STATISTICS DENMARK

Statistisk sentralbyrå
Statistics Norway

Statistiska centralbyrån
Statistics Sweden

MZ:2005:11

Mission Report

From a short-term mission on CSPro Training

June 10-26, 2005

TA for the Scandinavian Support Program to Strengthen the Institutional Capacity of the National Statistics, Mozambique

Julio Ortúzar
SERPRO S.A.
Benjamín 2935 Of. 302
Santiago, Chile 755-0048

Instituto Nacional de Estatística

Ref. no. 104.Moz.98

1 EXECUTIVE SUMMARY

The main objective of this mission was capacity building in the area of the use of CSPro, with specific attention to consistency checking and batch control. During the two week workshop that I would rather call ***on the job training***, most of the new CSPro features were reviewed, including the graphic interfaces to export data from CSPro to SPSS/STATA/SAS, and the production of quick tabulations aiming to an early check of the field work, permitting the early detection of problems that, otherwise, might have a negative impact on the data gathered.

Emphasis was paid to the overall planning and control of the survey operation. Although consistency checks are very important to produce sound and consistent data that will later facilitate the data analysis, there are other error types that will have a much greater impact in the accuracy of the data representation and require of different techniques to detect and correct them. Thus, after identifying these errors, time was also spent in designing strategies to unveiling and fixing them. The error types I'm referring to were mainly (i) misplacing PSUs (Primary Sampling Units) by gathering/entering the wrong id (e.g. the wrong province/district), (ii) missing PSUs and (iii) Duplication of PSUs. These three errors will tend to distort the data representation in different ways but the result will be the same: once the micro data are tabulated or converted into macro data, they are difficult to perceive although they are distorting or biasing the results in a degree that will strictly depend on the frequency these errors exist and how they compensate each other.

The progress made by INE's data processing staff was evident considering the short period of the mission. The progress should be measured in different areas as follows: (i) getting acquainted with the new CSPro futures, which we reviewed and applied immediately to the Labour Force Survey (LFS) that they are currently processing; (ii) analyzing how the CSPro features in general (new as well as old features) could be applied to different situations of the survey processing process to achieve specific results (e.g. linking of different questionnaires to perform consistency checks across them or simply to access information of a different questionnaire to use as filter for another one); (iii) making them conscious of the different type of errors that they should face to improve the quality of the survey data.

The progress was measured and observed by different INE's staff, demographers and different authorities including one expert from the Scandinavian Program. The evaluation test had to do with a demonstration of the current Data Entry application done by the INE IT staff, which included a large number of online consistency checks intra and across questionnaires, and more important, the online checking of PSU validity identification when compared to the sampling design data file. Furthermore, the same consistency checks were run in batch mode to test the already entered survey data to have an idea of the data quality. According to the audience, the results were impressive considering that the previous application didn't have more than range tests.

Notwithstanding the evident progress made, there is a long way to go in terms of capacity building. It would be pretentious to even think that in a two week seminar most of the theory related to statistical data processing was covered. However, at least they had the chance to experiment and get hands-on experience to different errors that can distort the data representation of the universe being studied. My recommendations about future steps follow:

1. It is of great importance to create working teams that combine staff with different skills and knowledge (i.e. demographers, sociologists, statisticians and information technology) to design strategies and plans for future studies. The integration of this team should aim to attack the various problems in the different phases of the survey/census starting by the questionnaire design, sample design –if applicable-, interviewers training, fieldwork, etc. The end-result of this team should be a written document where responsibilities are clearly established, and where all strategies to make possible an early detection of the numerous problems that for sure will arise is outlined.

