Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

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

Preparation for use of administrative data sources for Structural Business Statistics

Vesna Šehovac, Agency for Statistics of Bosnia and Herzegovina, Bosnia and Herzegovina

I. Introduction

Structural Business Statistics (in further text SBS) data in Bosnia and Herzegovina are collected annually in the form of survey. SBS currently passes through the data collection transformation phase. Through the Twinning Project, we are trying to use administrative data sources as well as analysing and improving the complete process of SBS data production.

Project activities are:

- Analysing current process collecting and statistical data editing;
- Analysing availability and quality of financial reports (administrative sources);
- Finding the best solution for using administrative data;
- Analyse available administrative data (timing, availability, quality);
- Deciding how and on which way to incorporate that data in our SBS database;
- Analysing AD data and creating procedures for controls;
- Controlling and identifying errors;
- Analysing data errors and defining basic principles of editing as well as defining editing rules;
- Creating automatic scripts for editing;
- Statistical data editing.

The transformation process from the previous state to the current state will be briefly presented in this paper

II. History

A. Secondary heading

SBS survey is conducted on a selected sample of companies in Bosnia and Herzegovina. An appropriate sample of enterprises from the business register is selected and they are asked for detailed information related to the data of business statistics. Action involves printing forms and sending to enterprises to gather the data. The forms are detailed and require exhaustive completion process from business entities. After completing the survey, the data go to two entity statistical centres - Statistical Institutes: the Federal Institute for Statistics and the Republika Srpska Institute for Statistics, as well as to the Agency for Statistics - Brčko District B&H, which does not belong to entity institutions but is under the autonomous District rule and Statutory Jurisdiction of the State.

After hard and demanding fieldwork, the data is entered into an application developed by the Agency for Statistics of B&H - BHAS, Application is in VB.NET technology and is based on a database that was made on the MS SQL Server. Regardless to the fact it is a web application, it is distributed to three physical locations at addresses of three Statistical Institutes in BH. Questionnaires from the field and imputing data was conducting separately (delivery and collecting questionnaires, entering data, maintaining, controlling) in three databases. After entering and validating the data on the field office the data are sent in a secure way to the Agency. BHAS unifies data in one database where final validation and calculation of variables are performed. After processing the data through SQL procedures: grossing up, calculating confidentiality, final validation, adding entrepreneurs and translating into Eurostat format, data are finally published and sent.

It is evident that the above mentioned demands great efforts and costs in the form of: preparation of forms, preparation of applications for each year and all three destinations, distribution of forms and applications, field work, manual entry, control and validation, contact of business entities in case of necessary amendments or edits, data transfer to central location, data unification, SQL data processing, variable calculations, grossing up.

III. Current activities

Businesses that already have a large number of administrative and bookkeeping reports in business terms that are legally necessary to be sent to the other administrative bodies are also all subjected of structural business statistics survey.

Due to the burden on business entities and statistical institutions and taking into account limited resources, we began to think about the use of administrative sources for collecting business statistics data as existing and reliable sources that our European partners have been using for many years. This effort is supported by one component of the Twinning project.

With the help of international experts, we will try to find the most suitable way to transfer to administrative sources, in order to:

reduce the burden on business entities filling out multiple forms with similar data,
reduce the burden on statistical institutions,
reduce costs,
accelerates data processing,
increases data quality,
get exact data instead of estimates.

Through the several projects missions of this year and efforts of statisticians in all three statistical institutions as well as the IT staff from the Agency for BHAS, the use of administrative resources will be initiated, communication with respective state agencies will be established, information-based transformation procedures on the database in order to operate and collaborate with survey data will be implemented, while the scope of the survey will only be reduced to large companies with 20 employees and more, and for other data administrative sources will be used. Since data for companies over 20 employees - from administrative sources as well as from SBS surveys will be available, it is possible to check the validity of the data and of one and the other source by comparing and performing the controls

The data will unite at the variable level. After merging data at a central location, the next step in processing data can be made. Performing predefined controls, checking consistency, communication with Entities NIs and Enterprises, making manual correcting and validation of data, will be explained in the rest of the document.

