 20 years, very little has been done to seriously apply this strategy to complex household surveys in developing countries. In fact, even in the most advanced national statistical agencies, paperless questioning has so far been restricted to relatively simple exercises, such as employment surveys and the collection of prices for the consumer price index.

25. A possible reason for this is that although paperless questioning lends itself well to interviews that follow a linear flow, with a beginning and an end, many household surveys conducted in developing and transition countries may require instead multiple visits to each household, separate interviews with each member of the household, or other procedures that are less strictly structured.

26. In spite of the absence of real empirical experience, certain observations about what needs to be taken into consideration in the design and implementation of a paperless questionnaire can be made:

- The data entry program interface will in some cases consist of a series of questions appearing one after the other on the computer screen, but in other cases it will need to reproduce the structure and visual format of the paper questionnaires, showing many data entry fields at the same time. This seems to be particularly important in the modules on expenditure and consumption, where the interviewer needs to "see" many consumption items simultaneously. The interface

must also allow for the possibility of marking questions in case of doubts, and it should also make it possible to return to the household for a second interview without repeating all questions.

- The questionnaire design process generally takes many months of work and involves many different people (subject-matter specialists, survey practitioners, etc.). With a paper questionnaire, the process is carried out by preparing, distributing, discussing and piloting various "generations" of the questionnaire until the final version is agreed upon. The equivalent steps for something that will never actually appear on paper still need to be defined.

- Interviewer training will need to be redesigned around the new technology. We know how to train interviewers to administer a paper questionnaire (theoretical sessions, simulations, mock interviews, training manuals, etc.,) but little work has been done to develop the equivalent techniques for a paperless survey.

- Finally, effective methods of supervision have to be developed. A large and rich set of procedures (visual inspection of the questionnaires, check-up interviews, etc.) has evolved for over a half-century to verify the work done by interviewers in the field. All these have been elaborated around the concept of a paper questionnaire and need to be re-engineered for paperless interviewing. It is very likely that the new technologies will offer completely different -- and possibly much more powerful -- options for effective supervision; for instance, most hand-held computers have voice recording capabilities that could be used to automatically record random parts of the interview along with the data files. By adding Global Position System (GPS) capabilities, it may also be possible to automatically record the time and place of the interviews. Again, the details have yet to be defined, field-tested and incorporated into the general scheme of survey fieldwork.

D. Quality control criteria

27. Regardless of the strategy chosen for quality control, the data on the questionnaires need to be subjected to five kinds of checks: range checks, checks against reference data, skip checks, consistency checks and typographic checks. Here, we revise the nature of these checks and the way they can be implemented under the various operational set-ups.

28. Range checks are intended to ensure that every variable in the survey contains only data within a limited domain of valid values. Categorical variables can have only one of the values predefined for them on the questionnaire (for example, gender can be coded only as "1" for males or "2" for females); chronological variables should contain valid dates, and numerical variables should lie within prescribed minimum and maximum values (such as 0 to 95 years for age.)

29. A special case of range checking occurs when the data from two or more closely related fields can be checked against external reference tables. Some common situations involve the following:

- *Consistency of anthropometric data*. In this case, the recorded values for height, weight and age are checked against the World Health Organization's standard reference tables. Any value for the standard indicators (height-for-age, weight-for-age and weight-for-height) that falls more than three standard deviations from the norm should be flagged as a possible error so that the measurement can be repeated.

- *Consistency of food consumption data*. In this case, the recorded values for the food code, the quantity purchased and the amount paid are checked against an item-specific table of possible unit prices.

30. Even when data are entered in centralized locations, it is generally convenient to detect and correct range errors in the initial data entry phase, rather than postpone this control for the editing phase, because range errors are often a result of the data entry operation itself rather than of interviewer mistakes. An error flag, such as a beep and a flashing field on the screen, may be set off when an out-of-range value is entered. If the error is merely typographical, the data entry operator can correct it immediately. It should, however, be possible to override the flag if the value entered represents what is on the questionnaire. In that case, an error report should be made so that the clerical staff can correct the error later by inspecting the questionnaire (or by the interviewer during a second interview, if the data are being entered in the field.) In the meantime, the suspect data item may be stored in a special format that registers its questionable status.

31. *Skip checks*. These verify whether the skip patterns have been followed appropriately. For example, a simple check verifies that questions to be asked only of schoolchildren are not recorded for a child who answered no to an initial question on school enrolment. A more complicated check would verify that the right modules of the questionnaire have been filled in for each respondent. Depending on his or her age and gender, each member of the household is supposed to answer (or skip) specific sections of the questionnaire. For instance, children less than 5 years of age should be measured in the anthropometric section but the questions about occupation are not asked of them. Women aged 15-49 years may be included in the fertility section but men may not be.

32. Sometime in the future, computer-assisted (paperless) interviews for surveys in developing countries may become common, and then the skipping scheme will possibly be controlled by the data entry program itself, at least in some cases. However, under the other operational set-ups (central data entry locations and data entry in the field), the data entry program should not actually follow the skip patterns on its own. For example, if the answer no is entered to the question, Are you enrolled in school?, the fields in which to enter data about the kind of school attended, grade in school and so on, should still be presented to the data entry operator. If there are answers actually recorded on the questionnaire, they can then be entered and the program will flag an incorrect skip. The supervisor or interviewer (or the centralized editing clerical staff) can determine the nature of the mistake at a later time. It may well be that

the no was supposed to be a yes. If the data entry program had automatically skipped the following fields, the error would not have been detected or remedied.

33. *Consistency checks*. These checks verify that values from one question are consistent with values from another question. A simple check occurs when both values are from the same statistical unit, for example, the date of birth and age of a given individual. More complicated consistency checks involve comparing information from two or more different units of observation.

34. There is no natural limit imposed on the number of consistency checks that can exist. Well-written versions of the data entry program for a complex household survey may have several hundred of them. In general, the more checks that are defined, the higher the quality of the final data set. However, given that the time available to write the data entry and data editing programs is always limited (usually about two months), expertise and good judgement are required to decide exactly which should be included. Certain consistency checks that are applicable in almost all household surveys have proved to be particularly effective and thus have become something of a de facto standard. These encompass:

- *Demographic consistency of the household*. The consistency between the ages and genders of all household members is checked with a view to kinship relationships. For example, parents should be at least (say) 15 years older than their children, spouses should be of different genders, etc.

- *Consistency of occupations*. The presence or absence of certain sections should be consistent with occupations declared individually by household members. For instance, the farming section should be present if and only if some household members are reported as farmers in the labour section.

- *Consistency of age and other individual characteristics*. It is possible to check that the age of each person is consistent with personal characteristics such as marital status, relationship to the head of the household, grade of current enrolment (for children currently in school) or last grade obtained (for those who have dropped out). For example, an 8-year-old child should not be in a grade higher than third.

- *Expenditures.* In this case, several different consistency checks are possible. Only in a household where one or more of the individual records show that a child is attending school should there be positive numbers in the household consumption record for items such as school books and schooling fees. Likewise, only households that have electrical service should report expenditures on electricity.

- *Control totals*. As said before, adding a control total wherever a list of numbers can be added is a healthy questionnaire design principle. The data entry program should check that the control total equals the sum of the individual numbers.

35. *Typographical checks*. In the early years of survey data processing, checking for typographical errors was almost the only quality control conducted at the time of data entry. This was generally achieved by simply having each questionnaire entered twice, by two different operators. These so-called double-blind procedures are seldom used nowadays, on the grounds that the other consistency controls that are now possible make them redundant. However, this may in some cases be wishful thinking rather a solid assumption.

36. A typical typographical error consists in the transposition of digits (like entering "14" rather than "41") in a numerical input. Such a mistake for age might be caught by consistency checks with marital status or family relations. For example, the questionnaire of a married or widowed adult age 41 whose age is mistakenly entered as 14, will show up with an error flag in the check on age against marital status. However, the same error in the monthly expenditure on meat may easily pass undetected, since either $14 or $41 could be valid amounts.

37. This emphasizes the importance of incorporating data management perspectives into the questionnaire design phase of the survey. Control totals, for instance, can significantly reduce typographical errors, because asking the interviewer to add up the figures with a pocket calculator is akin to entering them with double-blind data procedures. Check-digits can similarly be used for this purpose in some important variables. It is also possible to implement real double-blind methods for entering the data of certain parts of the questionnaire, but doing this for the whole questionnaire is both unnecessary and impractical – among other reasons, because modern data entry strategies are generally based on the work of a single data entry operator, not two different operators.

E. Data entry program development

38. The development of a good survey data entry and editing program is both a technique and a craft. The present section discusses some of the development platforms that are available today to facilitate the technical aspects of the process and some of the subtler issues related to the design of interfaces for the data entry operators and the future users of the survey data sets.

39. *Development platforms*. There are many data entry and editing program development platforms available in the market, but few of them are specifically adapted to the data management requirements of complex household surveys. A World Bank review conducted in the mid-1990s had found that at that time two DOS-based platforms were adequate: the World Bank's internally developed Living Standards Measurement Study (LSMS) package and the United States Bureau of the Census Integrated Microcomputer Processing System (IMPS) program. Both platforms have progressed since the review, in response to changing hardware and operating system environments. IMPS has been superseded by the Census and Survey Processing System (CSPro), a Windows-based application that provides some tabulation capabilities, besides serving in its primary role as a data entry and editing program development environment. The LSMS package has evolved towards LSD-2000, an Excel-based application that strives to develop the survey questionnaire and the data entry program simultaneously.

40. Both CSPro and LSD-2000 (or their ancestors) have proved their ability to support the development of effective data entry and editing programs for complex national household

surveys in many countries. These platforms are also easy to obtain and use. Almost any programmer -- in fact, almost anybody with a basic familiarity with computers -- can be expected to acquire in a couple of weeks the technical ability needed to initiate the development of a working data entry program.

41. *Design principles*. Unfortunately, development platforms cannot advise the programmers on just what data entry program needs to be developed. It may even be argued that the user-friendliness of the platforms risks making the development of inadequate data entry programs too easy. Confusing the mastery of the tools with the craft of putting the tools to good use is a mistake that survey managers should avoid by integrating both experienced programmers and subject-matter specialists in the development of the survey data entry and editing programs. Certain practical guidelines can be helpful in this regard:

- *Data entry screen design*. Data entry screens should look as much as possible like the corresponding pages of the questionnaire, but this rule has many exceptions. For example, if the questionnaire presents personal questions in the form of a matrix (with questions in rows and household members in columns, or the other way around), it is generally better to prepare a separate data entry screen for each person rather than to reproduce the paper grid on the computer screen. One reason for not reproducing the whole grid on the screen is that the number of respondents is variable. A stronger reason is that the statistical units observed are persons and not households.

- *Distinguishing between impossible and unlikely situations*. The data entry program should of course flag as errors any situations that represent logical or natural impossibilities (such as a girl's being older than her mother), but it should also react to situations that are not naturally impossible but very unlikely (such as a girl's being less than 15 years younger that her mother). Ideally, the data entry program should assess the severity of the errors and react differently depending on how serious they are, much as a human supervisor would if she or he was visually inspecting the questionnaire. This kind of "smart" programming is particularly important when data entry is integrated into field operations. Unfortunately, some programmers do not invest enough effort in this issue. A revealing sign is the tendency to always define the upper range of quantitative variables as "999…"(as many nines as the data entry field is long). The counterproductiveness of this practice is obvious: data entry fields should of course be long enough to input even the largest possible values, but the upper ranges should be small enough to flag unlikely values as possible outliers.