2. Control System Implementation. A survey and more important yet, a census, should be closely scrutinized and monitored by a control system where each small unit –PSU or Enumeration area (EA)- is followed up through the different phases of the survey/census operation. The control file, where all this information is stored, should accommodate at least the following information: (i) one record for each unit (EA or PSU) where all the related information will be stored; such information can be of two classes, pre-existing information derived from either the sample design, cartography and or pre-census in the case of a census, and actual information that is compiled from the actual data gathered. This allows statisticians and other subject matter specialist to analyze the validity of the actual data when compared to the expected figures. (ii) One bucket for each phase of the survey operation where information such as supervisor responsible for the specific phase and EA/PSU, date when finished, etc. can be stored. This simple information permits a strict follow up of the operation status at the same time it gives an answer to most of the errors outlined above related to PSUs and/or EAs. Reports based on the control file should also aim to the INE management providing a clear picture of the status or completeness rate of the different phases of the survey/census operation.

3. A very important output from the document mentioned in 1 above should consist of a detailed specification of the control file outlined in 2 and a complete set of editing rules or consistency checks that should be applied to the data entry or data capture application and to the batch editing application. Once these tasks have been completed, INE might rely on the quality of the data they will be analyzing in a later stage.

1. INTRODUCTION

This report has been prepared by Julio Ortúzar, Executive Director of SERPRO S.A. and refers to the mission carried out between June 10th and June 26th 2005. The main purpose of this mission was to conduct an advance CSPro seminar using as base for all practical examples two different survey questionnaires, the Labour Force Survey which was actually under way with nearly 50% of the data gathered already entered, and an Agricultural survey that is suppose to go to the field on August 2005.

The seminar's participant had had at least one CSPro seminar previous to this one and therefore, all of them had knowledge and experience with the software. Starting from this premise, the seminar included a review of all the new system's features with emphasis in the immediate practice applying the new concepts and features to their respective surveys.

According to the Terms of Reference, two were the main problems identified by the INE's IT staff: training in advanced techniques of CSPro oriented to expedite the online consistency checks and batch control.

The Labour Force Survey currently in the field has three questionnaires: (i) the Household Information including demographic characteristics of the members as well as the necessary links describing the household composition or structure; (ii) the Main questionnaire for people older than 6, where most of the Labour oriented questions are located; and (iii) the questionnaire for juvenile population (7-17 years of age) who did work during the last 7 days.

Most of the problems identified by the IT staff at INE were originated by the three questionnaires since they didn't know how to establish the link between a person in questionnaire 2 with the corresponding information in questionnaire number 1. Thus, it was difficult to perform consistency checks between information pertaining to different questionnaires and, furthermore, filters or skips using information from different modules were not being executed properly.

This report contains the views of the consultant(s), which do not necessarily correspond to the views of Danida or INE.

2. ACTIVITIES DURING THE MISSION

Unlike most other missions, this one was very specific and aimed to leave in INE's IT staff a solid knowledge of CSPro that would allow them in the feature to undertake the processing of any survey.