IV. Incorporating administrative sources data in SBS

A. Business integration model

The local environment is complicated by the fact that there are three entities involved in statistical production, with different sources of AD and three legislations to adhere to. The situation is demanding because there are 3 official statistical institutions in the statistical production. They have to be looked individually and together:

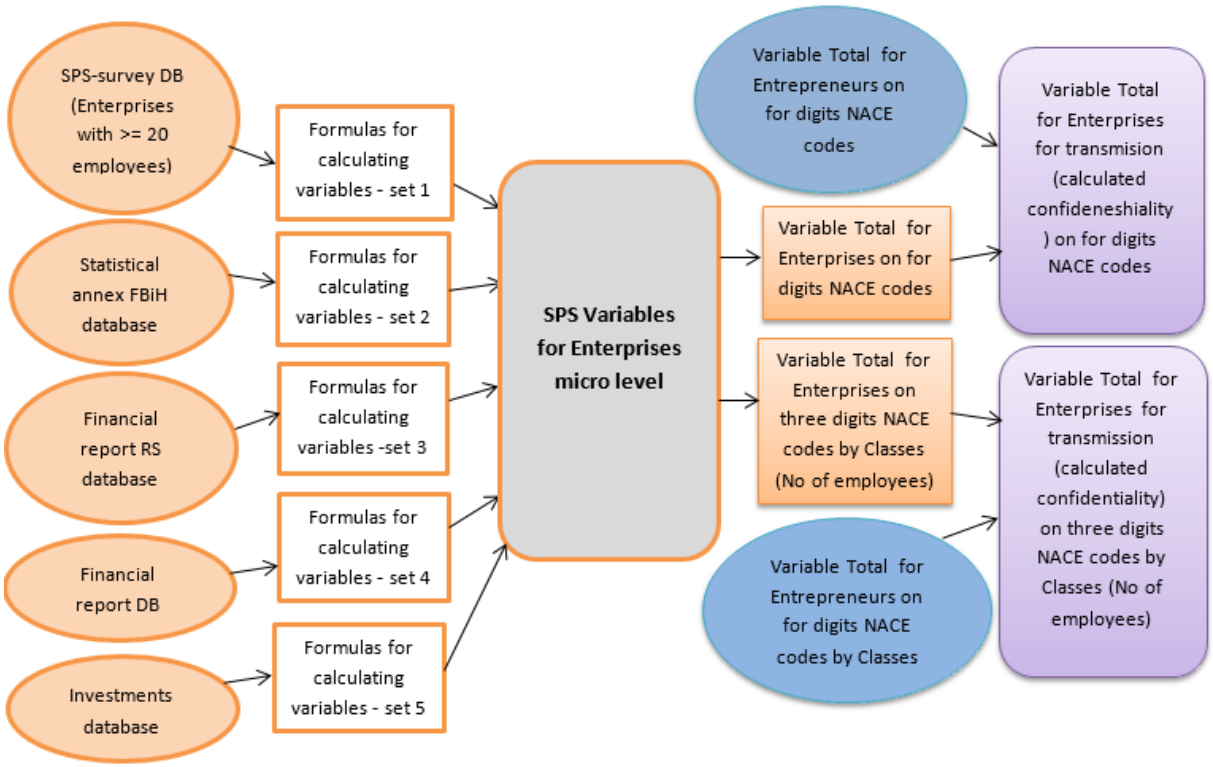
Agency for Statistics of BiH (BHAS) with Brčko District (BD) (a branch of BHAS)

Institute for Statistics of Federation of BiH (FIS)

Institute for Statistics of Republika Srpska (RSIS)

Annual financial reports are submitted to financial agencies, two in entities and one in the state level for BD.

