Spatial interpolation of apartment prices in Barcelona using Airbnb web scraped dataset[*]

Renato Vassallo[†]

February 13, 2023

Abstract

The purpose of this document is to to compare the results of several interpolation methods for spatial data of Airbnb apartment prices in Barcelona. Applying text mining techniques, an Airbnb web scraped dataset from 2022 is developed, and different interpolation methods are applied, including Voronoi approach, Nearest Neighbors (NN), Inverse Distance Weighting (IDW), and ensemble methods. Based on the RMSE value, the NN interpolation method with $n = 5$ yields much more accurate prediction values than the Voronoi or IDW methods. Areas with high Airbnb prices are located around L'Eixample, El Poblenou and Les Corts, while areas with low prices are in the vicinity of Porta and Sant Andreu neighborhoods. These results could be associated with the high student and business demand for accommodation near the city center.

Keywords: Spatial Analysis, Interpolation, Voronoi Polygon, Nearest Neighbors.

[*] This document is part of the final master's project for the Analysis of Spatial Data and Images course, taught by Professor Paula Moraga in January 2023.

[†] Correspondance Author: Renato Vassallo, Master in Data Science for Decision Making, Barcelona School of Economics, Phone: +34 647-581-560. Email: renato.vassallo@bse.eu.

1 Introduction

Today it has become easy to find information on apartments, hotels and lodgings, both short and long term. In this context, Airbnb has emerged as a platform business that provides and guides an opportunity to link two groups - the hosts and the guests. Anyone with an open room or free space can become a host on Airbnb and offer it to the global community. Price is one of the aspects that guests take into account the most when choosing where to stay. Some of the factors behind the prices could be associated with:

- Time of year
- The type of property: house, apartment, etc.
- The space that guests will have: entire accommodation, private room, etc.
- How many guests can fit comfortably in the space: number of beds and bedrooms.
- The main services: wifi, kitchen, if they allow pets, etc.

The objective of this document is to analyze the spatial distribution of the prices of apartments available on Airbnb for Barcelona towards the end of 2022. To do this, spatial interpolation methods are used to predict prices over a continuous space in the territory of Barcelona. After evaluating the performance of different models using cross-validation, we found that the prediction method that minimizes the RMSE is the Nearest Neighborhood method. Surprisingly, the evidence shows that the areas with high prices are not those closest to the sea, but are located in the center of the city (such as L'Eixample and El Born).
The document is structured as follows: Section 2 describes the data used and some statistics associated with apartment prices; Section 3 presents the empirical methods for the spatial interpolation of prices; Section 4 shows the results obtained; and finally Section 5 concludes.

2 Data

We work with a database obtained from www.insideairbnb.com, a website on which web scraped datasets of "snapshots" of cities are published. We decided to work with the files of Barcelona of the situation on 2022.
The initial data has 15778 observations for 75 features, among which there are numerical and text data. The pre-processing of this initial data is summarized in the following steps:

- Remove listings that have number of reviews equal to 0

- Conversion from strings/categorical to numerical features
- Data extraction from texts using NLP and regex techniques.
- Remove outliers using interquartile range (IQR) method.
- Data imputation (only when necessary).
- Filter for private rooms in rental units or entire rental units.
- Drop observations with remaining missing values.

After this process, we are left with a dataset of 7336 observations and 40 features. Figure 1 presents some descriptive statistics of the final dataset. It is observed that the price distribution is slightly skewed to the right, that is, on the right side of the graph, the frequencies of observations are lower than the frequencies of observations to the left side. To have a point of comparison, we've created an *is_top_*100 dummy that will assign a value of 1 if the listing is in the top 100 reviewed listings on Airbnb. The right panel shows that, on average, those top 100 ads have a higher price than the others (approximately 90 vs. 95 euros on average).