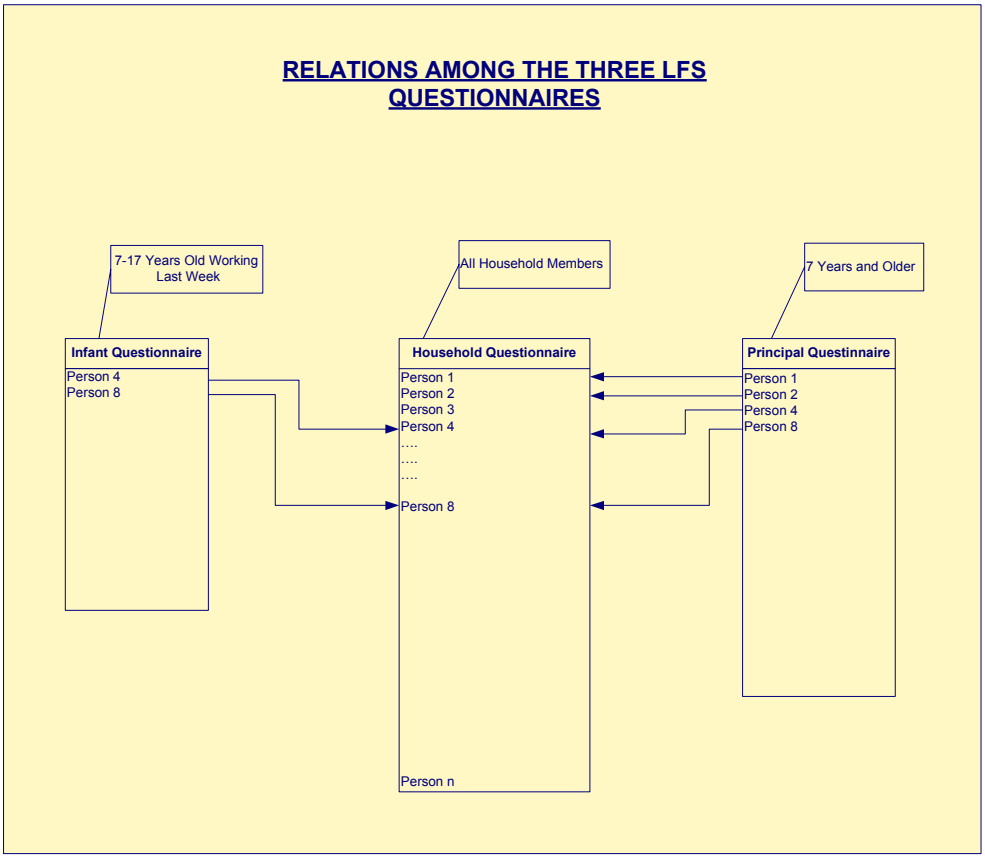
In the viewpoint of the consultant, this main goal was achieved although this is just one aspect of the statistical data processing, remaining other issues that are as important or more than this one that needs to be addressed when planning a survey or census operation. Thus, although the consultant spent most of the time concentrated in covering and practicing advance CSPro concepts enlightening them with practical exercises using their own survey's data, time was also devoted to cover other relevant topics in the processing of statistical data. However, given the limited time available, only a superficial introduction was given.

Basically, any survey or census operation should be controlled by some system that will provide information and with it, confidence to directors and managerial staff in charged of the operation regarding the information gathered, entered into a magnetic media and stored. In pursuing this primary objective, we would need to identify a primary area or small geographic unit to follow up through the different steps through which the statistical data will go through. Although we will limit the scope of this description to the data processing steps, the same infra-structure could well serve needs prior to DP like administrative steps. For our purposes, the unit that works best is the PSU (Primary Sampling Unit) since it has been defined by the sampler in terms of the geographic location, the number of expected households or units of analysis expected for each PSU, etc. Furthermore, the size of this unit seems right for (i) the amount of work that a Data Entry clerk can undertake as a daily load and (ii) it is the smallest area for which we have accurate information coming from the sample design, facilitating the control file creation and follow up of them. For a complete review of the methodology suggested, please refer to Appendix 2: "Survey Control and PSU follow up".

Among the broad CSPro concepts reviewed during the seminar, the following are the most important:

1. Relations between CSPro Objects. CSPro in more less the same manner as a relational data base permits links definitions between two tables (in CSPro called groups). These links or relations facilitate the reference between items of the related groups eliminating the need for indirections that not always are simple to understand. In the particular case of INE, the use of relations was a need given the three questionnaires the Labour Force Survey has. The graphic shown below will illustrate the current situation with the LFS. The Household questionnaire includes all the household members and the other two questionnaires include subsets of the existing individuals: the "Principal Questionnaire" includes all the members who are 7 years old or older, and the "Infant Questionnaire" includes only members between 7 and 17 years of age who worked during the last week prior to the survey. In both cases, the link is to the Household questionnaire (by mean of the person number) and consequently, to refer to the Principal questionnaire from the Infant questionnaire there was no direct link. However, having the common

link to the household questionnaire, it was possible to build the link from the Infant to the Principal questionnaire.



As the figure above illustrates it, the relations or the concepts behind the relations were a basic need to check the filter questions of each questionnaire –which in both cases involved variables of different questionnaires- but also for consistency checks between variables of the different modules.

