WP. 13
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (ii): Implementing editing strategies and links to other parts of processing

RELEASE OF MICRO SURVEY DATA. HOW TO DO THIS IDEALLY?

Supporting Paper

Submitted by Statistics Finland[1]

I. CLEANED DATA AND PRINCIPLES

1. Ideally, several requirements should be fulfilled when releasing micro survey data for a user. The first and most important one is that the data are *cleaned as completely as possible*. This requirement is not so simple as may be thought since the cleaned micro data should include at least the following characteristics:

? Descriptions of target population, frame population and updated frame population (if used), and how well the study population matches with the target population. This naturally requires informing about possible gaps or deficiencies of these populations.

? Units of the data set cover both respondents, non-respondents and ineligibles, or the data set includes the complete information of all of these. Units are identifiable in different forms including recognized codes, protected codes and codes for longitudinal matching purposes. The latest point is not needed if there is no purpose to make any longitudinal data later, nor taking some data from registers or other administrative sources.

? Auxiliary variables available, that is, those used in sampling designing, estimation and other cleaning operations and also possible variables that are not (yet) used in cleaning. These variables may be from individual level or from (micro-) aggregate level, and on the other hand, these may be derived from external sources or internal sources (e.g. Laaksonen 2002).

? Survey variables are of highest importance. The information about these should be available to the extent possible from the data set itself using good metadata tools of each particular software. The variables should, therefore, have the informative names and labels, and other complete descriptions should be easily available. Where this is not possible, additional computerized documents could be used (maybe manuals as well). Standardized flags should have been used to recognize the nature of each survey variable including the following cases: unit/item missingness (non-response, impossible value, not available, imputed value and the technique used for imputation), confidentiality (e.g. reduced accuracy due to confidentiality), quality (e.g. preliminary value, approximate value, surveyed value, 'true' value).

[1] Prepared by Seppo Laaksonen, email: Seppo.Laaksonen@Helsinki.Fi or Seppo.Laaksonen@Stat.Fi.

Note that the same initial variable may be included in different forms in the data set using different names and labels.

? A pattern of survey weights: initial and adjusted sampling weights (and methods of adjustments), weights for comparisons and other purposes (e.g. so that the estimates for a certain country group may be correctly done). If the data are based on multi-stage design, the information on each stage should be available including *psu*'s and *ssu*'s used. Also, all components of the inclusion probabilities would be useful.

? The data set is available in different data formats including those workable in general software packages such as SAS, SPSS and Stata, and also in such a format that can be easily converted to any other format (also in the future when the technology will be changed). These formats should include the full metadata information. If there are several versions (early, new) from the same data set, each of these should have been symbolized as understandably as possible.

2. When the data are cleaned and the complete cleaned data file has been created for a survey institution such as a national statistical institute (NSI), the next step is to convert this file to a format that can be released to a user. The purpose is to include in such a file as much of the initial information as possible. Naturally, due to confidentiality, some information will be eliminated from that ideal cleaned data set. The minimum is to exclude the confidential identifiers but to hold the encrypted identifiers. There is no intention to discuss here the details of statistical confidentiality methods but it is best if some values of a certain variable need to be protected, this variable will be included in the cleaned data set in both forms (with different names and labels). Hence when going to construct any released micro file, it is easy to pick up only the protected variables in such a file. It is also quite typical and maybe acceptable that the releasable file includes only respondents, thus non-respondents and ineligibles are missing but as mentioned the information about them should be included in weights and auxiliary variables (at least implicitly).

3. But the creation of a confidential data file is not enough although the file includes good metadata documentation so that this will be easily used by researchers. In the data environment, there should be much more information, and very valuable practical examples on data usage. These should include variance estimation exercises (and confidence intervals respectively) too. If there are specific patterns of variables that are not used as such but after further integration, a pedagogic example should have been attached. These will not be considered in detail here but some additional points will be mentioned in the next section. The key target is that a user can avoid the mistakes to the greatest extent when handling data. On the other hand, the data infrastructure should be such that it beckons to users. It should of course be noted that it is not clear for which level of a user such an infrastructure would be ideal to construct. This requires careful consideration but in general it is not rational to prepare it for a beginner.

II. EXPERIENCES FROM ROUND ONE OF THE EUROPEAN SOCIAL SURVEY

4. These questions will be discussed using the practice of the European Social Survey (ESS) as the key example. Currently, the information from round one of the ESS is available. This round was realized in 22 European countries (including Israel). Some more countries have participated in round two. There are a number of expert groups that have been working on this survey. The purpose has been to make all efforts to achieve high quality (also for cross-country comparisons). Some deficiencies have been naturally encountered, and when some problems have been observed in a certain country, these have been documented and users informed via the website and/or email. This is one important principle for the released data, thus information about specific data problems should be available for users immediately when observed. It is not always possible to correct data afterwards and hence the conclusion may have been to exclude some data or to use special codes.