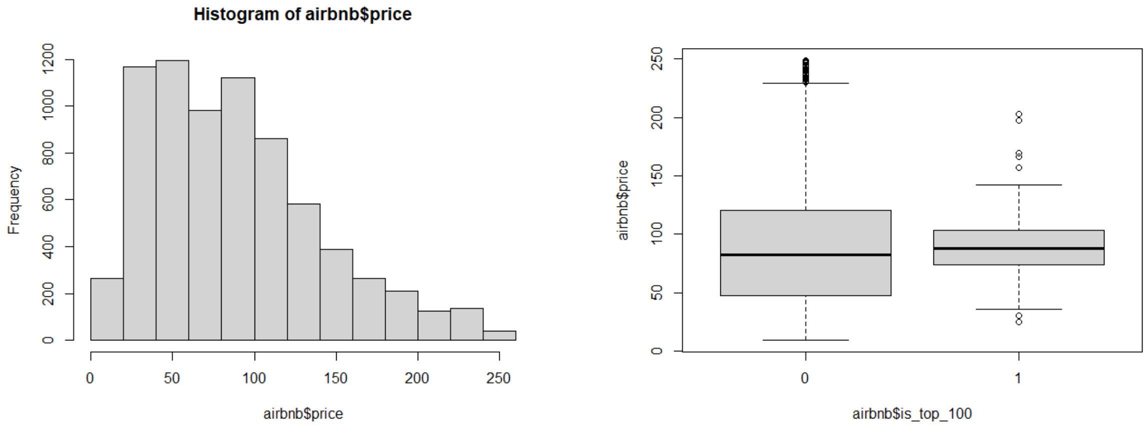
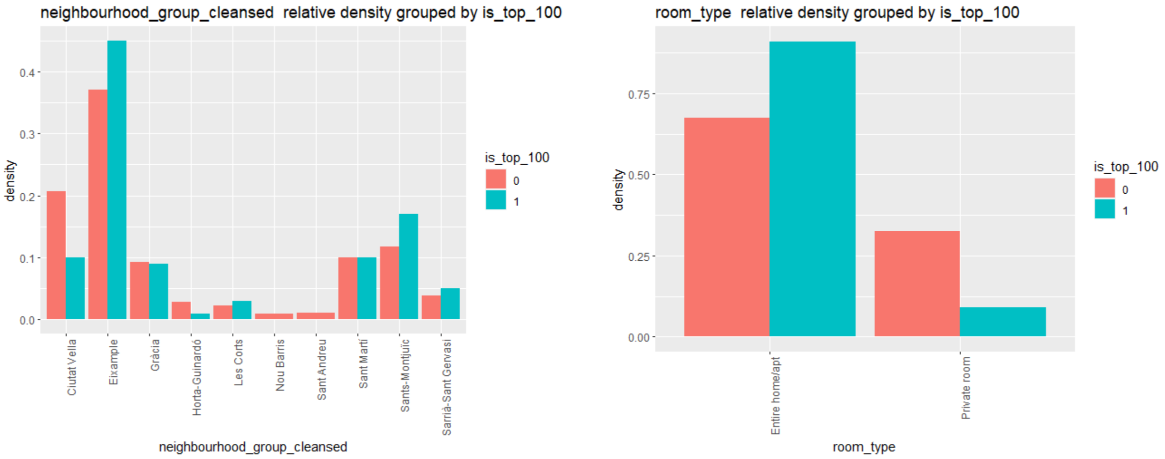


Figure 1: Descriptive statistics

Figure 2 characterizes the apartments by area, and it can be seen that the neighborhoods with the highest number of listings on Airbnb were L'Eixample, Ciutat Vella and Sants-Montjuic. On the other hand, we filter for private rooms in rental units or entire rental units.

From *insideairbnb* web page, we obtain a GEOJSON file that contains full list of Barcelona neighbourhoods with geospatial data that we will use to visualise information on the map.



(a)

(b)

Figure 2: Listings main characteristics

We use **Leaflet** R package and display listings from both groups using lat/long information coming from listing details dataset. Figure 3 give us an idea of geographical distribution with Red points being in Top 100 most popular listings.

