Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Neuchâtel, Switzerland, 18-20 September 2018)

Some quality indicators of editing and imputations in Household Budget Survey in Bosnia and Herzegovina

Prepared by Edin Šabanović, Agency for Statistics of Bosnia and Herzegovina, Bosnia and Herzegovina

I. Introduction

1. The production of official statistics is highly standardized in order to allow the production of reliable and comparable statistics and to facilitate statistics producers to modernize their production processes. According the Generic Statistical Business Process Model (GSBPM) the whole process of the production of official statistics is contained of eight phases, which are subdivided into several activities. GSBPM is a description of business processes within statistical surveys and it represents a general framework of all statistical activities needed to produce official statistics. Every producer is allowed to adjust this model according to its specifics in order to better define and monitor all processes and sub-processes.

2. Editing and imputations are very important activities within the GSBPM, they are time consuming, they use a significant part of survey budget and they can have a huge influence on survey estimates. For all these reasons editing and imputations are subjects of standardization, methodological harmonization and quality measuring. Approaching common standards in these processes is an important goal of international organisations dealing with statistics and national statistical institutes as main producers of official statistics in every country. All of these efforts are important primarily from the point of view of the quality of statistical products, but also from the aspects of resources that the editing and imputations use in relation to the workload of the entire survey, which are quite large. Both aspects of the standardization activity are particularly important in recent times and a lot of attention is paid to indicators of editing and imputation quality.

3. The Agency for Statistics of Bosnia and Herzegovina still does not have any editing model and editing procedures and methods are not standardized. They are rather defined within individual surveys and not as common generic activities with standardized framework and phases. Our attention is focused on editing and imputation processes and indicators of editing quality in the household budget survey in Bosnia and Herzegovina as one of most complex surveys in our practice, which is conducted in four years periodicity. The aim of this paper is to present some basic indicators of raw data, as measures of the quality of editing and imputation processes. The second section gives an overview of relevant statistical literature dealing with statistical data editing, imputations and the quality of these activities. The third section deals with editing and imputation processes as parts of the GSBPM and with indicators used to measure quality of these processes. In the fourth section we present basic indicators of editing quality, which are based on row data from 2015 household budget survey in Bosnia and Herzegovina. The last section is dedicated to final words and plans for future steps.

II. Literature review

4. Almost all statistical surveys are more or less regulated by different kind of regulations, among which the most common are laws, specific regulations, recommendations, guidelines, manuals, etc. The

same is valid for statistical activities as parts of statistical surveys. All these documents help national and international statistical offices to harmonize their surveys and to have comparable and results of high quality. The proces of statistical production is described by definition of the Generic Statistical Business Process Model, whose latest version is available in UN/ECE manual (UN/ECE, 2013). Within the survey activities, editing and imputations are in the central parts of data processing and data analysis phases. Definition, importance and impact of editing and imputation to the quality of official statistics is discusses in publications of UN Statistical Commission and UN Economic Commission for Europe (UNSC&UN/ECE, 2006). There were explained both, the evaluation of editing and imputation procedures and their role from the perspective of data users and data producers. In the Catalogue no. 12-587-X on survey methods and practices, Statistics Canada preseted definition and purposes of editing procedures as well as main categories of edits. It deals with places of editing and imputations in the GSBPM, design of edit rules, manual versus automated edits, guidelines for editing, constraints to editing and software packages for editing and imputing survey data (Statistics Canada, 2003). Editing and imputations are considered in relation to statistical processes that causes missing data in work of Rubin (1987) and Rancourt (2002). While Rubin deals with missing data patterns and missing data mechanisms, as a probabilistic models, which explains statistical relations between observed and missing data, Rancourt deals with the impact of editing interventions in terms of variance measures. The work of both authors stressed the impact of editing and imputations to the quality of survey results and the need for adjustment of editing and imputation methods to the pattern and mechanisms of missing data. The complete framework for editing and imputation is created within Eurostat project EDIMBUS (EDIMBUS, 2007), which was focused on editing and imputation in business surveys and realized by national statistical institutes of Italy, Netherlands and Swiss. This project resulted with recommended practice manual for editing and imputation processes and methods within European Union, where common practical and methodological framework for editing and imputation applications was given. This manual goes beyond business statistics and its methodological recommendations are valid for any other subject matter statistics. There is a very useful classification of editing and imputations indicators into three groups, depending of whether they measure impact of error detection, error treatment or error detection and treatment. Ollila et al. (2012) discussed on editing indicators from the point of view of international standards, Eurostat recommendations and finish practices. They deal with indicators of row data, indicators related to error identification and those who are related to error correction.

5. The literature review has highlighted the following elements:
   (i) Editing and imputations are very important activities within the GSBPM;
   (ii) Editing and imputations must be adjusted to patterns and mechanisms of missing data;
   (iii) These activities can have huge influence on the quality of official statistics due to close connection of survey phases;
   (iv) There are urgent needs for monitoring of those activities and for measuring and analysing their quality;
   (v) Indicators of editing quality should be a part of survey quality reports in order to better satisfy users' needs and to increase transparency of statistical production.

6. This paper is an attempt to give basic indicators of raw data, which give information about errors on data from household budget survey in Bosnia and Herzegovina. This is a first step in documenting editing and imputation procedures within data processing phases of this survey and it should lead to further work on measuring editing and imputation quality and to overall improvements of these survey phases.

4. Editing and imputation processes and indicators

A. Editing and imputation processes

7. Since all statistical activities within the GSBPM are very related eachother, each of them has an influence on the quality of succesive activities and, consequently, on the overal quality of survey estimates. So, the better harmonisation of survey processes, the higher quality of survey results. According the Generic Statistical Business Process Model, the production of official statistics consists of

eight main phases, each of them is divided in several activities (Scheme 1.)1. Editing and imputations are mostly located within phases 4 and 5 (Data processing and Data analysis), which implies that the majority of editing and imputations are done after data collection phase. But, in broader context, some activities on editing and imputations can be done also within other survey phases, for ex. during data collection where interviewers can do preliminary checks and correct some erroneous data, or where computer assisted data collection can identify possible errors in recorded data, etc.

