Automatic hotel prices collection on the Internet for the Tourism Survey in the Basque Country

Jorge Aramendi (j-aramendi@eustat.eus), Javier San Vicente (j-sanvicente@eustat.eus)

Keywords: Web scraping, Hotel Prices, Tourism

1. Introduction

EUSTAT, the Basque Statistical Office, aware that Big Data is a source of information for statistical offices, either for independent use or in combination with other more traditional sources, such as sample surveys or administrative records, carried out in 2016 a pilot study, with the collaboration of the ***Basque Country University (EHU-UPV)***, for the study of daily price series of hotel rooms and their possible use in our ***Hotel Occupancy and Prices Survey (ETR)*** [1].

The ETR estimates ***Average daily rates (ADR)*** for double room without taxes and extra services. The ADR is obtained as the weighted average of occupied rooms of nine different types of customers: tour operator, businesses, traditional travel agency, individuals, groups, Internet clients, on-line travel agency, online tour operator and others.

2. Methods

There are several websites that offer Hotel-reservations, so in our pilot, we decided to focus on only one of them to test our methods and tools. We chose *booking.com* to obtain the daily average prices for hotels and pensions from the Tourist Establishments Register of the Basque Country.

We used import.io [2] for web scraping, a program that allows creating scraping templates adapted to any particular type of structured web page. We have collected data for a period of 45 days and for each day we have collected all hotel room prices offered for the next seventh and fourteenth days

Booking.com offered data for 955 different tourism establishments for the Basque Country and one of the most critical phases has been the linkage with our register. The only common variable of both sources is the name of the establishment which has also varied in some units along these 45 days. In total we have obtained 95,984 records.

3. Results

3.1. Coverage

As shown in Table 1, the degree of full coverage is 70% and is directly proportional to the hotel category, so that it is 100% for 5-star hotels (H5), descending as it does category, limited to 79% in hotels of a star (H1). Similarly in the case of pensions: 67% coverage in the two stars (P2) compared to 43% in a star (P1). Availability of website by the establishment is a necessary requirement to bid on Booking; this is logically less common among lower category accommodation. Coverage by geographical area, of course, is conditioned by the distribution of the different types of establishment.

Table 1. Ratio of coverage of tourism register by geographical area and category of hotel with the scraped data.

Hotel Category	Gasteiz	Resto Alava	Rioja Alavesa	Bilbao	Area Metro. Bilbao	Bizkaia interior	Bizkaia costa	Donostia-San Sebastián	Area Metro.DSS	Gipuzkoa interior	Gipuzkoa costa	EUSKADI
H1	60%	86%	100%	50%	100%	56%	80%	100%	80%	77%	85%	78,8%
H2	100%	33%	67%	100%	60%	73%	91%	100%	100%	73%	100%	82,4%
H3	100%	100%	100%	84%	100%	100%	78%	100%	0%	78%	100%	90,2%
H4	100%	0%	100%	100%	100%	100%	100%	100%	100%	100%	80%	96,1%
H5	100%	–	100%	100%	–	–	–	100%	–	–	–	100%
P1	0%	14%	–	58%	44%	13%	28%	63%	57%	25%	54%	42,9%
P2	57%	0%	–	75%	60%	33%	44%	86%	86%	11%	69%	67,3%
TOTAL	66,7%	45,8%	83,3%	78,6%	71,0%	53,1%	63,6%	80,8%	77,8%	55,4%	79,0%	70,1%

Coverage has not been homogeneous in the study period. If we perform this analysis per day and type of data (7 or 14 days ahead) we see how the coverage has been declining in the period and, except specific days, 14 days view is always significantly higher than 7 days ahead. As the day of the week, holidays (Friday and Saturday) is found less coverage on holidays versus working both seven days and fourteen days ahead regardless of the category of hotel (Figure 1).

