Recognizing Patterns in the Price Time-Series of the Basque Country Hotels

Asier Badiola (a-badiolazabala@eustat.eus) , Ander Juarez (a-juarezmugarza@eustat.eus), Jorge Aramendi (j-aramendi@eustat.eus)
EUSTAT- Basque Statistics Office - www.eustat.eus

BACKGROUND

Different web browsers have been developed in the last years in order to try to find the best price on the market. That increase in the number of browsers has created more fluctuation and variability in the final price offered. In this study, we focus on the price of hotels and hostels located in the Basque Country with the data collected from the Internet using web scraping techniques.

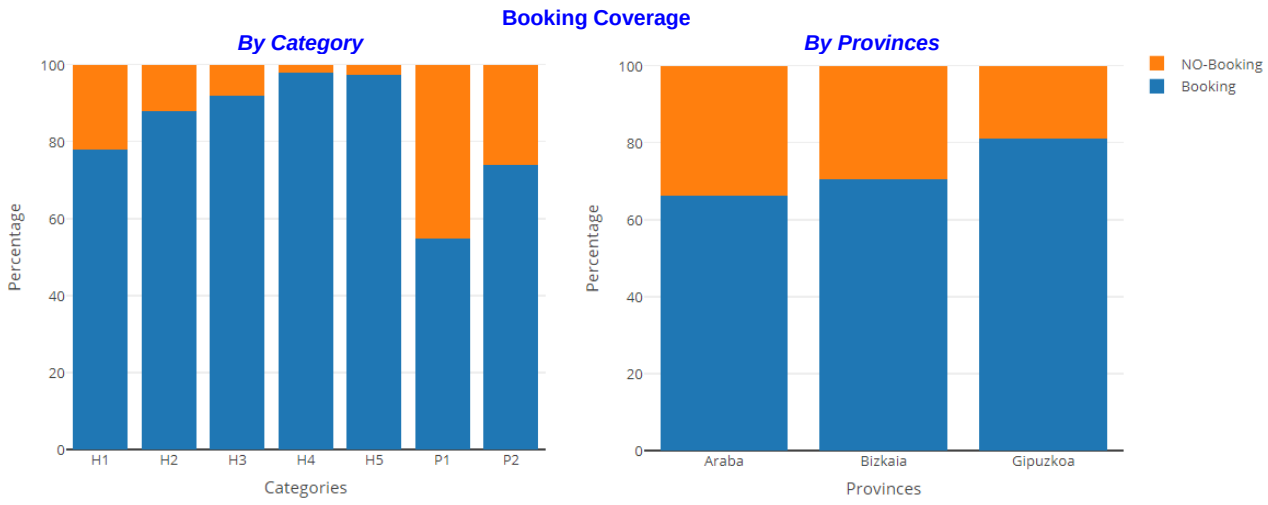
Data has been obtained from www.booking.com and per each day and each hotel/pension, 120 requests have been done. In other words, the price of the next 120 days has been collected daily. Once the 120 requests were done for each establishment and day, we took the median of all those prices for the posterior analysis (eventually EUSTAT is studying how to summarize those 120 prices into a relevant final price).

EUSTAT, aware that Big Data is an interesting source of information for statistical offices, has carried out a pilot study for the study of the daily series of hotel room prices and its possible use in the Survey of Tourist Establishments Receivers (ETR). In this poster EUSTAT presents the results of what is one its first case study in the field of Big Data.

OBJECTIVES

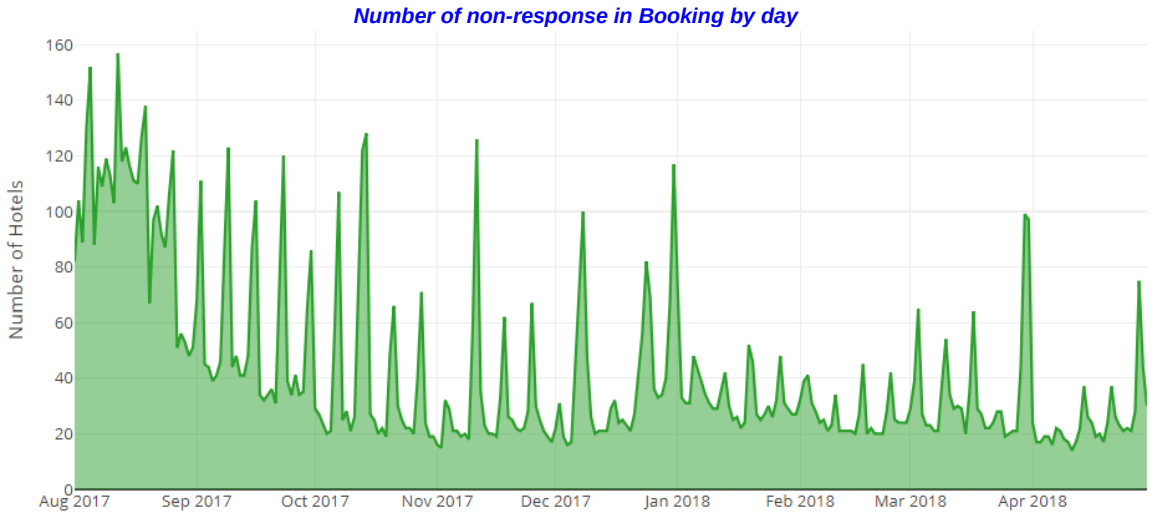
The objective is the clustering of the Basque Country hotel establishments, hotels and pensions, using the price time-series. The unit of analysis is the standard double room with bathroom, without breakfast and without VAT.