5. As earlier mentioned, the key tool for releasing ESS data is the website. It is possible that a certain country has included some additional variables in their survey and hence this additional

information is only available from this country. The ESS data are thus publicly available for everyone who desires to use them. This only requires a simple registration before obtaining access to all country files. This registration also means that some special information will be delivered to them (e.g. concerning data problems). Registration is therefore not only for controlling who are possibly using the data, but will be helpful for such a user.

6. Before starting to download the data, "users are obliged to read the ESS conditions of use, which accompanies the data files upon download." These consist of 8 aspects:

LEGAL ASPECTS

1) Restrictions
2) Confidentiality
3) Bibliographic Citation
4) Citation requirement
5) Disclaimer
6) Deposit requirement

USE OF DATA

7) Deviations from the questionnaire and the Data Protocol
8) Weighting

7. The latter two aspects are crucial from the research quality point of view since they are informed about specific data problems in countries and also explained why and how to use weights correctly in analysis. The data files include currently the two types of weights: design weights and population size weights, both averaged to one. No adjusted weights due to missingness (non-response, over-coverage) are available, although the need for these is obvious due to the large variability in non-response within most countries. The reason for this lack is the difficulty to create such adjustments for all participating countries, and if such weights would be only available for some countries, the cross-country comparability may be weakened. Nevertheless, an effort will be made to improve this matter in round two.

8. Data downloading can be done in different formats but SPSS and SAS are preferred. A user can choose a one-country file or to download the integrated file of all 22 countries. Also, a selection of variables may be done if all variables are not of interest. At the same time, a user will receive other useful material so that the work can be easily started. For example, he/she can obtain the full metadata of all classifications used and these may be inserted in outputs automatically. However, in order to understand these completely, it is necessary to look at the questionnaires that are available on the ESS website too.

9. The data file is rather informative even if the variable names (max 8 digits) may be awkward to understand, although they are quite logical. The variable labels are, fortunately, almost always very illustrative and a user can often understand a question without reading carefully the questionnaire. This helps greatly in the multivariate analysis for example.

10. A user has, however, to be careful with variable values since these are not always consistent with the variable names. For example, some scales in the questionnaire are of type "1 = Agree strongly, …, 5= Disagree strongly.' Nevertheless, the variable has a name based on a small value. This may be confusing. The questionnaire designers have considered it best to make the codes in this way, but the analyst has to think everything vice versa unless he/she makes a re-scale that may bring about other confusions. This is a typical problem in general in micro files. Fortunately most variables of the ESS are coded logically, for example many satisfaction variables have been coded as follows: "00 = Extremely dissatisfied, …, 10 = Extremely satisfied."

11. Moreover, the ESS uses specific codes for missing values:
7 or 77 = refused
8 or 88 = don't know
. = completely missing.
These do not provide difficult problems for a user although he/she has to create new variables that take missingness into account so that they match correctly to his/her analysis. Naturally, much transformations and refinements are needed as always for micro data analysis. The infrastructure of the ESS supports well these tasks, although a user is never very happy with these tools.

12. The data file also includes some fieldwork variables so that a user can, for example, analyse the interviewer effect and the interviewing time. This is not common in NSI files. Some item-non-response also occurs and this can thus be found from specific codes as mentioned above. Some missing values of type ' . ' are derived from the self-completed supplementary questionnaire so that some people do not answer any of these questions. It is possible for a user to evaluate the effect of this missingness on estimates.

III. COMMENTS

13. The ESS is a good example for releasing a sample micro file to researchers and to other users. It is thus completely free for users all over the world, which is not the case for all NSI files, but this could be much more common than it currently is. Most NSIs do not release any file in this format, thus using *public use files* (*PUF*). The most acceptable reason for this is data confidentiality, that cannot be solved well in many cases or the data will not be very applicable after such operations. But the main real reason is probably 'money.'

14. It is not the intention to propose that all non-confidential sample survey files of NSIs should be freely available on the web as is the case for the ESS, but this line should be much more common in the future. In order to be successful, this strategy requires a very careful data construction and excellent metadata tools for supporting the quality data analysis. The ESS has succeeded well, although there have not been so many proper users as expected until now. This is mainly due to the lack of competent researchers. Thus although there are excellent tools for using the data, a user should be a good survey methodologist, and at the same time interested in the subject-matters of the survey, and the number of such experts is not very high.

