

TWINNING CONTRACT

BA 15 IPA SR 01 17

Support to the reform of the statistics system in Bosnia and Herzegovina



MISSION REPORT

**Activity 2.5.5:
Further analysis on regular survey on tourism statistics**

**Component 2: Business Statistics
Sub-component 2.5: Tourism Statistics**

**Mission carried out by
Else Marie Rasmussen, Statistics Denmark
Ivana Brozović, Croatian Bureau of Statistics**

21-24 May 2019

Version: Final

Expert contact information

Else Marie Rasmussen

Statistics Denmark

Copenhagen, Denmark

Tel: +45 39173362

Email: emr@dst.dk

Ivana Brozović

Croatian Bureau of Statistics

Zagreb, Croatia

Tel: +358 014806272

Email: brozovici@dzs.hr

Table of contents

Executive Summary	4
1. General comments	4
2. Assessment and results	5
3. Conclusions and recommendations	10
Annex 1. Terms of Reference	12

List of Abbreviations

BHAS	Agency for Statistics of Bosnia and Herzegovina
BiH	Bosnia and Herzegovina
CBBH	Central Bank of Bosnia and Herzegovina
EC	European Commission
EU	European Union
FBiH	Federation of Bosnia and Herzegovina
FIS	Institute for Statistics of Federation of Bosnia and Herzegovina
MS	EU Member State
RSIS	Institute for Statistics of Republika Srpska
RTA	Resident Twinning Adviser
ToR	Terms of Reference

Executive Summary

The purpose of this mission was to analyze the overall data collection process for the second data collection wave and analyze the data quality. The data presented had a low percentage of missing values, and the initial conclusion on the raw data is that the data quality is good.

Furthermore, the purpose was to discuss data editing as well as agree upon a suitable grossing-up procedure. Finally, the two IT applications for data entry were presented.

It is recommended by the MS experts that the grossing-up procedure chosen is similar to the procedure already used in other households surveys such as LFS and HBS. In Croatia and Denmark, the grossing-up procedures are for CATI and web surveys and the reporting unit is the individual, not the household.

Furthermore, the MS experts recommend that the merged data is analyzed in more detail and the suggested methods of outlier detection and imputation are applied. The imputation should only be done on the merged data.

Finally, it is recommended that the modalities of “Don’t know” and “Don’t want to answer” are removed from the questions on type of accommodation and the reason not to participate in tourism, but still kept in the questions on expenditure.

1. General comments

This mission report was prepared within the EU Twinning Project “Support to the reform of the statistics system in Bosnia and Herzegovina”. It was the fifth mission to be devoted to 2.5: *Tourism Statistics* within Component 2: *Business Statistics* of the project.

The purposes of the mission were:

- Follow up from the previous mission

Prepared by the MS experts

- Presentation on grossing-up and aggregation procedures

Prepared by the BC experts

- Revision of the questionnaire and the interview guide
 - Initial analysis of the data quality (response rate, standard error, outliers and similar) in order to choose the suitable imputation method
 - Presentation on IT application
-
- Analysis of overall data collection process
 - Presentation of work done on data editing and processing (non-response treatment, outlier checking, etc.)
 - Agreement on weighting and grossing up procedures
 - Presentation on methods and procedures for calculation of variables and preparation of output tables

The consultants would like to express their thanks to all officials and individuals met for the kind support and valuable information which they received during the stay in Bosnia-Herzegovina and which highly facilitated the work of the consultants.

This views and observations stated in this report are those of the consultants and do not necessarily correspond to the views of EU, BHAS, FIS, RSIS, CBBH, Statistics Denmark, INSEE, Statistics Finland and Croatian Bureau of Statistics.

2. Assessment and results

The main focus of this mission was to analyze the collected data for the second data collection wave, to discuss and decide on data editing and validation methods and furthermore agree upon a suitable grossing-up and aggregation procedure.