8. Editing can be defined as the application of checks to identify missing, invalid or inconsistent entries that point to data records that are potentially in error (Statistics Canada, 2003, p. 202). This is a statistical procedure allowing statisticians to identify non-sampling errors due to measurement errors, non-response or data processing errors. By editing procedures, statistical data are checked in terms of individual values and mutual compatibility between the values for different variables. Statistical editing aims to identify and localize missing, invalid or inconsistent data values, while imputations refer to replacing detected edit failures by some plausible values. Editing and imputation are carried out in subsequent phases in order to get complete and internally coherent data set for the analysis.

9. Editing procedures are mostly preceded by complex automated verifications performed by a computer program after the data have been captured. Computer programs incorporate edit rules, which are prepared on the basis of (Statistics Canada, 2003, p. 203):

(i) expert knowledge of the subject matter;
(ii) other related surveys or data;
(iii) the structure of the questionnaire and its questions, and
(iv) statistical theory.

10. There are three main categories of edits, which originated from the editing definition: validity, consistency and distribution edits. Validity edits check the the syntax of responses and for missing values, while consistency edits verify that relationships between questions are respected. At the other hand, distribution edits detect outliers with respect to the distribution of the data.

11. As already said, edits can be performed during and after data collection. Edits during the data collection may be performed by respondents, interviewers, supervisors or statistical staff and their aim is to identify needs for improvements of data collection or needs for more training, to detect obvious errors and perform immediate follow-up with the respondent and to clean-up raw data.

12. There are several constraints to editing, which should be taken in consideration when planning and performing its activities. Main constrtaints are (Statistics Canada, 2003, p. 207):

a) available resources (time, budget and people);
b) available software;
c) respondent burden;
d) intended use of the data;
e) co-ordination with imputation.

B. Indicators of the quality of editing and imputation

13. Editing procedures are time and cost consuming statistical activity. According international researches, in a typical statistical survey, the editing consume up to 40% of survey costs (Nordbotten, 2006, p. 3). Since the timeliness and relevance are equally important principles in producing official statistics, it is very important to optimize editing procedures when planning statistical surveys. One of the possible approaches toward the optimization of resources for editing is the application of selective editing. It helps avoiding over-editing the data, which can introduce editing bias and lead to poor use of survey budget.

[1] UNECE (2013). *The Generic Statistical Business Process Model GSBPM*. Version 5.0, available on: http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0

Quality Management/Metadata Management							
1. Specify needs	**2. Design**	**3. Build**	**4. Collect**	**5. Process**	**6. Analyse**	**7. Disseminate**	**8. Evaluate**
1.1. Identify needs	2.1. Design outputs	3.1. Build collection instruments	4.1. Create frame and select sample	5.1. Integrate data	6.1. Prepare draft output	7.1. Update output system	8.1. Gather evaluation inputs
1.2. Consult and confirm need	2.2. Design variable description	3.2. Build or enhance process components	4.2. Set up collection	5.2. Classify and code	6.2. Validate outputs	7.2. Produce dissemination products	8.2. Conduct evaluation
1.3. Establish output objectives	2.3. Design collection	3.3. Build or enhance dissemination components	4.3. Run collection	5.3. Review and validate	6.3. Interpret and explain outputs	7.3. Manage release of dissemination products	8.3. Agree an action plan
1.4. Identify concepts	2.4. Design frame and sample	3.4. Configure workflows	4.4. Finalise collection	5.4. Edit and impute	6.4. Apply disclosure control	7.4. Promote dissemination products	
1.5. Check data availability	2.5. Design processing and analysis	3.5. Test production system		5.5. Derive new variables and units	6.5. Finalise outputs	7.5. Manage user support	
1.6. Prepare business case	2.6. Design production system and workflow	3.6. Test statistical business process		5.6. Claculate weights			
		3.7. Finalise production system		5.7. Calculate aggregates			
				5.8. Finalise data files			

Sheme 1. Phases and activities of the Generic Statistical Business Process Model (GSBPM)

14. Although there are advantages and disadvantages of this method, selective editing is an approach, which treats only critical edit failures in order to get appropriate survey estimates and which balances between resources and statistical accuracy and relevance.

15. The quality of official statistics is also regulated by the principle 4 (Commitment to quality) of the European Statitics Code of Practice. The overall quality of the statistical products depends on qualities of its sub-products and production processes and activities. The specific caracteristics of the statistical production is mutual-dependency of all processes within the statistical survey.

16. For all these reasons, editing and imputation must be considered as statistical activities in the chain of several survey activities, whose quality depends of the previous activities, but which also has an impact to the quality of succeeding activities and to the final product. A good approach to editing and imputation should consider following three phases: planning of editing and imputation processes, performing of editing and imputation procedures and evaluating the quality of editing and imputation processes and results. This chapter deals with the third phase-evaluating the quality of editing and imputation processes. We will discuss the development of the quality of data in different stages of editing processes and its impact on the overall quality of the final data.

17. One approach to measuring the quality of editing and imputations processes divides quality indicators into three groups (Ollila, Ahti-Miettinen, Oinonen): (i) Indicators of raw data; (ii) Indicators related to error identification, and (iii) Indicators related to error correction. Other similar classification of quality indicators groups them into following three groups (EDIMBUS, 2007): a) Indicators measuring impact of error detection; b) Indicators measuring impact of error treatment and c) Indicators measuring impact of error detection and treatment.