- *Error reporting language*. Some of the quality control criteria included in the data entry program may report on the errors detected by relatively simple means that are either self-explanatory or require little training to be understood. For instance, the LSMS data entry program reports range-checking errors by showing blinking arrows pointing up "↑↑↑" or down "↓↓↓" along with the offending value, depending on whether it is considered to be too low or too high. However, the most complex consistency controls require much clearer and more explicit

reporting. For instance, a check on the demographic consistency of the household could eventually produce a text such as "Warning: Lucy (ID Code 05, a girl 9 years old) is unlikely to be the daughter of Mary (ID Code 02, a woman 21 years old)", ideally on a printout rather than on the computer screen only. This kind of "smart and literate" programming may take longer than seemingly simpler alternatives (such as using error codes), but it will save many hours of fieldwork and field staff training, and it will also free the programmers themselves from the burden of writing an error codebook.

- *Variable codes*. A complex household survey typically contains hundreds of variables. The programmers in charge of the data entry program will need to refer to them by means of codes, according to the specific conventions of the development platform used. It is important that a rational and simple coding system be selected for this purpose from the beginning of the data entry program development process, because this will facilitate the communication between members of the development team, and also because it will save time in the subsequent steps of preparation and dissemination of the survey data sets. Finding a good coding system, however, can be harder than it seems. The process may start easily enough, with the first few variables getting codes such as "AGE", "GENDER" and so forth, but may soon become unmanageable, as finding adequate mnemonic codes becomes harder. A good option is to simply refer to the section and question numbers on the questionnaire, without any intent to make the codes self-explanatory (for example, if "Age" and "Gender" are variables 4 and 5 of Section 1, they could be coded as "S1Q4" and "S1Q5", respectively).

- *Data entry workloads*. When data entry is integrated into field operations, the most natural work unit for data entry is the household. This is because under these conditions the data entry operator always has only one or just a few questionnaires to work with, and also because consistency controls and error reporting are conducted on a household-by-household basis. In a central data entry location, the workloads can be blocks of 10-20 households (such as survey localities or PSUs). The idea is that: (a) the block should be entered by a single data entry operator in a single computer in at most a couple of days; and (b) the corresponding pack of questionnaires should be easily stored and retrieved at all times.

F. Organization and dissemination of the survey data sets

42. The structure of the survey data sets must reflect the nature of the statistical units observed by the survey. In other words, the data from a complex household survey **cannot** be stored in the form presented in table XV.1 directly below, that is to say, as a simple rectangular file with one row for each household and columns for each of the fields on the questionnaire.

Table XV.1. Data from a household survey stored as a simple rectangular file

	Variable 1	Variable 2	…	Variable j	…	Variable m
Household 1			…		…	
Household 2			…		…	
…	…	…	…	…	…	…
Household i			…	Datum i,j	…	
…	…	…	…	…	…	…
Household n			…		…	

43. Such a structure (also known as a "flat file") would be adequate if all of the questions referred to the household as a statistical unit, but as discussed before, this is not the case. Some of the questions refer to subordinate statistical units that appear in variable numbers within each household, such as persons, crops, consumption items and so forth. Storing the age and gender of each household member as different household-level variables would be both wasteful (because the number of variables required would be defined by the size of the largest household rather than by the average household size) and extremely cumbersome at the analytical stage (because even simple tasks such as obtaining the age-gender distribution would entail laboriously scanning a variable number of age-gender pairs in each household).

44. Both the CSPro and LSD-2000 platforms use a file structure that handles well the complexities that arise from dealing with many different statistical units, while minimizing storage requirements, and interfacing well with statistical software at the analytic phase.

45. The data structure maintains a one-to-one correspondence between each statistical unit observed and the records in the computer files, using a different record type for each kind of statistical unit. For example, to manage the data listed on the household roster, a record type would be defined for the variables on the roster and the data corresponding to each individual would be stored in a separate record of that type. Similarly, in the food consumption module, a record type would correspond to food items and the data corresponding to each individual item would be stored in separate records of that type.

46. The number of records in each record type is allowed to vary. This economizes the storage space required, since the files need not allow every case to be the largest possible.

47. In principle, only one record type is needed for each statistical unit, although sometimes more than one record type may be defined for the same unit for practical reasons. For instance, questions on education and health may be stored in two separate record types, even if the statistical unit is the person in both cases.

48. Each individual record is uniquely identified by a code in three or more parts. The first part is the "record type", which appears at the beginning of each record. It tells whether the information is, for example, from the cover page, or the health module, or for food expenditures. The record type is followed - in all records - by the household number. In most record types, a third identifier will be necessary to distinguish between separate statistical units of the same kind within the household, for instance, the person's identification number or the code of the expenditure item. In a few cases, there will be only one unit for the level of observation and thus the third identifier will be unnecessary. For example, housing characteristics are usually gathered for only one home per household. In a few cases, there may be an additional, fourth code. For example, the third identifier might be the household enterprise, and the fourth code would apply to each piece of equipment owned for each enterprise.

49. After the identifiers, the actual data recorded by the survey for each particular unit follow, recorded in fixed-length fields in the same order as that of the questions in the questionnaire. All data are stored in the standard American Standard Code for Information Interchange (ASCII) format.

50. The survey data sets need to be organized only as separate flat files (one for each record type) for dissemination, because the fixed-length field format of the native structure is also adequate for transferring the data to standard Database Management Systems (DBMSs) for further processing, or to standard statistical software for tabulation and analysis. Transferring the data to DBMSs is very easy because the native structure translates almost directly into the standard database format (DBF) that is accepted by all of them as input for individual tables (in this case, the record identifiers act as natural relational links between tables.)

51. Dissemination also requires that the structure of each record type be properly documented in a so-called Survey Codebook, which needs to be given to any user interested in working with the data sets. The codebook should clearly specify the position and length of each variable in the record. For categorical variables, it should also specify the encoding. The figure XV.1 below presents a page of the Nepal Living Standards Survey codebook (the encodings of certain variables were abridged).

Figure XV.1. Nepal living standards survey II

Record Type 002 **Section 1, Part A1: Household Roster**

VARIABLE	CODE	RT	FROM	LENGTH	TYPE		
Household		2	4	5	QNT		
ID CODE	IDC	2	9	2	QNT		
1 Name	Q01	2	11	24	TYP		
D Ethnicity	Q01A	2	35	3	QLN	Chhatri	001
						Brahmin Hill	002
						...	...
						Others	102
2 Gender	Q02	2	38	1		Male	1
						Female	2
3 Relationship	Q03	2	39	2		Head	1
						Spouse	2
						Child	3
						...	...
						Other relative	11
						Servant/Servant's relative	12
						Tenant/Tenant's relative	13
						Other Perbon non-related	14
4A District born in	Q04A	2	41	2	QLN	Taplejung	01
						Panchthar	02
						...	...
						Other country	93
4B District born U/R	Q04B	2	43	1	QLN	Urban	1
						Rural	2
5 Age	Q05	2	44	2	QNI		
6 Marital status	Q06	2	46	1	QLN	Married	1
						Divorced	2
						Separated	3
						Widowed	4
						Never married	5
7 Spouse in list?	Q07	2	47	1	QLN	Yes	1
						No	2
8 ID Code of Spouse	Q08	2	48	2	QNT		
9 Months at home	Q09	2	50	2	QNT		
10 Member or not?	Q10	2	52	1	QLN	Yes	1
						No	2

52. Both the CSPro and LSD-2000 platforms permit producing the survey codebook as a by-product of the data entry program development process. LSD-2000 also provides interfaces to convert the data entry files into DBF files and to transfer the data into the most commonly used statistical software (Ariel, CSPro, SAS, SPSS and Stata). This emphasizes the importance of defining a variable encoding system carefully at the data entry program development phase: if this is done well, the survey analysts will be able to immediately use the survey data when the data sets become available.

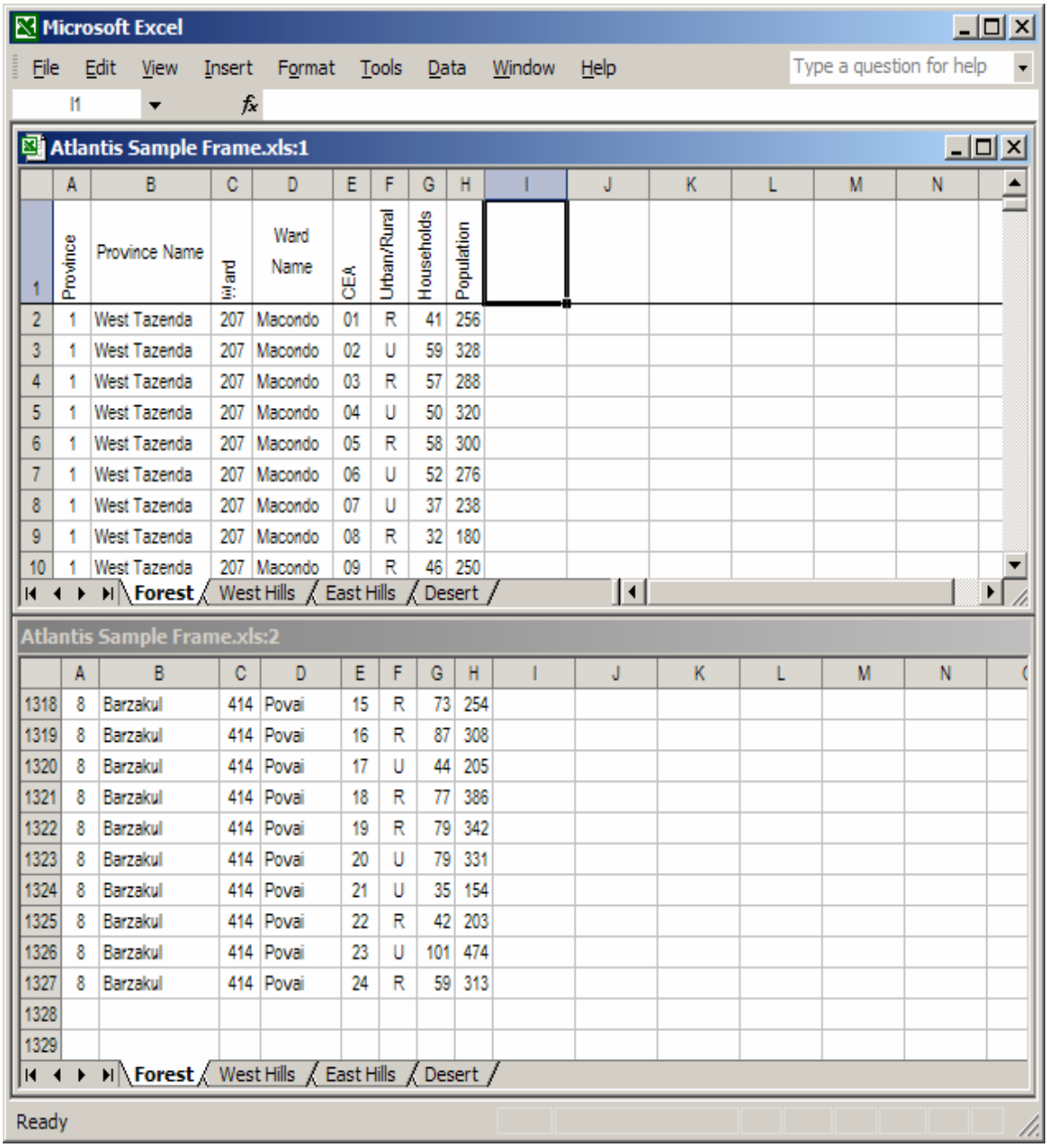
G. Data management in the sampling process

53. The present section discusses the role of data management in the design and implementation of household survey samples. It contains recommendations for the computerization of sampling frames and for conducting the first stages of sampling selection, including practical methods for implicit stratification and sampling of primary sampling units (PSUs) with probability proportional to size (PPS). The development of a database with the penultimate sampling units as a by-product of the prior sampling stages is discussed, emphasizing its role as a management tool while the survey is fielded, and how its contents can

be updated with field-generated information (such as the results of the household listing operation and the data on non-response) in order to generate the sampling weights to be used at the analytical stage.