During the fifth mission, the following was achieved:

- Results on data collection process for the second data collection wave is analyzed
- Data editing and processing (non-response treatment, outlier checking, etc.) presented
- Weighing and grossing-up procedure agreed
- Methods and procedures for calculation of variables agreed
- Input provided to the ToR of next activity

Revision of questionnaire for the second data collection wave

All suggestions mentioned in the report from the previous mission were implemented in the questionnaire for the second wave of data collection with the exception of the suggestion to have separate questions on private and business trips. This suggestion may be implemented for the next reference year if deemed feasible.

The questionnaire was revised in time for the second data collection wave, and there is an overall satisfaction with the final questionnaire.

Data collection process

Since the previous mission, the second data collection wave covering the reference period of November-December 2018 has taken place. The first data collection wave covering the reference period of January-October 2018 was in November/December, and the whole reference year 2018 is now covered. For future surveys, it is recommended by the experts to have a quarterly or biannual data collection and thus four or two data collection waves for a full reference year. This however depends on the funds allocated. The most important is to have the same reference period length due to comparability issues and memory bias.

The initial results and experiences from the second data collection wave and from the data entry process are presented below.

FIS experience

The sample size for the second data collection wave was 1700 households, and the response rate was 84.4 per cent. The overall non-response rate was 15.6 per cent and the refusal rate was 5.3 per cent. There was a very high completion rate.

The data has not yet been fully analyzed, but an initial finding is an issue with the modalities “Don’t know” or “Don’t want to answer” in Question 9 regarding reasons not to participate in personal trips with at least one overnight stay. A high proportion of respondents selected these two options. For a future reference period, it is possible to remove these two modalities, as they are not to be disseminated to Eurostat. Furthermore, the MS experts recommend in general to have as few of these open modalities as possible. They tend to be used too much and do not encourage respondents to answer correctly.

Another issue during the data collection was the question on the number of children, as some respondents would not provide this. They regarded it a very sensitive question. It is recommended by the MS experts to obtain this information through registers or other administrative sources, if it is possible.

FIS is very satisfied with the interviewers and the process thus far despite budgetary constraints and limited time. It is difficult for them to plan the budget for the next data collection in advance.

RSIS experience

RSIS had a sample of 1700 households. They obtained 1368 responses and thus a response rate of 80.5 per cent and a 19.5 per cent non-response rate. There were no problems during the data collection and there was an overall satisfaction with the process. RSIS had difficulties hiring good interviewers. It was not possible to get the same interviewers for the first and the second data collection wave. This issue seems to be general regarding the data collection in all three entities.

Brcko experience

Brcko had a sample of 650 households. There were 453 responses and 197 non-responses, which translates into a response rate of 69.7 per cent. 80 per cent of the questionnaires were fully completed and 20 per cent partially completed. The non-response rate is lower (30 per cent) than in the first data collection wave (38 per cent).

They were not so satisfied with the interviewers in the second data collection wave compared with the first data collection wave, and they had difficulty finding good interviewers. There is a need to establish a permanent unit of interviewers. They verified data by calling the respondents and checking anomalies. Sometimes, there was a big difference in the response rate depending on the interviewer. Brcko made random controls in order to check if the interviewers did their job properly. It was not the same interviewers in the first and second data collection wave, so there is a problem with a lack of consistency, as it would be preferred to have the same interviewers.

In general, the second data collection wave went well. There was a significant time pressure, but everything was done in time. Some basic logical errors were removed during the data entry, and the data has been cleaned.

Data entry and editing

Data collected from the paper questionnaires for Brcko and FIS was entered into a Blaise IT application while data from RSIS was entered into an IST IT application. The data was subsequently merged into one database in Blaise.

The data editing process so far:

1. Initial cleaning process where the controllers check the interviewers.
2. Data entry into IT application. The logical controls were the same for all the entities. The logical controls functioned as warnings (soft validation).
3. After data entry logical controls were applied to the micro data.
4. Error listings were generated in Blaise.
5. Checking the error manually by looking at the specific questionnaire
6. Need for agreement on common validation rules so the error checks can be corrected automatically in a statistical programme as R or SAS (look at rules from Eurostat)

A list on validation controls was made in the previous mission. This list was used to generate an error list. In the table on the next page some common errors are listed alongside data editing suggestions from the MS experts.

