Chapter 12

Summary and key tasks of survey data cleaning

Phases in surveys
A summary of the survey actions by a long list of steps and tasks whose order is roughly such as in practice

A. Aims of the study in which one or more surveys or registers are decided to use for getting necessary empirical data. Thus a survey can be a crucial tool for this purpose. The connections between the whole study and a particular survey are determined in a best way before going to the next step.

B. Specification of the survey design as well as possible. This consists of all the following phases of this summary, but in some sense roughly. The most difficult tasks should already be investigated quite well but some flexibility is good to leave in the design since it cannot be known well in advance how everything works finally. Possible problems should have tried to predict or anticipate.

C. Determining *the target population*: This should be made as precisely as possible, and it should be realistic to achieve. It is good first to consider *a population of interest* and *target group of the intended survey* that may be ideal but not realistic in its best meanings. Note that if you wish to generalize your results, you have to decide what the target population is. If you have no any generalization target, you do have no need to go forward in this summary.

D. In order to approach to the target population it is needed get one or more *frame populations* or sampling frames. These are the lists or registers that consist of the units of the target population at the last stage of the survey design. It is good to take such auxiliary variables from each frame that may be useful in further steps of the survey. If a frame includes directly all study units, many things are easier but it is still good to notice that the frame is not up-to-date for the survey time or period, but it is more or less old. Hence it is good to try to get the updates of the frame or the frames. These are called *updated frame population(s)*. It is good to try to update the auxiliary variables at the same time.

E. *Decide whether to use sampling or not*. Even though the whole population has been tried to survey, similar requirements as sample-based surveys are good to follow. Some survey methods are mandatory to exploit if some missingness occurs as it is the case in all proper surveys.

F. Decision about sampling principles, probability based sampling or non-probability based sampling. The phases of this summary are for probability based sampling but many similar requirements are good to follow in case of non-probability based sampling as well as possible. Case studies could also be used but their generalization to the target population level is difficult but they are at least useful in developing proper sample surveys, and as qualitative studies.

G. Deciding the sampling design for the survey may be a big effort since it requires many things to be taken into account. Its basic target is first to decide what is a reasonable effective sample size and using this it is possible to calculate the required gross sample size for the whole survey and its sub-populations.

H. Planning the data collection including: survey mode (single mode, multi-mode or mixed-mode), data collection tools (mail, phone, face-to-face, web), confidentiality questions during the data collection and when publishing the data and results, cross-sectional or longitudinal/panel, budget and estimation of costs.

I. Questionnaire designing and the questionnaire itself should be ready at this stage but it should be started from the beginning of the survey. If the survey is new, this may be the most demanding task. A part of questionnaire designing, a pre-test or a pilot survey is good to perform, and the survey process should be tested at the same time if possible.

J. *Sampling and the creation of the first-order sampling design data file.* This file consist of the gross sample units and sampling design variables, and as many other useful auxiliary variables (macro and micro) as possible.

K. Decision about the time and the length of the fieldwork. The time should be such when potential respondents are most willing to participate. If the length is short, there is less opportunity to flexibly change something. If the length is long, such as 3 months, it is possible to use for example ‘responsive design’ that aims at motivating such gross sample groups that have not been well participated in the first half of the fieldwork period. The long fieldwork period gives also opportunity to correct mistakes found, to some extent.

L. Data entry as much as possible during the fieldwork. This gives opportunity to check the data in certain basic meanings, so-called data pre-editing. If data entry is manual, the same basic editing can be done but this takes more time and requires more resources.