54. *Organization of the first stage sampling frame.* The first-stage sampling units for many household surveys are the census enumeration areas (CEAs) defined by the most recently available national census. Creating a computer file with the list of all CEAs in the country is a convenient and efficient way to develop the first-stage sampling frame. Except in countries where the number of CEAs is massive (such as Bangladesh with over 80,000), the best way to do this is with a spreadsheet program such as Excel, with one row for each CEA, and columns for all the information that may be required. It must include the full geographical identification of the CEA and a measure of its size (such as the population, the number of households or the number of dwellings). It is generally more convenient to create a different worksheet for each of the sample strata. Figure XV.2 below shows how a first-stage sampling frame could look in the "Forest" stratum of a hypothetical country (the Excel screen has been split into two windows to show the first and last CEAs simultaneously).

Figure XV.2. Using a spreadsheet as a first-stage sampling frame



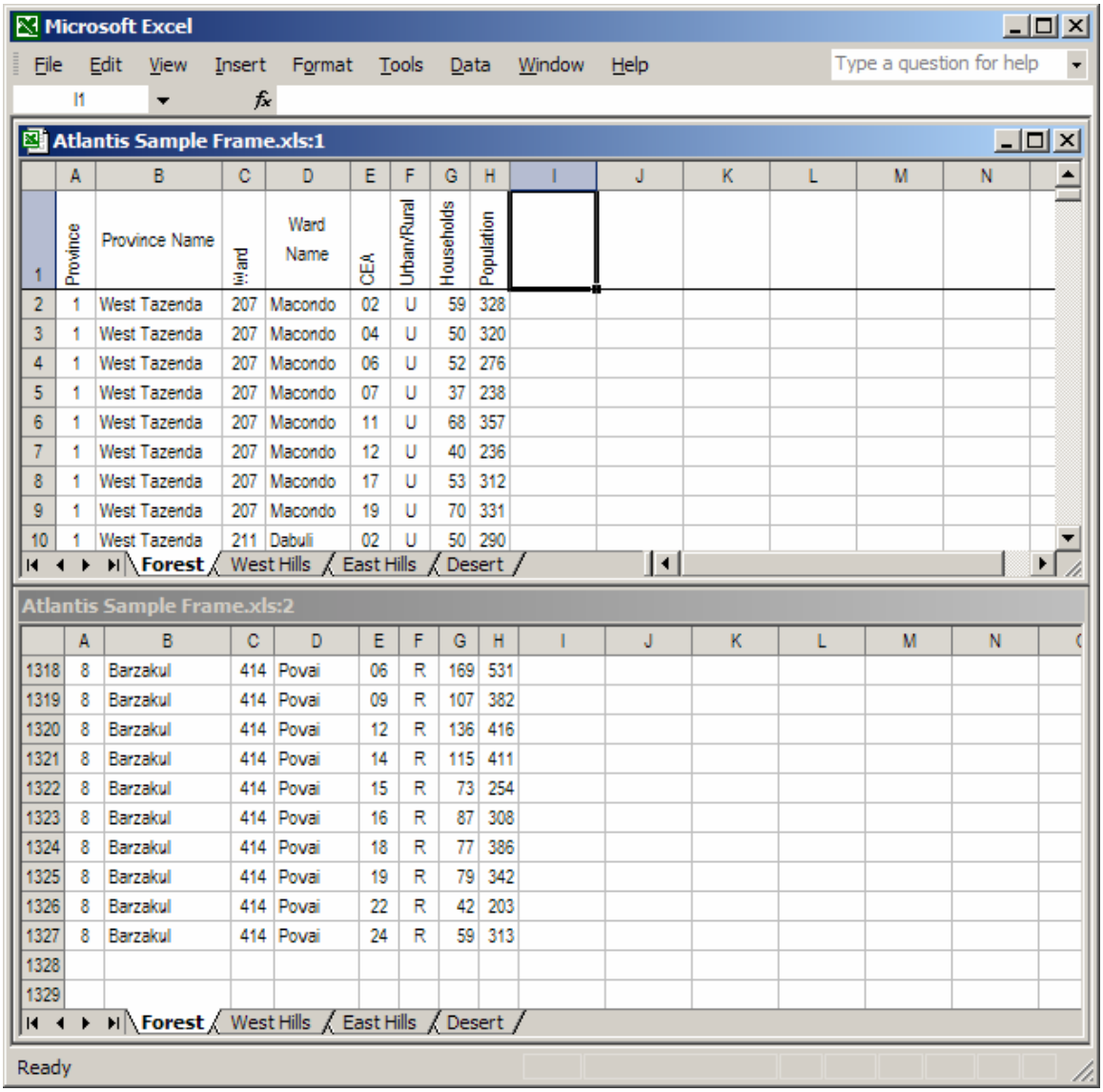
55. In this example, the 1,326 CEAs in the Forest stratum are identified by means of the geographical codes and names of the country's administrative divisions (provinces and wards) and by a serial number within each ward. The sampling frame also contains the number of households and the population of each CEA at the time of the census, and indicates whether the CEA is urban or rural.

56. Before proceeding with the next steps of sampling selection, it is critical to verify that the sampling frame is complete and correct by checking the population figures with the census totals published by the statistical agency. It is also important to verify that the size of all CEAs is sufficiently large to permit their use as primary sampling units. If the sample design calls for

penultimate-stage clusters of, for example, 25 households each, it will not be possible to meet that requirement in CEAs of fewer than 25 households. In that case, small CEAs should be combined with geographically adjacent CEAs to constitute primary sampling units. This process may be tedious if the quest for neighbouring CEAs has to be conducted by hand, by continuous reference to the census maps. However, since statistical agencies often assign the CEA serial numbers according to some geographical criterion (the so-called serpentine or "spiral" orderings), so that the CEAs that are neighbours in the spreadsheet are also neighbours in the territory, it is generally possible to make the combinations automatically in the spreadsheet. In our example, every CEA has over 30 households, so no grouping is needed. It should be noted, however, that the illustration above is somewhat unrealistic for effectuating this procedure because urban and rural CEAs are mixed in the numerical listing, a situation not likely to be encountered in an actual country. In other words, grouping of adjacent CEAs by computer cannot be effected when urban and rural CEAs are scattered in the list rather than grouped together.

57. Another step preceding the first sampling stage is deciding if the sampling frame needs to be sorted by certain design criteria in order to implicitly stratify the sample within each of the explicit strata. Administrative divisions are almost always used for this purpose but, in some cases, another criterion -- that is to say, urban/rural stratification -- may be considered even more important. Assuming that in our example, this is the case in respect of the urban/rural classification, the sampling frame needs to be sorted by urban/rural, then by province, after that by ward, and finally by the CEA serial number. This can be easily done with the "sort" command provided by the spreadsheet program in figure XV.3.

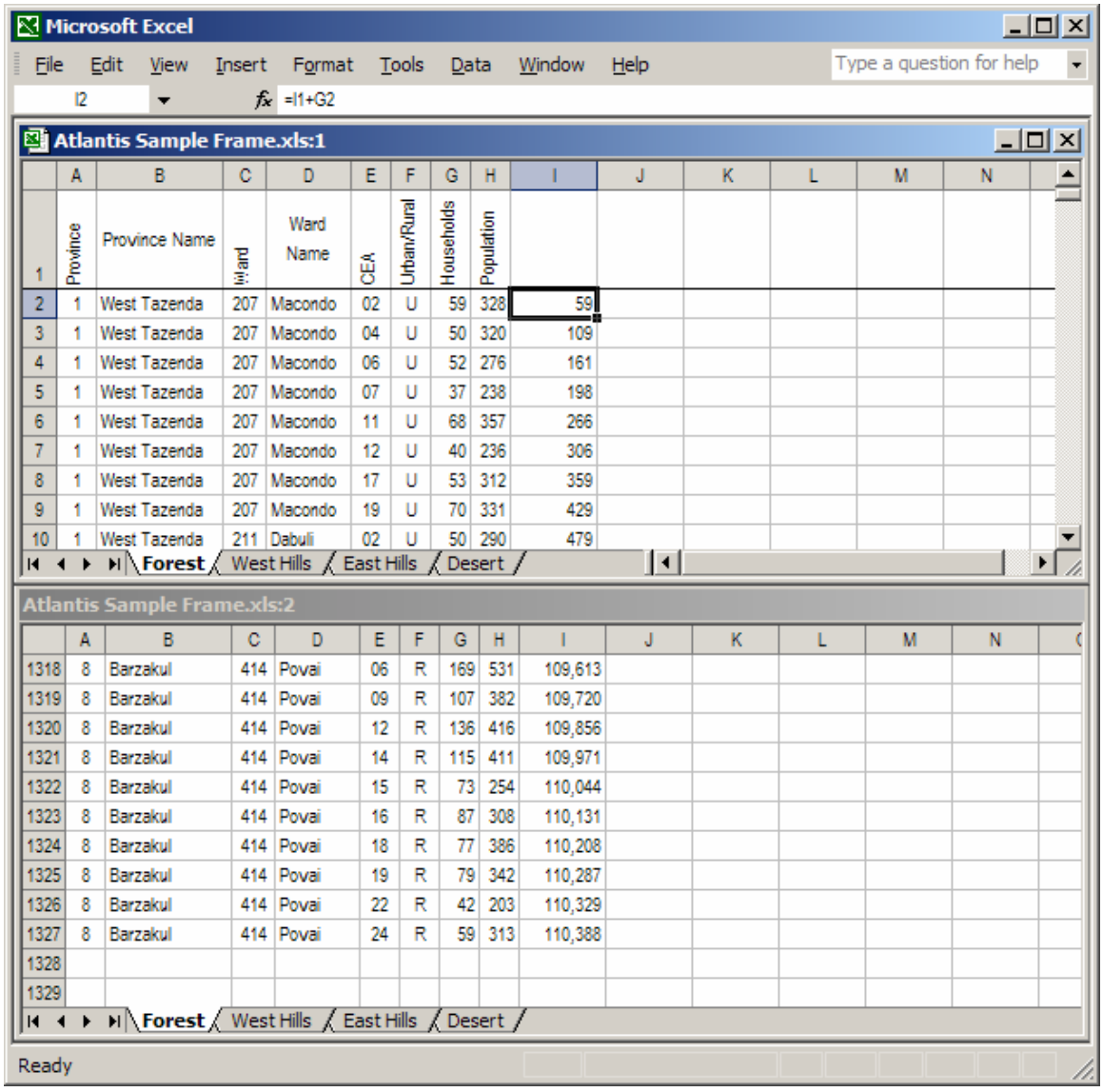
Figure XV.3. Implementing implicit stratification



58. *Selecting primary sampling units with probability proportional to size*. Most household surveys select the primary sampling units using probability proportional to size (PPS). When it is available in the sampling frame, the number of households in the CEA is generally used as a measure of its size, but in some cases the population or the number of dwellings can be used instead. We will now illustrate the PPS procedure, assuming that the design calls for the selection of 88 CEAs with probability proportional to the number of households (column **G** of the worksheet) in the Forest stratum (see figure XV.4).

