Post-enumeration Survey-A Measure of the Quality of the 2013 Census of Population, Households and Dwellings in Bosnia and Herzegovina

Edin Šabanović, edin.sabanovic@bhas.ba, Agency for Statistics of Bosnia and Herzegovina

ABSTRACT

Bosnia and Herzegovina has conducted Population and Housing Census in the period 1-15 October 2013. After 22 years, the country will get basic data on its population. But, the Census was heavily politicized and main political parties, as well as respondents treated it as a more political than statistical activity. All wanted to have as much as possible people of their ethnic group. Therefore it was enumerated a lot of people who are not resident population.

In order to measure the quality of Census data, Post-enumeration survey (PES) was conducted in the period 02-10 November 2013. It will provide indicators of the census quality in terms of its coverage and content. On the basis of international methodology, specific methods of census and PES units matching were developed. Statisticians did not count with so huge over-enumeration of non-residents, wich could make problems in producing quality indicators.

The paper will give an overview of main characteristics of the census and the PES in Bosnia and Herzegovina; it will present methods of matching inhabitants, households and dwellings and deal with possible problems in producing indicators of census quality. It will for sure be a challenge for political authorities, but also for statistical institutions in the country.

Key words: census, post-enumeration survey, quality, coverage, content

**Background**

The last population and households census data available for Bosnia and Herzegovina was from the census conducted in 1991 within former Yougoslavia. One year later the war started and during the period of four year there were large population movements across the country and tousends od people left their homes and became replaced persons or refugees. About one million inhabitants went abroad, wheile thousands of other residents changed the residence within the country. Mostly rural areas were distroyed and abandoned and people went for safety reasons to towns in Bosnia and Herzegovina or emigrated to other countries. Emigration to other countries was continued after the end of the war because people tried to find abroad better living conditions. Apart of these population movements also new housing construction in the post-war period was considerable. Because of these events, 1991 census data have quickly become outdated and unusable.

In the situation where basic population census data were missing, it was very hard to provide reliable statistics and statistical institutions in the post-war Bosnia and Herzegovina faced a lot of problems, first in terms of sampling of households and individuals and of calculations of national account indicators. On the other hand, Bosnia and Herzegovina in the post-war period did not have neither the population nor the households registers. The postal address registers, if available, were incomplete and the same situation was with the elections lists. There was an urgent need for a new census of population, households and dwellings.

**2013 Population, Households and Dwellings Census in Bosnia and Herzegovina**

Although many efforts were made within the Agency for Statistics of Bosnia and Herzegovina in order to conduct a new census of population, households and dwellings in census-year 2001 and afterwards, because of many reasons (more political than statistical and financial), it happened only in 2013. The census was conducted in a treditional way by using face-to-face interviewing method. The main purpose of BiH Census was to enumerate all census units and to calculate resident population, households and dwellings on the midnight of 30 September 2013. According to the census preliminary results, 3.79 mil. persons, 1.16 mil. households and 1.62 mil. dwellings were enumerated. The quality of this results will be estimated by indicators calculated from the post-enumeration survey, which was conducted two weeks after the census.

**Post-enumeration survey**

The Post-enumeration survey (PES) was conducted in 240 randomly selected enumeration areas (EAs) as a replication of the census enumeration. Data from the PES and the Census will be compared through: (1) Data integration of electronic files on basic Census and PES units (persons, households and dwellings) with establishing appropriate record linkage system; (2) Using of dual estimation system for estimating PES and Census populations and (3) Calculation of appropriate measures and indicators for coverage and quality responses.

Record linkage is a necessary “intermediate” phase in the statistical production process of evaluating Census with PES data and a quality of this process has a strong influence on comparison of the data. The aims of the linking procedure are: (1) to link same persons/households/dwellings, (2) to have minimum errors in the linking procedures; (3) to have a clear status of linking for all units in PES and Census: linked or not-linked and (4) to prepare data for further phases of analysis.

**4. Data Integration and Record Linkage in PES and Census**

Generally, data integration (DI) is the combination of data from different sources about the same or a similar entities. The action allows a construction of a unique record of data collected in different sources for that entity. Record linkage as one of the techniques of data integration, consists of identifying pairs of records, coming from either the same or different data files, which refer to the same entity, on the basis of agreement of common indicators.

