Working Paper

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(The Hague, Netherlands, 24-26 April 2017)

Data preparation process analysis of the structural survey of the Swiss population census

Prepared by Christian Panchard and
Daniel Kilchmann, Swiss Federal Statistical Office, Switzerland

I. INTRODUCTION

1. The structural survey of the federal population census is part of the Swiss register and survey combined census system. Since 2010 a sample of about 300'000 persons is selected each year.

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

3. The aim of the data preparation process analysis is to gather a deeper knowledge about data quality, the impact of the SDPP on results and how the impact evolves during the SDPP between its phases.

4. FSO started to implement quality indicators into the SDPP of the structural survey based on the findings of the project with the University of Applied Sciences Northwestern Switzerland, documented by Hulliger and Berdugo [2015] and presented at the last Work Session, Kilchmann [2015].

5. The data preparation process analysis was implemented by the FSO for the structural survey 2014 and the structural survey 2013 was used as baseline to analyse the evolution of the process quality.

6. 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 . Section IV shows the results of the analysis for the year 2014 and section V gives an overview of the evolution between 2013 and 2014. Preliminary conclusions and next steps are closing this contribution in sectionVI.

II. Data preparation process of the structural survey

A. The structural survey

7. The census's structural survey (CSS) is performed every year with a sample size of about 300'000 persons (sampling rate of roughly 5%) where the households are coordinated negatively from one sample to the other.

8. 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.

B. Data preparation process of the CSS

9. The statistical data preparation process (SDPP) of the CSS was split up in several phases and several data archives were produced according to the recommendations of the EDIMBUS-RPM, Luzi, O. et al. [2007].

10. Four states of CSS data are used for this analysis:

(0) D_0 containing raw data.
(1) D_1, data after changes due to telephone recalls.
(2) D_2, data after deterministic imputation.
(3) D_3, data after nearest neighbour imputation and deterministic post-treatment.

11. The data set D_3 corresponds to the data used for publication.

12. Information about changes to data values and units were saved in special tables, and flag variables were created throughout the whole process.

III. Analysis of the CSS-SDPP

A. Settings

13. The general aims of statistical data preparation, as described in the EDIMBUS-RPM, 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.

14. The aims of the analysis of the CSS-SDPP are defined as follows:

(1) Evaluation of data quality.
(2) Evaluation of the impact on results of individual treatments or whole phases.
(3) Detection of potential improvements to the process design.
(4) Highlighting of possible questionnaire design problems.
(5) Monitoring over time.
(6) Evaluation of the chosen implementation strategy for the CSS.

15. 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].

16. The levels of indicators used for this analysis are

(1) Global indicators: for the whole data set (all observations, all variables).
(2) Variable group indicators: for single questions of the questionnaire (all observations).

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

18. In this settings imputation stands for imputation for missing values and changes to existing values.

19. 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:

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

where i denotes the index for observation and j the index for variables.

20. The response indicator r_{ij}, structurally missingness indicator b_{ij} and the imputation indicator g_{ij} are calculated for each version of the data set D_0, D_1, D_2 and D_3.

21. For the analysis of data set D_3, it is interesting to refer back to the original missingness and to import the response indicator r_{ij} from the data set D_0 which is then denoted by r_{0ij}.

22. Still for the analysis of data set D_3, a joint imputation indicator g_{13ij} can be defined as

$$g_{13ij} = max(g_{1ij}, g_{2ij}, g_{3ij}) = 1 - [(1 - g_{1ij})(1 - g_{2ij})(1 - g_{3ij})]$$

where g_{kij} denotes the g_{ij} of D_k, meaning that g_{13ij} takes value 1 if any imputation was made during the SDPP and 0 otherwise.

23. The indicators are implemented in a R package 'sdap', so that they can be used for every occurrence of the CSS-SDPP and also for the SDPP of other surveys. This package, deveopped by Hulliger and Berdugo of Applied Sciences Northwestern Switzerland, can be installed from source R-forge http://R-Forge.R-project.org with the instruction:

`install.packages("sdap", repos="http://R-Forge.R-project.org")`.

B. Indicators

24. A can denote a group of variables or only one single variable corresponding to an individual tick or to a numeric variable of the questionnaire. In the application it usually refers to all variables building a question of the questionnaire.

25. **Item response rate**

The item response rate gives the proportion of non missing values in a data set:

$$IRR(A) = \frac{\sum_i w_i \left[\sum_{j \in A} max(r_{ij}; b_{ij})\right]}{\sum_i w_i \left[\sum_{j \in A} 1\right]}. \tag{1}$$

26. **Item response ratio**

The item response ratio measures the proportion of the total based on respondents:

$$IRO(A) = \frac{\sum_i w_i \left[\sum_{j \in A} r_{ij}(1 - b_{ij})\ \hat{y}_{ij}\right]}{\sum_i w_i \left[\sum_{j \in A} (1 - b_{ij})\ \hat{y}_{ij}\right]}, \tag{2}$$

where $\hat{y}_{ij}$ denotes the value after imputation. For unimputed units this will be the original value. The indicator excludes units which are structurally missing. These values had been given a special code during data preparation. We use the convention that $NA \cdot 0 = 0$ to avoid problems when summing over observations with $r_{ij}(1 - b_{ij}) = 0$.