18. Indicators of raw data evaluate the quality of data editing and imputation processes in the beginning phase of these activities. They are used for describing raw data and giving initial pieces of information about errors and their possible effects on variables and results. In this editing phase techniques and measures of descriptive statistical analysis are used such as tabulations, measures of central tendency, distributional analysis of data and data visualisation. On the basis of this analysis, several indicators of raw data can be calculated. In order to allow definition of these indicators and on the basis of above mentioned references, we will introduce formal notation in order to describe statistical data set, variables, observations and other relevant values. In following chapters, we will define several indicators of raw data.

19. Let Y be a data matrix of dimension (nxk), which is a result of data collection within statistical survey, where n denotes number of statistical units or observations and k is number of variables. Let y_{ij} (i=1, 2, ..., n; j=1, 2, ..., k) is a value of variable j for statistical unit i and which may be observed or missing. We define the response indicator in a Rubin`s way, R_{ij}, for observation y_{ij}

$$R_{ij} = \begin{cases} 1, & if\ y_{ij}\ \ is\ observed \\ 0, & if\ y_{ij}\ \ is\ missing \end{cases} \qquad (1)$$

Additionally, x denotes auxiliary variable, which originates from the obtained data set or from other source, while w_i denotes survey weight for observation i.

20. Indicators of raw data cover two groups of indicators: (a) indicators for missingness of data and (b) indicators for impact of observation.

In defining indicators for missingness of data, we will consider only item non-responses, while unit non-responses are not a subject of our attention.

21. Indicators of missingness of data consist of following indicators:

(i) Unweighted item rasponse rate:

$$I_1 = \frac{\sum_i R_{ij}}{n} \quad (2)$$

(ii) Weighted item rasponse rate:

$$I_2 = \frac{\sum_i w_{ij} R_{ij}}{\sum_i w_{ij}} \quad (3)$$

Indicators I_1 and I_2 are calculated for each variable j. Weighted rasponse rate calculates proportion of responses in variable j for the whole population and it is especially useful in complex sample designs or when calibration of weights is performed.

(iii) Weighted rasponse rate for variable y_j proportioned with auxiliary variable x:

$$I_3 = \frac{\sum_i w_{ij} x_i R_{ij}}{\sum_i w_{ij} x_i} \quad (4)$$

Indicator I_3 is useful when variable y_j is correlated with variable x.

(iv) Number of observations with at least one missing value:

$$I_4 = \sum_i \left(1 - \prod_j R_{ij}\right) \quad (5)$$

(v) Rate of observations with at least one missing value:

$$I_5 = \frac{\sum_i \left(1 - \prod_j R_{ij}\right)}{n} \quad (6)$$

(vi) Missingness proportion:

$$I_6 = \frac{\sum_j \left(1 - R_{ij}\right)}{k} \quad (7)$$

(vii) Average proportion of missing values:

$$I_7 = \frac{\sum_i \sum_j \left(1 - R_{ij}\right)}{nk} \quad (8)$$

Indicator I_6 is calculated for each observation i and it is usually calculated for a subset of variables $s<k$. Indicator I_7 is calculated for a complete data set.

(viii) Ratio of item non-response estimated and survey weight estimated totals of x:

$$I_8 = \frac{\sum_i w_{ij}^* x_i R_{ij}}{\sum_i w_i x_i} \qquad (9)$$

(ix) Proportion of variation of item-non response estimated and survey weight estimated totals of *x*:

$$I_9 = \frac{\sum_i w_{ij}^* x_i R_{ij} - \sum_i w_i x_i}{\sum_i w_i x_i} \qquad (10)$$

In both indicators I_8 and I_9 item adjusted weights w_{ij}^* for variable *j* are defined with the equation (11):

$$\sum_i w_{ij}^* R_{ij} = \sum_i w_i \qquad (11)$$

The last two indicators are used to estimate change that item non-response has on the total of variable y_j if *x* and *y* are correlated.

22. The second group of indicators based on raw data are indicators for impact of observation. These indicators are used for measuring significance and this group consists of following four indicators:

(i) Significance of each individual observation y_{ij} in sum of variable y_j:

$$I_{10} = \frac{y_{ij}}{\sum_i y_{ij}} \qquad (12)$$

(ii) Significance of each individual observation y_{ij} subgroup *q*:

$$I_{11} = \frac{\sum_{i \in q} y_{ij}}{\sum_i y_{ij}} \qquad (13)$$

(iii) Significance of each weight w_i in sum of weights:

$$I_{12} = \frac{w_{ij}}{\sum_i w_i} \qquad (14)$$

This indicator identifies statistical units that may have a huge impact to survey results due to high value of their weights.

(iv) Significance of each weighted observation $w_i y_{ij}$ in estimate of total variable y_j:

$$I_{13} = \frac{w_{ij} y_{ij}}{\sum_i w_i y_{ij}} \qquad (15)$$

The last indicator expresses the true impact of observation to total survey estimate.

23. All above defined indicators are very easy to calculate because they represent ratios or proportions to totals. They could be used as inputs for selective editing in order to optimize editing and imputations processes. The potential of these indicators is in their ability to describe raw data, to preliminary detect errors and to evaluate significance of observations in case of incomplete data sets.

5. Indicators of raw data from Household budget survey in Bosnia and Herzegovina

24. In this chapter most important indicators of raw data, which are defined in previous chapter, will be presented. Indicators are calculated from the raw data set from 2015 Household budget survey (HBS) in Bosnia and Herzegovina. Data set contains 7,702 observations (households) with 135 consumption variables. For the sake of simplicity of the analysis for the purpose of this paper, only variables belonging to COICOP division 1 (Food and non-alcoholic beverages) are taken into consideration. The selection of indicators to be calculated was limited by the absence of information about indicator variables e_{il}, f_{ij} and b_{ij}, as they were defined in (EDIMBUS, 2007, p.63). This absence came from the imperfection of data editing and imputations documentation and imperfection of survey instruments within 2015 HBS, which did not allow us to recognize edit failures, detection of erroneous units or structurally missing values. For all these reasons, the number of calculated indicators is relatively small and every absence of data for variable Y is treated as missing data, which should be a subject of editing and imputation.

