Spatial influence in Basque Country's hotel prices

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

BACKGROUND

Despite being a small territory, there is a huge variety of places people could visit at the Basque Country. Starting from Donostia's beaches, passing through Bilbo with the Guggenheim or the mythic San Mames and finishing with Gasteiz, the Green Capital. That is why we thought that it would be interesting to analyze if those differences are shown in the tourism and, thus, in those place's hotel's prices.

In order to achieve that, the data was collected from two main sources. The Basque Country's hotel's prices were collected thanks to a personalized web scrapping technique developed by the Basque Statistics Office, Eustat. Moreover, complementary information was collected from the Directory of Tourist Establishments of Eustat. This complementary information varies from the category of the establishments to the specific coordinates, passing through the occupation or the size of the hotels.

OBJECTIVES

This project had 2 main objectives: to study the influence of the location in the hotel prices using spatial correlation indices and to develop an intuitive and accessible interactive chart which shows the prices and their evolution over the time.

OUTLIER DETECTION

As the prices were automatically taken from Booking with a web scraping technique, and, though some previous pre-process was done, some outliers were present in our data. Once an early study was done, some points were discovered which, despite they seemed to be outliers, they were not. Those prices were related to the different festivities or events of the different territories and cities.

Anomalies example
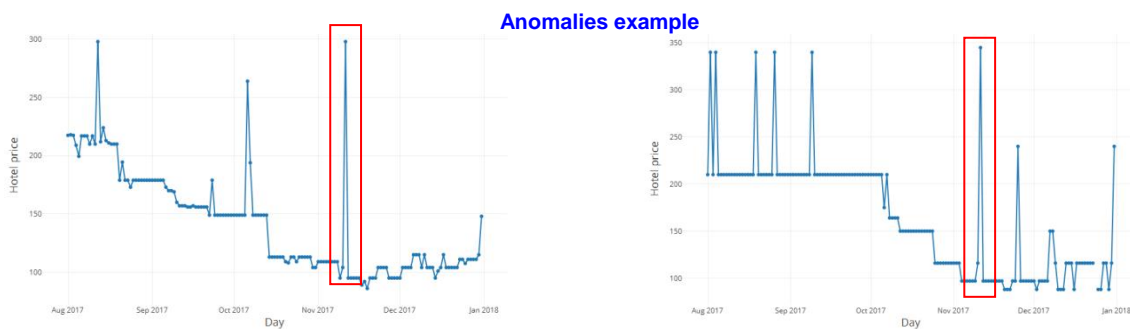
Taking those points into account and before starting to find the real outliers, the data was scaled to remove the influence of their category and their location, making them comparable. With that objective, the series were transformed to show the price variation instead of the price, isolating the strange prices.


Scaling data
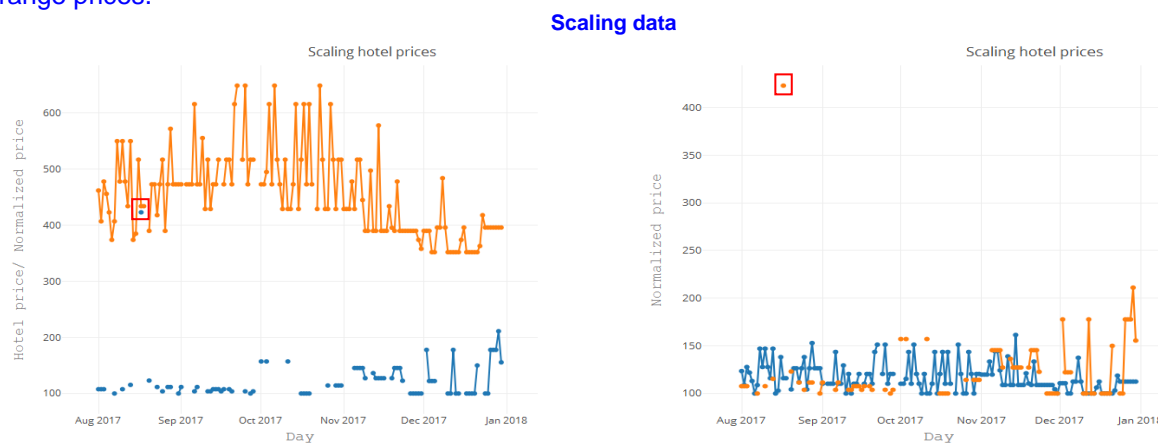
NA INPUTATION

The price of some establishments in some punctual days could not have been scraped either owing to a mistake in the scraping algorithm or because Booking did not allow booking a single night in those hotels in those days (sometimes it only allows booking more than one night).

