Working paper 27

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (iv): Evaluation and feedback

Analysis of the data preparation process of the structural survey of the federal population census

Prepared by Daniel Kilchmann, Swiss Federal Statistical Office,
Beat Hulliger, University of Applied Sciences Northwestern Switzerland, Switzerland

I. INTRODUCTION

1. The structural survey of the federal population census is part of the Swiss register and survey combined census system. Each year since 2010 a sample of about 250'000 persons is selected.

2. Item non-response, inconsistencies and outliers are detected and treated during the statistical data preparation process (SDPP). This process was designed based on the recommendations of the EDIMBUS-RPM, Luzi, O. et al. [2007], and is therefore split in several phases.

3. The SFSO launched a project to analyse this SDPP with the aim of gathering a deeper knowledge about the impact of the SDPP on results and how the impact evolves during the SDPP between its phases.

4. Based on the findings of the project a better understanding should be gained whether the chosen conceptual framework (EDIMBUS) and process design (SFSO-SDPP) are appropriate. Furthermore, the findings should help to select useful indicators, calculated during the statistical data preparation process, to achieve the aims mentioned above.

5. The data preparation process of the structural survey of the Swiss census is shortly outlined in section II. Section III gives a short overview of the aims of the analysis and the basic indicators under investigation. Preliminary conclusions and next steps are closing this contribution in section IV.

II. Data preparation process of the structural survey

A. The structural survey

6. Since 2010 the SFSO has moved from a classical census to a census based on administrative data combined with sample surveys. The census's structural survey (CSS) is performed every year with a sample size of about 250'000 persons (mean sampling rate of roughly 3%) and is the most prominent of the samples of the new census system in terms of size, periodicity and of its variables.

7. The CSS consists of a person and a household questionnaire. The first one covers labour market, language, religion, education, migration and commuting of the person. The second one focuses on household composition, household member characteristics and dwelling variables.

8. About 25% of the respondents fill in a electronic questionnaire where edit rules are implemented and 75% send back paper questionnaires which are scanned at another federal office. An insignificant part of the sample responds by telephone.

9. The structural survey 2013 was used for the analysis as it was the most recent structural survey available and its underlying processes had reached a consolidated status.

B. Data preparation process of the CSS

10. The redesign of the census system was a good opportunity to apply the findings of the EDIMBUS-RPM to its data preparation processes. Therefore, the data preparation process of the CSS was split up in several phases and several data archives were produced according to the recommendations of the EDIMBUS-RPM, see figure 1 in appendix A. Note, however, that the EDIMBUS-RPM was designed for business surveys while the CSS is a typical person-household survey.

11. In order to allow in depth analysis of the CSS-SDPP more data archives than the minimal number stated in the EDIMBUS-RPM were produced.

12. To achieve the aims of the analysis project, changes between raw data (A_0), data containing the changes due to telephone recalls (A_1), data after deterministic imputation ($A2$) and data after nearest neighbour imputation and deterministic post-treatment ($A3$) are analysed.

13. The dataset A_0 contained originally only alphanumeric variables due to the questionnaire scanning. Hence, it was inevitable to transform some variables to allow the comparison of A_0 with A_1, A_2 and A_3. In the following we will call A_0 the transformed data set.

14. The data set $A3$ corresponds to the data used for publication except for the derived variables which were created inside the macro data preparation phase.

15. Information about changes to data values and units were saved in special tables (not shown in the process flow 1) or flag variables were created throughout the whole process.

16. Due to a thorough development of a nearest neighbour imputation SAS-macro called NNE based on NIM, Bankier, M., Lachance, M. and Poirier, P. [2000], and finalized in 2014, the first three surveys (CSS2010-CSS2012) were published with missing categories. These surveys were then imputed for the first three-year pooled data published in 2015. All CSS ongoing from the CSS2013 were fully imputed for their first publication.

