WPJ WebEx Meeting

Data and time: 1 March 2019, 12:00 – 13:20 CET

Participants:

PL	Marek Cierpiał-Wolan, Dorota Jasiukiewicz, Łukasz Zadorożny, Piotr Szlachta, Sebastian Wójcik, Teresa Matuła
NL	absent
IT	Mascia di Torrice + 2 people
SK	Boris Frankovič, Martina Naňáková
PT	Rui Alves
BG	Galya Stateva, Kostadin Georgiev
EL	Christina Pierrakou, Eleni Bisioti, Maria Laftsidou
DE	absent

1. Opening and agenda

Marek Cierpiał-Wolan welcomed all partners at the third WebEx meeting. He briefly presented the agenda of the meeting, then gave the information received during the WebEx meeting of the leaders of 27-02-2019. He described the literature template received from the WPK work package and information on the use of the WIKI platform. Another topic raised by Marek was the date and manner of preparation of the first deliverable "ESSnet Methods for web scraping, data processing and analyses".

2. Methodology of web scraping - summary

Piotr Szlachta started by thanking partners for sending all the comments regarding the script for the hotels.com portal and providing data for analysis. Next he started a short presentation answering the most frequently asked questions received by email regarding the web scraping process. The main issues concerned the preparation of own solutions for the web scraping process, changes in the code when a partner does not have right staff, common structure for collected data. The

next part of the presentation was a description of new functionalities for the script web scraping data of the hotels.com portal and presentation of the web scraping process for the booking.com portal.

At the end of the presentation, Piotr briefly described the next actions that will be taken in task 1a such as providing instructions on how to use both scripts and adaptation to own needs. He also discussed the preparation of a common template for checking the completeness and quality of the collected data. He also mentioned a proposal of Hesse colleagues about additional technical WebEx meeting in March.

After Piotr's presentation, the leader of the package (PL) asked partners for remarks. Eleni Bisioti (EL) asked about localization of data gathered by scripts and made a suggestion about setting up the language to English. She also mentioned that they have issues with getting emails with codes due to email restrictions in their organization. Piotr mentioned that all portals are using local setting for languages and currencies but will check possibility of changing scripts to be more universal. He will also look for a solution for sending codes to partners. During the discussion, Galya Stateva (BG) commented about the script received by Bulgaria so far and asked about the preparation of scripts for another portals. Piotr said all new scripts will be parametrized and easily adopted to needs of each partner. The next speaker was Boris Frankovič (SK) who asked how many scripts were sent and when new ones will be available. Piotr assured that new scripts would be sent out the following week and asked all partners to share they own scripts. The last question concerning web scraping was asked by Mascia di Torrice (IT). The issue with portal developers potentially blocking web scraping scripts was raised. Piotr said that all script prepared by Poland are using best practices when it comes to making the whole process as not invasive as possible.

3. Data sources catalogue – summary

Dorota (PL) summarized the joint work on the data sources catalogue.

- All countries delivered information on the existing and potential data sources that eventually could be useful for the purposes of WPJ according to the agreed template of detailed table. Polish team could have completed tables for each partner country.
- These country tables were converted into some specific tables. The table with information about each type of source, taking into account the frequency and basic information provided was developed.

- Provided data sources have been pre-grouped into thematic areas according to data they include.
- All data sources have been assigned as useful for estimating the demand, supply side of tourism or for both sides, some sources are not assigned to any of these sides (it was difficult to determine this at the stage of source inventory but it will probably be possible during the implementation of the task under point 1c).

These tables were consulted with all countries after previous meeting. Some countries completed the missing information, some added information about new identified sources.

In general, the proposals of the developed tables were accepted. Mascia from Italy pointed out that maybe WATER CONSUMPTION and WASTE PRODUCTION should be considered useful both for estimation of demand and supply side statistics. Dorota explained that the reason of allocation of these data sources to the supply side was that the basic information received on the basis of these sources will concern the number of people/tourist consuming water/producing waste but of course these data will be then used for demand side of tourism statistics estimations. The same situation will be in the case of sources on the electricity consumption. All these data sources were moved to the group of sources useful for estimating both sides of tourism statistics.

Dorota mentioned also about the remarks of the colleague from Portugal. He stated that all information, including variable names and data sources, should be presented in English or should have bi-lingual version (English and native language). Another remark of Portugal concerned the column “short/middle/long term source”, which could be more specific (for example: “Data available from [year]”). Future availability of sources should also be taken into consideration and information such as data ownership and publishing policy could be added.