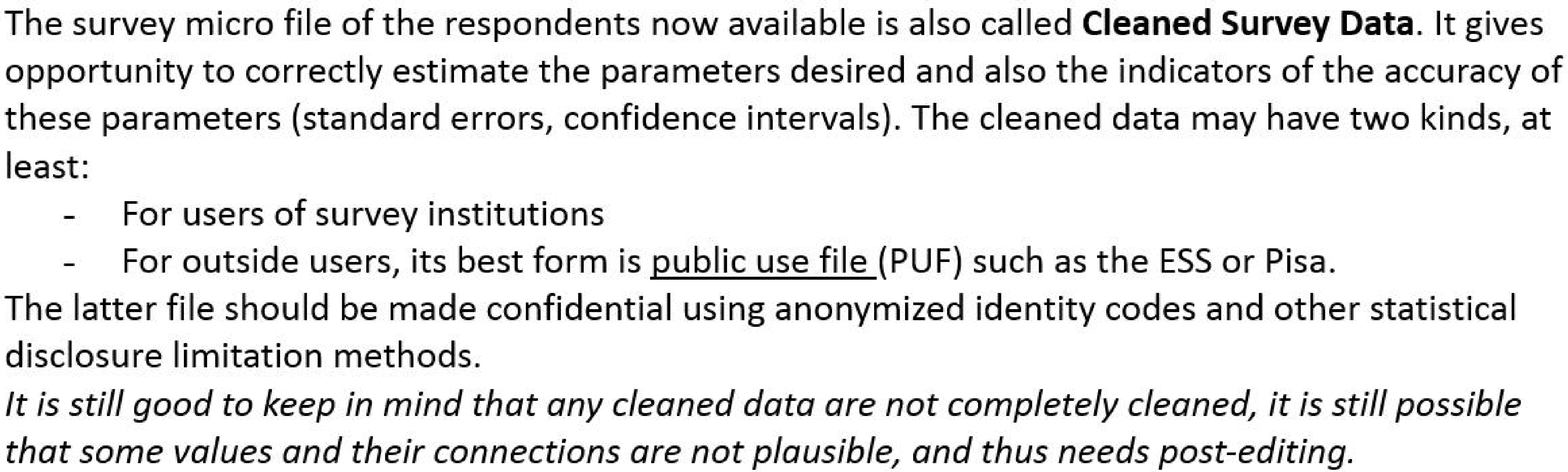
M. Completing the sampling design data file. The most important new variable from the fieldwork is the survey outcome, that is, who have responded completely, and who not? Naturally, reasons for nonresponse and other fieldwork experiences are also documented in the same file, if possible. The sampling design data file may be completed with other auxiliary variables from registers and other administrative sources and from the macro statistics. It is possible to include some auxiliary information found by interviewers as well even though it is not common.

N. *Completing statistical editing,* thus how plausible are individual values and their relationships with core variable values. If they are not plausible, try to correct for them. It is also good to code missing values and their reasons as well as possible (e.g. the value ‘zero’ is a real value but the missing value has a code such as ‘-1’ or ‘99’).

O. *Imputation of such missing and deficient values* that are good to replace with best possible fabricated values or proxies so that the core estimates of the survey will be more precise than without imputation, that is, if they are left missing.

P. Constructing the sampling weights for the respondents, best possible using sampling design variables and other auxiliary variables. The respective weights are needed although the entire target population is attempted to survey but some missingness has been occurred.

Q. Now the survey micro data file is basically ready but in order to use it well by all who will have access into it, the file should be in a good electronic formats (SAS, SPPS, Stata, R, Excel, etc.). The file should have a good meta data so that each user easily knows what each variable means, and what their values (categories, ranges) mean. Para data variables of the survey process and fieldwork are also useful. The survey information that cannot be included in the file, is documented in another way, and should be publicly available.



The survey micro file of the respondents now available is also called **Cleaned Survey Data**. It gives opportunity to correctly estimate the parameters desired and also the indicators of the accuracy of these parameters (standard errors, confidence intervals). The cleaned data may have two kinds, at least:

- For users of survey institutions
- For outside users, its best form is public use file (PUF) such as the ESS or Pisa.

The latter file should be made confidential using anonymized identity codes and other statistical disclosure limitation methods.

It is still good to keep in mind that any cleaned data are not completely cleaned, it is still possible that some values and their connections are not plausible, and thus needs post-editing.

At this stage, you and your team is entitled to celebrate.