2. Batch Control. There seems to be a misunderstanding in the TOR about the meaning of the term "Batch Control" since it is associated to a batch operation running the consistency checks. The truth of the matter is that the term comes from the time when CSPro was unable to launch other(s) CSPro applications and therefore, the DOS batch files were used to control the whole survey operation through menus where each option launched a different CSPro application. However, since the moment when CSPro was able to launch other CSPro or any Windows applications, the DOS batch files were replaced by the much simpler CSPro applications that expedite and facilitate the dynamic generation of menus and parameters passed directly to CSPro applications. Along with this facility, several new commands or system functions have been developed to simplify the interaction between the developer and the dynamic generation of menus. The result, an important breakthrough in the process of hiding from the end user the less friendly direct CSPro interface replacing it by a more user friendly menu-driven set of applications where there is no leeway for end user errors in the administration or misinterpretation of the various applications.

An intensive review of the techniques for dynamic generation of menu-driven applications was performed. Application examples for both surveys were developed and tested as the concepts were discussed, leaving no space for misunderstandings or doubts.

3. Consistency Checks. According to the TOR, this was a very important topic to be analyzed and discussed. However, it is important to point out that the CSPro techniques to implement consistency checks is a rather simple process once the application developer has a clear idea of the various consistency rules that need to be applied to the survey data. In other words, the application developer might be a CSPro expert but if there is luck of clarity in what need to be checked, there won't be an adequate consistency check process.

The errors source or errors origin is an important consideration at the time of planning the editing process. Errors originated by the data entry operator can be eliminated or at least minimized by the verification process. They can also be greatly minimized by the online consistency checks and adequate error messages. However, errors originated during the interview –by the interviewer or the respondent- can only be detected by a well thought set of consistency rules that normally require a deep and thorough knowledge of the questionnaire(s). Errors originated in the field/interview should never be fixed by the DE operators; their errors fixing capabilities should be strictly limited to their own errors, leaving the others for more qualified specialist. Based on this premise, there will be two instances for error detection: (i) the online error checking during data entry, where the operator will fix only errors originated entering the data and (ii) the same editing process but executed after data have been entered, also known as batch secondary edit process aiming to produce an error listing for small areas (PSU) to be analyzed by supervisors and/or subject matter specialists that will be responsible for "fixing" the inconsistencies. Errors showing up on these listings should be restricted to those originated in the field since presumably those originated by the DE operators should have already been fixed.

The last rationale brings up another important issue that has to do with the type of data collection/data entry that wants to be implemented. Although the centralized data entry is simpler to implement, fixing errors that are originated during the interview is far more complex and might introduces some distortion in the data. On the other hand, field distributed data entry means a more sophisticated preparation of the survey operation, requiring a different infrastructure in terms of replacing normal DE machines (PCs) by more expensive laptops. The trade-off comes in terms of better quality data since re-visits to households are possible when the number of errors or importance of them justify it. The ultimate refinement in data capture is the CAPI operation since all the interview is mainly carried out by the CAPI application. This means that all errors detected by the CAPI application can be fixed immediately with the direct cooperation of the respondent. In theory, once the interview is finished, the data are "error free" and no other future application will unveil new errors. The price that has to be paid for a CAPI operation however is relatively high since a laptop/notebook per interviewer is needed. Additionally, the preparation of a CAPI application requires a higher degree of expertise and sophistication. It's convenient to point out that SERPRO, under a sub-contract with the US Aid for International Development is currently making the feasibility study for implementing CSPro (the run time component of the interactive system) in Pocket PCs. The software should have the same power and features of the PC version, having the advantages of the Pocket PC hardware: (more than 8 hours of battery life, dust resistant, rapid battery recharge, low weight, touch-screen, wireless built-in connectivity and low price). Our feasibility study

should be finished by the first week in August and we will make sure a copy of it is sent to INE. The software conversion should be ready for beta test early next year.