25. Unweighted item response rates (I_1) for all consumption variables from Diary of purchase are shown in table 1, while basic descriptive statistics and graphical distribution of these variables are presented in table 2 and graph 1.

Table 1. Unweighted item response rates, 2015 HBS in BiH

No.	Variable name	Variable label	Unweighted item response rate (%)
1	DA_A_01	Rice_Diary_Amount	52.31
2	DA_A_02	Wheat flour_Diary_Amount	48.64
3	DA_A_03	OFlour_Diary_Amount	18.35
.	.		.
.	.		.
.	.		.
.	.		.
133	DA_T_02	OtherMed_Diary_Amount	4.97
134	DA_U_01	Lotto_Diary_Amount	9.22
135	DA_U_02	PetFood_Diary_Amount	4.54

Source: Author`s calculation.

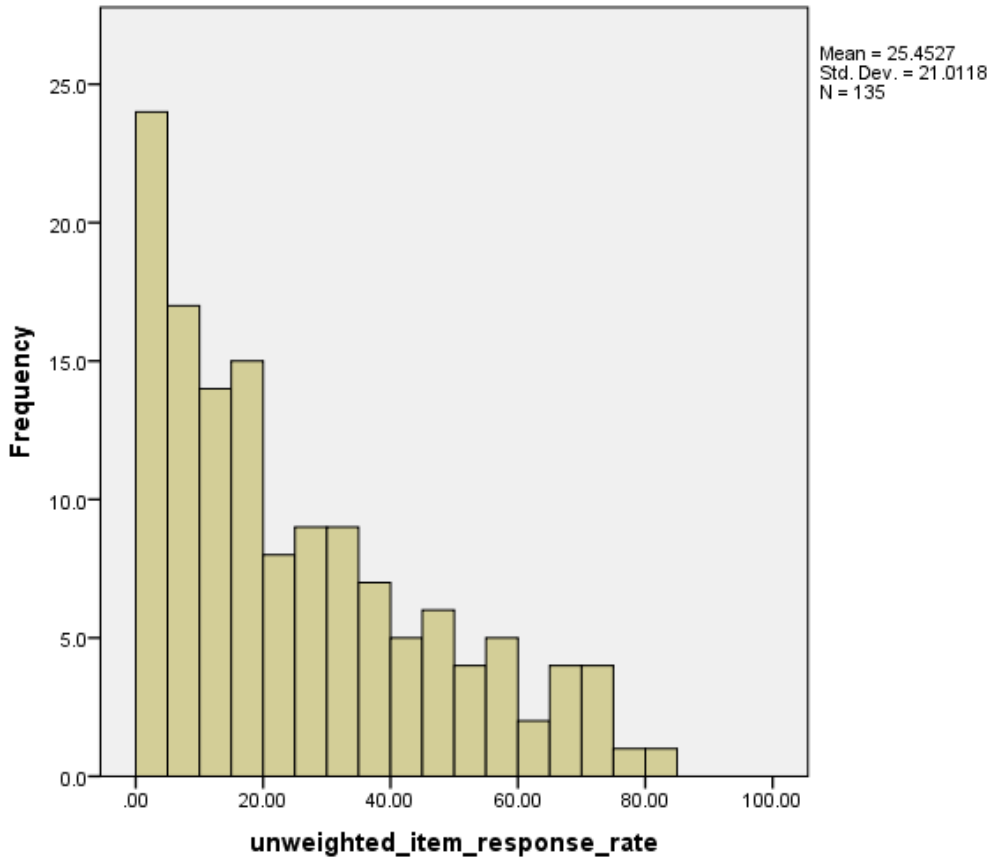
Table 2. Descriptive statistics of unweighted item response rates, 2015 HBS BiH

N	Valid	135
	Missing	0
Mean		25.4527
Median		19.2900
Mode		6.37[a]
Minimum		0.44
Maximum		82.00
Percentiles	1	0.5120
	5	1.7060
	25	8.1100
	50	19.2900
	75	37.7000
	95	69.5520

a. Multiple modes exist. The smallest value is shown

Source: Author`s calculation.

Graph 1. Histogram of unweighted item response rates, 2015 HBS BiH



26. Similar calculations are made for weighted item response rates (I_2). They are presented in following two tables and graph.

Table 3. Weighted item response rates, 2015 HBS in BiH

No.	Variable name	Variable label	Unweighted item response rate (%)
1	DA_A_01	Rice_Diary_Amount	51.77
2	DA_A_02	Wheat flour_Diary_Amount	48.58
3	DA_A_03	OFlour_Diary_Amount	18.60
.	.		.
.	.		.
.	.		.
.	.		.
133	DA_T_02	OtherMed_Diary_Amount	5.10
134	DA_U_01	Lotto_Diary_Amount	9.36
135	DA_U_02	PetFood_Diary_Amount	4.59

Source: Author`s calculation.