59. First, create a new column in the spreadsheet, with the cumulated size of the CEAs. Enter the formula **=I1+G2** in cell **I2** and copy it all the way down to the last row of column **I** (notice that the last row in column **I** will contain the total number of households in the Forest stratum (110,388).

Figure XV.4. Selecting a PPS sample (first step)

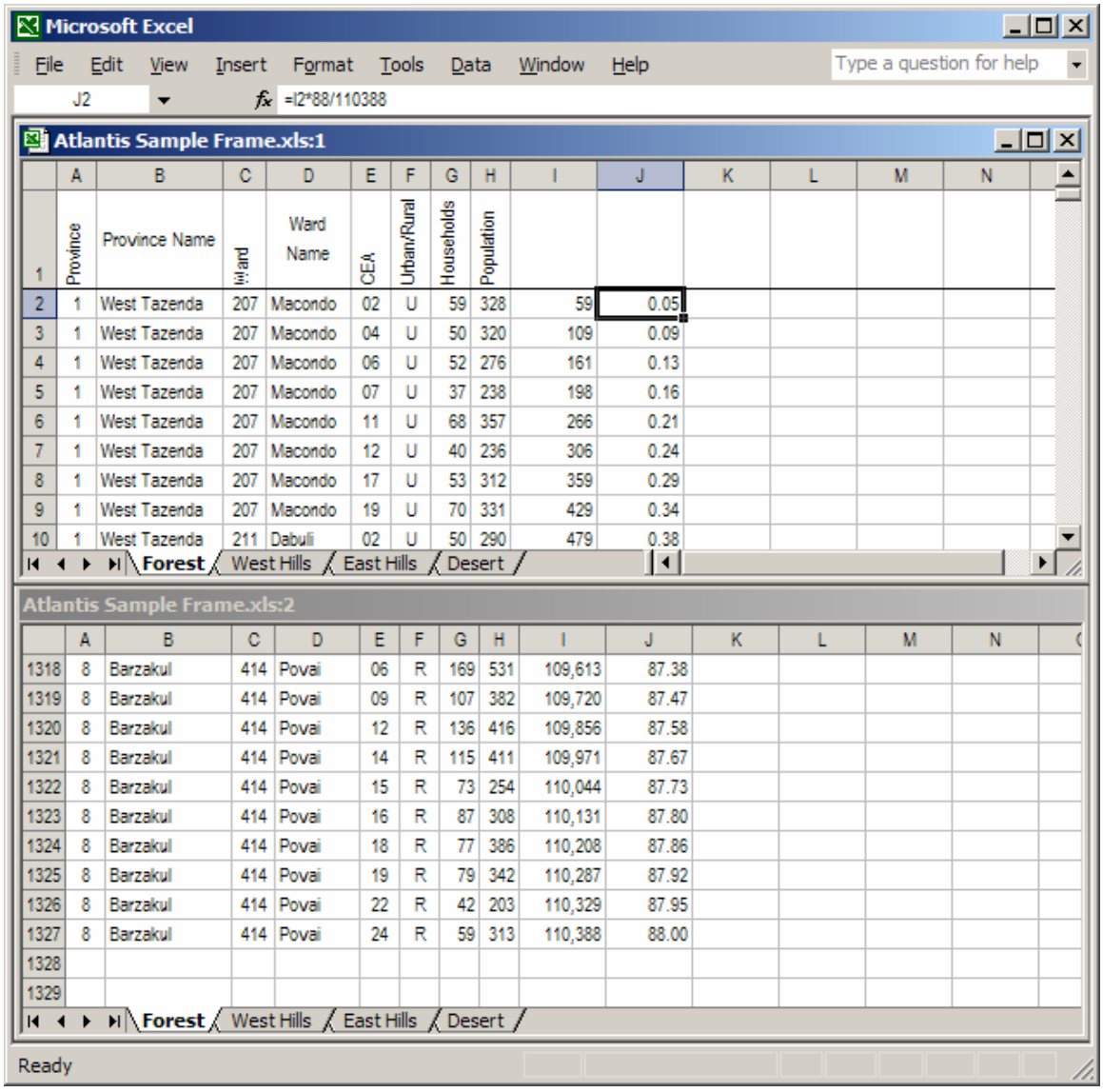


	A	B	C	D	E	F	G	H	I
1	Province	Province Name	Ward	Ward Name	CEA	Urban/Rural	Households	Population	
2	1	West Tazenda	207	Macondo	02	U	59	328	59
3	1	West Tazenda	207	Macondo	04	U	50	320	109
4	1	West Tazenda	207	Macondo	06	U	52	276	161
5	1	West Tazenda	207	Macondo	07	U	37	238	198
6	1	West Tazenda	207	Macondo	11	U	68	357	266
7	1	West Tazenda	207	Macondo	12	U	40	236	306
8	1	West Tazenda	207	Macondo	17	U	53	312	359
9	1	West Tazenda	207	Macondo	19	U	70	331	429
10	1	West Tazenda	211	Dabuli	02	U	50	290	479

	A	B	C	D	E	F	G	H	I
1318	8	Barzakul	414	Povai	06	R	169	531	109,613
1319	8	Barzakul	414	Povai	09	R	107	382	109,720
1320	8	Barzakul	414	Povai	12	R	136	416	109,856
1321	8	Barzakul	414	Povai	14	R	115	411	109,971
1322	8	Barzakul	414	Povai	15	R	73	254	110,044
1323	8	Barzakul	414	Povai	16	R	87	308	110,131
1324	8	Barzakul	414	Povai	18	R	77	386	110,208
1325	8	Barzakul	414	Povai	19	R	79	342	110,287
1326	8	Barzakul	414	Povai	22	R	42	203	110,329
1327	8	Barzakul	414	Povai	24	R	59	313	110,388
1328									
1329									

60. Second, create another column with the scaled cumulated size of the CEAs, multiplying the values in column **I** by the scaling factor 88/110,388 (the idea is to have a column that grows from zero to the number of CEAs to be selected, proportionally to the size of the CEAs; see figure XV.5). Enter the formula **=I2*88/110388** in cell **J2** and copy it all the way down to the last row of column **J**:

Figure XV.5. Selecting a PPS sample (second step)

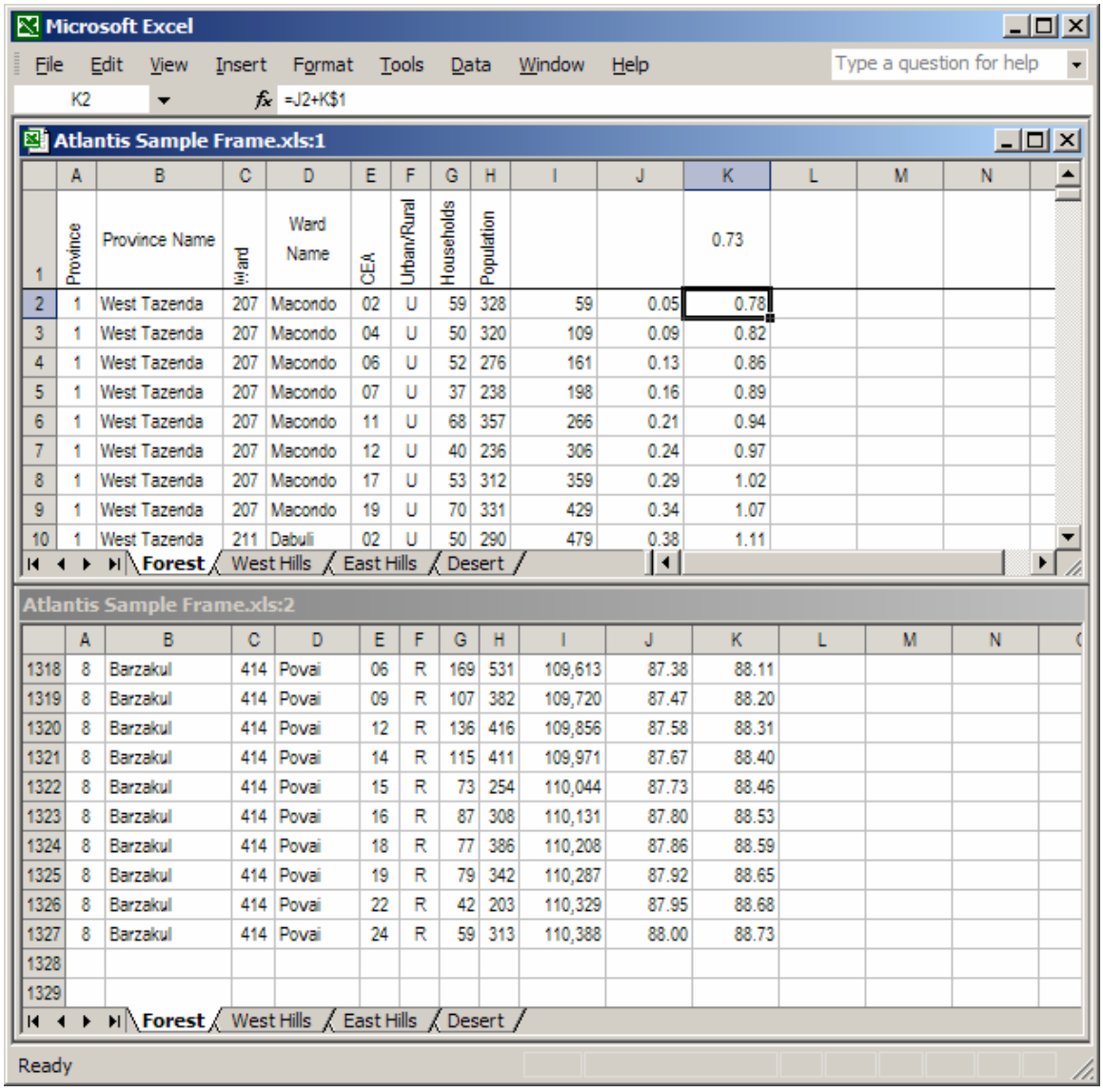


	A	B	C	D	E	F	G	H	I	J
1	Province	Province Name	Ward	Ward Name	CEA	Urban/Rural	Households	Population		
2	1	West Tazenda	207	Macondo	02	U	59	328	59	0.05
3	1	West Tazenda	207	Macondo	04	U	50	320	109	0.09
4	1	West Tazenda	207	Macondo	06	U	52	276	161	0.13
5	1	West Tazenda	207	Macondo	07	U	37	238	198	0.16
6	1	West Tazenda	207	Macondo	11	U	68	357	266	0.21
7	1	West Tazenda	207	Macondo	12	U	40	236	306	0.24
8	1	West Tazenda	207	Macondo	17	U	53	312	359	0.29
9	1	West Tazenda	207	Macondo	19	U	70	331	429	0.34
10	1	West Tazenda	211	Dabuli	02	U	50	290	479	0.38