4. Exporting data to SPSS, SAS and STATA. The new EXPORT graphic interface was shown and reviewed in detail. Using the CSPro RELATIONS, examples of exporting in one record variables pertaining to the three questionnaires of the LFS were developed. Generally speaking, exporting data to any of the three main statistical analysis systems shouldn't be a problem when using the new CSPro graphic interface.

5. Tabulations using the current graphic interface. Although the current CSPro cross-tabs graphic interface is very basic, two objectives were accomplished with its review and analysis: (i) the easy generation of simple tables and (ii) an introduction to the new powerful CSPro cross-tab graphic interface that will be liberated around September of this year. It is important to point out that the new cross-tab module will make possible the generation of complex tables along with some statistics like central tendency (mean, median and mode), dispersion (minimum, maximum, variance, standard deviation, standard error) and others like proportions, percentiles, percents, etc. It is highly recommended that the INE IT staff take a two week workshop on this important CSPro component since with this module, the CSPro training will be completed.

3. RECOMMENDATIONS

The main problems identified in the processing of the LFS data comes from the lack of a detailed plan to detect the different types of errors that are likely to be imbedded in the statistical data. The production of such plan should be the result of a multi-disciplinary group where the IT staff should be part of it but where Demographers, Sociologists and in general, subject matter specialists should contribute to the production of a conceptual plan where no detail is left at random.

The plan should aim to unveil and repair or fix errors at two different levels requiring of different strategies: (i) Those problems that tend to distort the universe being studied, that mainly are caused by the completenes of such universe but can also be affected by missplacing small areas in different geographic/administrative areas. It has already been suggested a CONTROL application that aims to prevent these type of errors and should be part of the DP planification in any survey/census. (ii) A detailed specification of all consistency checks to be performed at the ”unit of study” level. The unit of study might be a household, a farm or any other unit being analyzed depending on the survey type. This specification list requires a deep and thorough knowledge of the questionnaire and the inter-relations between the different modules of the questionnaire and therefore, the cooperation of the multi-disciplinary group is a valuable contribution.

The above mentined precautions aim to produce a survey/census data file that has been rigorously and methodically treated step by step to prevent errors that might distort the statistical results we expect to obtain. The next step might be the production of those statistics in terms of crosstabulations and statistical parameters or indicators that statisticians and policy makers need to really fulfill the ultimate goal which should be the research and analysis of the data gathered. To achieve this final goal, it would be highly desirable that the IT staff get the proper trainning in the CSPro new module that will soon be released. The completion of this training would leave the INE IT staff capacitated in all areas of a survey data processing.

Appendix 1: Terms of Reference.

TERMS OF REFERENCE

For a short-term mission
On
Training on Data Processing for LFS to local IT staff
12 June - 26 June 2005

Within the Scandinavian Assistance to Strengthen the Institutional Capacity of INE/Mozambique 2003-2007

Consultants: SERPRO S.A.
Main Counterpart: Mr. Tomás Bernardo IT Deputy Director at INE

1.1 Background

The Integrated System of Household Surveys, based on Core Welfare Indicators Questionnaire (CWIQ) and rotate modules, started in 2000. The Labour Force Survey (LFS), ending by 2005, is the first Cycle of this new approach. The work on the LFS shows that the National Institute of Statistics (INE) is becoming fairly competent in the area of surveys planning and logistics. But it also demonstrates weaknesses in planning and execution of the associated data processing. In this area, INE acutely needs to strengthen its competence in processing the data. (IDS 2003 suffered many of the same problems, and by increasing the skill level it is hoped to avoid repeating this in the future). The present TOR proposes a mission aiming to rectify this by offering training in CS Pro to the involved data processing staff.

Besides the immediate relevance to the LFS, the mission should help execution of the informal sector survey, financed by funds from a recently started Italian project, due to be executed in 2005. One component of this is planned to be a survey to households that may be incorporated in the LFS, at least with few basic questions. The other component includes activities related to the preparation of the Population Census 2007

1.2 Main reasons for the mission

INE lacks deep experience in data processing, acutely in working with CS Pro in the context of the labor force survey. Assistance is needed to build capacity in this area.