17. **Implementation phase**: During the development of the NNE-macro several loops in the process were performed between the macro and the micro phase. All of them were triggered by analysis on a macro level and leading on the one hand to changes of deterministic imputation rules and on the other hand to modifications to the NNE and its input parameters during its development stage.

18. No loops were anymore necessary once the changes in the micro data preparation phase were performed. Therefore, it was decided not to include these loops in the analysis of the CSS-SDPP.

19. No issues in published results were detected by now, which means that the analysis of the CSS-SDPP is not based on know problems but only on the basis of the aims described in the following section.

III. Analysis project of the CSS-SDPP

A. Aims

20. The general aims of statistical data preparation as described in the EDIMBUS-RPM, Luzi, O. et al. [2007] are to evaluate the input data quality of the survey, to detect problems of the data collection and preparation process and to provide data fit for use.

21. Based on these general aims, and as already summarized in the introduction the aims of the analysis project were defined as follows:

(1) Evaluation of the impact on results of individual treatments or whole phases.
(2) Detection of potential improvements to the process design.
(3) Highlighting of possible questionnaire design problems.

22. Furthermore, the CSS has mostly categorical variables where each response category is coded by a binary variable to allow for multiple responses. Hence a single question of the questionnaire corresponds to a group of binary variables called *response group*.

23. Categorical variables played seldom a major role on indicator lists, e.g. Ehling, M. et al. [2007] and Luzi, O. et al. [2007]. Therefore, the study should also enhance the use of indicators and their interpretation for categorical variables.

24. There was no baseline SDPP for comparison at hand, as the CSS-SDPP was the first of its kind allowing to evaluate the impact in depth. Therefore, it is expected that the findings of this project will lead to a baseline for forthcoming comparison studies.

25. However, there was no true and complete dataset available to test the SDPP. Hence, indicators requiring true values are not considered.

26. It is planned to implement the indicators in a R package to allow to reuse them for further occurrences of the CDD-SDPP and also for the SDPP of other surveys.

B. Evaluation of the impact on results of individual treatments or whole phases

27. Information about the impact of the data preparation process on results is crucial with respect to the expected impact and above all with respect to the analysis.

28. A quite different impact from what was expected based on previous SDPP of the same or similar surveys might be an indicator for problems throughout the whole survey process, including questionnaire design, data collection and preparation processes. This first set of indicators should establish a baseline for future comparisons.

29. As for the impact of the SDPP with respect to the analysis, the project should give the users suitable measurements to allow a better understanding of the data and results based on them.

30. The indicators under investigation are based on the list of indicators of the EDIMBUS-RPM and cover the ones of the standard Eurostat quality reports Quality team of Eurostat [2014]. Some extensions, particularly for categorical variables, are also considered.

C. Setting

31. Due to limited resources the SFSO delegated most of the work of the analysis project to Prof. Beat Hulliger of the University of Applied Sciences Northwestern Switzerland (FHNW).

32. The project started in April and will run until the end of October.

33. The preparation of the data sets for the analysis was quite cumbersome with some unexpected surprises, although lots of data archives and metadata was available. One major problem was that the data archives did not always fit the theory of the SDPP and therefore some information had to be gathered in different data sets than expected.

34. The following paragraphs are based on a preliminary intermediate report from the FHNW.

D. Levels of indicators

The levels of indicators may be structured into

(1) Global indicators: for the whole data set (all observations, all variables)
(2) Subset indicators: for subsets of the data (all variables)
(3) Group indicators: for all observations (groups of variables)
(4) Observation indicators: for single observations (all variables)
(5) Variable indicators: for single variables (all observations)
(6) Subset-group indicators: for subsets of variables and groups of variables

35. When referring to a subset of variables the notion *group* is used while when referring to a subset of observations the notion *subset* is applied.

36. At this stage hierarchical data sets (like persons within households or local units within a business) are not considered but the levels of indicators might be extended to such data.