	A	B	C	D	E	F	G	H	I	J
1318	8	Barzakul	414	Povai	06	R	169	531	109,613	87.38
1319	8	Barzakul	414	Povai	09	R	107	382	109,720	87.47
1320	8	Barzakul	414	Povai	12	R	136	416	109,856	87.58
1321	8	Barzakul	414	Povai	14	R	115	411	109,971	87.67
1322	8	Barzakul	414	Povai	15	R	73	254	110,044	87.73
1323	8	Barzakul	414	Povai	16	R	87	308	110,131	87.80
1324	8	Barzakul	414	Povai	18	R	77	386	110,208	87.86
1325	8	Barzakul	414	Povai	19	R	79	342	110,287	87.92
1326	8	Barzakul	414	Povai	22	R	42	203	110,329	87.95
1327	8	Barzakul	414	Povai	24	R	59	313	110,388	88.00
1328										
1329										

61. Third, enter a uniformly distributed random number between 0 and 1 in the topmost cell of a new column and add it to all rows of column **J**, to create a new column with the randomly shifted scaled cumulated size (see figure XV.6). It is possible to select random numbers automatically within the spreadsheet, but it is better to select this random shift externally (using a table of random numbers, for instance) to prevent the system from selecting a different sample whenever the workbook is recalculated. Enter, for instance, the random number 0.73 in cell **K1**, then enter the formula **=J2+K$1** in cell **K2** and copy it all the way down column **K**.

Figure XV.6. Selecting a PPS sample (third step)

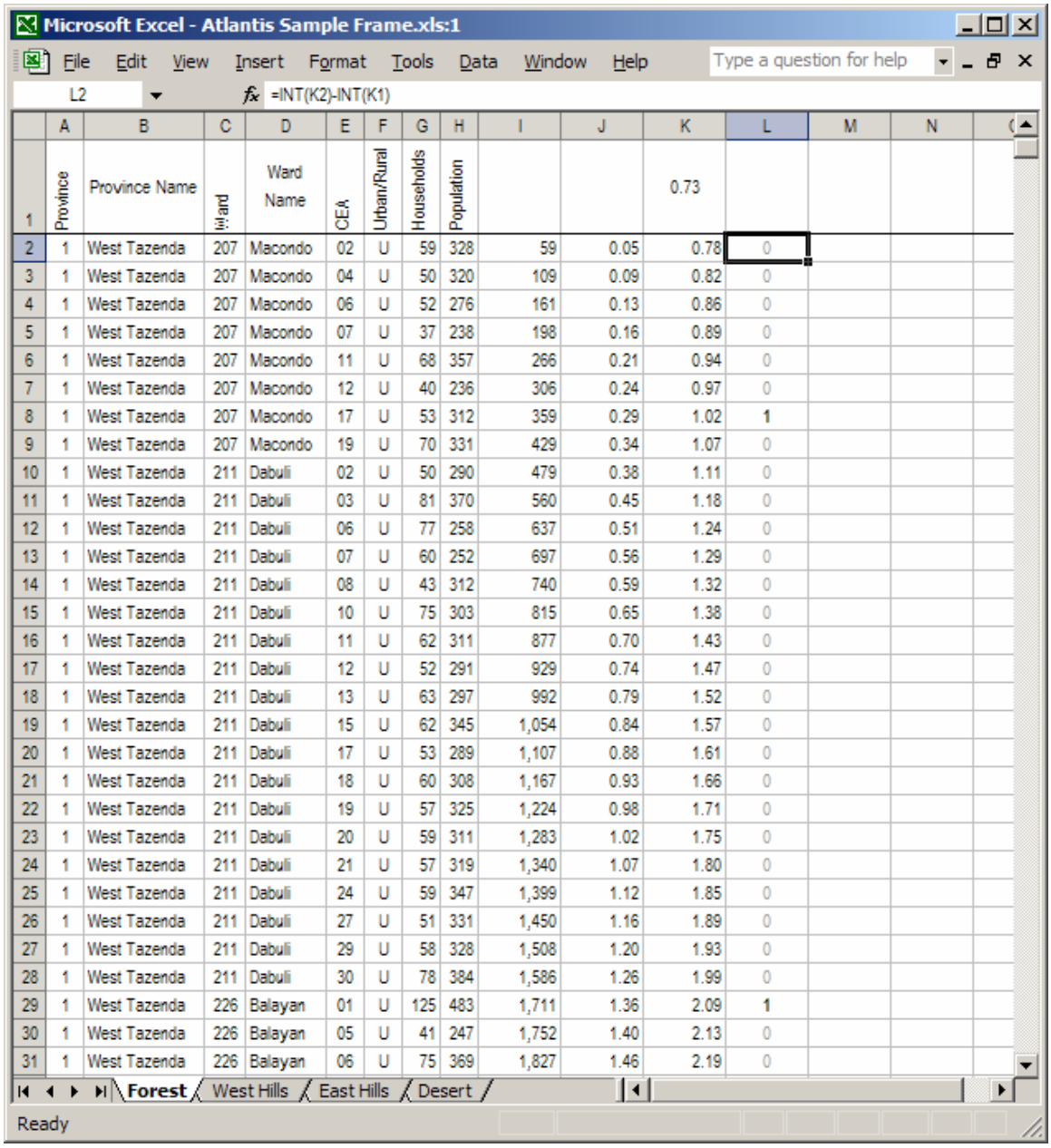


	A	B	C	D	E	F	G	H	I	J	K
1	Province	Province Name	Ward	Ward Name	CEA	Urban/Rural	Households	Population			0.73
2	1	West Tazenda	207	Macondo	02	U	59	328	59	0.05	0.78
3	1	West Tazenda	207	Macondo	04	U	50	320	109	0.09	0.82
4	1	West Tazenda	207	Macondo	06	U	52	276	161	0.13	0.86
5	1	West Tazenda	207	Macondo	07	U	37	238	198	0.16	0.89
6	1	West Tazenda	207	Macondo	11	U	68	357	266	0.21	0.94
7	1	West Tazenda	207	Macondo	12	U	40	236	306	0.24	0.97
8	1	West Tazenda	207	Macondo	17	U	53	312	359	0.29	1.02
9	1	West Tazenda	207	Macondo	19	U	70	331	429	0.34	1.07
10	1	West Tazenda	211	Dabuli	02	U	50	290	479	0.38	1.11

	A	B	C	D	E	F	G	H	I	J	K
1318	8	Barzakul	414	Povai	06	R	169	531	109,613	87.38	88.11
1319	8	Barzakul	414	Povai	09	R	107	382	109,720	87.47	88.20
1320	8	Barzakul	414	Povai	12	R	136	416	109,856	87.58	88.31
1321	8	Barzakul	414	Povai	14	R	115	411	109,971	87.67	88.40
1322	8	Barzakul	414	Povai	15	R	73	254	110,044	87.73	88.46
1323	8	Barzakul	414	Povai	16	R	87	308	110,131	87.80	88.53
1324	8	Barzakul	414	Povai	18	R	77	386	110,208	87.86	88.59
1325	8	Barzakul	414	Povai	19	R	79	342	110,287	87.92	88.65
1326	8	Barzakul	414	Povai	22	R	42	203	110,329	87.95	88.68
1327	8	Barzakul	414	Povai	24	R	59	313	110,388	88.00	88.73
1328											
1329											

62. The sample is defined by the rows where the integer part of the shifted scaled cumulated size change. In this example, the shifted scaled cumulated size changes from 0.97 to 1.02 for CEA number 17 in ward number 207 (Macondo) of province number 1 (West Tazenda), implying that this is the first CEA to be selected in the sample. The value changes again, from 1.99 to 2.09 in CEA number 01 of ward 226 (Balayan) of the same province, so that this is the second CEA selected. The selected sample can be flagged automatically by entering the formula **`=INT(K2)-INT(K1)`** in cell **`L2`** and copying it all the way down column **`L`**. The sample is defined by the rows with a non-zero value in column **`L`** (see figure XV.7).

Figure XV.7. Selecting a PPS sample (fourth step)



Microsoft Excel - Atlantis Sample Frame.xls:1

File Edit View Insert Format Tools Data Window Help — Type a question for help

L2 =INT(K2)-INT(K1)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Province	Province Name	Ward	Ward Name	CEA	Urban/Rural	Households	Population			0.73			
2	1	West Tazenda	207	Macondo	02	U	59	328	59	0.05	0.78	0		
3	1	West Tazenda	207	Macondo	04	U	50	320	109	0.09	0.82	0		
4	1	West Tazenda	207	Macondo	06	U	52	276	161	0.13	0.86	0		
5	1	West Tazenda	207	Macondo	07	U	37	238	198	0.16	0.89	0		
6	1	West Tazenda	207	Macondo	11	U	68	357	266	0.21	0.94	0		
7	1	West Tazenda	207	Macondo	12	U	40	236	306	0.24	0.97	0		
8	1	West Tazenda	207	Macondo	17	U	53	312	359	0.29	1.02	1		
9	1	West Tazenda	207	Macondo	19	U	70	331	429	0.34	1.07	0		
10	1	West Tazenda	211	Dabuli	02	U	50	290	479	0.38	1.11	0		
11	1	West Tazenda	211	Dabuli	03	U	81	370	560	0.45	1.18	0		
12	1	West Tazenda	211	Dabuli	06	U	77	258	637	0.51	1.24	0		
13	1	West Tazenda	211	Dabuli	07	U	60	252	697	0.56	1.29	0		
14	1	West Tazenda	211	Dabuli	08	U	43	312	740	0.59	1.32	0		
15	1	West Tazenda	211	Dabuli	10	U	75	303	815	0.65	1.38	0		
16	1	West Tazenda	211	Dabuli	11	U	62	311	877	0.70	1.43	0		
17	1	West Tazenda	211	Dabuli	12	U	52	291	929	0.74	1.47	0		
18	1	West Tazenda	211	Dabuli	13	U	63	297	992	0.79	1.52	0		
19	1	West Tazenda	211	Dabuli	15	U	62	345	1,054	0.84	1.57	0		
20	1	West Tazenda	211	Dabuli	17	U	53	289	1,107	0.88	1.61	0		
21	1	West Tazenda	211	Dabuli	18	U	60	308	1,167	0.93	1.66	0		
22	1	West Tazenda	211	Dabuli	19	U	57	325	1,224	0.98	1.71	0		
23	1	West Tazenda	211	Dabuli	20	U	59	311	1,283	1.02	1.75	0		
24	1	West Tazenda	211	Dabuli	21	U	57	319	1,340	1.07	1.80	0		
25	1	West Tazenda	211	Dabuli	24	U	59	347	1,399	1.12	1.85	0		
26	1	West Tazenda	211	Dabuli	27	U	51	331	1,450	1.16	1.89	0		
27	1	West Tazenda	211	Dabuli	29	U	58	328	1,508	1.20	1.93	0		
28	1	West Tazenda	211	Dabuli	30	U	78	384	1,586	1.26	1.99	0		
29	1	West Tazenda	226	Balayan	01	U	125	483	1,711	1.36	2.09	1		
30	1	West Tazenda	226	Balayan	05	U	41	247	1,752	1.40	2.13	0		
31	1	West Tazenda	226	Balayan	06	U	75	369	1,827	1.46	2.19	0		

Forest | West Hills | East Hills | Desert

Ready

63. The list of all sampling units selected in the first stage should be transferred to a separate worksheet that will become a fundamental tool for the management of the survey. The survey managers can, for instance, add columns to record the particulars of all major activities in each PSU (expected and actual dates of fieldwork and data entry, identification of the responsible team, etc.).