According the Fellegi [2], rapid development of tools for data integration in statistical systems and businesses is due to the intersection of the rising needs of new information, using benefits of IT technology and establishing big data bases. In the European statistical system, there is a growing interest for DI in the last twenty years. Two big ESS projects [1]for data integration and integration of survey and administrative data were carried out in period 2008-2011. Data integration from the point of view of application in PES in ESS, is especially interesting in efforts of ISTAT [4], Swiss Statistics [5] and ONS [6] in the censuses round of 2000. There was a variety of different approaches within NSI in Europe in the 2010-11 rounds of censuses. Mainly three ways: (1) continuation with a conventional census, (2) replacement of the census with an alternative system based primarily on administrative records of registers or (3) hybrid models which involve some form of census enumeration coupled with increased use of non-census sources. It means extensive use of data integration, but also great challenges in evaluating data with more integrating aproach than the classical PES.

Data linkage procedures may take several forms, for instance: (1) One-to-one match, as person to person; (2) Many-to-one, as persons in one HH, or persons in EA, or in one town (e.g. geo-coding) and (3) De-duplication process of removing duplicates in the field (e.g. one person enumerated two times with same resident status as present in two different dwellings).

In the case of BIH, data integration is done among Census and PES electronic files of enumerated statistical units and a record linkage with a deterministic approach is used as a statistical method of data integration. PES and Census records are linked for same units, not for similar ones.Theoretically there is no uncertainty in linking: either a pair of records agrees on the unique identifier or they don`t. In the practice, a problem arises when the quality of the variables is not good enough to guarantee that the value of the unique identifier is correct and unique. In this case the problem is that some true matches are not detected and some false matches are interpreted as true ones. In our work mostly the one-to-one match is used, but sometimes also many-to-one match is used, for identifying and linking households or dwellings. De-duplication is also procedure used in the phase of preparing data for record linkage.

**5. Mathematical Model of Record Linkage with Deterministic Approach**

As the record linkage methods with deterministic approach for data integration is used, records are matched in smaller or wider groups (blocks), but only records which satisfy prior defined rules for unique identification on same census unit, are linked.

For simplicity, a model of record linking of two data files (as we have in BiH case) is explained.

Let **A** and **B** be two data sets. Data set **A** composed of **nA** units, while data set **B** composed of **nB** units. Common variables, available on both **A** and **B**, are **X1, X2,..., Xk**, **k≥1.** Let **Ω** be the Cartesian product of all possible pairs of records of the two data sets:

**Ω** **= A** **X B** **= {(a,b): a∈ A, b∈B}** and **Ω** has **nAxnB** units.

The set **Ω** can be split in twodisjoint sets**:**

**M**–set of links **: M= {(a,b): a=b}** and **U**–set of non-links: **U= {(a,b): a≠b}**

**M U U=Ω= A** **X B** and **M ∩U=Ø.**

The objective of record linkage is to determine pairs in **M.** The comparison of unit **a** in **A** with those in **B** is done by „key variables“. For every pair **(a,b**) the key variables are compared and the comparison function is defined as **y**ab=f(**X**Aa, **X**Bb) for registering how different are the values of the matching variables on compared units **a** and **b**.

Let **y**=(y1, y2, y3,... yk) be a vector of comparison of **k** key variables with:

**y**h = **f**(XahA, XbhB ), where h=1,2,...k.

Comparison function gives possibilities of „discrimination“of the different pairs of units between those that are linked and belong in set **M,** or not linked and are in set **U**.

There are many comparison functions for record linkage, but the simplest form evaluates equality: **yh={1,** ifXahA=XbhB; **0**, otherwise}.

In the case of BiH, set **A** is the data set of records on enumerated persons in PES, set **B** is the data set of records on enumerated persons in Census, a key variable for matching persons can be personal identification number (PIN) and comparison function is equality. Records from **A** (PES) with the same PIN as records of **B** (Census) are linked persons among PES and Census. Comparison function takes a role of decision on the status of linkage.

**6. Phases of Record Linkage**

Record linkage is a complex procedure that can be decomposed in many different phases. One of the possible classifications is grouped in three main phases: (1) Pre-processing, (2) Record linkage and (3) Analysis. It is worth noting that record linkage by itself is only a small part of overall data integration process. Gill [3] estimated that ¾ of efforts is in preparing input files, only 5% in carrying out the linkage and 1/5 in checking the results of the linkage.