27. **Imputation rate**

The imputation rate measures the proportion of changed values in a data set:

$$IMR(A) = \frac{\sum_i w_i \left[\sum_{j \in A}(1 - b_{ij}) g_{ij}\right]}{\sum_i w_i \left[\sum_{j \in A}(1 - b_{ij})\right]}. \tag{3}$$

And in the case of refering back to the initial respondents:

$$IMRR(A) = \frac{\sum_i w_i \left[\sum_{j \in A} r_{0ij}(1 - b_{ij}) g_{ij}\right]}{\sum_i w_i \left[\sum_{j \in A}(1 - b_{ij})\right]}. \tag{4}$$

For the analysis of the final data set D_3, g_{13ij} can be used instead of g_{ij}.

28. **Imputation ratio**

The imputation ratio measures the impact of changed values on the total in a data set:

$$IMRO(A) = \frac{\sum_i w_i \left[\sum_{j \in A}(1 - b_{ij}) g_{ij}\ \hat{y}_{ij}\right]}{\sum_i w_i \left[\sum_{j \in A}(1 - b_{ij})\ \hat{y}_{ij}\right]} \tag{5}$$

And in the case of refering back to the initial respondents:

$$IMROR(A) = \frac{\sum_i w_i \left[\sum_{j \in A} r_{0ij}(1 - b_{ij}) g_{ij}\ \hat{y}_{ij}\right]}{\sum_i w_i \left[\sum_{j \in A}(1 - b_{ij})\ \hat{y}_{ij}\right]} \tag{6}$$

For the analysis of the final data set D_3, g_{13ij} can be used instead of g_{ij}.
The imputation may be a consequence of an 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.

29. **Structural missingness rate**

The structural missingness rate measures the proportion of structurally missing values:

$$SMR(A) = \frac{\sum_i w_i b_{ij}}{\sum_i w_i}. \tag{7}$$

The unweighted version of this indicator is needed to understand the evolution of the indicators due to the imputation of structurally missing values.

IV. Results of the analysis

30. The variables rent, status in employment, completed education, current activity status and main language were selected for the following analysis to illustrate the results of the indicators for the

CSS 2014. Only the weighted version of the indicators was used for the variable rent, because it takes the households which were excluded from the household analysis into account.

A. Item response rate

31. The item response rate is near to 100% for all variables, which means that there is a large proportion of non missing values in the data set (see table 1). Rent has a lower response rate because it is not easily retrievable for some people, or even unknown, when it is directly deduced from the wages. The indicators increase constantly during the SDPP with the exception of rent because the rent of tenants was set to structurally missing between D_1 and D_2.

TABLE 1. IRR

Data set	D_0	D_1	D_2	D_3	D_3 *with* r_{0ij}
Status in employment	0.973	0.983	0.983	1.000	0.970
Completed education	0.968	0.992	0.995	1.000	0.971
Current activity status	0.960	0.993	0.996	1.000	0.963
Main language	0.990	0.997	1.000	1.000	0.993
Rent	0.872	0.929	0.892	1.000	0.884

B. Item response ratio

32. The weighted item response ratio of rent is about 80.7% in D_3 with respect to the initial respondents. It means that the part of the total based on imputation is not negligible, which is due to the lower response rate and outlier detection.

C. Imputation rate

33. The imputation rate is near to 0% for all variables, which means that the proportion of changed items is low (see table 2). Most imputations were performed before D_1 and between D_2 and D_3. The higher values for rent are due to setting of structurally missings and recoding between D_2 and D_3, showing a slight potential of optimization of the SDPP for this variable.

TABLE 2. IMR

Data set	D_1	D_2	D_3	D_3 *with* r_{0ij}	D_3 *with* g_{13ij}	D_3 *with* r_{0ij} *and* g_{13ij}
Status in employment	0.003	0.001	0.005	0.003	0.009	0.004
Completed education	0.006	0.001	0.005	0.004	0.012	0.007
Current activity status	0.010	0.001	0.006	0.006	0.017	0.011
Main language	0.002	0.001	0.002	0.002	0.005	0.003
Rent	0.074	0.199	0.175	0.044	0.260	0.062

D. Imputation ratio

The imputation ratio for rent confirmed the findings made above (see table 3). About 5.8% of the total is due to imputations based on outlier detection and fixing of structural errors like scanning errors. However, this figure probably over-estimates the real impact because the whole amount $\hat{y}$ is taken into account of the changes even for small changes.