Table 1. Most common data errors

Error	Suggestion from MS experts
<ul style="list-style-type: none"> • YES to a trip, but no expenditure given. • YES to rented accommodation, but no accommodation expenditure given. • YES to transport, but no transport expenditure given. • YES to package trip, but no package trip expenditure. 	The simple solution is to mark the expenditure as imputed and use the mean imputation method: Using averages grouped by destination (domestic/foreign), purpose (private/business), duration or grouped by background variables such as age, gender, education. The trip characteristics can be combined with the background variables.
The total expenditure is given, but expenditure items are missing.	<ul style="list-style-type: none"> • If the total expenditure is known, but some expenditure items are unknown or missing, then you can estimate the items by applying the average share by similar respondents or/and similar trips. • If you miss just the expenditure for one item, impute for the difference between the total cost and the sum of the expenditure for items • If missing data on expenditure for more items - the difference between the total expenditure and the sum of the reported expenditure is calculated, and this difference is allocated to items that are proportional to the number of respondents who reported all costs
Missing data on resident or non-resident.	If they filled out the questionnaire, they should be marked as resident.
Trips outside of reference period.	If a trip has ended outside the reference period, it should be filtered out.
YES to private trips, but no destination given.	If no destination is given, you can choose the most probable answer by defining some criteria such as “If Transport=By plane” AND duration > 6 nights, then destination=MOST visited foreign destination with more than > 6 nights and transport by air. In other words, the most frequent answer given grouped by relevant variables = Mode imputation.
Expenditure in Q24 and Q25 differs.	If the expenditure differs, choose the expenditure estimate in Q25 and estimate the expenditure items by applying the shares for the expenditure items for similar trips or respondents.
Number of trips is given, but not all of the trips are listed with characteristics.	It is possible to use the listed trip characteristics as a proxy for the missing trip characteristics. This is however not recommended in a large scale.
YES to same-day visit, but no purpose of the same-day visit listed.	Use mode imputation or logical rule such as: If trip is done alone, it could be business, if done with household members, then it is private.
YES to trip, but no overnights stays filled in.	Mode imputation grouped by destination and purpose.

FIS experienced an issue with the modalities “Don’t know” and “Don’t want to answer” in the question regarding reasons not to participate in tourism. There was a high share of these two

answers in the data, even though the interviewers were instructed not to read these two modalities aloud. The question already has a modality called “Other reasons”, and it is not mandatory to send data on “Don’t know” and “Don’t want to answer” to Eurostat. Thus, the MS experts recommend to remove these two modalities in the questionnaire for the next reference year. Furthermore, we recommend to mark the modalities as missing values in the already collected data and impute the mode value.

Outlier detection

According to Eurostat, outliers should be detected by using the mean and standard deviation method:

$$\text{If } \left| \frac{x - m}{s} \right| > 3, \text{ then the records are detected as outliers}$$

With x = the value, m = the average and s = the standard error.

This method should be applied on:

- Q19: Expenditure on transport (done separately for transport by air, waterway and other)
- Q20/Q4: Expenditure on accommodation divided by number of nights spent (done separately for rented and non-rented accommodation)
- Q21: Expenditure during the trip on food and drinks in cafés and restaurants
- Q22: Other expenditure during the trip
- Q23: Expenditure on durables and valuable goods

All the records with expenditure equal to zero are excluded from the detection. This outlier detection should only be used as a warning tool. When the outliers have been detected, there should be a manual plausibility check. Outliers are not necessarily wrong. Outlier detection is done prior to imputation. Some suggest a minimum of 5 observations in each cell in the micro data. It is recommended to use the same principles for a minimum threshold as in similar household surveys such LFS and HBS. The minimum threshold recommended by Eurostat for disseminating aggregate tables is 20 observations. If it is 21-49, then the estimate should be flagged.

Unbundling of package trip expenditure

There was a short discussion on how to treat unbundling of package trip expenditure. The unbundled expenditure is relevant for the Balance of Payment and TSA purposes. The package trip expenditure can be unbundled by using auxiliary information on the expenditure items of a similar non-package trip to the same destination and/or with the same duration.

