STATISTICS DENMARK

Statistisk sentralbyrå
Statistics Norway

Statistiska centralbyrån
Statistics Sweden

MZ:2015:07

Mission Report

for a short-term mission on technical matters of INCAF/IOF

28 September – 17 October 2015

Anne Abelsæth

INSTITUTO NACIONAL DE ESTATÍSTICA

Ref: Contract DARH/2008/004
October, 2014

Anne Abelsæth
Statistics Norway
Anne.Abelsaeth@ssb.no

Contents

ACTIVITIES DURING THE MISSION

During the first days of the mission, the Consultant focused on checking the data from fourth quarter, as the other scanstat consultants, David Megill and Lars Lundgren, needed this for their work. Errors due to flaws in the CSPro application were corrected.
During the rest of the mission, a new system for an ongoing INCAF was planned and started working on. The aim is to make a system that is simpler yet more robust than the current IOF system, and at the same time focusing on making it simple to add or remove new questionnaires. The consultant worked closely with Angelo Intimane and Antonio Nhamuave of the IOF team and other INE staff. The formal counterpart was Arão Balate, Direcção de Censos e Inquéritos.

She also collaborated with her Scanstat consultant colleagues, Lars Lundgren (via email) and David Megill.

INE had expected that the consultant did more data cleaning than she ended up doing. This was due to a misunderstanding and the consultant regrets that very much: She is first and foremost a programmer, not a data analysts, and she did not have the competense to do cleaning at the level expected. (The cleaning done so far in this project, are only corrections due to errors in the software, not traditional data cleaning using statistical methods)

Data cleaning of fourth quarter data

A number of batch jobs were created to automatize the data cleaning. These batch programs are to be found in the folder INCAF\MAIN\BTCH\Batch jobs for 4th quarter\. They are given names starting with a number, which is according to the order they should be run.

1. **Final status:** Due to an error in the application, some interviews were not given a final status when the interview ended. The batch 1_fix_Final_And_PessoasNoAf_Status.bch used the following algorithm to decide the final status. If the following criteria were fulfilled, the interview was considered complete:

   a. There is at least one person in the person roster
   b. There is at least some data in the HABITACAO section
   c. At least one of the sections about natural disasters has at least one field filled out.

2. **Number of persons in the household:** (Same application as in bullet point 1). As the household ID is re-used when there are new families in a dwelling, the pessoas_no_af variable is not always correct. The first batch counts the actual number of people currently living in the household, as this is important to calculate the weights

3. **Sorting:** The household data file is sorted in ascending order, as this is important for the next batches to work properly.

4. **Cleaning of emprego data, first step:** The first batch to fix emprego takes care of the specific errors:
   a. Cases without data is deleted (In some cases there are occurences of incomplete data even though the resultado_final indicates otherwise)

b. Names and age of respondent is not always correct (again, probably due to the reuse of household IDs). Used the data in Agregado Familiar to impute age, as this seemed to be correct.

5. **Emprego, second step:** After taking care of the special cases in previous step, the age was imputed from agregado familiar in the rest of the cases where they differed.