6*.1. Pre-processing*

*Harmonization of Data Files*

Harmonization of data files is important starting phase for successful data integration and is a very time consuming task that can be avoided by a good statistical organization: a centralized system of definitions for units, populations, variables, variable classifications and a good established meta-data system. Harmonization consists of (1) Harmonisation of units (definition of units, reference period and target populations); (2) Harmonisation of variables (definition of variables, classifications, derivations of variables) and (3) Harmonization on other operational aspects (using rough or clean data, adjustment for measurement errors (accuracy) and adjustment for missing data).

In the case of PES and Census, harmonization of units and variables is totally applied. In PES are used the same definitions of units, reference period, definitions of variables, and classifications as in Census. For the third point on using rough or clean data, there was not a unique approach in the statistical systems and big discussions were made. In BiH a compromise in using rough and clean data is made with the main aim to increase the number of successfully linked records.

*Selection of Matching Variables*

In the selection of matching variables in DI, following general recommendations are to be respected: (1) to select variables with high level of identification power and quality; (2) number of matching variables and some of their characteristics influence the identification of links and (3) before selecting matching variables, perform some exploratory analysis on the distribution of the variable, e.g. number of categories, missing data, etc.

For deterministic record linkage, unique identifiers are created combining some of the common variables (linkage keys): pairs that agree on all the keys correspond to matches. Obviously, a very strict check must be done in the phase of data pre-processing, to harmonize the matching variables and to prevent errors in the linkage procedure. Comparison of the records can be performed on the basis of common variables recorded in files. Generally, as the number of the variables gets larger, the higher is the capability to identify matches correctly. It is preferable to have matching variables that are able to distinguish as much as possible the units (high discriminative power). The variable with larger number of categories has the higher discriminative power. Lack of discrimination does not affect only variables with a few categories, but the same happens also when a variable has a large number of categories, but a few of these categories are much more frequent than others. This is the case of the variables *name* and *surname* for which some categories are much more frequent than others. It would be useless to select them as matching variables.

Matching variables used in BiH are: (1) for persons: PIN and date of birth; (2) for households: number of linked persons in PES or Census, number of persons in PES and Census, data about head of HH, number of dwelling floor and addresses; (3) for dwellings: number of HHs, building-identification code, data about dwellings and buildings and addresses.

 *Selection of Blocking Variables*

In order to reduce the number of pairs, when dealing with large datasets, and improving the performances of the whole record linkage procedure, it is necessary to filter records considering only those pairs that agree on a block variable. Techniques like blocking, sorting, filtering and clustering can be implemented to reduce the search space (AxB or AxA, in case of de-duplication) and removing pairs that are undoubtedly non matches. Choosing *day of birth* and *gender* as block variables decreases the number of candidate pairs and the number of comparisons. On the contrary, when using *gender* on its own, records are split in just two very large subsets. Variables like the *phonetic codes* of *name* or *surname*, or each field of the *date of birth* (day, month and year) can be considered as a good choice for blocking aims.

In the BIH the main blocking variable is EA. The first search and links are among the records of Census and PES with the same EA code. EAs, building-identification code, part of date of birth and addresses were used as blocking variables for persons. For households they were building-identification code and addresses, while for dwelling were used number of HHs, building-identification code, data for dwellings and buildings and addresses.

*Collection of Information Connected with Census and PES*

Developing a good understanding of the sources of data is essential to have meaningful results in the analysis of integration of data. Regular analysis of daily Census and PES field-work reports was done in order to indicate cases with eventual bigger non-response or refusals in some areas or by some groups and to understand ehaviour of public in the period before and during the field-work.

*Manual Preparation of Census and PES Data for Data-entry Purposes*

In this phase, the check of consistency of identifications in all PES questionnaires was done in order to have clean cases for data-entry. Also recheck of aggregated data in paper tables for census units and for cases of non-response was performed.

While during the manual preparation of the pilot PES materials also some editing of the data was done, during the manual preparation of the main PES materials only sorting of questionnaires and recheck of the aggregated data were performed.