These comments can still be taken into account, especially because Portugal is responsible for the next task (1c), related to the analysis of data sources.

The Excel file with all tables should now finally be examined by the colleagues from Bulgaria.

Because this task (inventory and general characteristics of data sources) is almost finished, it is time to move on to the next task – task 1c.

Next part was presented by Łukasz Zadorożny (PL). During his presentation Łukasz discussed task 1c and presented the characteristics of the collected data sources.

As he said, the available data sources were collected and cataloged, however, the structure of databases and their content are still unknown.
In task 1c, each country that sent its data sources had to specify the contents of databases, including the type of variables they contain.
Łukasz gave an example of Health insurance data that is in data sources catalog. On the basis of the content, there is a need to identify the variables taking into consideration their types: either numeric or string.
Moreover, he added that only when the sources would be described in the way he presented, it would be possible to go to the next stages of the project such as variable possible mappings. In conclusion, every source will need to be described in this way and it is essential for each country to describe its own variables.

4. Methods of combining data – overview

Sebastian Wójcik (PL), head of Mathematical Statistics Division, presented a proposition of flash estimates of occupancy of tourist accommodation establishments statistics. It involves the following steps:

- merging outputs tables from Survey On Tourist Accommodation Establishments (10 or more bed places) into one database of time series,
- selecting crucial variables for flash estimates,
- merging outputs of web scraping,
- cleansing and formatting data,
- trimming outliers,
- preliminary analysis,
- generating set of statistics (converting daily data into monthly data),
- time-series model for crucial variables from Survey On Tourist Accommodation Establishments.

R-script was prepared to run few steps of the aforementioned procedure. He presented some outcomes of preliminary analysis of web scraping data:

- Distribution of prices of tourist accommodation establishments turned out to be right-skewed and bimodal. Such properties may be observed month-by-month.
- Mean and median prices, although they belong to two different classes of statistics, preserved the same pattern of changes. Correlation coefficient between them amounted to 0.97.

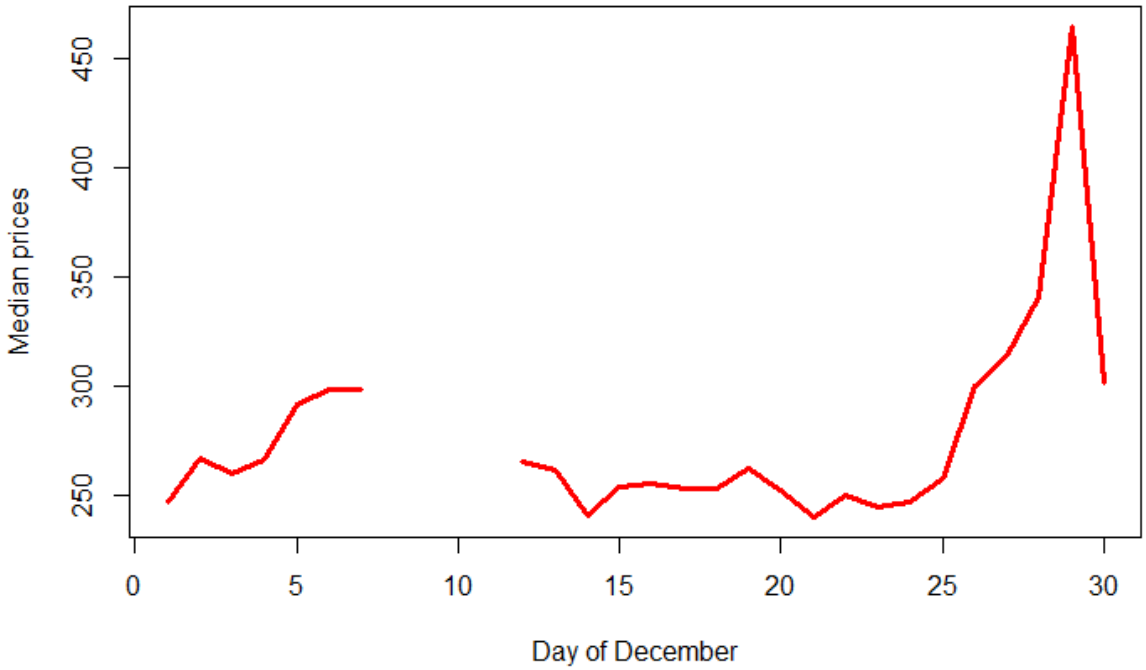
A preliminary analysis was carried out on Survey On Tourist Accommodation Establishments data. Sebastian Wójcik (PL) presented a correlogram of crucial variables. It turned out that variables forms groups with respect to resident and non-residents. Within each group all of the variables are positively, highly and significantly correlated.