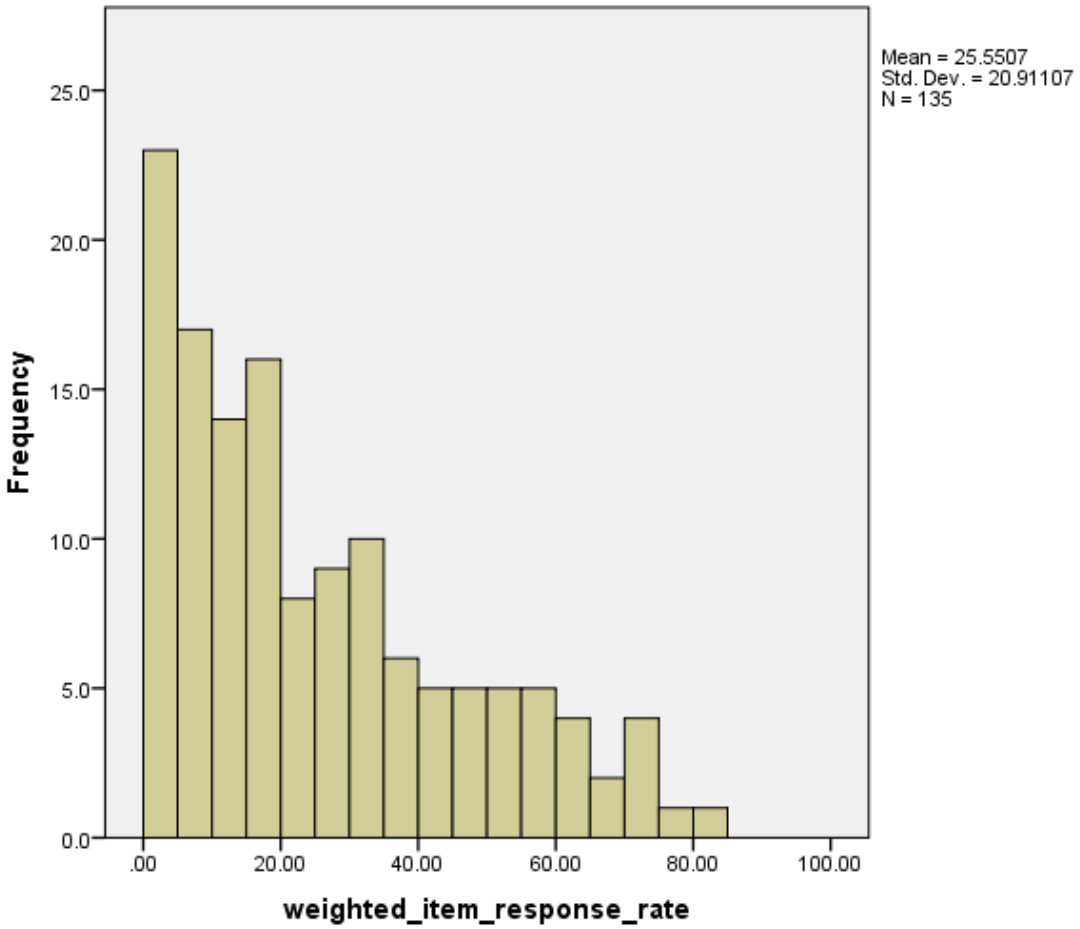
Table 4. Descriptive statistics of weighted item response rates, 2015 HBS BiH

N	Valid	135
	Missing	0
Mean		25.5507
Median		19.4900
Mode		2.40
Minimum		0.46
Maximum		81.48
Percentiles	1	0.5176
	5	1.7820
	25	8.2100
	50	19.4900
	75	37.0800
	95	69.9680

Source: Author`s calculation.

The average item response rate is about 25.5% showing low response for consumption variables in the diary of purchase. The consumption of cigars had the lowest response rate, while the consumption of coffee shown the highest response rate. Low response rates are caused, inter alia, by impossibility to define structurally missing values, as already said in the introduction to this chapter.

Graph 2. Histogram of weighted item response rates, 2015 HBS BiH



27. Weighted response rates for consumption variables proportioned with auxiliary variable x (indicator I_3) are calculated by using household size as an auxiliary variable, which is correlated with consumption variables and available for all observations. Results are shown in following tables and graph:

Table 5. Weighted response rates for consumption variables proportioned with household size, 2015 HBS in BiH

No.	Variable name	Variable label	Unweighted item response rate (%)
1	DA_A_01	Rice_Diary_Amount	55.14
2	DA_A_02	Wheat flour_Diary_Amount	51.44
3	DA_A_03	OFlour_Diary_Amount	19.93
.	.		.
.	.		.
.	.		.
.	.		.
133	DA_T_02	OtherMed_Diary_Amount	5.54
134	DA_U_01	Lotto_Diary_Amount	10.41
135	DA_U_02	PetFood_Diary_Amount	5.21

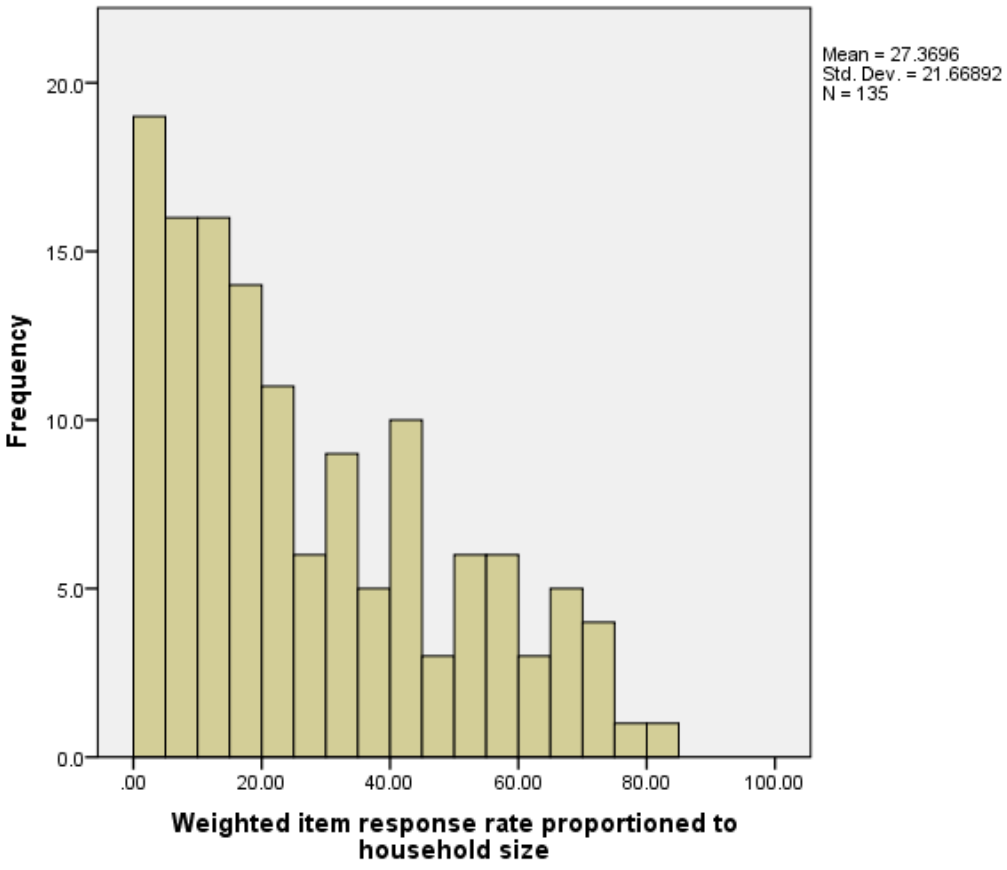
Source: Author`s calculation.