*Data Entry of Census and PES*

Census questionnaires were scanned, while PES forms were entered manually through data entry application prepared in VisualBasic.net. In the main PES data entry was done twice in order to avoid data entry errors. The final database was edited and used in process of matching of persons/households/dwellings. Control forms are entered in electronic data files with all comments of enumerators/controllers that can help in linking data, especially when some problems and discrepancies appear.

*Derived Variables and De-duplication*

Some additional variables, as resident status of person, number of enumerated persons in dwelling, number of enumerated persons in HH, number of females and males in HH and others are derived with the main aim to increase the matching and linking appropriate units and every personal record has these data. So, the dwellings have data for number of enumerated persons, number of HHs, number of linked persons and others.

It is highly recommended at first to check all personal files and to investigate people with same personal identifications and resident status: as PIN, data of birth and parts of name. For duplicate cases, the PES team has to decide which case is right and which one is wrong. Wrong cases are excluded from personal files and put in separate files.

*6.2. Record Linkage*

 *Record Linkage of Persons*

First step is linkage for persons from Census and PES data is using the 13-digits PIN number and/or name and/or surname. But, partial data on PIN number and its combination with name and surname of persons and addresses are also used. Matched persons get a status code “linked”, but with some differentiation of the codes in order to know exact way of finding the linked persons (1–linked after matching by 13-digits PIN number, name and surname, 2-linked after matching by 13-digits PIN number and just name, 3-linked after matching by 13-digits PIN number and just surname, 4–linked after matching by 7-digits PIN number part, name and surname, 5–linked after matching by 6-digits PIN number, name and surname, 6–linked after matching name, surname and address, 7–linked after matching name, surname, address and household members names and surnames). Persons enumerated in PES but not found in Census, and enumerated in Census but not found in PES get status code “unlinked”, but again with differentiation of codes in order to know exact situation and number of cases (9 –found in PES, not found in Census, 96–found in Census, not found in PES, 97–found in Census, found in PES too but refused enumeration in PES). Persons that are excluded from PES get status code “out of scope”, with small differentiation of codes (95–persons from collective dwelling, 98–excluded from PES for different reasons). All persons that are not matched by automatic matching are manually checked and matched using same criteria used for automatic matching.

*Record Linkage of Households*

First step is linking households is the elimination of empty dwellings or dwellings with temporary present persons that do not build a household.

Households whose all members are linked in a process of linking the persons, are linked directly. Number of persons from the Census, number of persons from PES, number of linked persons in PES, number of linked persons in Census were used as derived variables. Households with linked relevant mentioned numbers are matched with status “linked”.

Households that are merged in PES but exist as separated households in Census, or households that are merged in Census but exist as separated households in PES are matched using above mentioned numbers of persons and numbers of linked persons. The identified households get status “linked”, but have different codes in order to distinguish the cases.

Households with same number of linked persons from the Census, number of linked persons from PES and number of persons in PES, as well as the households with same number of linked persons from the Census, number of linked persons from PES and number of persons in Census are matched with the status “linked”, but with different codes in order to track the matching rules. Households with number of persons from the Census different from the number of persons from PES and different from the number of linked persons in PES and number of linked persons in Census are checked automatically or eventually manually by the names and surnames of the members and linked with status “linked” (or eventually “unlinked”). Households whose members are enumerated in PES but not enumerated in Census, as well as the households whose members are enumerated in Census but not in PES are matched with status “unlinked” and different codes of matching.

Records that in the questionnaire do not have indications of empty households, but do not have any members, are manually checked in the paper questionnaires and matched in accordance with the situation found in Census. Records that in the questionnaire have an indication of empty household, as well as single households with member enumerated also as a part of other household are excluded with the status “out of scope” but again with different codes.

*Record Linkage of Dwellings*

First step is determination of the starting basis of the dwellings, which was the basis of dwellings enumerated in the KP3 questionnaire, but without duplication of dwellings (these with two or more households enumerated in one dwelling).

Dwellings with linked and same households, as well as the dwellings with different households in PES and Census but still linked, are matched directly with matching status “linked”. Dwellings with different households in PES matched to one dwelling with also different households in Census, as well as dwellings with different households in Census matched to one dwelling with also different households in PES, are manually checked and eventually linked or unlinked in dependence of the found situation and/or errors. Dwellings with merged or split households are matched and linked in case the number of dwellings with those households is the same in PES and Census, or unlinked in case the number of dwellings with those households is not the same in PES and Census. Dwellings that are empty and have complete identification data like address and/or ID-building number in both PES and Census, are matched with status “linked”. Dwellings that are empty but have no complete identification data like address and/or ID-building number or dwellings that are at same time classified as empty and as individual dwellings for living, are not matched and have status “not solved”. The resultas of record linkage were summerized in appropriate frequency tabels indicating all above described categories.