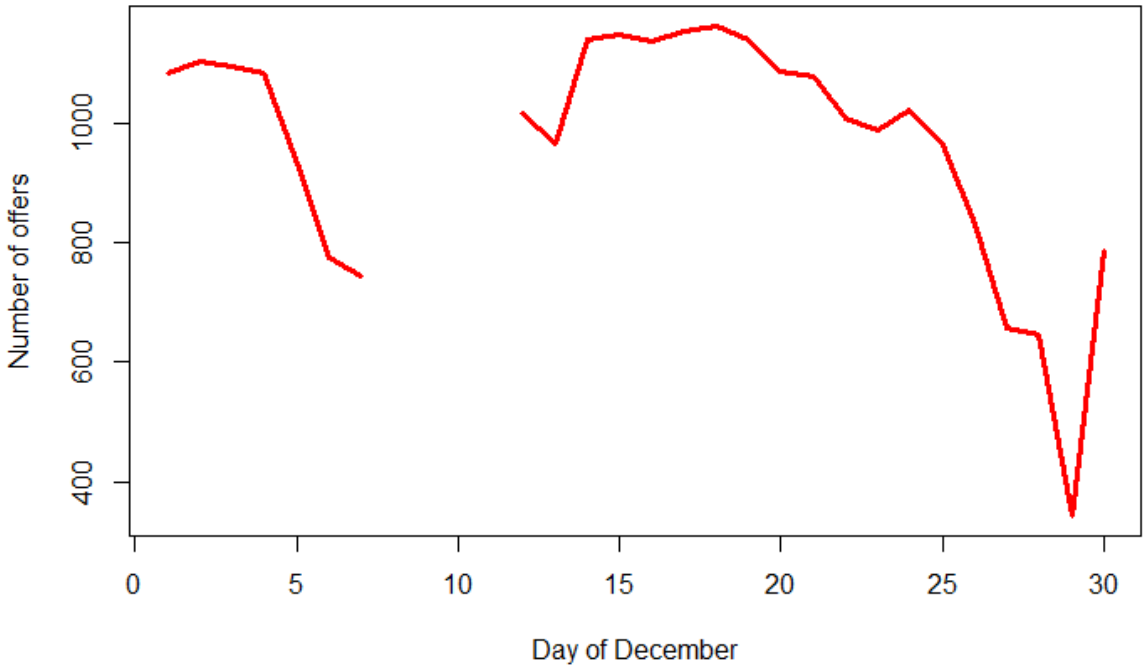
Maria Laftsidou (EL) suggested we should include seasonality in our analysis. Sebastian Wójcik (PL) answered that last correlogram was generated from whole 2018 data – not only from December and January data as in a case of web scraping analysis.

Christina Pierrakou (EL) asked what would be a key to link data from official statistics and web scraping. Sebastian Wójcik (PL) answered that in the case of our flash estimates proposition the last stage of estimation is based on time-series model. Thus, we aim to generate several monthly statistics from web scraping data to use it as explanatory variables in the time-series model. If there is any other idea of a closer data linkage and more direct approach we will examine several methods of flash estimates.

Rui Alves (PT) wondered if he could receive R-script and provide it to acquaintances from his tourism department and check what results they would obtain with that script. Sebastian Wójcik (PL) affirmed that R-script would be prepared soon, maybe the following week and then sent to all of the WPJ participants.

Mascia di Torrice (IT) wondered if we can scrape website pertaining to demand side of tourism industry. Sebastian Wójcik (PL) stated that at this moment we have not found relevant web source. Nevertheless, we hope that the demand side of tourism can be observed in indirect way. According to law of demand we should observe a decreasing number of offers of tourist accommodations and increasing level of prices whenever the demand increases. Some signs that the idea could be useful were discovered by computing, e.g. median prices in December. It turned out that prices shifted a lot due to increased demand just before New Year’s Eve, while the number of offers decreased (plots generated after WebEx meeting).





Mascia di Torrice (IT) asked about the key for combining web scraping statistics and survey statistics. For flash estimates we propose to use time-series model in the final stage of estimation. To this end it is only required to convert daily web scraping data into several monthly statistics. Hence we obtain two sets of time series of the same frequency. For other tasks we wish to use at least time-space combining, which will involve temporal and spatial key. In the most desirable situation closer matching will be used.

5. Any remaining issues

There were no remaining matters and the meeting was closed.