



7 th WORKSHOP ON LABOUR FORCE SURVEY METHODOLOGY

DATA PROCESSING AND DATA QUALITY

Madrid, Spain, 10 – 11 May 2012

A. Data processing: Processing of data with panel design and multi-mode data collection. Quality Controls and checks of data

A3–Quality controls and checks in data-editing procedures

Brigitte Hirschbichler - Austria

Theme A – Data processing (3)

Quality controls and checks in data-editing procedure

By Brigitte Hirschbichler, Statistics Austria

The data-editing procedure is divided in several single steps. The very first step thereof is the data collection in Blaise, followed by export of the Blaise data and transformation into a text-file (dg1) and several steps including plausibility checks and imputation, resulting in the final data-file for sale (dg8). The abbreviation “dg” for the different data-files stands for “Datengeneration” = generation of data set. The addition of the numbers 1 to 8 explains the location within the data-management-process from the raw data-file to the final data-file.

Within this process from data collection to final data set different methods can be used to improve the quality of data. Plausibility analysis, quality controls and checks of data are made at different phases in the editing procedure. Therefore the basic structure of relevant sub processes shall be described with regard to plausibility checks and quality controls. In this context the data editing process has to meet input requirements concerning automation, transparency and expandability.

Using computer-assisted interviewing provides the possibility of plausibility checks directly during the interview. These checks or signals can be used to minimise implausible answers or type errors. In the Austrian LFS the software Blaise is used for computer assisted interviewing. CAPI (computer-assisted personal interviewing) is used in the first wave since April 2006 and mostly CATI (computer-assisted telephone interviewing) is used for the second to the fifth waves since 2004. Apart from checks during the interview more analysis and controls take place after the interview, embedded in the overall structure of the data management.

After dg1, which is a text-file, the data management is programmed with the statistic software spss. The entire data-management-programming in spss is designed in a way to allow extensions and modifications easily. Altogether about 450 spss-Programmes constitute the whole process. These programmes have different prefixes depending on whether the syntax refers to a global procedure (meta.), a procedure at a lower level (proz.), the coding of a variable (var.), plausibility checks (pla.) or the imputation of a variable (imp.). Thus it is possible to implement changes for example in the plausibility checks for a variable without the need to change the entire process. For traceability reasons different files at different phases of the process are saved.

First the Blaise-Database, where information is stored on household-level, is translated (by Manipula) into person-level (dg1.txt). Next a Spss import-syntax is generated (by Cameleon). Wherewith dg1.txt is converted into the first Spss-file within the procedure = dg2.sav. Based on dg2.sav the var.sps syntaxes are executed. Data generation 3 includes missing-values for each variable and correct filter (routing). Data generation 4 is generated after all plausibility procedures (pla.sps) have been completed. Then the imp.sps syntaxes are executed and dg5.sav (and dg6.sav - but this file is no longer stored) is generated. After this step dg7.sav is generated. Plausibility checks after imputation and if applicable last modifications or corrections are made at this phase. Data generation 7 is the final internal data-file including weights. Based on dg7.sav the data-file for Eurostat is generated. Data generation 8 is the final external data-file, which can be purchased from Statistics Austria. The difference between dg7 and dg8 is, that for external users (dg8) specific information are removed or aggregated for data privacy reasons.

One main problem relating to plausibility controls has to be considered: If the result of a plausibility check, where two or more variables had been involved, is negative (= implausible), it is not deducible which of the variables contains the untrue respectively the implausible value. Within an automated plausibility procedure it is not possible to decide each time anew from case to case which value is true and which one has to be deleted. There is also the possibility to phone back the corresponding case, but regarding that the final-data-file should be completed in due time, this option can only be used in exceptional cases. The problem of not knowing which value is true and which one is not true can only be solved basically in the course of the interview. Therefore signals in Blaise are adjusted, if major inconsistencies occur during data-editing. To control for implausible answers already during the interview is an important part of the plausibility procedure.

The next important part is related to routing errors. Here a hierarchical approach has been chosen. Important variables for the routing (age, military or civilian service, employment status) are, with the exception of some minor checks, considered correct. The checks for consistent answers of a respondent take place afterwards. New checks are developed continuously which makes them more detailed in the course of time. Since the programming is on variable-level later changes can be implemented any time.

The last plausibility checks are made based on the final data-file (dg7). After these controls (and possibly last corrections) everything should be correct.

The plausibility procedures are thus embedded mainly at three steps within the process from interview to final data-file: During the interview, between data

generations three and four (dg3 and dg4) and before data generation seven (dg7) is finished.