Grossing-up method

There was also a presentation on the Danish and the Croatian grossing-up procedures followed by a discussion on the possibilities. The Danish and Croatian grossing-up procedures are not directly applicable to this survey, and it is therefore recommended to use the same grossing-up procedure as for the household surveys LFS and HBS. It was suggested to invite sampling experts from the MS countries, if deemed necessary.

Presentation on IT application

The two IT applications for data entry were demonstrated by BC experts.

IST application

The IST application is used for data collected in RSIS. Soft validation is used in the application, and it gives warnings. Logical controls for first data collection wave have been implemented, but logical controls for second data collection wave are still missing. Filters are not integrated. Data is entered manually from the filled paper questionnaires. No scanning solutions are available. The application records every time data is edited, but every time you enter the system, it is registered as an editing session, and the record is therefore not valid.

IST works with SQL, and it is possible to develop enquiries. You can generate Excel files with data on the operators that have entered the data. IST can also generate tables with aggregate numbers, and more controls can be implemented in the system. Even though there are missing answers you can still proceed with the data entry. This application is opposite Blaise where missing answers stops the data entry. The application was made for PAPI. It will look different if it is for CAPI.

Blaise application

The Blaise application is used for data collected in Brcko and FIS. Furthermore, it has the merged data from all three entities. Validation controls and filters are integrated, so it is applicable with CAPI. The application generates output files for further analysis and a list of errors.

It is recommended by the MS experts to use the same IT application for data entry instead of two applications. This will minimize cost and time consumption and increase data comparability between the different entities.

Data analysis

There was a presentation on the merged data, which consists of four data sets in SPSS. One for same-day visits, one for multiple trips, one for socio-demographic variables and one for participation. It needs to be clarified whether this analysis was for the whole reference year or only the second data collection wave. If it is the latter, then the first and second data collection wave should be merged before proceeding with the data analysis.

Some of the variables were discussed in more detail:

- The question on type of destination (Q7) had a higher percentage of missing than the previous questions, but this is to be expected, as this question is rather difficult.
- There were very few missing values (3) in the question regarding transport (Q8). It was recommended to produce a table on transport by destination for imputation purposes. It is also recommended to change the wording “automobile” to “rented motor vehicle” and “private motor vehicle”.
- The questions on tour operator for transport (Q9) and online booking of transport (Q10) had a missing share of 1.1 per cent.
- The question on accommodation (Q11) also had a very low number of missing values (5).
- The question on tour operator for accommodation (Q12) had a bit higher share of missing. It is only relevant for rented accommodation. In the questionnaire, it says “transport”, and it should to be changed to “accommodation”.
- It is recommended to make a table on Q15 regarding the number of persons the expenditure is for.

- The total expenditure of package trip (Q17) had a higher missing value percentage of 30.4 percent, but this is to be expected regarding the more sensitive expenditure questions.
- It is recommended to make tables on the specific expenditure items also.
- The total cost (Q24) does not match the expenditure items summed up in 27.5 per cent of the cases.
- Same-day visits: 4846 visits. There is a very low missing share of max. 10 per cent, which is very good. The missing share on same-day visit expenditure was 16.7 per cent, and this is still very low considering the type of question.
- The income question had a missing share of 14.7 per cent. This question is optional according to the EU regulation.

3. Conclusions and recommendations

The initial data analysis shows that the data quality is good with a very low percentage of missing values. It is not clear whether the data from the first and the second data collection wave has been merged prior to the analysis. It should be prioritized to merge the data from the two data collection waves before further analysis. There is a need for a more extensive data analysis.

There was an issue with the modalities “Don’t know” and “Don’t want to answer”, because a large number of respondents used these modalities instead of answering the question on reasons not to participate in tourism. It is therefore recommended to remove these modalities from the questions on type of accommodation and reasons not to participate in tourism and only keep them in the expenditure section due to the higher sensitivity of these questions.

There was a presentation and discussion on outlier detection, finding missing values and subsequent imputation, and it was suggested by the MS experts to apply suitable and simple methods such as mean and mode imputation. This should be done before the next mission. The imputation should only be done on the merged data set.