TABLE 3. IMRO

Data set	D_1	D_2	D_3	D_3 *with* r_{0ij}	D_3 *with* g_{13ij}	D_3 *with* r_{0ij} *and* g_{13ij}
Rent	0.061	0.013	0.174	0.045	0.251	0.058

E. Structural missingness rate

The structural missingness rate is near to 40% (see table 4), which means that those people are not active (status in employment) or are tenants (rent).

TABLE 4. SMR

Data set	D_0	D_1	D_2	D_3
Status in employment	0.403	0.394	0.391	0.379
Rent	0.452	0.470	0.406	0.412

V. Evolution of the indicators

34. Table 5 gives the unweighted indicators between the years 2013 and 2014 for the data set D_1 and their difference Δ. Low differences mean that the CSS-SDPP is stable over time.

35. A first idea to detect unusual evolutions of Δ, which would highlight special circumstances to the process manager, is to calculate a Wald type confidence interval, see Agresti [2013], page 71:

$$[\Delta \pm 1.96 \times \sigma(\Delta|CSS_{2013}, CSS_{2014})], \qquad \text{with}$$

$$\sigma(\Delta|CSS_{2013}, CSS_{2014}) = \sigma(i_{2014} - i_{2013}) = \sqrt{\frac{i_{2014}(1-i_{2014})}{n_{2014}} + \frac{i_{2013}(1-i_{2013})}{n_{2013}}}$$

where i = IRR, IMR, IMRO or SMR, and n = sample size.

36. The detection of unusual evolutions of the indicators is still under investigation, in particular those concerning the weighted indicators.

VI. Conclusions and outlook

37. The analysis of the CSS-SDPP developed for 2013 has been repeated for 2014.

TABLE 5. Indicators of D_1

		2013	2014	Δ	CI	
IRR	Status in employment	0.983	0.983	0.000	-0.001	0.000
	Completed education	0.993	0.992	-0.001	-0.001	-0.001
	Current activity status	0.994	0.993	-0.001	-0.001	-0.000
	Main language	0.998	0.997	-0.000	-0.001	-0.000
IMR	Status in employment	0.003	0.003	0.000	0.000	0.000
	Completed education	0.006	0.006	0.000	-0.001	0.000
	Current activity status	0.010	0.010	0.000	-0.001	0.000
	Main language	0.002	0.002	0.000	0.000	0.000
SMR	Status in employment	0.392	0.394	0.002	-0.001	0.004

38. For variables other than rent then quality of the CSS-SDPP can be considered as good. There is a large proportion of non missing values in the data and this proportion increases up to the value 100% (see table 1). On the other hand the proportion of changed items stays close to 0% during the CSS-SDPP (see table 2).

39. For variable rent, the indicators show a slight optimization potential. For example, improving the questionnaire could hopefully result in an increase of the response rate and at the same time in a decrease of the number of extreme values which had to be treated as outliers (see table 3). A better scanning quality would result in less corrections of structural errors. Furthermore, recoding might be avoided by designing the SDPP slightly differently. On the other hand, there is no real need to adapt the SDPP because the effects of the above mentioned "problems" were marginal and procedures for treating structural errors and for recoding are purely automatic.

40. As seen in table 5 the CSS-SDPP is stable over time. The evolution monitoring is still under investigation, particularly with respect to weighted indicators.

41. The applied indicators showed to be useful to evaluate input and output data quality. An additional indicator to the imputation ratios, like the "relative average imputation impact", D.33 in Luzi, O. et al. [2007], might improve the assessment of the effective impact due to changes, where only real changes are accounted for and not the whole value $\hat{y}$ like in the imputation ratios.

42. Based on several states of the CSS-SDPP, the indicators highlight easily which part of the SDPP has the biggest impact. That information might be beneficial when the SDPP should be revised.

43. Some minor potential improvements to the questionnaire and the SDPP could be detected. However, there is no urgent need to make changes to the existing CSS-SDPP.

44. The evolution monitoring is still under investigation, especially for weighted indicators.

45. Based on the above findings one can conclude that the implementation strategy chosen for the CSS-SDPP, following the recommendations of the EDIMBUS-RPM, was successful and showed its power to control the process and the possibilities of monitoring.

References

A Agresti. *Categorical Data Analysis*. Wiley, 2013. ISBN 2012009792.

Beat Hulliger and Juan D. Berdugo. Analysis of the statistical data preparation process of the swiss structural survey 2013. Technical report, Swiss Federal Statistical Office, 2015.

D. Kilchmann. Analysis of the data preparation process of the structural survey of the federal population census. *Working paper No. 27 presented a the Conference of European Statisticians, Work Session on Statistical Data Editing, Budapest, Hungary*, 2015. URL http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2015/mtg1/WP_27_Switzerland_SDPP_analysis.pdf.

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.