37. The formulae are given for the weighted versions, i.e. with a positive weight w_i per observation. Unweighted versions are obtained by setting $w_i = 1, \quad \forall i \in S$, where S denotes the sample.

38. It is assumed that the response indicator r_{ij}, the indicator for structurally missing data b_{ij} and the imputation indicator g_{ij} only assume values 1 or 0, say they are dummy variables with the following meanings:

$$
\begin{aligned}
r_{ij} &= \begin{cases} 1 & \text{item response for variable } j \\ 0 & \text{item non-response.} \end{cases} \\
b_{ij} &= \begin{cases} 1 & \text{structurally missing NA} \\ 0 & \text{structurally non-missing} \end{cases} \\
g_{ij} &= \begin{cases} 1 & \text{imputed item for variable } j \\ 0 & \text{not imputed item.} \end{cases}
\end{aligned}
$$

39. The response indicator r_{ij} is calculated with the raw data set A_0 and is not changed afterwards. The structurally missingness indicator b_{ij} is calculated for the raw data set A_0 but may be changed

afterwards if the filtering question is imputed. The imputation indicator g_{ij} is calculated for each version of the data set as the change from the last version and therefore exists for A_1 - A_3. The overall change from the raw data to the final data set may be derived from the changes between the versions.

40. A subset of observations is denoted in the following by S_1 and A_1 denotes a group of variables.

41. **Unit response rate**

$$URR(S_1, A_1) = \frac{\sum_{i \in S_1} w_i \left(1 - \prod_{j \in A_1} (1 - \max(r_{ij}, b_{ij}))\right)}{\sum_{i \in S_1} w_i}, \tag{1}$$

where $w_i = 1$ for the unweighted version URR and w_i a sampling weight for the weighted version $WURR$. Remember that A_1 might refer to one question of the questionnaire where each response category is coded by a binary variable.

42. **Item response rate**
The item response rate for an observation i and a variable group A_1 is

$$IRR(i, A_1) = \frac{\sum_{j \in A_1} \max(r_{ij}, b_{ij})}{\sum_{j \in A_1} 1} \tag{2}$$

We use $\max(r_{ij}, b_{ij})$ for the (rare) case where $r_{ij} = b_{ij} = 1$.

43. The item response rate for a subset of observations S_1 and a variable group A_1 is

$$IRR(S_1, A_1) = \frac{\sum_{i \in S_1} w_i [\sum_{j \in A_1} \max(r_{ij}, b_{ij})]}{\sum_{i \in S_1} w_i [\sum_{j \in A_1} 1]}, \tag{3}$$

where $w_i = 1$ for the unweighted and w_i a weight for the weighted item response rate.

44. The item response rate for one variable j over a subset S_1 is

$$IRR(S_1, j) = \frac{\sum_{i \in S_1} w_i \max(r_{ij}, b_{ij})}{\sum_{i \in S_1} w_i}. \tag{4}$$

45. This indicator corresponds to the weighted response rate Luzi, O. et al. [2007][1]. Thus the indicator can be calculated for subsets and groups.

46. **Item response ratio** The item response ratio should measure the impact of the missingness on the total of a variable in the final data set. For the responding units the imputed value $\hat{y}_{ij}$ is used (after controls and imputations). Of course, in many observations this will be the original value. The indicator should not count those imputed values which should be structurally missing. These values had been given a special code during imputation. We use $r_{ij}(1 - b_{ij})$ as a factor to ensure that if $r_{ij} = 1$ and at the same time $b_{ij} = 1$ the factor is 0. We also use the convention that $NA \cdot 0 = 0$ to avoid problems when summing over observations with $r_{ij}(1 - b_{ij}) = 0$.

$$IRO(S_1, j) = \frac{\sum_{i \in S_1} w_i \, \hat{y}_{ij} r_{ij} (1 - b_{ij})}{\sum_{i \in S_1} w_i \, \hat{y}_{ij} (1 - b_{ij})} \tag{5}$$

47. This indicator is not aggregated over variable groups normally. However, it may be reasonable to build derived variables (like summing over components of a total) and calculate the indicator for derived variables.