 *Reconciliation Visits*

Field reconciliation visit is an integral part of the PES methodology aimed at resolving apparent discrepancies between Census and PES enumeration. If there is a number of units that cannot be matched, field reconciliation visits are meant to resolve these differences.

The reconciliation visits operation consists of follow-up visits to households in the PES sampled EAs. The purpose of the reconciliation visits is to collect relevant information in order to determine the final match status of unsolved cases identified during initial matching, specifically to: (1) resolve the final match status for “unmatched” cases; (2) determine whether households and/or persons enumerated in the census but not in the PES were correctly or erroneously enumerated in the census; (3) determine whether households and/or persons enumerated in the PES but not in the census were correctly or erroneously enumerated in the PES; (4) clarify doubtful cases or cases with insufficient or unclear information; (5) clarify cases with unclear or insufficient person presence and (6) investigate EAs where boundary or enumeration quality problems are suspected.

 *Finalysing the Status of Record Linkage*

After reconciliation visits, the linking process is finished using the found results from the field. As the Census and PES questionnaires include questions to identify the status of the persons regarding a resident member of the household or temporary present member in the household, comparison of the status of matched persons in PES and Census is done and the persons are matched due to this status too.

 *Analysis*

Analysis of the results is according to the methodology of PES, described in details in the part about coverage control and control of the content.

**7. Concluding Remarks**

Due to the fact that it had not implemented for years, the Population, Households and Dwellings Census in Bosnia and Herzegovina is a challenge, primarily for bosnian population and political parties and not so much for statistical offices that treat it as a job. It was a reason why main political parties had to much influence to this statistical activity during the preparation phase, as well as during the field work. This influence was primarily directed to respondents, which were motivated to register even in the case that they are not residents, but, in some disguised form, also to statistical offices. At the end, the number of enumerated persons is too big. In that case, the post-enumeration survey`s task has become more sensitive and more complicated.

The Agency for Statistics of Bosnia and Herzegovina applied international PES methodology and best practice in order to calculate indicators of census quality in terms of covarage and content. Compared to the practices used in countries from region, our PES questionnaires were not designed to enable re-writing census data on PES forms after the PES field work in order to be matched. Our data integration is done among Census and PES electronic files of enumerated statistical units and a record linkage with a deterministic approach is used as a statistical method of data integration. Results from the pilot post-enumeration survey are promising and together with the implementation of appropriate methods of data cleaning in Census and PES, are a good basis for realistic census figures.

**References**

[1] ESSnet Statistical Methodology, <http://www.cros-portal.eu/content/isad-finished>

[2] Fellegi I.P. (1997), Record Linkage and Public Policy-A Dynamic Evolution, Record Linkage Techniques–1997, Proceedings of an International Workshop and Exposition, Arlington (USA) 20-21

[3] Gill (2001).*Methods for Automatic Record Matching and Linking and their use in National Statistics*, National Statistical Methodological Series No 25, National Statistics, United Kingdom

 [4] Nuccitelli A. (2004), Integrating the 2001 population Census and Post-Enumeration Survey data: the Italian experience, European Conference on Quality and Methodology in Official Statistics (Q2004), Wiesbaden (Germany)

[5] Swiss Statistical Office (2004), Coverage Estimation for the Swiss Population Census 2000, Neuchatel (Switzerland)

[6] Abbott O. (2008), 2011 UK Census Coverage Assessment and Adjustment Methodology, European Conference on Quality and Methodology in Official Statistics (Q2008), ONS-UK

[7] United Nations Secretatiat, Department of Economic and Social Affairs, Statistics Division (2010), Post Enumeration Surveys-Operational Guidelines

[8] Agency for Statistics of Bosnia and Herzegovina (2014), Preliminary Results of the 2013 Census of Population, Households and Dwellings in Bosnia and Herzegovina, First Release

[9] Agency for Statistics of Bosnia and Herzegovina (2014), Manual for Linking Data Working Paper