DATA DESCRIPTION



The problem of non-response rates appear due to many problems in the web-scraping methods: changes in the web structure, accommodation offers, connection errors…

Number of non-response in Booking by day



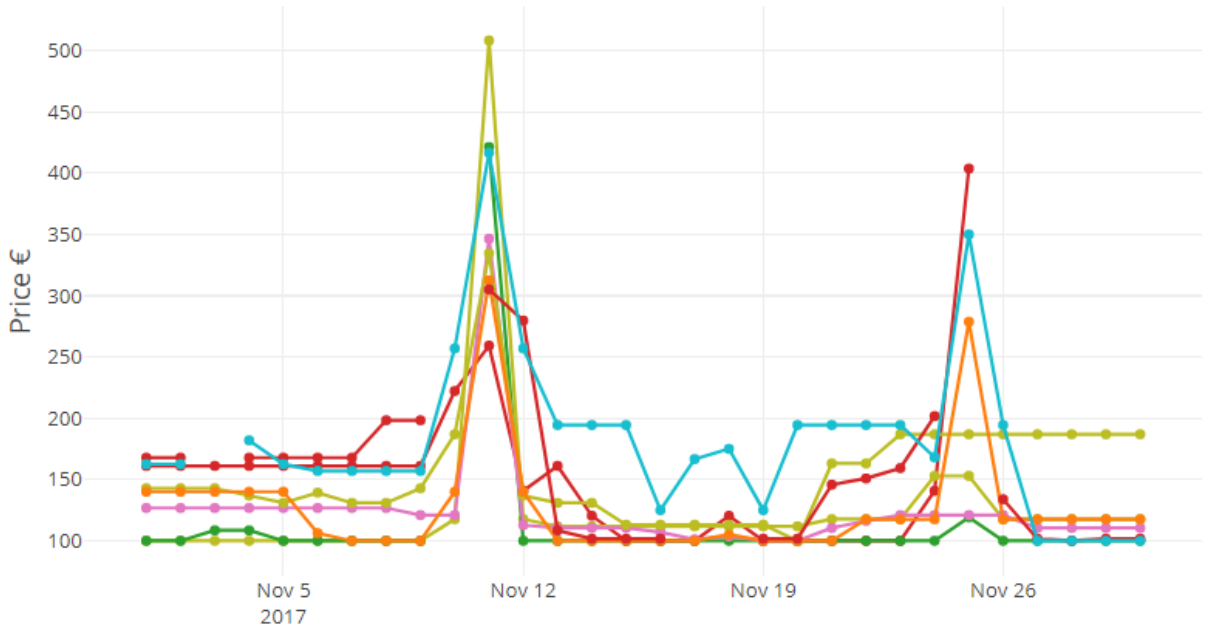
METHODOLOGY

This methodology is structured in 3 main parts: **outlier detection**, **data imputation** and **clustering**.

Outlier detection:

After testing with many R packages such as *outliers* and *tsoutliers*, we realized that comparatives between different time-series must be done to distinguish an outlier from a special event, creating a personalized outlier detector algorithm.

Some hotel's prices during Nov 2017



Data imputation:

Every (or almost every) time-series clustering algorithm needs complete time-series (without any lost value) in order to calculate the distance among them, which makes data imputation compulsory. For that staff *imputeTS* package has been used. The best results have been given by *na.seadec* method.

```
library(imputeTS)
na.seadec(ts(x, frequency = 7), algorithm = 'kalman')
```

Clustering:

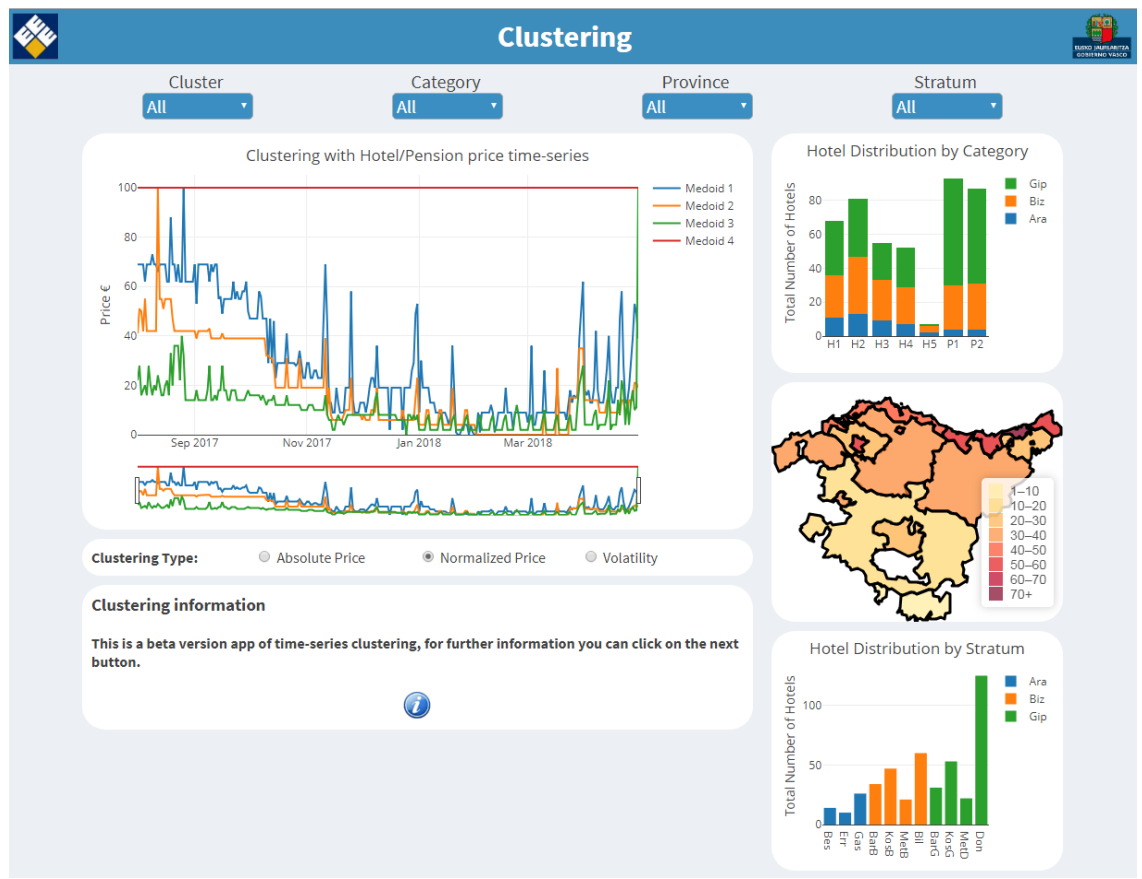
3 types of different clusterings have been done:

- Clustering with absolute prices.
- Clustering with normalized prices (in order to analyze the trend).
- Clustering in relation to volatility of prices.

For the first 2 types, many algorithms have been tried from *TSclust*, *TSdist* and *dtw* (R) packages (finally the Euclidean distance was used), and for the last one, *close-to-close volatility estimator* (or *close/close*) was used.

RESULTS

Results have been satisfactory and have been shown using firstly with *Shiny* (a package from R) and then with *JavaScript*.



The interactive app (beta version) can be found in:

- https://asbaza.github.io/

CONCLUSIONS

It has been shown that the establishments with biggest price volatility are located in Donostia-San Sebastián and are above all pensions. On the other hand, the ones with lowest volatility are most of them in the inner places and outside the capitals.

If we focus on the price, more than the 30% of the establishments of Donostia-San Sebastián and Rioja Alavesa are grouped in the 3 most expensive clusters, due to the effect of tourism. From the opposite position, more than the 70% of the hotels and pensions of Araba (excluding Rioja Alavesa), Bizkaia (excluding Bilbao) and inner cities of Gipuzkoa are grouped in the 2 cheapest clusters.

REFERENCES

[1] EUSTAT. Research and Development Activities Survey, data and documents. http://en.eustat.eus/estadisticas/tema_101/opt_0/ti_Hotel_establishments/temas.html

[2] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2014). An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated.

[3] Usue Mori, Alexander Mendiburu and Jose A. Lozano (2016). Distance Measures for Time Series in R: The *TSdist* package. R journal 8, 2, 451-459.

[4] Moritz Steffen *imputeTS*: Time Series Missing Value Imputation (2017). R package version 2.5. https://CRAN.R-project.org/package=imputeTS