1.3 Benefactors of the mission

The mission will benefit INE and also users of statistics on labor force and employment issues.

1.4 Objectives of the mission

The overall objective of the mission is to assist INE in built capacity in the use of CS Pro, in particular batch control and consistency checks. The mission also should identify weak spots in our data processing pipeline and, if needed, propose changes to procedures and/or further training needs.

1.5 Expected results

- Local Staff trained in more advanced usage of CSPro.
- A document with methodologies describing procedures for data processing using CSPro.
- Problems of the LFS 2004/05 data processing resolved.

1.6 Agenda for the mission

The agenda will be specified on the first day of the meeting in Maputo.

1.7 Terms of Reference

- Train IT staff members in Advanced CSPro programming, namely in:
- Batch control (Join different types of questionnaires)
- Consistency check of multiple responses

1.8 Tasks to be done by INE to facilitate the mission

- Supply the consultants with information on the Integrated LFS regarding problems found during data processing.
- Ensure the availability of the involved INE staff.
- Prepare and supply the consultant with relevant documents and information, such as documentations on how the LFS data entry is being carried out
- Supply good working conditions for the consultant

1.9 Consultant and Counterpart

Consultants: SERPRO???
Main counterparts at INE: Mr. Tomás Bernardo IT Deputy Director at INE

1.10 Timing of the mission

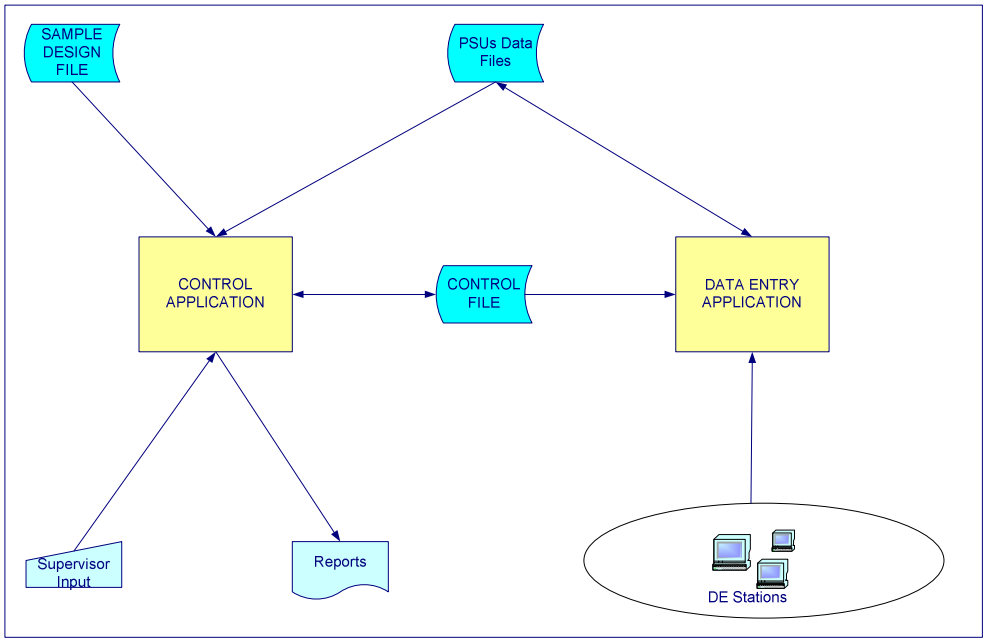
12 June - 26 June.

1.11 Report

The consultants will prepare a draft report of the mission to be agreed with INE. The structure of the report should be according to Danida-format (See Article 3 in the contract). The main content should be the methodologies and procedures of data processing.

The Counterpart has to ensure that the final printed report includes a comprehensive version in Portuguese, if the main report is in English/Spanish and, at least, a summary in English if the main report is in Portuguese.