The grossing-up procedures were presented, and it was clear, that the Danish and Croatian procedures are not directly applicable to this survey due to different data collection methods and reporting units. It is therefore recommended to use an already existing and standardized grossing-up procedure for a household survey such as LFS or HBS.

The sample size for the second data collection wave was reduced compared to the first data collection wave to approx. 4050 households. Due to the high data quality and response rate, it is feasible to reduce the sample size even further in future surveys.

The two IT applications for data entry were demonstrated by BC experts. The IST application is used by RSIS, while the Blaise application is used by Brcko and FIS. Furthermore, the Blaise application is also used for the merged data set. For future surveys, it is recommended by the MS experts to use the same IT application in all three entities, if possible. This will minimize cost and time consumption and increase data comparability between the different entities.

What to do before the next mission for the BC Counterpart

Action	Deadline	Responsible person
Clean data from both data collection waves in entities	June 2019	BHAS, RSIS and FIS
Clean and merge data from the two data collection waves into one	June 2019	BHAS
Data quality analysis on merged data (response rate, missing values, anomalies)	September 2019	BHAS, RSIS and FIS
Outlier detection followed by imputation on merged data	September 2019	BHAS, RSIS and FIS
Grossing-up procedure	October 2019	BHAS, RSIS and FIS
Presentation on macro validation	September/October 2019	MS experts
Presentation on analysis of aggregate and micro data	September/October 2019	MS experts
Aggregate data	October 2019	BHAS, RSIS and FIS

Annex 1. Terms of Reference**Terms of Reference****EU Twinning Project BA 15 IPA ST 01 17****Component 2: Business Statistics****Sub-component 2.5: Tourism Statistics****21-24 May 2019****Hosting institution: BHAS, Zelenih beretki 26, Sarajevo****Activity 2.5.5: Further analysis on regular survey on tourism statistics****1. Mandatory result and benchmarks for the component**

Mandatory result:

- New indicators on demand-side tourism statistics, in accordance with EU Regulation 692/2011 (Annex II – National tourism) produced and made available to users by 8th project quarter

Benchmarks:

- Plan for development of demand-side tourism statistics produced by 2nd project quarter
- Questionnaire for a regular survey prepared by 2nd project quarter
- Criteria for an IT application defined by 5th project quarter
- First results of survey analyzed by 6th project quarter
- Indicators on demand-side tourism statistics compiled by 7th project quarter
- Indicators on demand-side tourism statistics made available to users by 8th project quarter
- Methodological document on demand-side tourism statistics developed by 8th project quarter
- Quality report for tourism statistics developed by 8th project quarter

2. Purpose of the activity

- Follow up from the previous mission
- Prepared by the MS experts
 - Presentation on grossing-up and aggregation procedures
- Prepared by the BC experts
 - Revision of the questionnaire and the interview guide
 - Initial analysis of the data quality (response rate, standard error, outliers and similar) in order to choose the suitable imputation method
 - Presentation on IT application

- Analyses of overall data collection process
- Presentation of work done on data editing and processing (non-response treatment, outlier checking, etc.)
- Agreement on weighting and grossing up procedures
- Presentation on methods and procedures for calculation of variables and preparation of output tables

3. Expected output of the activity

- Results on data collection process in 2018 analysed
- Data editing and processing (non-response treatment, outlier checking, etc.) presented
- Weighting and grossing up procedures agreed
- Methods and procedures for calculation of variables agreed
- Input provided to the ToR of next activity

4. Participants

Agency for Statistics of BiH (BHAS)

Alen Bajramović

Azra Bander Demirović

Institute for Statistics of Federation of BiH (FIS)

Dženana Vreto

Edina Mehidić

Nikola Radović

Institute for Statistics of Republika Srpska (RSIS)

Jelena Glamočika

Sladana Nikić

MS Experts

Else Marie Rasmussen, Statistics Denmark

Ivana Brozović, Croatian Bureau of Statistics

Twinning Project Administration

Katja Møller Hjeltvang, RTA

Đemka Šahinpašić, RTA Assistant

Adisa Okerić-Zaid, Interpreter