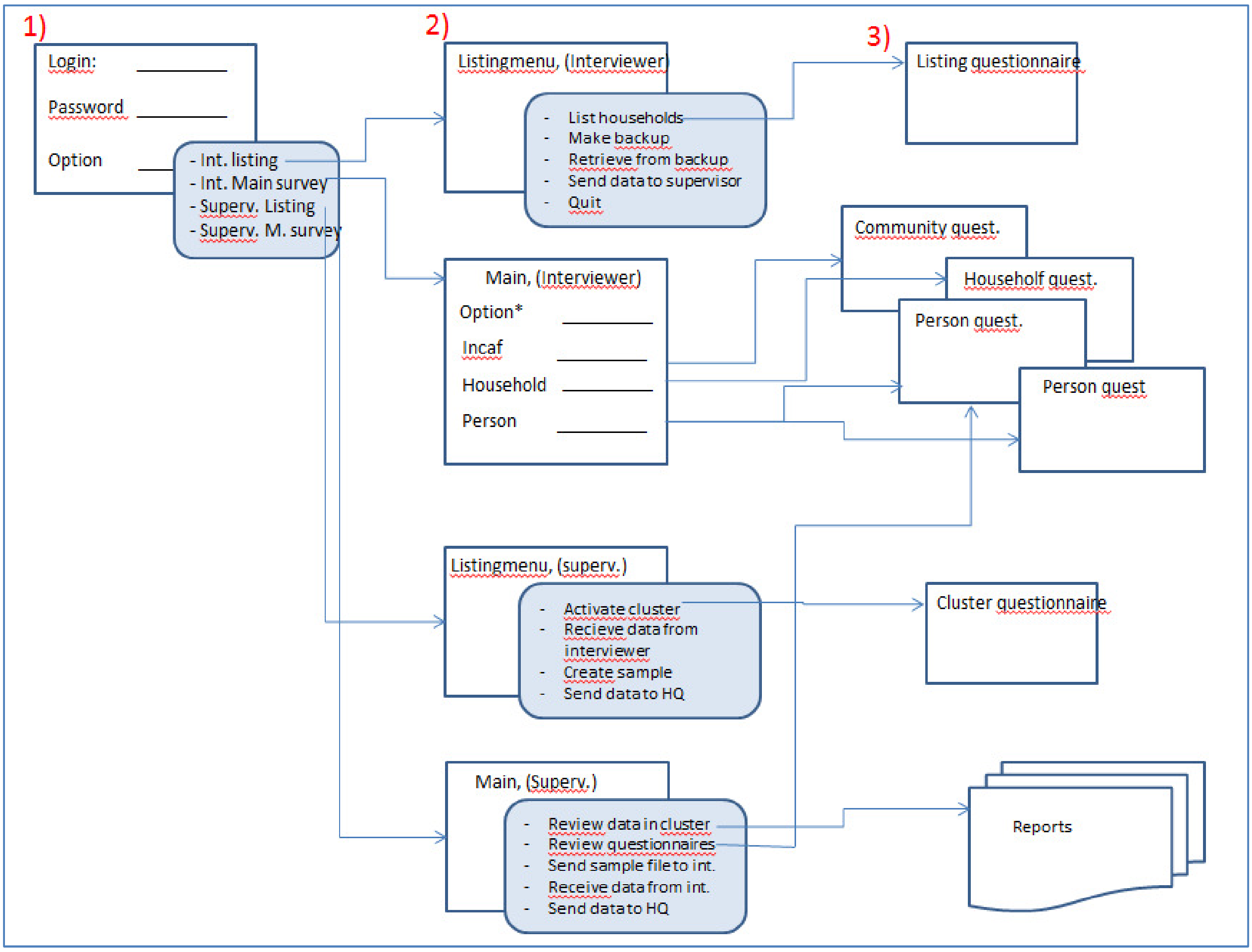
The final data files are named HH_repaired.srt and EMP4_ages _fixed.dat. They are handed over to the IOF staff together with the programs.

Also, a number of summary batch programs were made by request of Davil Megill. These are to be found in the folder INCAF\MAIN\BTCH\Batch jobs for 4th quarter\summary batches

The new INCAF application

The main foci of the improved INCAF application are 1) to make it easy to add new questionnaire, or to remove those that are not relevant anymore, 2) to make it easier for INE staff to maintain the application and 3) to incorporate a listing module in it.

As the tasks for the supervisor and interviewers on the whole are quite different, it is recommended to split the application(s) into 4 separate sub modules: interviewer- and supervisor menus for listing, and interviewer- and supervisor menus for the main surveys according to the following outline:



In the above diagram, the starting point in the upper left corner (marked 1)), is common for Interviewers and supervisors. The menu displayed once the user has been verified, however, is different according to the role of the user, and the four different applications are outlined in under 2). The actual questionnaires are at the column marked 3).