[1] Page 67, formula (D.20)

48. **Imputation rate**
The imputation rate measures the proportion of changed items in a data set. It may be defined for an individual observation as

$$IMR(i, A_1) = \frac{\sum_{j \in A_1} g_{ij}}{\sum_{j \in A_1} 1}. \tag{6}$$

49. The aggregate over a subset of observations S_1 is

$$IMR(S_1, A_1) = \frac{\sum_{i \in S_1} w_i [\sum_{j \in A_1} g_{ij}]}{\sum_{i \in S_1} w_i [\sum_{j \in A_1} 1]} \tag{7}$$

50. For a group A_1 of a variable, made up by the dummy variables of its categories, IMR corresponds to the imputation rate of Quality team of Eurostat [2014]. The IMR takes into account that imputations for multiple response groups may differ for each member variable (category).

51. **Imputation ratio** The imputation may be a consequence of a inconsistency or of a missing value. It may also happen that a missing value is imputed, particularly when $b_{ij} = 1$, imputed or original, i.e. for structurally missing items. We then use the convention that $NA \cdot g_{ij} = 0$ whatever g_{ij} is. The case that $b_{ij} = 1$ and $g_{ij} = 1$ may occur if the filtering question has been changed due to an inconsistency. Therefore, only if $\hat{y}_{ij} \neq NA$ and $g_{ij} = 1$ its product is counted in the numerator.

$$IMRO(S_1, j) = \frac{\sum_{i \in S_1} w_i \hat{y}_{ij} g_{ij}}{\sum_{i \in S_1} w_i \hat{y}_{ij}} \tag{8}$$

52. The aggregated imputation ratio of the response group A_1 results in a mean imputation ratio for that group.

$$MIMRO(S_1, A_1) = \frac{\sum_{j \in A_1} \sum_{i \in S_1} w_i \hat{y}_{ij} g_{ij}}{\sum_{j \in A_1} \sum_{i \in S_1} w_i \hat{y}_{ij}} \tag{9}$$

53. **Original value share of a total** The original value share of a total is the part of the total of a variable of a response group which has not been changed.

$$OVS(S_1, j) = \frac{\sum_{i \in S_1} w_i \hat{y}_{ij} (1 - g_{ij})}{\sum_{i \in S_1} w_i \hat{y}_{ij}} = 1 - IMRO(S_1, j) \tag{10}$$

54. Note that $(1 - g_{ij}) = 1$ only if $r_{ij} = 1$ and $b_{ij} = 0$, meaning that a observation may only be counted in the nominator if it belongs to the respondent set. Values which should be structurally missing ($b_{ij} = 1$) but actually are not missing ($r_{ij} = 1$) will be imputed as missing finally and therefore are not counted in (8), (9) and (10).

55. Aggregation of $OVS(S_1, j)$ is useful for a response group A_1 because $\hat{y}_{ij}$ is 0 or 1 only. This leads to the mean share of unchanged response items of a categorical variable, say a question of the questionnaire.

$$MOVS(S_1, A_1) = \frac{\sum_{i \in S_1} w_i \hat{y}_{ij} (1 - g_{ij})}{\sum_{j \in A_1} \sum_{i \in S_1} w_i \hat{y}_{ij}} = 1 - MIMRO(S_1, A_1) \tag{11}$$

56. **Imputation indicators for responded items** The imputation rate due to changes in items which were not missing could serve as an indirect measure of the overall impact of the edit rules, i.e.

of the error detection and localization followed by imputation. With the convention that $\frac{0}{0} = 0$, such an imputation rate for responses could be

$$IMRR(i, A_1) = \frac{\sum_{j \in A_1} g_{ij} r_{ij} (1 - b_{ij})}{\sum_{j \in A_1} r_{ij} (1 - b_{ij})}. \tag{12}$$