The function *na.seadec* from the *R* package *imputeTS* was used in order to infer those lost values and the ones marked as outliers in the previous analysis.

NA imputation with seasonal decomposition
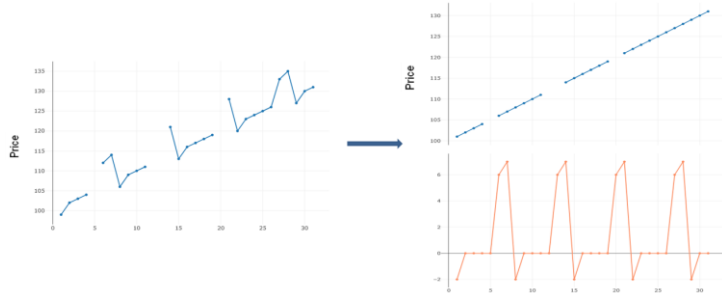
SPATIAL KORRELATION ANALISIS

Different correlation indexes were checked. In addition, the indices were modified to take into account not only each hotel price but also the influence of the surrounding hotel prices. All studied indexes have the following structure:

$$\Gamma = \lambda \sum_i \sum_j a_{ij} p_{ij}.$$

Here the element a_{ij} reflect the spatial relation between the hotels i and j while the elements p_{ij} depends on the values of each hotel. In order to represent the spatial relation between the hotels it was decided to take $a_{ij}=1/d_{ij}$ were d_{ij} is de distance between the hotels I and j. The following indexes are the ones that had been tested:

Moran
$$I^* = \frac{1}{A_{Tot}} \sum_i \sum_j a_{ij} \frac{(x_i - \overline{x_A})}{s_A^2} \frac{(x_j - \overline{x_A})}{s_A^2}.$$

Geary
$$C^* = \frac{1}{2} \frac{1}{A_{Tot}} \sum_i \sum_j a_{ij} \frac{(x_i - x_j)^2}{s_A^2}.$$

Lebart
$$L^* = \frac{1}{A_{Tot}} \sum_i A_i \frac{(x_i - y_i)^2}{s_A^2}.$$

η_1
$$\eta_1^* = \frac{1}{A_{Tot}} \sum_i A_i \frac{(y_i - \overline{x_A})^2}{s_A^2}.$$

η_2
$$\eta_2^* = \frac{1}{A_{Tot}} \sum_i \sum_j a_{ij} \frac{(y_i - y_j)^2}{s_A^2}.$$

From all those indexes, the chosen one for the final testes has been the *Geary* correlation index.
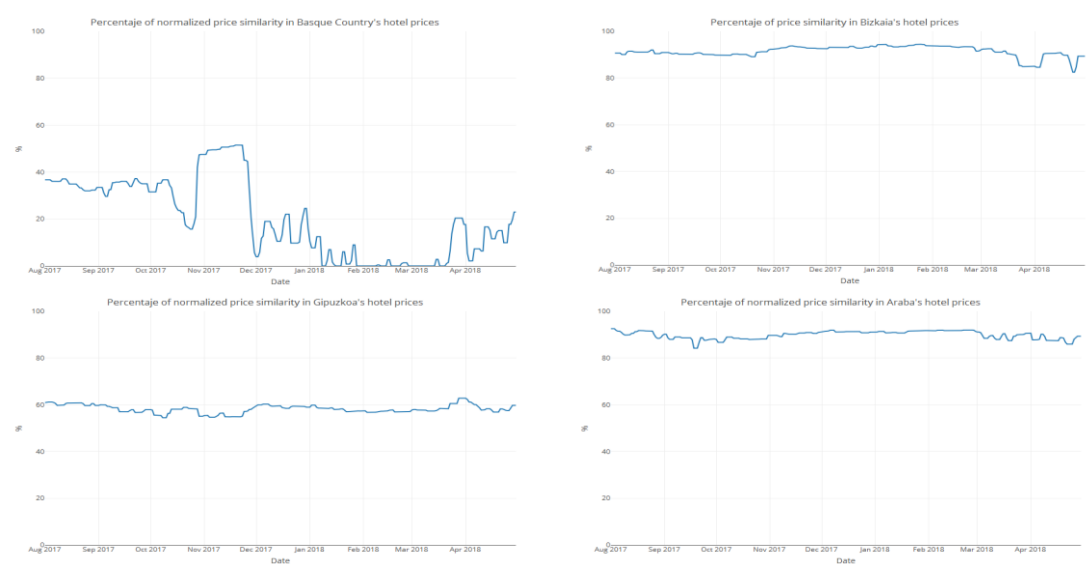
INTERACTIVE CHART