The selected data model is mixed mode. Data collection process after including administrative sources looks like this:



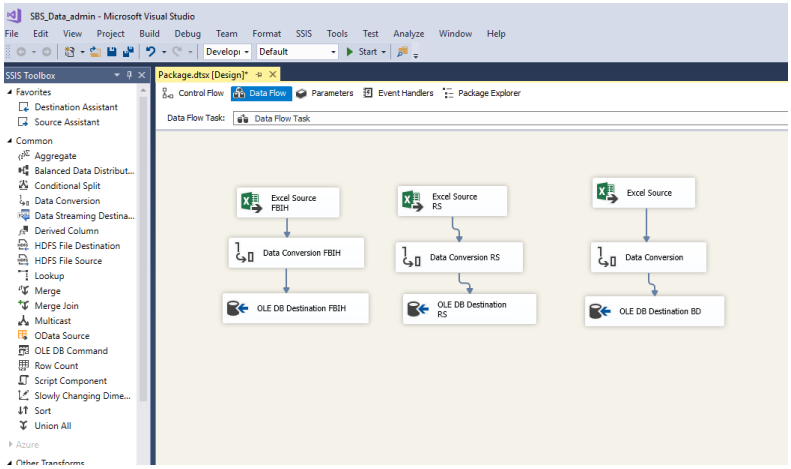
Picture 1. - Business integration model schema

Prior to the twining project, only the SBS survey was used for the enterprises and estimation of SBS variables for entrepreneurs was made. It has now been decided to change method of production, only large companies with 20 employees are working through the existing system (SBS survey), but for other companies with fewer than 20 employees will use administrative sources from authorized entity financial institutions. Now, it is also posibble to validate the quality of a large enterprise survey data - because there are two different sources for the same data, so that in the future, we will go completely to administrative sources if they are to be found as a reliable sources. It would also reduce the burden on the administration of enterprises that now have to fill in more similar forms with the same or similar data.

B. Analysing data quality through comparison data from survey and AD

The analysis of variables derived from the administrative sources Pilot processing for the 2016 reference year was continued, in order to consolidate the differences and inconsistencies related to population of companies covered with SBS 2016 survey. Comparisons were made between variables derived from SBS survey and variables calculated from administrative sources.

At the beginning it was necessary to equalize the sources of data. Our country is structure of two entities and one district, which results with three administrative sources with completely different data formats. It was necessary to unify completely different data sources but with the same data. The decision was made to perform the alignment at micro level, all the data collected in Excel on the level of the company will be in the calculation of the variables. All the data was transferred to the database on MS SQL, although three different scripts were being programmed that translate data into a common SBS variable. Thereafter, at the micro level, we get identical table data table with variables that can be united at the state level.

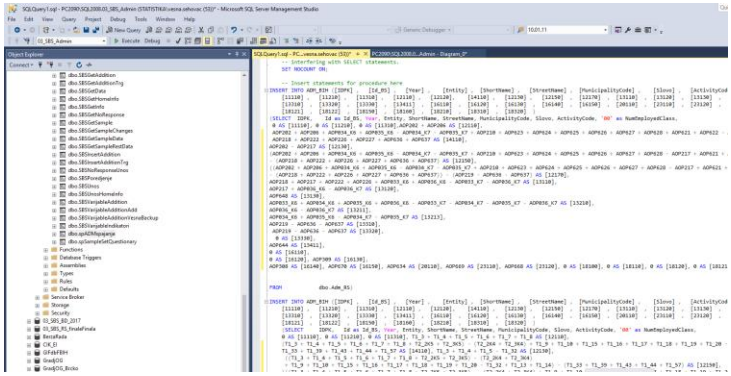


Importing and unifiying entities data in SSIS Project in Visual Studio

Picture 2.. - Transformation administrative data source files in MS SQL data tables