64. The worksheet will be used, in particular, to compute the selection probabilities and the corresponding raising factors (or weights) required for obtaining unbiased estimates from the sample. This summary worksheet does not need to be separated by stratum. It is better to put all selected PSUs in a unique worksheet, specifying the stratum in one of the columns. In our

example, the "sample" worksheet for the first 19 of the 88 selected CEAs is presented in figure XV.8.

Figure XV.8. Spreadsheet with the selected primary sampling units

Microsoft Excel - Sample.xls

File Edit View Insert Format Tools Data Window Help Type a question for help

I1 fx Stratum

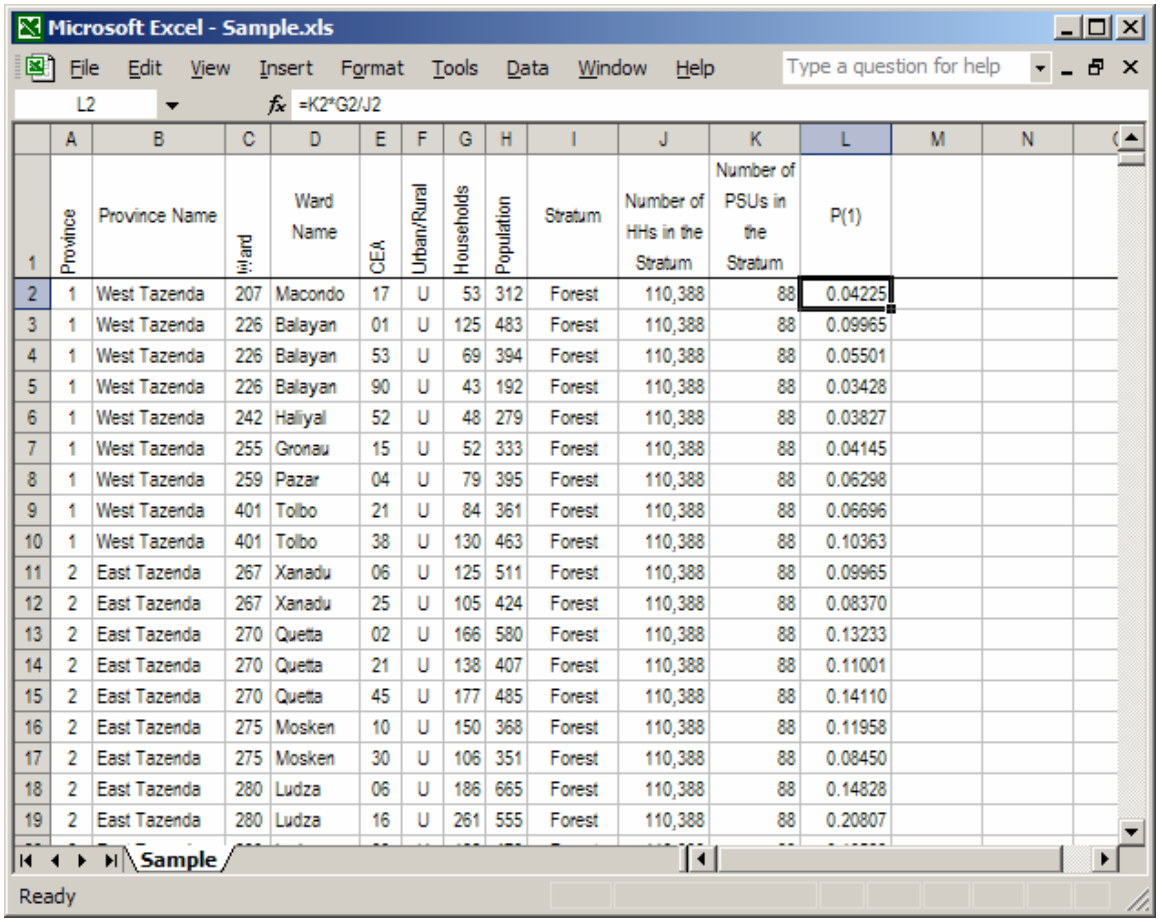
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Province	Province Name	Ward	Ward Name	CEA	Urban/Rural	Households	Population	Stratum					
2	1	West Tazenda	207	Macondo	17	U	53	312	Forest					
3	1	West Tazenda	226	Balayan	01	U	125	483	Forest					
4	1	West Tazenda	226	Balayan	53	U	69	394	Forest					
5	1	West Tazenda	226	Balayan	90	U	43	192	Forest					
6	1	West Tazenda	242	Haliyal	52	U	48	279	Forest					
7	1	West Tazenda	255	Gronau	15	U	52	333	Forest					
8	1	West Tazenda	259	Pazar	04	U	79	395	Forest					
9	1	West Tazenda	401	Tolbo	21	U	84	361	Forest					
10	1	West Tazenda	401	Tolbo	38	U	130	463	Forest					
11	2	East Tazenda	267	Xanadu	06	U	125	511	Forest					
12	2	East Tazenda	267	Xanadu	25	U	105	424	Forest					
13	2	East Tazenda	270	Quetta	02	U	166	580	Forest					
14	2	East Tazenda	270	Quetta	21	U	138	407	Forest					
15	2	East Tazenda	270	Quetta	45	U	177	485	Forest					
16	2	East Tazenda	275	Mosken	10	U	150	368	Forest					
17	2	East Tazenda	275	Mosken	30	U	106	351	Forest					
18	2	East Tazenda	280	Ludza	06	U	186	665	Forest					
19	2	East Tazenda	280	Ludza	16	U	261	555	Forest					
20	2	East Tazenda	280	Ludza	38	U	132	473	Forest					

Sample

Ready

65. *Selection probabilities and sampling weights*. The first-stage selection probabilities P(1) can be easily computed in the "sample" worksheet by multiplying the number of households in the sample PSU by the number of PSUs selected in each stratum (columns **G** and **K** in figure XV.9 below) and dividing the result by the total number of households in the stratum (column **J**). This is written as the formula **=K2*G2/J2** in cell **L2,** copied all the way down column **L**.

Figure XV.9. Computing the first-stage selection probabilities



Microsoft Excel - Sample.xls

File Edit View Insert Format Tools Data Window Help — Type a question for help

L2 =K2*G2/J2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Province	Province Name	Ward	Ward Name	CEA	Urban/Rural	Households	Population	Stratum	Number of HHs in the Stratum	Number of PSUs in the Stratum	P(1)		
2	1	West Tazenda	207	Macondo	17	U	53	312	Forest	110,388	88	0.04225		
3	1	West Tazenda	226	Balayan	01	U	125	483	Forest	110,388	88	0.09965		
4	1	West Tazenda	226	Balayan	53	U	69	394	Forest	110,388	88	0.05501		
5	1	West Tazenda	226	Balayan	90	U	43	192	Forest	110,388	88	0.03428		
6	1	West Tazenda	242	Haliyal	52	U	48	279	Forest	110,388	88	0.03827		
7	1	West Tazenda	255	Gronau	15	U	52	333	Forest	110,388	88	0.04145		
8	1	West Tazenda	259	Pazar	04	U	79	395	Forest	110,388	88	0.06298		
9	1	West Tazenda	401	Tolbo	21	U	84	361	Forest	110,388	88	0.06696		
10	1	West Tazenda	401	Tolbo	38	U	130	463	Forest	110,388	88	0.10363		
11	2	East Tazenda	267	Xanadu	06	U	125	511	Forest	110,388	88	0.09965		
12	2	East Tazenda	267	Xanadu	25	U	105	424	Forest	110,388	88	0.08370		
13	2	East Tazenda	270	Quetta	02	U	166	580	Forest	110,388	88	0.13233		
14	2	East Tazenda	270	Quetta	21	U	138	407	Forest	110,388	88	0.11001		
15	2	East Tazenda	270	Quetta	45	U	177	485	Forest	110,388	88	0.14110		
16	2	East Tazenda	275	Mosken	10	U	150	368	Forest	110,388	88	0.11958		
17	2	East Tazenda	275	Mosken	30	U	106	351	Forest	110,388	88	0.08450		
18	2	East Tazenda	280	Ludza	06	U	186	665	Forest	110,388	88	0.14828		
19	2	East Tazenda	280	Ludza	16	U	261	555	Forest	110,388	88	0.20807		

Sample

Ready

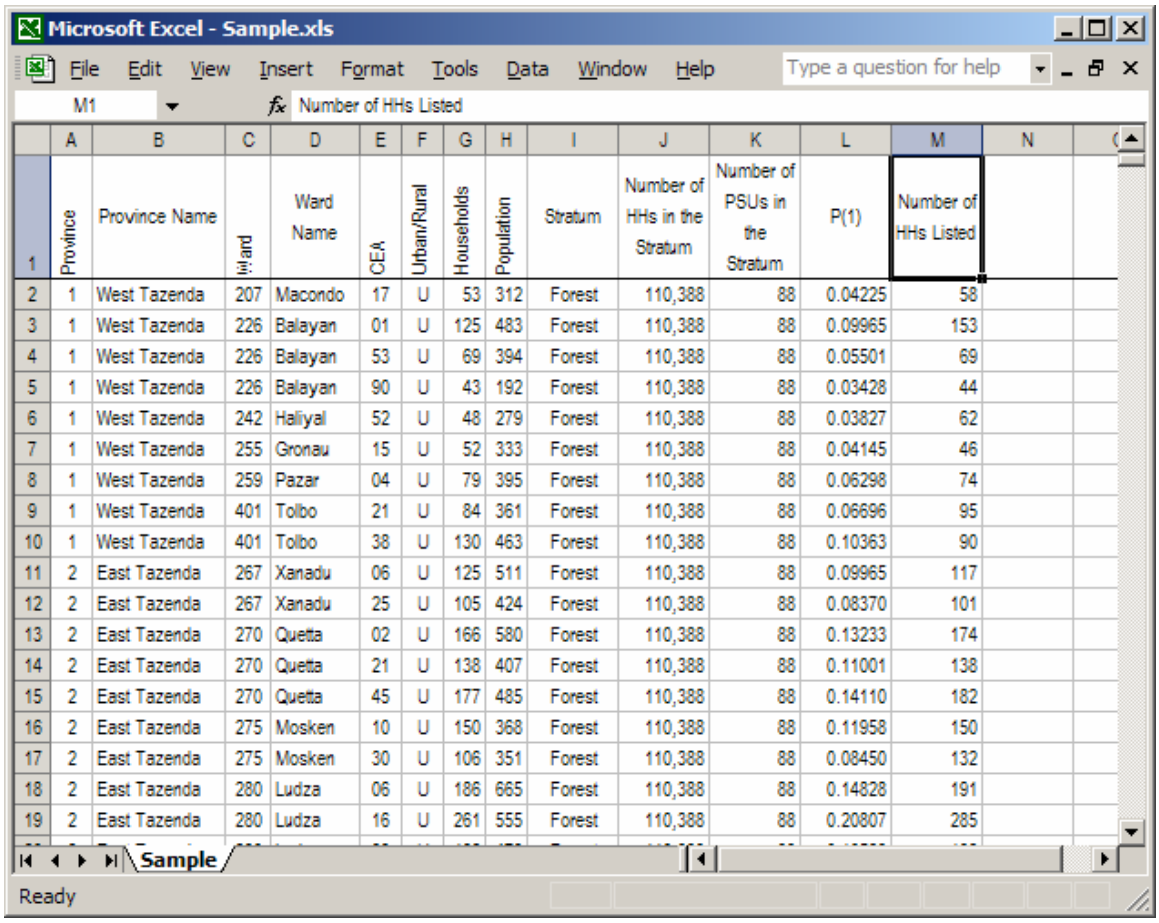
66. The selection probabilities in the subsequent stages depend, of course, on the particulars of the sampling design. We will illustrate the computations for a two-stage sampling design with a fixed number of households selected with equal probability in each PSU in the second stage. This sampling design is in fact one of those most commonly used in practice. The number of households per PSU selected in the second stage may vary across strata; but in the hypothetical country of our example, we will assume 12 households per CEA in all strata.