The main menu application of the interviewer needs further explanation: The ”option” button will have choices for doing an interview, but also for backing up data, sending data to supervisor, receive sample from supervisor etc.
If the interviewer chooses to do interviews, there will be three different ”levels” of questionnaires, and all levels can have several questionnaires. In addition it will be predefined what quarter(s) each of the questionnaires are due. The levels are:

- community level (ex. Collecting prices on the markeds or similar). The interviewer only chooses which cluster to continue working with, and he is immeadately taken to the corresponding questionnaire.
- household level: the households of the cluster is displayed for the interviewer to choose from. Questionnaires at this level could be ”Agregado familiar” (like in the IOF), but also agriculture questionnaire, cosumption etc.
- Person level: After choosing household, the relevant person from the household is chosen. Questionnaires here might be childrens- or womens questionnaires, employment, etc.

The levels doesn’t necessarily have to be the above mentioned: changing just the words displayed for the interviewers, it would be easy to adept them to business surveys or any other kinds of surveys.

Every questionnaire have ”metadata” attached to it using the structure of a two dimensional array, where the second dimension contains the following information

1) The name of the questionnaire (text to be seen by the interviewer in the menu, example: "INICIAR QUESTIONÁRIO DO AGREGADO FAMILIAR”
2) The name of the actual application to start up ("AGREG_FAMILIAR")
3) Short name of questionnaire,which also is the name of the folder where it’s found and the name of the folder containing data ("AGR")
4) Level/Type of questionnaire (1 = community, 2 = household, 3 = personal)
5) If questionnaire is at person level: What kind of person (1 = adults 18 or older, 2 = women 15 t0 49, 3 = children between 5 and 17. More groups can be added as needed)
6) Which quarter the questionnaire is to be asked (1 – 4, or 5 = always)
7) A marker to show whether the interviewer has chosen to interview this questionnaire: 0 until chosen, 1 later. This is only for technical use in the program, and should always be 0 at start-up.

So far, only an outline of the system is made. Most of it is still not finished.

FINDINGS AND RECOMMENDATIONS – further work on the application

As stated in the previous mission report:

”The IOF application will work fine for the rest of the IOF period, and most of the errors are probably fixed by now. It is, however, unecessary complex and complicated: it takes experienced CSPro programmers to update or change it. Even a ”simple” thing as to add a new questionnaire is probably still a bit too complicated for an average programmer. So, if the system is to be used in surveys after the IOF, it is highly recommended to restructure and simplify it.”

This still applies. The consultant has the following specific issues that should be taken into the new solution:

1. A template for each type of questionnaire (community, household and person level) should be available whith descriptions of how to program the things that are common for each of them (how to retrieve the parameters from the pff file, how the ”FIM”-button needs to work, how to look up data from other relevant surveys etc.)

2. More data checks should be programmed into the questionnaires than currently are in the IOF.

3. In today’s IOF, when a dwelling has a new household in it, the ID of the household is re-used, and this is the reason for a lot of the errors in the data. In the new INCAF, Ids should never be reused.

4. The bluetooth software to transfer data between interviewer and supervisor seems to be discontinued: The consultant could not find any documentation about it on the internet, nor did she find a version of the client that was newer than 2006. It works for now, but in a longer run, it should be considered to change to a system where documentation is available.
   As of today, CSPro does not support bluetooth data exchange between windows tablets/laptops, but the consultant expects this to change, as it is supported for Android devices today.

5. If there are questions dependent of – or using classifications, look-up files rather than hard coded valuesets could be used (For instance lists of products in consumer questionnaires or employment codes). This would also make it possible to dynamically set meassure types and validity ranges to further ensure data quality (The consultant believes the lack of these things is the main reason for bad quality in IOF’s expenditure and consumer data.)

6. The US census bureau programming standard for CSPro (http://www.csprousers.org/downloads/workshops/general/CSProProgrammingStandards.pdf) should be adapted to make programs easier to read and more understandable.

APPENDIX 1. Persons met

Instituto Nacional de Estatística (INE)

Arão Balate, Director, *Direcção de Censos e Inquéritos*
Cristóvão Muahio, Chief, *Departamento de Metodologia e Amostragem*
Antônio Nhamuave, Programmer
Angelo Intimane, Programmer

Scanstat
Lars Carlsson, Resident Advisor
Lars Lundgren, Household Surveys Consultant
Davil Megill, Sampling Consultant