57. Note that imputing a missing value when $b_{ij} = 0$ and $r_{ij} = 1$ is counted in the numerator.

58. This indicator may also be aggregated over subsets S_1 yielding

$$IMRR(S_1, A_1) = \frac{\sum_{i \in S_1} w_i [\sum_{j \in A_1} g_{ij} r_{ij} (1 - b_{ij})]}{\sum_{i \in S_1} w_i [\sum_{j \in A_1} r_{ij} (1 - b_{ij})]} \tag{13}$$

59. For the imputation ratio a similar restriction to responded items may be useful.

$$IMROR(S_1, j) = \frac{\sum_{i \in S_1} w_i \hat{y}_{ij} g_{ij} r_{ij} (1 - b_{ij})}{\sum_{i \in S_1} w_i \hat{y}_{ij} r_{ij} (1 - b_{ij})} \tag{14}$$

60. Note that imputing a missing value for a responding item ($r_{ij} = 1$) when $b_{ij} = 0$ is not counted in the numerator and denominator due to the convention that NAs are treated as 0.

61. Several indicators for measuring the imputation impact on the distribution of categorical variables are still under investigation.

E. Core set of indicators

62. Tentatively we may use the set of indicators in table 1 as a core set of indicators. The indicators are mainly used to judge the quality of a final data set and to study the effect of the statistical data preparation process. There are now 5 basic indicators of which 2 may take on a further parameter to restrict the indicator to responded items. In addition all indicators may come in a weighted and unweighted form.

63. Weighting for the variables seems not useful. Rather one would investigate single variables to see exceptional behaviour.

64. The application level is also shown in the table.

TABLE 1. Proposal for core set with application level

Description	Name	Global	Subset	Observation	Group	Variable
unit response rate	URR	x	x		x	x
item response rate	IRR	x	x	x	x	x
item response ratio	IRO		x		x	x
imputation rate (responded)	IMR(R)	x	x	x	x	x
imputation ratio (responded)	IMRO(R)		x			x

F. Detection of potential improvements to the process design

65. Analysing the SDPP thoroughly might show potential improvements to diminish cost intensive procedures or changing the order of procedures.

66. The costs are not available in detail, therefore, assumptions will have to be made. This is certainly a drawback for suggesting important changes in the SDPP.

G. Highlighting of possible questionnaire design problems

67. Problems showing during the SDPP are sometimes due to a bad questionnaire design. Hence, item response rates and indicators showing a strong involvement in inconsistency edit rules by variable might be useful to highlight unexpected response behaviours.

IV. Conclusions and outlook

68. The analysis of the statistical data preparation process of the structural survey of the federal population census system is still at the beginning. Therefore, a limited list of indicators was under study until now and it has to be evaluated whether these are sufficient to meet the aims of the analysis.

69. The indicators described in section III will be tested and implemented in a R-package. Based on the results of the tests it is planned to try to detect potential improvements to the process design and highlight possible questionnaire design problems in a further step.

70. The publication of certain indicators to inform the users of the data will have to be discussed once they are tested thoroughly.

Appendix A.