15. On the other hand, since the system is open, it offers everyone a chance to look at the material of the survey. For example, a journalist who interviewed me in Finland on my ESS research outputs, looked at the ESS website at my request, especially at the questionnaires, and saw concretely what were asked from interviewees. She and many others have no interest and ability to use micro data personally but they can understand the feasibility of the survey that is also valuable. This experience thus suggests that the NSIs should at least open their file content (meta data) to all who could be interested in it. This is not so simple either, since it requires a careful documentation. It is not enough to document the file, also the survey/sampling design and the quality of the survey are extremely important to cover.

16. The ESS is not therefore just a unique public use file. In the United States, this has been a tradition over a number of years. For example, Google gives easily a lot of PUF findings such as the National Practitioner Data Bank (NPDP). I looked at their web, and observed that the access system is about the same as that in the ESS. I need not spend much time to download a rather big data file from this bank (more than 300,000 records). The variables were described reasonably and the metadata tools of the file were quite similar to those in the ESS. The study design of this example is not so complex as the survey/sampling design of the ESS and consequently this U.S. case does not require the respective information within the file. Some instructions to correctly use the data are available. If a user nevertheless meets problems, they recommend him/her to contact data specialists listed on the web.

17. Google also gave the link to the Household Internet Use Survey of Statistics Canada. This is not so public as the two examples mentioned above, since it requires ordering the data and paying a fee. The website fortunately includes some information about this survey covering the following sections.

Content note, History note, Microdata specifications, Subjects, Communications, Categories, Internet, Keywords, Free access, Bilingual products, Contacts, Sponsors

18. The section **Microdata specifications** also includes necessary survey design information such as reference year, target group, exclusion of the population and the total weighted number of the population. From the web no details of the sampling design could be foudn, estimation procedures and many other things that are available on the ESS web and at rather detailed level. This Statistics Canada case resembles many U.S. examples.

19. Using the Google search 'Public Use File and Survey' no examples from the first 5 pages outside the U.S. or Canada could be found. The reason may be in terminology since the ESS really is such a file and even more public than many examples of the U.S. and Canada. Eurostat is working in this direction but currently any free micro PUF's are not available from their website.

20. In this meeting, it would be useful to discuss this topic further if not politically then technically, since these questions must have been solved in some way in each NSI. The basic point is that the microdata file must have been constructed in such a format that includes all the elements for creating a PUF. This could mean that first we create a cross-sectional file including all units available (respondents, non-respondents, ineligibles). This file should also include all the information for creating the other types of necessary files. So, an NSI data owner only needs to choose one of the following alternatives (these buttons are just examples) and the file required for this purpose is available in some seconds:

1. Use the file as such
2. Create a file of the unit respondents
3. Create a file for an in-house subject-matter researcher
4. Create a file for an out-house subject-matter researcher
5. Create a file for an in-house methodologist
6. Create a file for an out-house methodologist
7. Create a PUF file

In each case it should be possible to choose sub-data only, based on the selection of records and/or variables.

21. In addition, it is necessary that the system match/link files cross-sectionally, hierarchically and longitudinally. This is no easier than if the question is from one PUF only, but if the above system is working, the new problems arising are not so difficult to pass over. In addition, this system requires the good non-confidential identifiers that may sometimes be very difficult to construct for longitudinal linking/matching files. It will not be easy to establish hierarchical and longitudinal public use files respectively. This question could be a good topic for future UNECE work sessions to discuss.

BIBLIOGRAPHY

Jowell, R. and the Central Co-ordinating Team (2003). European Social Survey 2002/2003: Technical Report, London: Centre for Comparative Social Surveys, City University.

Gabler, S. & Häder, S. & Laaksonen, S. & Lynn, P. (2004). Methods for achieving equivalence of samples in cross-national surveys. *Working Papers from Institute for Social and Economic Research 2004-09*. http://www.iser.essex.ac.uk/pubs/.

Laaksonen, S. (2002). Need for High Level Auxiliary Data Service for Improving the Quality of Editing and Imputation. Paper for the UNECE Work Session on Data Editing in Helsinki, 27-29 May. Available. among others, on the UNECE website: http://www.unece.org/stats/documents/2002/05/sde/8.e.pdf

Some Websites:

Main site of the ESS:
www.Europeansocialsurvey.com
Data archive of the ESS:
http://ess.nsd.uib.no/

The National Practitioner Data Bank:
http://www.npdb-hipdb.com/publicdata.html

Household Internet Use Survey of Statistics Canada:
http://www.statcan.ca:8096/bsolc/english/bsolc?catno=56M0002XCB