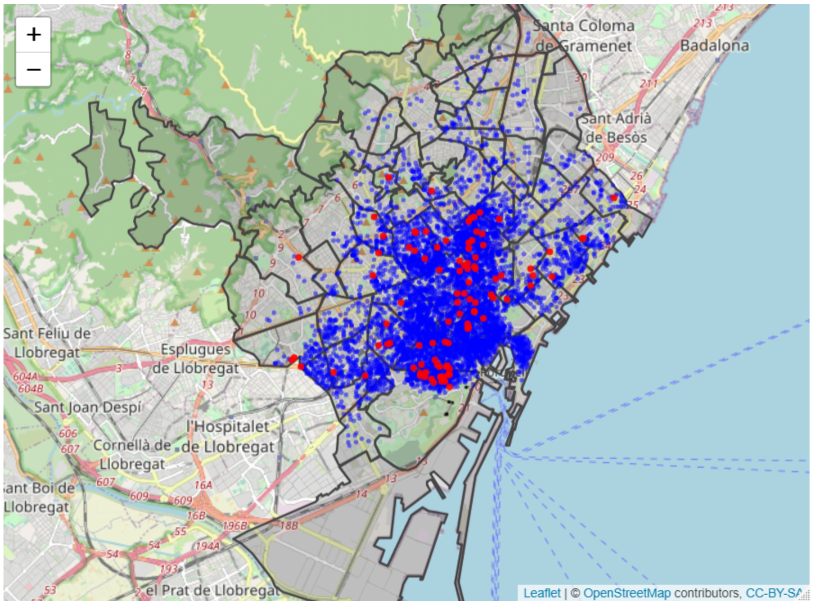
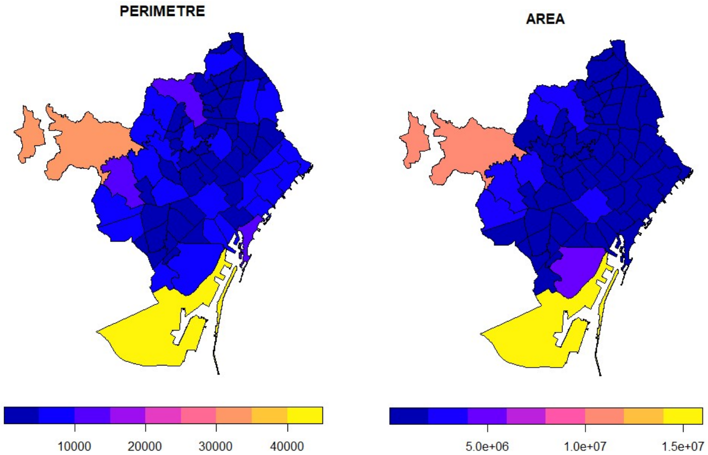


Figure 3: Neighbourhoods geospatial data

3 Methods

We begin by visualizing the maps that determine the administrative borders of the neighborhoods of Barcelona. Figure 4 shows the 73 neighborhoods and the colors denote the perimeter and area associated with these zones. For this, the standard *plot* command is used and the variable to be represented is specified.



(a) Perimeter (b) Area

Figure 4: Map of Barcelona by neighborhoods

We can make a better plot with a base map. We create a variable *map* denoting the study region with the union of these polygons. This is the region where we will predict the variable of interest price per square meter.

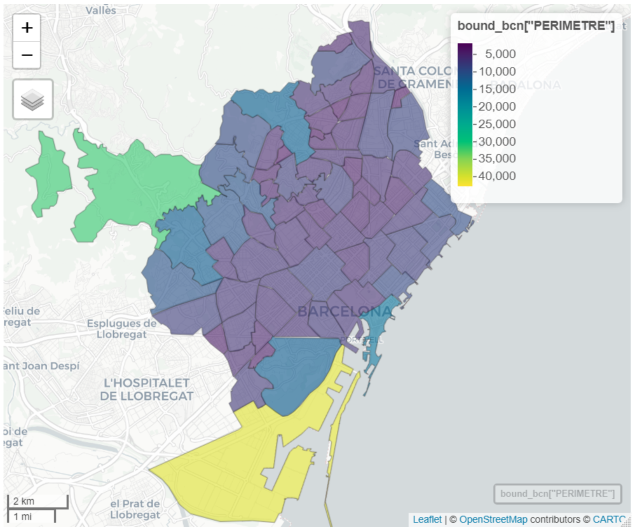


Figure 5: Map of Barcelona using **mapview**

Given that we have close to 8,000 observations, we group these variables by neighborhood to reduce the dimensionality of the data without losing representativeness and characteristics within each neighborhood. Figure 6 shows the average price by neighborhood. This will be our input to make the predictions about the continuous space of Barcelona.