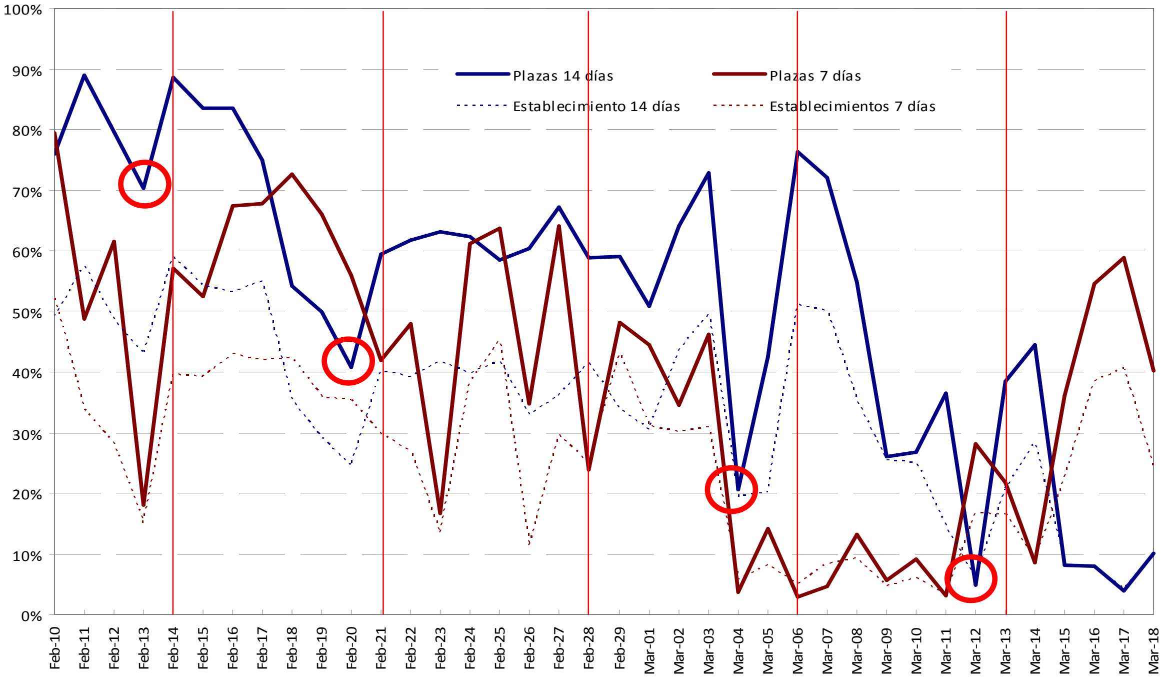

Figure 1. Coverage ratio in offered beds and establishment by days in advance for the 45 days of the pilot

3.2. Comparison of results

Prices analyzed correspond to the months of February and March 2016 and practically have no differences over the terms of the consultation (7 or 14 days ahead). Except in 5-star hotels, where the price to 14 days ahead the average price obtained is reduced.

On average, the average ADR of the survey for the type of customer internet is 9.3% lower than the result of scraping on the web for the same dates with a range of differences ranging from -3.1 % for 1 star hotels to -20.7%.

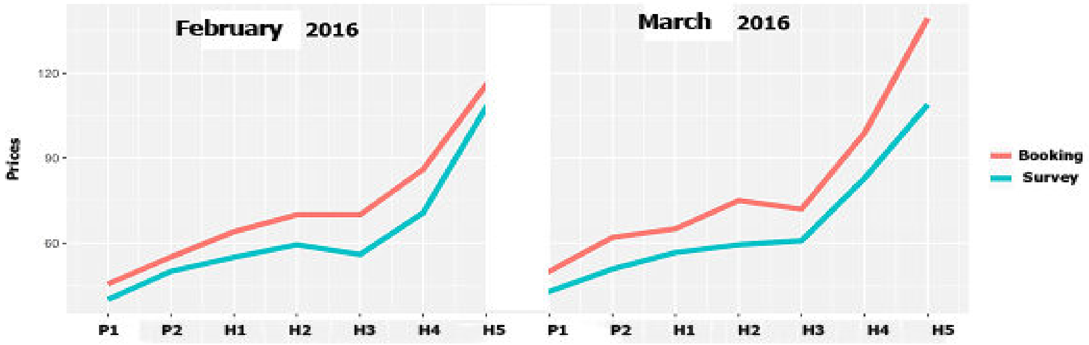

Figure 2. Comparison between the scraped average price and the average ADR by hotel category and month

4. Conclusions

As for the coverage of our tourism register, it is very high in hotels of higher categories and reasonable in the case of lower categories.

The comparison of both sources, presents differences that required more research; issues like the definition/components of price in one source to another, the way in which the average daily rate is obtained in the ETR survey - with a question on average prices - the type of hotels and their prices not captured with scraping or Booking performance in determining the number of places and their prices.

Based on the experience gained in Big Data in this pilot, Eustat has launched a new and more ambitious project in June 2017. In this new project, to determine the ADR, Eustat is scrapping the Booking website 3 times a day for each of the 120 previous days. So, each hotel, agrotourism and apartment could have a maximum of 360 prices for a particular day of reservation. By early August, more than 8 million valid records have been collected.

References

[1] EUSTAT. Hotel Occupancy and Prices Survey (ETR), data and documents.

http://en.eustat.eus/estadisticas/tema_101/opt_0/tipo_1/ti_Hotel_establishments/temas.html

[2] https://www.import.io