C. Mathematical and logical controls of AD

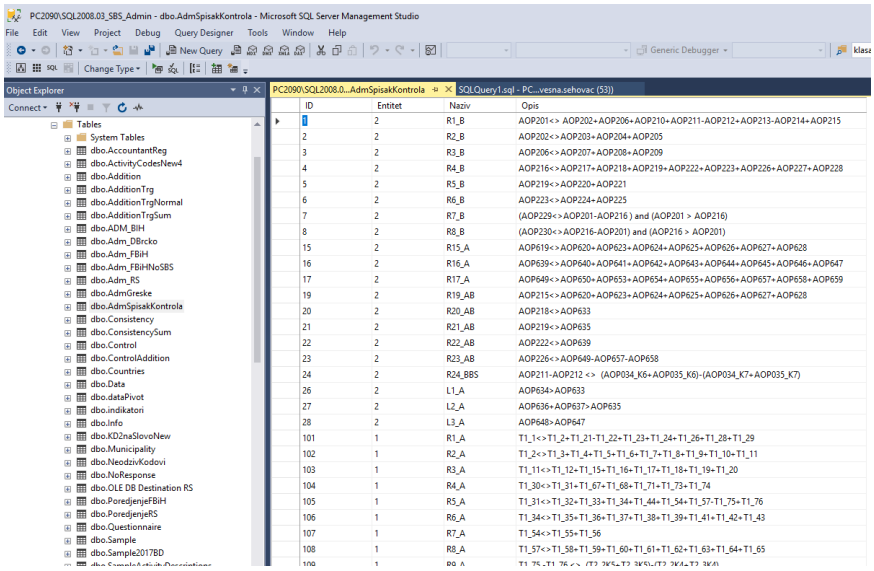
After transforming administrative sources files in MS SQL DB Tables and prior to the unification of the data on variable level (from all of the three entity tables from administrative sources), have to be checked on errors. Checking is performed automatically on the micro level. The result of that procedure determine the quality and later data treatment.



Completing transformations and calculations of the variables programaticaly within MS SQL scripts

Picture 3. - Stored procedure for calculating variables

There is a list of micro level controls that should identify inconsistency between the columns and presented in matematical condition form. Data errors will be checked in the tables of administrative sources of both entities and the district. Program automatically select and identify common mistakes defined in table with definition of error, tagging them and handling them through a table in MS SQL that is user friendly, flexible and editable. The script takes one by one definitions of errors from that table and identifies it in data table if exists, and places it in a table with data errors. That table uniquely identifies micro data row and type of error.



Picture 4. - Table with rules and descriptions

An analysis and comparison of the results obtained by SBS survey with the results obtained from administrative sources at the activity section level and companies size class. Generally, the difference in total is minimal, but at the level of individual activities for particular variables, higher deviations of results have been noted. In particular, the analysis was performed on variables on inventories, e.g. total purchases and gross margin on goods for resale, and their influence on value added.

Additionally, the analysis of reports from mathematical and logical controls of AD has led to a conclusion that most errors are related to the comparison of data from the Income statement and data from the Accounting Annex (Accounting Annex created for statistical needs).

A set of variables derived from Income statement were compared with the variables derived from the same items in the accounting annex, for the companies that were the subject of analysis. Variables derived from Income statement were also compared with those derived from the same items in the Balance Sheet. The main focus was on turnover, inventories and personnel costs. Some inconsistencies have been identified, indicating that the Income statement and Balance sheet data are more reliable and better than the same items in the Annex.

Furthermore, some inconsistencies can be eliminated by editing some of the bad quality items on administrative source for which the error is very clear and unambiguous automatic editing rules can be defined. Set of editing rules, which specify or constrain the admissible values, should be implemented by computer program which should both detect and correct the data. In general, Income statement should be treated as a priority AD source in regards to additional statistical / accounting annex.

Differences between AD and annex with minimal amounts are suggested to be ignored because they do not affect the result and are most likely caused by rounding. Wherever possible, it is necessary to improve AD quality by incorporating controls on the AD forms themselves and by defining uniform and consistent user instructions in the next period. Generally, for the future data analysis and attempting to eliminate inconsistency at unit level, it is necessary to take into account the principle of matching revenues with expenses. Nevertheless, despite some inconsistencies, administrative data can be recommended in SBS production, with focusing on further auditing of major business entities, and enhancing and extending the data editing rules list related to individual administrative items. Furthermore, as a result of the analysis, the first set of data editing rules for the AD quality improvement were created at the mission. Since the definition of editing rules is demanding and complex work, and

assumes a lot of analysis, modelling, simulations and calibrations, for the future development of editing rules, priority should be given to chosen priority items and then step by step expand the list of rules.

D. **Treatments of Errors:**

The data collected by the statistical institute inevitably contain errors. In order to obtain statistical production of sufficient quality, it is important to detect and deal with these errors.