Table 6. Descriptive statistics of weighted response rates for consumption variables proportioned with household size, 2015 HBS in BiH

N	Valid	135
	Missing	0
Mean		27.3696
Median		20.7900
Mode		0.50[a]
Minimum		0.50
Maximum		82.80
Percentiles	1	0.5504
	5	2.0220
	25	9.5600
	50	20.7900
	75	42.6000
	95	69.9220

Source: Author`s calculation.

Graph 3. Histogram of weighted response rates for consumption variables proportioned with household size, 2015 HBS in BiH



The extreme values of this indicators are again present by the consumption of cigars (lowest value) and coffee (highest value).

28. Since there is no possibility to indicate structurally missing values, all empty cells are considered as missing values and therefore the number of observation with at least one missing value (indicator I_4) is 7,702 (each household has at least one missing value). Consequently, the rate of observations with at least one missing value (indicator I_5) is 100%. In this way, these two indicators show worse situation than it is.

29. Missingness proportion (indicator I_6) is calculated for each observation i (i=1, 2, …, 7,702). Individual proportion and descriptive statistics for these variables are shown in following tables and graph.

Table 7. Missingness proportion by households, 2015 HBS BiH

Household	Missingness proportion:
1	90.37
2	88.15
3	77.78
.	.
.	.
.	.
7700	74.81
7701	65.19
7702	74.81

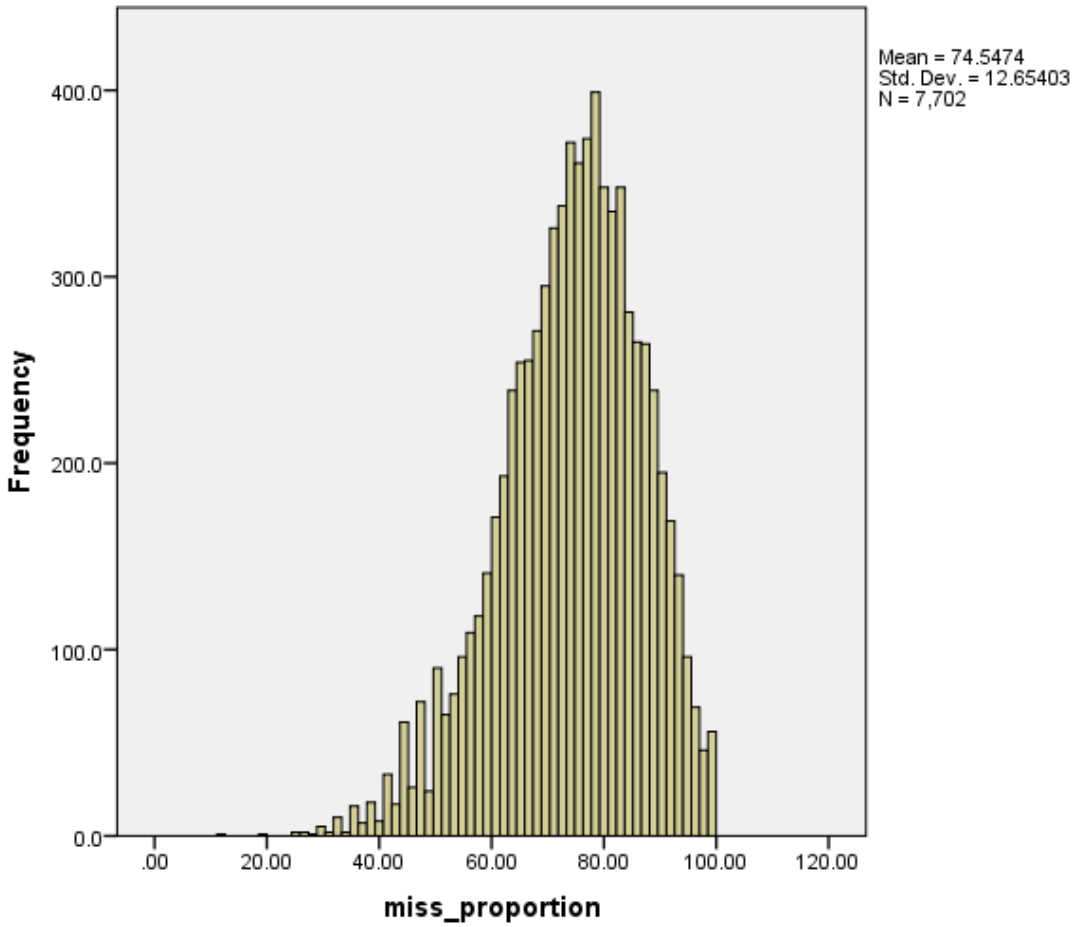
Source: Author`s calculation

Table 8. Descriptive statistics of missingness proportion, 2015 HBS BiH

N	Valid	7702
	Missing	0
Mean		74.55
Median		75.56
Mode		76.30
Minimum		11.85
Maximum		100.00
Percentiles	1	40.74
	5	51.11
	25	66.67
	50	75.56
	75	83.70
	95	93.33

Source: Author`s calculation

Graph 4. Missingness proportion by household, 2015 HBS BiH



Average proportion of missing values (indicator I_7) is 74.55% and it is considered as very high.