APPENDIX 2: Survey Control and PSUs Follow-Up



Under this topic a methodology to look for the correct representation of the universe being studied is described. We have identified a PSU as an ideal set of micro-data to become the unit to be monitored and followed up and the block diagram above identifies the different elements that will depict our methodology.

As the diagram shows it, there are two different applications that are linked by a common data file, the "Control File", performing complementary activities that are itemized below:

CONTROL APPLICATION

One common error in the processing of statistical data is the misrepresentation of the universe being studied either by duplication of small geographic or administrative areas or simply the omission of them. In any of those situations, the resulting error can have an important impact in the results since we are over-populating some areas and we are under estimating figures in the second case. The prevention of misrepresenting the universe being studied is rather simple but it is necessary to stick to an organized plan like the one outlined in the following paragraphs.

This application should perform at least the following tasks:

- Creation of the Control file. Using the basic information coming from the sample design, the application will generate the Control File. The Control File will have one record per PSU, using the required geographic/administrative areas (i.e. Province, District, etc.) to which the PSU belong to as part of the key or direct identification. Based on the

sample design data file, the following information should be auto-generated by the application:

- ✓ Complete PSU Identification;
- ✓ Expected number of Unit of Analysis by PSU;
- ✓ Weight or expansion factor;
- ✓ Other relevant information;

Besides the information listed above, at the time the Control file is generated by the application, all the remaining components of the record should be initialized to their default initial values. Thus, this is a task that should be executed at the beginning of the operation but never once the file has been updated with actual information.

- Reception of the PSU: After the information has been gathered in the field, the collection of individual questionnaires (PSU) is sent back to the central office (if DE is centralized) or to the field team headquarters (if DE is field distributed). CAPI (Computer Aided Personal Interview) data capture requires a slightly different follow up and is not covered here. The data items that can be captured at the reception of the PSU might be:
  - ✓ Date when the reception took place;
  - ✓ Manual counting of the number of units (households or other units) for the PSU;

  The reception date as well as the actual count of units will permit to elaborate reports about missing PSUs for certain areas as well as advance reports oriented to administrators of the survey operation.
- Assignment of PSU to primary DE clerk: The DE Supervisor should assign each PSU to a specific DE operator/clerk. The assignment is recognized by capturing the following items:
  - ✓ Date when PSU assignment was performed for primary DE;
  - ✓ DE operator's code or identification;

  Associated to the above described items should be the date when the DE for the PSU was completed as well as the number of units (households or other unit) entered, although the timing when this information is placed in the Control file is different.
  - ✓ Date when PSU primary DE is finished;
  - ✓ Number of units entered (counted directly by application);
- Assignment of PSU for secondary entry: Although CSPro is able to perform the data verification on line –meaning that the entered data are in the background and verification of data in foreground- we suggest that data be entered into two different files and later compared by using a CSPro utility. The reason for doing so is to have a better error statistics allowing the supervisor to judge the quality of the data entry operators' job. The information derived from this assignment is similar to the primary DE. Note that both assignments can be done in parallel since they are two independent files.
  - ✓ Date when PSU assignment was performed for secondary DE;
  - ✓ DE operator's code or identification;
  - ✓ Date when PSU secondary DE is finished;
  - ✓ Number of units entered (counted directly by application);
- Item by Item comparison of primary and secondary data file guided by common data dictionary: This operation is performed automatically by a

CSPro utility producing a complete report of the differences found. The application developer can include/exclude specific variables from the compare process or simply include them all. The report is useful to rate the quality of DE operators allowing an early detection of those that are potentially dangerous for a safe operation. Once this task shows identical files for the primary and secondary DE, the DE is stamped as finished by adding the date when the PSU DE was approved.

Having the Control Data File updated and the Data Entry Application tuned to work harmoniously with the Control file, we can avoid (i) duplication of PSUs since they have to be specifically assigned by the DE supervisor to a specific DE operator; (ii) the omission of PSUs since the report system provided by the Control Application should clearly indicate what PSUs are missing at any time during the survey operation.