Experience in the field of error correction has led to the assumption that uncommon influential errors, with high impact on estimates, affect only a small number of enterprises. These are the companies that occupy the highest shares in production. They have big differences and deviation from other data and should be treated selectively and in a small number of cases manually. It is also recommended to contact such companies for the identification of faults and their treatment in the best possible way.

Other observations that are not contaminated or contain errors that have little impact on estimates and need to be addressed in the fastest and most effective way to reduce the cost of editing, while retaining the desired level of assessment quality.

Automatic editing usually implies that the data matched with the set of predefined limitations: so called editing or editing rules. The data file is checked for the errors per record. If the record fails one or more editing rules, the method produces a list of fields that can be found so that all the rules are met. We focused on identifying common errors based on logical and mathematical relationships between columns that are defined in mathematical formulas and established safe editing methods where is possible. The goal of automatic editing is to accurately detect and treat errors and missing values in data files in a fully automated way, without human intervention.

An initial set of rules for automatic editing has been established, which will eventually be upgraded. It was found that the order of execution of the editing operation is crucially important, and that if it was improperly executed it could cause distortion of the image of the obtained reports.

The data will be edited at the micro level and the original tables will be kept in history table for security and later analysis and comparison.

Editing Code	**Conditions**	**Editing rules**	**Entity**
Edit traffic for companies in Activities Area L			
EDPROM Description	If AOP623<= AOP215 where Section is L Revenue from rental is less than or equal to other business income	AOP206 = AOP206 add AOP623 Revenue from sales of profits is increased for rental income.	RS
EDAOP215 Description	If AOP623< = AOP215 where Section = L Revenue from rental is less than or equal to other business income	AOP215=AOP215 -AOP623 Reduce other business revenue for lease revenue (continued previous edit.)	RS
EDPROM Description	Section =L If the activity of the company in the Section L	T1_7 = T1_7 + T1_15 Revenues from the sale of domestic market effects are increased for rental income.	FBIH
EDT1_15 Description	Section =L If the activity of the company in the Section L	T1_11 =T1_11 - T1_15 Reduce other business revenue for lease revenue (continued previous edit.)	FBIH
EDT1_15 Description	Section =L If the activity of the company in the Section L Editing of material costs and purchase value	T1_15=0 Revenue income set to zero (continued previous edit.)	FBIH

EDTRNV Description	For all activities other than G, if any AOP202=0 and AOP217≠ and AOP036_K6=0 Revenue from the sale of goods is 0, the purchase value of the goods sold is different from 0 and the stock of merchandise goods at the end of the year are 0.	AOP218= AOP218 + AOP217 Increase the cost of raw materials and materials for the purchase value of the goods sold.	 RS
EDITAOP217 Description	For all activities other than G, if any AOP202=0 and AOP217≠ and AOP036_K6=0 Revenue from the sale of goods is 0, the purchase value of the goods sold is different from 0 and the stock of merchandise goods at the end of the year are 0.	AOP217=0 The purchase value of the goods sold set to zero (continued previous edit.)	 RS
EDTRNV Description	For all activities other than G, if any T1_3=0, T1_4=0, T1_5=0, T1_32≠ 0 and T2_3K5=0 Revenue from the sale of goods (related legal entities on the domestic and foreign markets) is 0, the purchase value of the goods sold is different from 0 and the stock of merchandise goods at the end of the year is 0.	T1_33 = T1_33 + T1_32 Increase the cost of raw materials and materials for the purchase value of the goods sold.	 FBIH
EDT1_32 Description	For all activities other than G, if any T1_3=0, T1_4=0, T1_5=0, T1_32≠ 0 and T2_3K5=0 Revenue from the sale of goods (related legal entities on the domestic and foreign markets) is 0, the purchase value of the goods sold is different from 0 and the stock of merchandise goods at the end of the year is 0.	T1_32=0 The purchase value of the goods sold set to zero (continued previous edit.)	 FBIH

Table 1. - Editing rules

E. Estimation of missing data for entrepreneurs

Regarding improvement of the new production method of SBS, we updated new SBS production application and integration of administrative data. Data flow was analysed in details, as well as key cooperation with SBR. Also, experience with treatment of the part of the population without main administrative data source was presented from Croatia. BHAS already has certain experience with unit non response estimate and assess SBS variables of entrepreneurs. The assessment is based on regression method using available data from SBR for entrepreneurs.