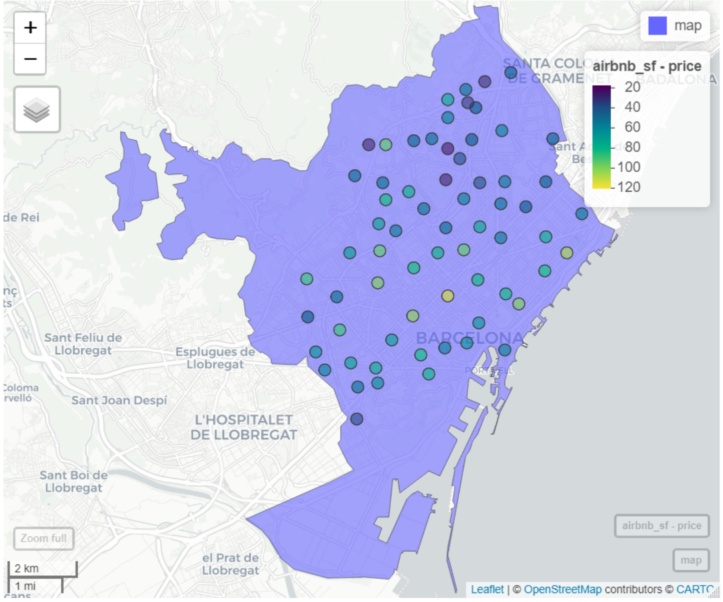


Figure 6: Geostatistical data for prices in Barcelona

For the interpolation of the values of the variable of interest, we consider 4 classical methods: (i) Voronoi polygon; (ii) Nearest Neighbors; (iii) Inverse Distance Weighting (IDW); and, (iv) Ensemble method. The results are shown in the next section.

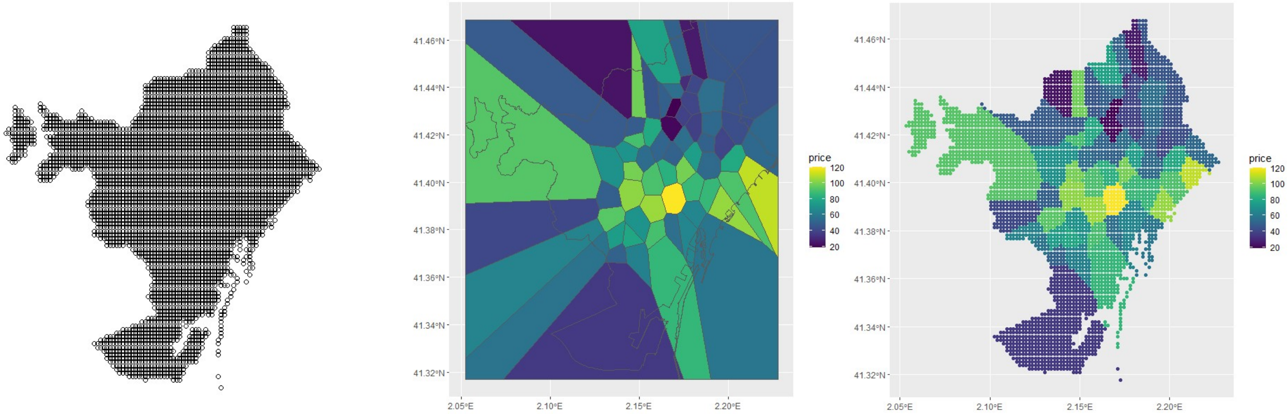
4 Results

Our goal is to make predictions of price per square meter of Airbnb listings continuously in space within the boundary of Barcelona. We decide to predict at locations forming a fine grid within Barcelona. We can use the *st_make_grid*() function of sf to create a grid with prediction locations over the bounding box of an *sf* object.

4.1 Voronoi polygons

We develop a simple prediction using Voronoi polygons (Voronoi, 1908), so we'll create these and see what they look like with the original data points also displayed.
Figure 7a shows the grid within the Barcelona boundaries, which has been generated considering a number of columns and rows equal to $n = c(100, 100)$. Figure 7b shows the Voronoi diagram

for the subset of prices considered in the previous section, assuming constant values in each of the grid polygons. Finally, Figure 7c results from the intersection of both described components.



(a) Grid within Barcelona

(b) Voronoi diagram

(c) Predicted values

Figure 7: Voronoi method

Note that the Voronoi polygons contain all of the attribute data from the original points, just associated with polygon geometry now.