67. This sampling stage generally requires that a household listing operation be conducted in each of the selected PSUs. The household listings do not need to be computerized, because the selection of the households to be visited by the survey can be carried out by hand from the paper listings. However, there are many advantages of having the listings entered into computer files (for instance, if the PSUs selected in the first stage constitute a so-called master sample that will be used for various surveys, or for various rounds of a panel survey).

68. The number of households actually found in each of the sampled PSUs at the time of the listing operation will generally be different from the "number of households" originally recorded by the census in the first-stage sampling frame. **A column should be appended to the "sample" worksheet to record the number of households listed**. If the listing forms are computerized, this column can be filled programmatically (using Excel macros, for instance.) Otherwise, filling

in this column as a part of the household listing operation should become a top priority of the survey managers. In figure XV.10, the frame "number of households" and the "number of households listed" appear, respectively, in columns **G** and **M.**

Figure XV.10. Documenting the results of the household listing operation



Microsoft Excel - Sample.xls

File Edit View Insert Format Tools Data Window Help — Type a question for help

M1 — fx Number of HHs Listed

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Province	Province Name	Ward	Ward Name	CEA	Urban/Rural	Households	Population	Stratum	Number of HHs in the Stratum	Number of PSUs in the Stratum	P(1)	Number of HHs Listed	
2	1	West Tazenda	207	Macondo	17	U	53	312	Forest	110,388	88	0.04225	58	
3	1	West Tazenda	226	Balayan	01	U	125	483	Forest	110,388	88	0.09965	153	
4	1	West Tazenda	226	Balayan	53	U	69	394	Forest	110,388	88	0.05501	69	
5	1	West Tazenda	226	Balayan	90	U	43	192	Forest	110,388	88	0.03428	44	
6	1	West Tazenda	242	Haliyal	52	U	48	279	Forest	110,388	88	0.03827	62	
7	1	West Tazenda	255	Gronau	15	U	52	333	Forest	110,388	88	0.04145	46	
8	1	West Tazenda	259	Pazar	04	U	79	395	Forest	110,388	88	0.06298	74	
9	1	West Tazenda	401	Tolbo	21	U	84	361	Forest	110,388	88	0.06696	95	
10	1	West Tazenda	401	Tolbo	38	U	130	463	Forest	110,388	88	0.10363	90	
11	2	East Tazenda	267	Xanadu	06	U	125	511	Forest	110,388	88	0.09965	117	
12	2	East Tazenda	267	Xanadu	25	U	105	424	Forest	110,388	88	0.08370	101	
13	2	East Tazenda	270	Quetta	02	U	166	580	Forest	110,388	88	0.13233	174	
14	2	East Tazenda	270	Quetta	21	U	138	407	Forest	110,388	88	0.11001	138	
15	2	East Tazenda	270	Quetta	45	U	177	485	Forest	110,388	88	0.14110	182	
16	2	East Tazenda	275	Mosken	10	U	150	368	Forest	110,388	88	0.11958	150	
17	2	East Tazenda	275	Mosken	30	U	106	351	Forest	110,388	88	0.08450	132	
18	2	East Tazenda	280	Ludza	06	U	186	665	Forest	110,388	88	0.14828	191	
19	2	East Tazenda	280	Ludza	16	U	261	555	Forest	110,388	88	0.20807	285	

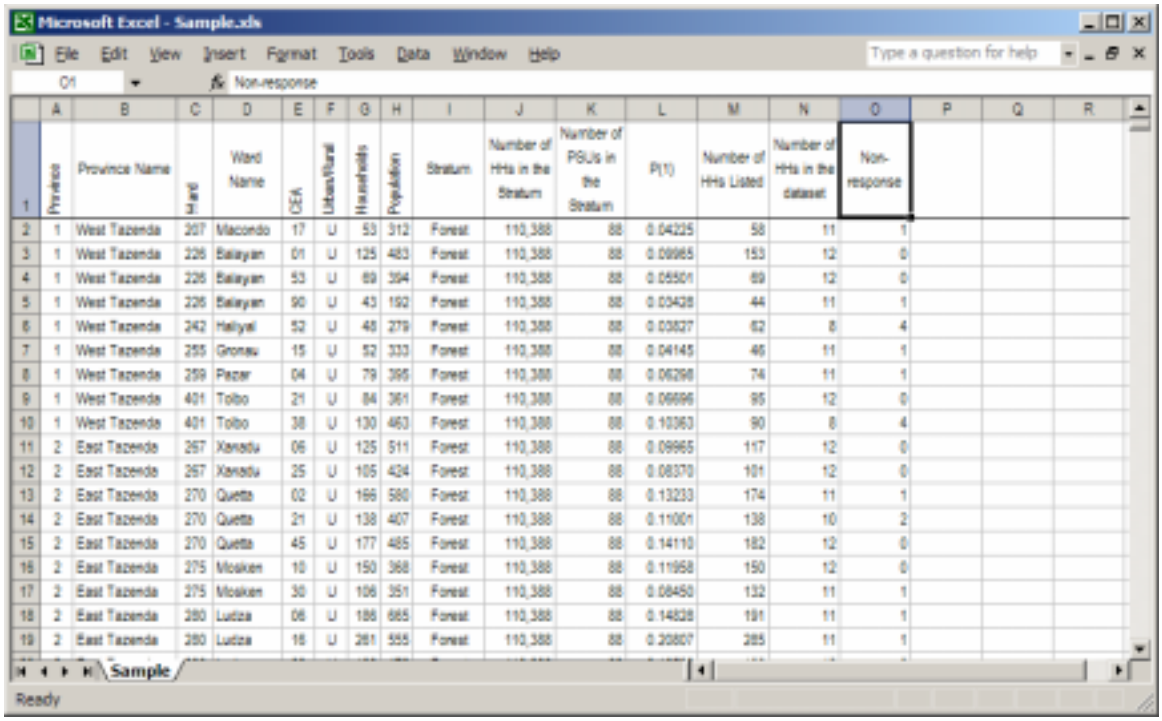
Sample

Ready

69. As fieldwork and data management operations are completed, additional columns should be added to the "sample" worksheet, to record, **on a per-PSU basis**, the number of households for which useful information is actually recorded in the survey data sets, as well as the number of households for which information is not available for various reasons. The standard non-response reasons for adding a "useless questionnaire" column are extensively discussed elsewhere in the present publication (see, for instance, chapter VIII, and section F of chapter XXII for refusal, dwelling vacant, etc.). A column for "useless questionnaire" may need to be added also when the survey is unable to integrate computer-based quality controls into field operations. This is unfortunately a common outcome of centralized data entry techniques.

70. Continuing with the example presented in figure XV.11 below, we will simplify the situation, assuming that two additional columns are added to the "sample" worksheet, for the "number of households in the data sets" and for total "non-response".

Figure XV.11. Documenting non-response



Microsoft Excel - Sample.xls

File Edit View Insert Format Tools Data Window Help

Type a question for help

O1 Non-response

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Province	Province Name	Ward	Ward Name	CEA	Urban/Rural	Households	Population	Stratum	Number of HHs in the Stratum	Number of PSUs in the Stratum	P(1)	Number of HHs Listed	Number of HHs in the dataset	Non-response			
2	1	West Tazenda	207	Macondo	17	U	53	312	Forest	110,388	88	0.04225	58	11	1			
3	1	West Tazenda	226	Balayan	01	U	125	483	Forest	110,388	88	0.09965	153	12	0			
4	1	West Tazenda	226	Balayan	53	U	69	394	Forest	110,388	88	0.05501	69	12	0			
5	1	West Tazenda	226	Balayan	90	U	43	192	Forest	110,388	88	0.03428	44	11	1			
6	1	West Tazenda	242	Haliyal	52	U	48	279	Forest	110,388	88	0.03827	62	8	4			
7	1	West Tazenda	255	Gronau	15	U	52	333	Forest	110,388	88	0.04145	46	11	1			
8	1	West Tazenda	259	Pezar	04	U	79	366	Forest	110,388	88	0.06298	74	11	1			
9	1	West Tazenda	401	Tolbo	21	U	84	361	Forest	110,388	88	0.06696	95	12	0			
10	1	West Tazenda	401	Tolbo	38	U	130	460	Forest	110,388	88	0.10363	90	8	4			
11	2	East Tazenda	267	Xanadu	06	U	125	511	Forest	110,388	88	0.09965	117	12	0			
12	2	East Tazenda	267	Xanadu	25	U	105	424	Forest	110,388	88	0.08370	101	12	0			
13	2	East Tazenda	270	Quetta	02	U	166	580	Forest	110,388	88	0.13233	174	11	1			
14	2	East Tazenda	270	Quetta	21	U	138	407	Forest	110,388	88	0.11001	138	10	2			
15	2	East Tazenda	270	Quetta	45	U	177	485	Forest	110,388	88	0.14110	182	12	0			
16	2	East Tazenda	275	Moskan	10	U	150	368	Forest	110,388	88	0.11958	150	12	0			
17	2	East Tazenda	275	Moskan	30	U	106	351	Forest	110,388	88	0.08450	132	11	1			
18	2	East Tazenda	280	Ludza	06	U	186	665	Forest	110,388	88	0.14828	191	11	1			
19	2	East Tazenda	280	Ludza	16	U	261	555	Forest	110,388	88	0.20807	285	11	1			