30. Indicators I_{10} and I_{13} measure significance of each unweighted and weighted individual observation y_{ij} in total of variable y_j, respectively. Descriptive statistics for each variable is presented in following tables:

Table 9. Descriptive statistics of unweighted significance of individual observation in total of variable, 2015 HBS BiH

Variable	N		Mean	Median	Mode	Minimum	Maximum	Percentiles					
	Valid	Missing						1	5	25	50	75	95
DA_A_01	4029	3673	0.0248	0.0187	0.02	0.01	0.32	0.0085	0.0102	0.0145	0.0187	0.0289	0.0562
DA_A_02	3746	3956	0.0267	0.0298	'.03	0.00	0.24	0.0017	0.0026	0.0096	0.0298	0.0327	0.0630
DA_A_03	1413	6289	0.0708	0.0316	0.02[a]	0.01	0.76	0.0111	0.0158	0.0221	0.0316	0.0711	0.3000
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
DA_U_02	383	7319	0.2611	0.1138	0.23	0.00	4.55	0.0091	0.0241	0.0637	0.1138	0.2720	0.9058
DA_U_02	710	6992	0.1408	0.0908	0.04	0.01	2.09	0.0182	0.0182	0.0454	0.0908	0.1748	0.4540
DA_U_02	350	7352	0.2857	0.1924	0.04[a]	0.02	2.50	0.0208	0.0406	0.0832	0.1924	0.3473	0.8507

Source: Author`s calculation

Table 10. Descriptive statistics of weighted significance of individual observation in total of variable, 2015 HBS BiH

Variable	N		Mean	Median	Mode	Minimum	Maximum	Percentiles					
	Valid	Missing						1	5	25	50	75	95
DA_A_01	535013	498439	0.0250	0.0187	0.02	0.01	0.32	0.0085	0.0102	0.0149	0.0187	0.0289	0.0564
DA_A_02	502088	531363	0.0262	0.0298	0.03	0.00	0.24	0.0017	0.0026	0.0091	0.0298	0.0326	0.0630
DA_A_03	192237	841215	0.0688	0.0316	0.03	0.01	0.76	0.0111	0.0158	0.0221	0.0316	0.0671	0.3000
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
DA_U_02	52693	980758	0.2588	0.1183	0.23	0.00	4.55	0.0159	0.0262	0.0637	0.1183	0.2629	0.9104
DA_U_02	96728	936724	0.1380	0.0908	0.04	0.01	2.09	0.0182	0.0182	0.0454	0.0908	0.1635	0.4540
DA_U_02	47424	986028	0.2737	0.1862	0.04	0.02	2.50	0.0208	0.0374	0.0822	0.1862	0.3369	0.8320

Source: Author`s calculation

The lowest weighted significance of individual observation in total of variable is shown by the consumption of coffee (0.0157), while the highest value of this indicator is present by the consumption of cigars (20.9144), which is consistent with values of previous quality indicators for these two variables.

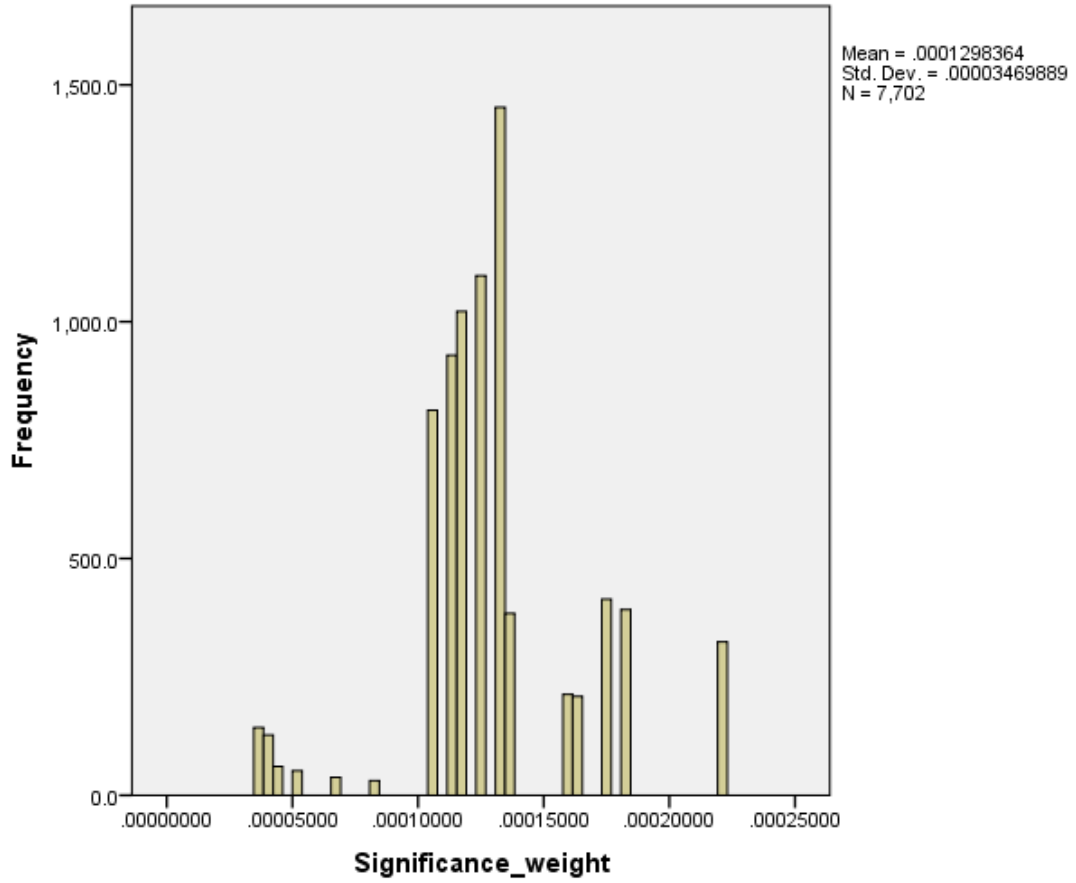
31. Significances of each weight in the sum of weight (indicator I_{12}) are presented in the table 11 and graph 5.