In order to develop the interactive chart the software used has been R. This software, together with the integrated development environment *RStudio* and the packages *leaflet* and *Shiny,* has let us develop an intuitive interactive chart.

Used packages:
Shiny
Leaflet

RESULTS

Analyzing the modified Geary's correlation index's results, we noticed that despite in a general view it seemed that the Basque territories were very different; these dissimilarities were mainly caused by the high prices and price variety of Gipuzkoa's establishments and, more specifically, Donostia's establishments. Taking those hotels apart we saw that Bizkaia's and Araba's hotels were not so different in overall.
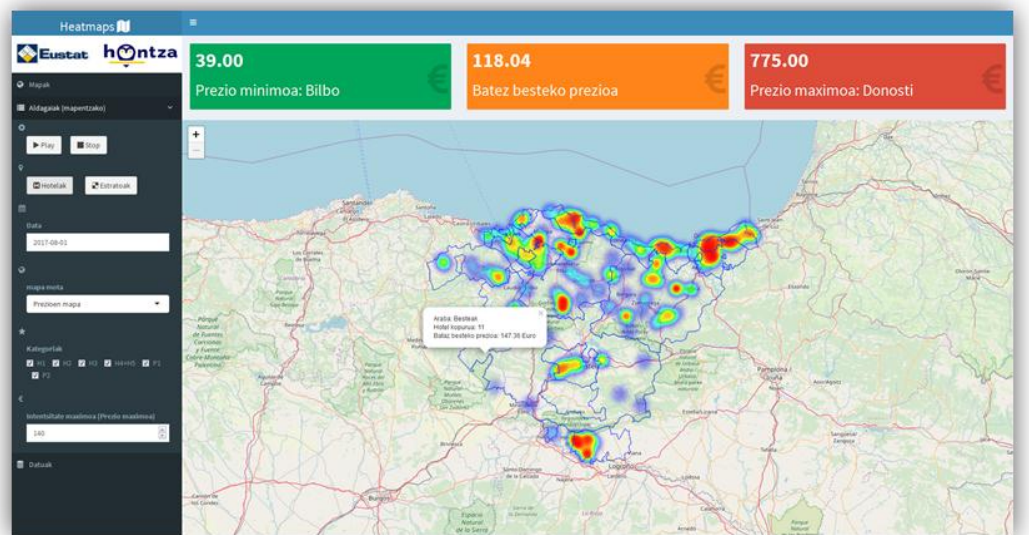
As it can be seem in the chart we got, apart from the main heatmap additional information is given. Firstly, the minimum price, the mean price and the maximum price of the day that is being visualized are given as well as where are located. Apart from that, the data shown can be filtrated by the hotels categories and the different stratum can be displayed with further information.

CONCLUSIONS

An interactive chart has been developed which allows following the Basque Country's hotel establishment's prices and watching the behavior of those. Moreover, thanks to the modified *Geary's* spatial index we noticed that the establishments of the Basque Country have a strong spatial relationship with the exception of the Gipuzcoa's ones, more precisely, with de Donostia's ones. This is a direct consequence of the high tourist influx that Donostia has.

A possible next step could be to make a continuous chart filling the empty places with estimated values via krigging for example. In addition, we could go a step further and make a model that infers the possible establishment number and prices on the towns without already existing hotels.

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] Anselin, L. (1995). Local indicators of spatial association-LISA. Geographical analysis, 27(2),93-115.

[3] Concepción, Morales, Eduardo René (2010). Contribución a la segmentación de imágenes mediante algoritmos de clasificación automática. Euskal Herriko Unbertsitatea/ Universidad del País Vasco.

[4] Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. shiny: Web Application Framework for R, 2017. R package version 1.0.5.