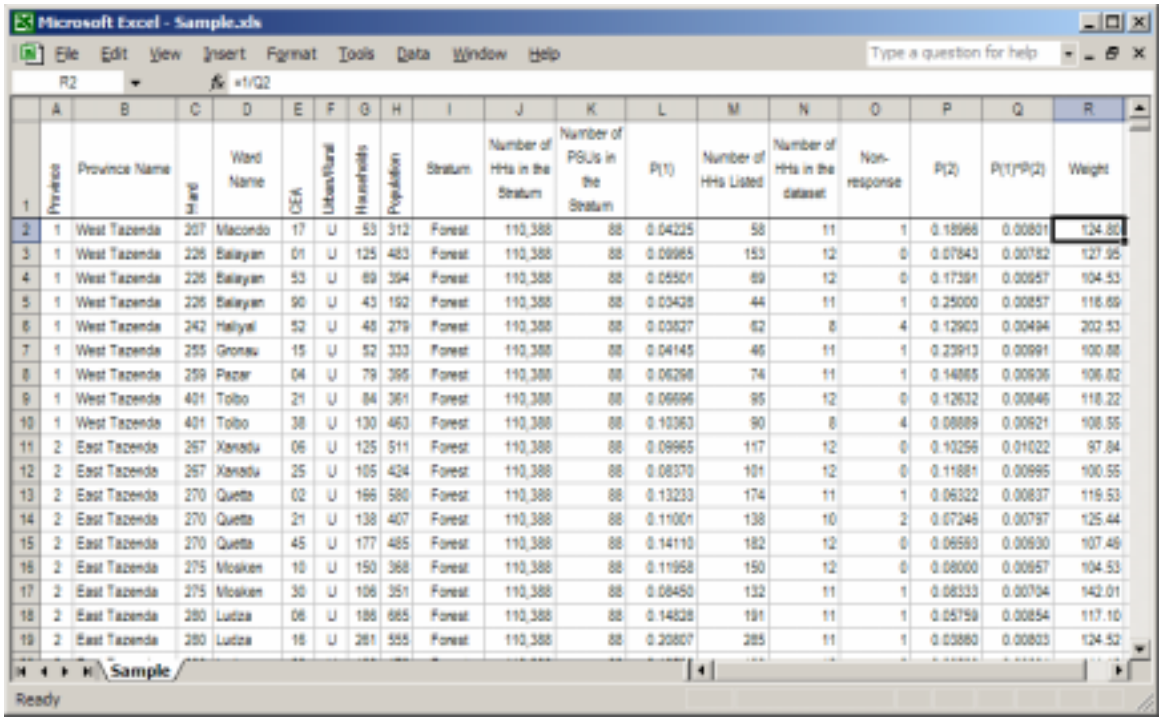
Sample

Ready

71. Although there are no universally accepted models for non-response, a very common assumption is that the "useful" households in the final data sets are in fact an equal-probability sample of all the households listed in their respective PSUs (see chaps. II and VIII for an extensive discussion). Under this hypothesis, the probability P(2) of selecting each of these households in the second stage can be computed by simply dividing the number of useful households by the number of households listed. The total selection probability of each household in the PSU is the product P(1)*P(2) and the sampling weight is the inverse of that probability.

72. These formulae can be easily implemented in the spreadsheet (see figure XV.12). Write formula **=N2/M2** in cell **P2**, formula **=L2*P2** in cell **Q2** and formula **=1/Q2** in cell **R2**; then copy them all the way down columns **P**, **Q** and **R.**

Figure XV.12. Computing the second-stage probabilities and sampling weights



Microsoft Excel - Sample.xls

File Edit View Insert Format Tools Data Window Help

R2 =1/Q2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Province	Province Name	Ward	Ward Name	CEA	Urban/Rural	Households	Population	Stratum	Number of HHs in the Stratum	Number of PSUs in the Stratum	P(1)	Number of HHs Listed	Number of HHs in the dataset	Non-response	P(2)	P(1)*P(2)	Weight
2	1	West Tazenda	207	Maccondo	17	U	53	312	Forest	110,388	88	0.04225	58	11	1	0.18966	0.00801	124.80
3	1	West Tazenda	226	Balayan	01	U	125	483	Forest	110,388	88	0.09965	153	12	0	0.07843	0.00782	127.95
4	1	West Tazenda	226	Balayan	53	U	69	394	Forest	110,388	88	0.05501	69	12	0	0.17391	0.00957	104.53
5	1	West Tazenda	226	Balayan	90	U	43	192	Forest	110,388	88	0.03428	44	11	1	0.25000	0.00857	116.69
6	1	West Tazenda	242	Haliyal	52	U	48	279	Forest	110,388	88	0.03827	62	8	4	0.12903	0.00494	202.53
7	1	West Tazenda	255	Gronau	15	U	52	333	Forest	110,388	88	0.04145	46	11	1	0.23913	0.00991	100.88
8	1	West Tazenda	259	Pezar	04	U	79	366	Forest	110,388	88	0.06298	74	11	1	0.14865	0.00936	106.82
9	1	West Tazenda	401	Tolbo	21	U	84	361	Forest	110,388	88	0.06696	95	12	0	0.12632	0.00846	118.22
10	1	West Tazenda	401	Tolbo	38	U	130	460	Forest	110,388	88	0.10363	90	8	4	0.08889	0.00921	108.55
11	2	East Tazenda	267	Xanadu	06	U	125	511	Forest	110,388	88	0.09965	117	12	0	0.10256	0.01022	97.84
12	2	East Tazenda	267	Xanadu	25	U	105	424	Forest	110,388	88	0.08370	101	12	0	0.11881	0.00995	100.55
13	2	East Tazenda	270	Quetta	02	U	166	580	Forest	110,388	88	0.13233	174	11	1	0.06322	0.00837	119.53
14	2	East Tazenda	270	Quetta	21	U	138	407	Forest	110,388	88	0.11001	138	10	2	0.07246	0.00797	125.44
15	2	East Tazenda	270	Quetta	45	U	177	485	Forest	110,388	88	0.14110	182	12	0	0.06593	0.00930	107.49
16	2	East Tazenda	275	Moskan	10	U	150	368	Forest	110,388	88	0.11958	150	12	0	0.08000	0.00957	104.53
17	2	East Tazenda	275	Moskan	30	U	106	351	Forest	110,388	88	0.08450	132	11	1	0.08333	0.00704	142.01
18	2	East Tazenda	280	Ludza	06	U	186	665	Forest	110,388	88	0.14828	191	11	1	0.05759	0.00854	117.10
19	2	East Tazenda	280	Ludza	16	U	281	555	Forest	110,388	88	0.20807	285	11	1	0.03860	0.00803	124.52

Sample

Ready

73. The probability-based weights computed in this way apply to all households in each PSU. Some survey practitioners may use "post-stratification" techniques to further adjust these weights in order to ensure that the survey estimates match certain known population distributions (such as age and gender distributions, or total consumption figures obtained from sources external to the sample survey itself). These adjustments are made with specialized software directly in the survey data sets, not in the sampling spreadsheets, and they generally operate on a per-household or per-person basis rather than on a per-PSU basis.

H. Summary of recommendations

74. This chapter has aimed at shedding some light on the relevance of incorporating data management criteria at every stage of a survey, as opposed to deeming it a matter integral only to the last analytical phases. One of the clearest cases in point are the Living Standards Measurement Study surveys, which have taken it upon themselves to design their questionnaires, plan and carry out field operations, and deal with data entry and processing in such a way as to allow the data to be properly managed even before any of them are collected. The guiding principles behind that effort constitute the core of this chapter; and even as they take on different characteristics according to the specific application in a given country, those principles may still be condensed and codified as follows:

(a) Survey data management begins with questionnaire design, and within it deals with:

(i) *Proper identification of the statistical units*. The recommendation is to use a simple or upgraded three- or four-digit serial numbers for the survey's PSUs, and then

a two-digit serial number for each household within it, plus proper serial identification of each subordinate unit within the household;

(ii) *Built-in redundancies*. The design of the questionnaire should include deliberate redundancies, intended to detect mistakes of the interviewer or data entry errors. Examples of this are a bottom line for totals or adding a check digit to the codes of important variables.

(b) During field operations, the following should be taken into account:

(i) *Operational strategies for data entry and data editing*. It is recommended that countries give careful consideration to the option of entering all data in the field. This may be done through a data entry operator working in a fixed location other than that of the surveyed households, by an operator joining the rest of the interviewing team and entering data directly to a laptop computer in each household or by the as yet not properly researched paperless interview method using a palmtop (though this needs more research). Entering all data in the field versus centralized entry will go a long way towards ensuring quality and consistency;

(ii) *Quality control criteria*. The data on the questionnaires needs to be subjected to five different control mechanisms: range checks, checks against reference tables, skip checks, consistency checks and typographical checks;

(iii) *Data entry technology*. According to a 1995 World Bank review, two reliable data entry and editing platforms suitable for complex household surveys were the World Bank's internally developed LSMS package and the IMPS program of the United States Bureau of the Census. Their updated versions are LSD-2000 and the CSPro, respectively. Allowing for existing expertise and other factors affecting each country's own set of conditions, there are a few basic guidelines that should be taken into account when designing data entry and editing tools: exceptions aside, computer screens should resemble their corresponding questionnaire sections; data entry programs should discern impossible and unlikely situations and specifically flag each; error-type reporting language and expressions should be colloquial and easily understood;

(iv) *Organizing and disseminating the survey data sets*. For these purposes, flat files are not suitable, since they do not deal properly with subordinate statistical units (persons, crops, consumption items, etc.) within the household. A structure with a different record type for each kind of statistical unit is to be preferred.

(c) Finally, data management may also prove instrumental in implementing the sampling protocol, by guiding it through its main stages: organization of the first-stage sampling frame, usually created from the latest available set of census enumeration areas (CEAs); selection of primary sampling units with probability proportional to size, measured by the number of households, dwellings or the size of the population; and calculation of selection probabilities and the corresponding sampling weights.

References

Ainsworth, M., and J. Muñoz (1986). *The Côte d'Ivoire Living Standards Survey: Design and Implementation*. Living Standards Measurement Study Working Paper, No. 26. Washington, D.C.: World Bank.

Blaizeau, D. (1998). Seven expenditure surveys in the West African Economic and Monetary Union. In *Proceedings of the Joint International Association of Survey Statisticians/International Association for Official Statistics (IASS/IAOS) Conference on Statistics for Economic and Social Development.* Aguascalientes, Mexico: International Statistical Institute.

_____, and J.L. Dubois (1990). *Connaître les Conditions de Vie des Ménages dans les Pays en Développement.* Paris: Documentation française.

Blaizeau, D, and J. Muñoz (1998). *LSD-2000. Logiciel de Saisie des Données: Pour Saisir les Données d'une Enquête Complexe.* Paris: Institut national de la statistique et des études économiques.

Grosh, M. and J. Muñoz (1996). *A Manual for Planning and Implementing the Living Standards Measurement Study Survey*, Living Standards Measurement Study Working Paper, No. 126. Washington, D.C.: World Bank.

Muñoz, J. (1989). Data management of complex socioeconomic surveys: from questionnaire design to data analysis. In *Proceedings of the 47th Session of the International Statistical Institute.* Paris: International Statistical Institute.

_____ (1996). Cómo mejorar la calidad de la información: opciones para mejorar la organización del trabajo de campo, el sistema de entrada de datos, el análisis de consistencia y el manejo de la base de datos. In *Reunión de Iniciación del Programa para el Mejoramiento de las Encuestas de Condiciones de Vida en América Latina y El Caribe*. Asunción: Inter-American Development Bank.

_____ (1998). Budget-Consumption Surveys: New Challenges and Outlook. In *Proceedings of the Joint International Association of Survey Statisticians/International Association for Official Statistics (IASS/IAOS) Conference on Statistics for Economic and Social Development*. Aguascalientes, Mexico: International Statistical Institute.

United States Bureau of the Census. CSPro Census and Survey Processing System, available from http://www.census.gov/ipc/www/cspro/.