Table 11. Descriptive statistics of significance of each weight in the sum of weights, 2015 HBS BiH

N	Valid	7702
	Missing	0
Mean		0.000130
Median		0.000127
Mode		0.000105
Minimum		0.000037
Maximum		0.000222
Percentiles	1	0.000038
	5	0.000068
	25	0.000115
	50	0.000127
	75	0.000138
	95	0.000184

Source: Author`s calculation

Graph 5. Significance of each weight in the sum of weights, 2015 HBS BiH



The lowest significance of weight is shown by households enumerated in rural area of Brcko district BiH in third quarter 2015 (0.0037%), while the highest significance of weight is present by households enumerated in urban areas of Federation BiH during the second quarter 2015 (0.022%). The average significance of weight is 0.013%.

6. Conclusion and future steps

32. In this paper we have presented several basic indicators of raw data as measures of editing quality in the very beginning phase of editing processes. We have classified these indicators in three groups according their function and according the phase in editing processes within the generic statistical business process model. These indicators are to be calculated immediately after the data collection phase. Use of these indicators, help statisticians to monitor the quality of data processed in different sub-stages of data processing within statistical survey and it contributes to the overall data quality.

33. It is important to mention that not all presented indicators are suitable for every kind of statistics. But, among them, there are few standard indicators that always should be calculated and used in measuring the quality of data collected in every kind of statistics.

34. The applicative part of the paper was produced on the basis of raw consumption data from 2015 HBS BiH. Only consumption variables from COICOP division 1-Food and non-alcoholic beverages were used. The number of indicators calculated is smaller in comparison to the number of indicators defined, because of the lack of specific information on edit failures, detection of erroneous units or structurally missing values. This lack of information is caused by imperfection of editing and imputation documentation and the absence of filter questions in the Diary of purchase, which are needed for definition of structurally missing values.

35. For all above mentioned reasons, indicators of raw data have shown quite huge nonresponses, which means poor quality of data collected. But, also for the same reasons, obtained results must be considered as overestimation of poor quality of data. Calculated indicators of raw data certainly must be used for the first evaluation of designs of previous household surveys in Bosnia and Herzegovina from the point of view of measuring quality of data collected and the quality of data processing methods, such as editing and imputation methods. This analysis is the first step in measuring the quality of raw data and it must be significantly extended and improved in nearest future. Main massages from this analysis are related to questionnaire design, which must allow definition of structurally missing values, and to urgent need for better documentation of all processes and sub-processes within editing and imputation stages. This is a starting point for establishing appropriate editing model in the Agency for Statistics of Bosnia and Herzegovina, as a standard for editing and imputation procedures in statistical surveys. In the end, it will allow calculation of more indicators of the quality of editing and imputation methods and it will contribute to the transparency and the overall quality of statistical production.

7. References

De Waal, T., Pannekoek, J. & Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation.* Wiley

EDIMBUS (2007). *Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys*. Project Report, ISTAT, CBS, SFSO.

Nordbotten, S. (2006). *Evaluating Efficiency of Statistical Data Editing: General Framework.* In Statistical Data Editing, UNSC&UN/ECE, Vol. No. 3, Impact on Data Quality

Pauli Ollila, Outi Ahti-Miettinen, Saara Oinonen (2012). *Outlining a Process Model for Editing With Quality Indicators.* UNECE, Conferene of European Statisticians, Work Session on Statistical Data Editing, Oslo, Norway

Pauli Ollila, Outi Ahti-Miettinen, Saara Oinonen. *Indicators with Respect to Process Model for Editing in Statistic Finland.* Statistics Finland
http://home.lu.lv/~pm90015/workshop2012/papers/W2012_CP_OINONEN_SAARA.pdf

Rancourt, E. (2002). *Using Variance components to measure and evaluate the quality of editing practices*. Conference of European Statisticians, UN/ECE Work Session on Statistical Data Editing, Helsinki.

Rancourt, E. (2005). *Assesing and Dealing with the Impact of Imputation through Variance Estimation*. Working paper No. 10, Conference of European Statisticians, UN/ECE Work Session on Statistical Data Editing, Ottawa, 2005.

Rubin D. B. (1976). *Inference and Missing Data*. Biometrika, Vol. 63, No. 3.

Rubin D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley

Statistics Canada (2003). *Survey Methods and Practices*. Catalogue no. 12-587-X, ISBN 978-1-100-16410-6

Statistics Finland (2007). *Quality Guidelines for Official Statistics.* 2nd Revised Edition, Helsinki. http://www.tilastokeskus.fi/meta/qg_2ed_en.pdf

UN/ECE (2013). *The Generic Statistical Business Process Model GSBPM*. Version 5.0
http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0

United Nations Statistical Commission and United Nations Economic Commission for Europe (2006). *Statistical Data Editing*. Vol. No. 3, Impact on Data Quality