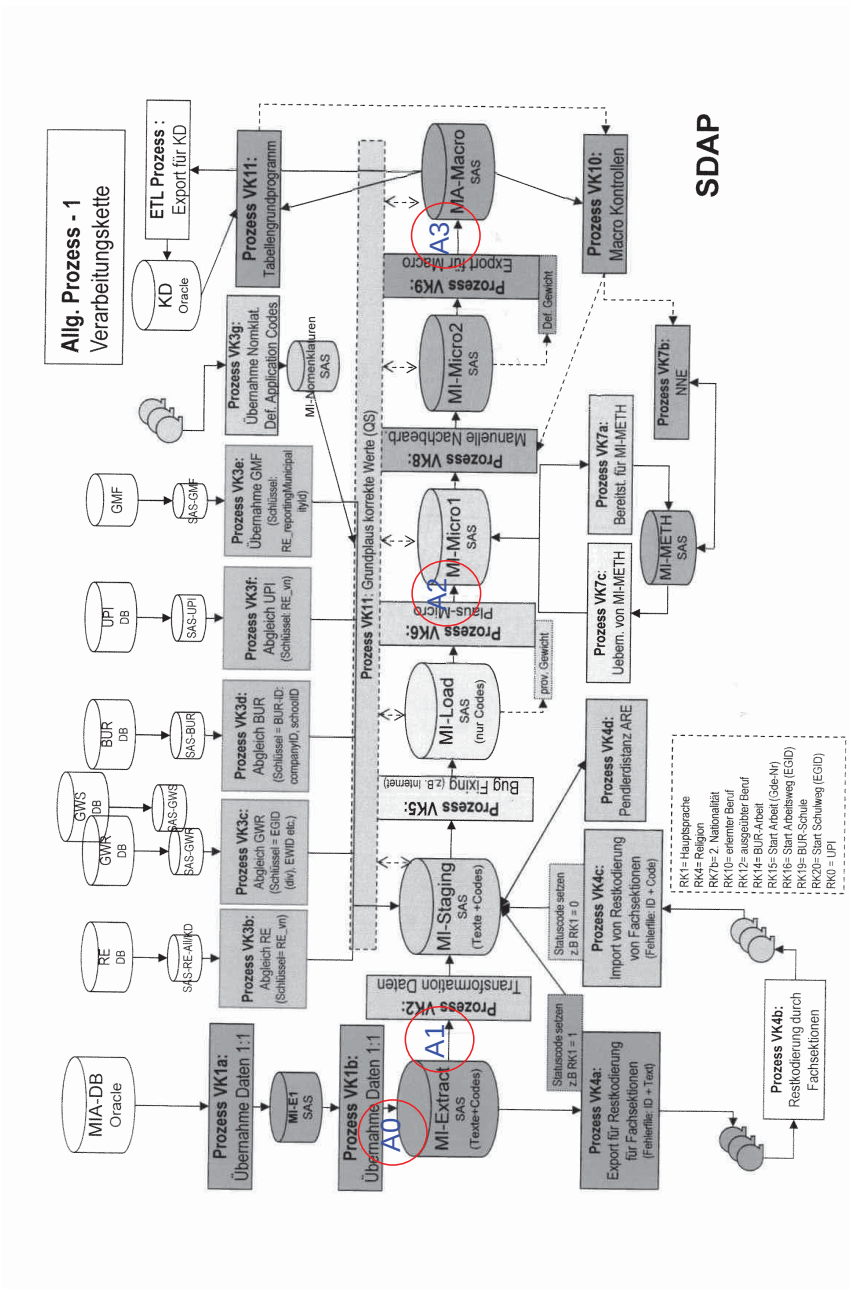

FIGURE 1. Data preparation process of the structural survey of the census.

References

Bankier, M., Lachance, M. and Poirier, P. 2001 canadian census minimum change donor imputation methodology. *Working paper, UNECE work session of Statistical Data Editing, Cardiff*, 2000. URL http://www.unece.org/stats/documents/2000.10.sde.htm.

Ehling, M. et al. *Handbook on Data Quality Assessment Methods and Tools*. European Commission, Eurostat, 2007.

Luzi, O. et al. *EDIMBUS-RPM*. Eurostat, August 2007. URL http://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended+Practices-for-editing-and-imputation-in-cross-sectional-business-surveys-2008.pdf.

Quality team of Eurostat. *ESS Guidelines for the Implementation of the ESS Quality and Performance Indicators (QPI)*. European Commission, Eurostat, 2014. URL http://ec.europa.eu/eurostat/documents/64157/4373903/02-ESS-Quality-and-performance-Indicators-2014.pdf/5c996003-b770-4a7c-9c2f-bf733e6b1f31.