But, in this way estimation is performed only on aggregate level. Therefore, it is recommended to move from estimation on aggregated level to estimation on unit level, based on ratio estimator – the share of a variable in turnover of the same stratum population (for which variables are calculated from available sources). Since there is still no turnover information in SBR for some units, the 3-step evaluation model is recommended:

Step 1 – estimation of turnover where it's missing:
Calculation of coefficient A = average „turnover" per employee in NACE and size class
- use for estimation of turnover at unit level
Estimation: unit turnover = unit number of employees * coefficient A

Step 2 – estimation of other financial variables:
Calculation of coefficient B1-n = share of every variable in turnover in NACE and size class
-use for estimation of financial variables on unit level
Estimation: unit variable = unit turnover * coefficient B1-n

Step 3 – estimation of variables of personnel costs
Calculation of coefficient C1-n = average personnel costs per employee in NACE and size class
-use for estimation of variables Personnel costs, Wages and salaries and Social security cost
Estimation: unit personnel costs = unit number of employees * coefficient C1-n

The following procedure is recommended to apply:

Considering that Xs represents „total turnover" for specific stratum, and representing each variable in stratum as Ys, a ratio is obtained by: $\sum_{i=n}^{n}$ (Ys / Xs).

Turnover existing in SBR-a is represented by Xe.

Each missing variable is represented by Ye and the estimation on unit level is done by applying the following formula:

$$Y^e = X^e \times \sum_{i=n}^{n} \frac{Y^s}{X^s}$$

Where:

e = individual company variable
S = stratum variable
i = 1, ..., n = represents all companies with observed variables and that belong to the same stratum for which we want to estimate the variables.

Stratum is defined taking into account the activity classification (NACE Rev 2., 4-digit level).

In the case of non-existing minimum 3 units in some stratum, an interactive process is taking place based on higher NACE level.

V. CONCLUSION

The aim of this paper was to present current status of SBS production in BiH. It could be summarized as survey based SBS with harmonized forms by FIS, RSIS and BD. The survey forms contain identical data fields but differences are found in accounting terms (chart of accounts) – depending on the entity receiving the form. Statistical production is performed by all three institutions using one in-house developed application integrating both soft and hard controls. The questionnaires with outliers are checked by methodologists.

Eurostat data tables are transmitted for Annexes I-IV on a regular basis since 2012 and the procedure is documented and followed on schedule. Entrepreneurs SBS data was added in 2014 as reference year. Annex VIII and Inwards Foreign Affiliates (iFATS) data are also submitted to Eurostat on a regular basis. At the end of the process data is merged, controlled, validated, edited, and imputed., the calculation of confidentiality, marking, conversion of data into Eurostat format and sending is carried out. We produced result in this manner (survey without administrative sources) until last year.

Based on the pilot phase results, the entities should evaluate how much and which parts of SBS should be moved from survey to AD sources. Considering the majority of EU countries still use a mix of the two methods, it is probable that BiH will fall into this mixed-mode category.

For Editing and imputing we chosed selected and automated way using MS SQL stored procedures, as well as manual editing before and after, for errors with large impact (and with communication with enterprises).

This research is currently at an annual level, and we hope that transition to administrative sources and automatic editing and imputation of data will lead to better quality data, reduction of burden on the both side (NI as well as Enterprises), lower costs, faster to handle, bigger efficiency and better performance. This is the way how we move forward with SBS production.

Keywords: Administrative sources (AD); Business Statistics; Quality; Variables; Reduction of Burden, Enterprises, Entrepreneursđ

Abbreviations:
AD - administrative source
BHAS - Agency for Statistics of Bosnia and Herzegovina
FIA - Financial Information Agency of FBiH
APIF - Agency for Intermediary, IT and financial services of RS

Entities:
FBH - Federation of BiH
RS - Republic of Srpska
DB - District Brčko