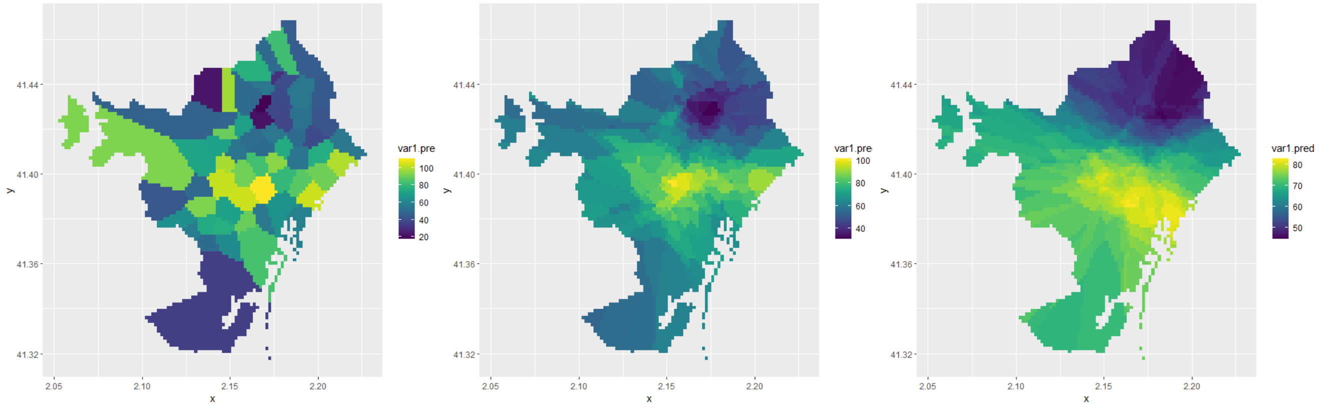
4.2 Nearest Neighborhoods

In the nearest neighbors interpolation, values at unsampled locations are estimated as the average of the values of the closest sampled locations. We can consider a specific number of closest sampled locations or neighbors. Specifically:

$$z = \frac{\sum_{i=1}^{k} z_i}{k}$$

where z is the value to be interpolated, z_i is the value corresponding to neighbor i, and k is the number of neighbors considered. To estimate the values of unsampled locations we use the **gstat** package, and we consider an intercept only model.

The results can be seen in Figure 8. As the number of neighborhoods increases, the interpolation becomes smoother and more continuous within the allotted space. When $n = 1$ we observe that the predictions are very similar to those obtained using the Voronoi polygon. However, when $n = 20$ we have a more continuous distribution where the yellow color is distinguished more intensely over the region of the center of Barcelona. On the contrary, in the northern part of the city, a darker color predominates, denoting lower prices.



(a) $n = 1$

(b) $n = 5$

(c) $n = 20$

Figure 8: Nearest Neighbors method

4.3 Inverse Distance Weighting (IDW)

In IDW, values at unsampled locations are estimated as the weighted average of values from the rest of locations with weights inversely proportional to the distance between the unsampled and the sampled locations. Specifically:

$$z = \frac{\sum_{i=1}^{n} z_i (1/d_i^{\beta})}{\sum_{i=1}^{n} (1/d_i^{\beta})} = \sum_{i=1}^{n} z_i w_i,$$

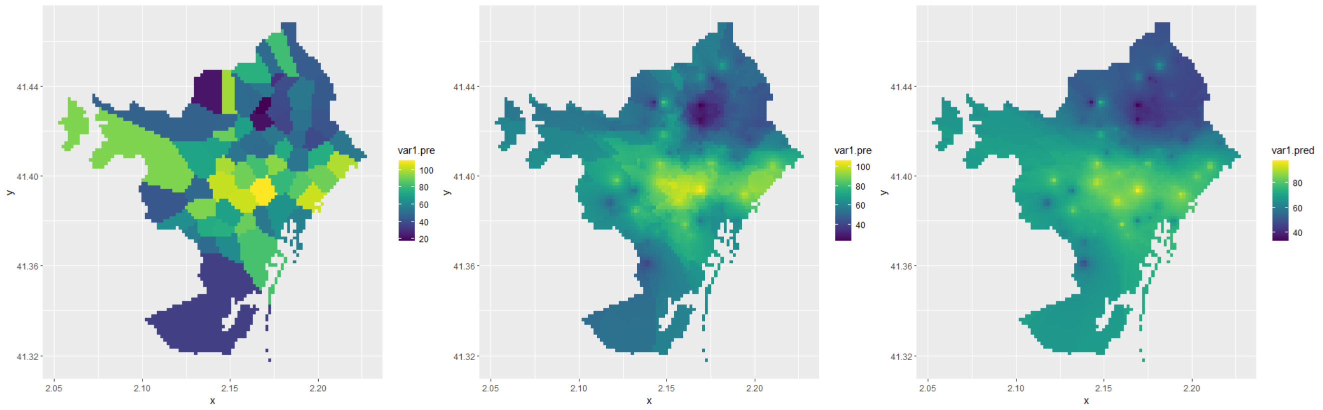
where z is the value to be interpolated, n the number of sampled locations, z_i is the value at location i, and d_i correspond to the distance between location i and the location where we want to predict. Here, weights are given by $w_i = \frac{1/d_i^{\beta}}{\sum_{i=1}^{n}(1/d_i^{\beta})}$. β is the distance power that determines the degree to which nearer points are preferred over more distant points. If $\beta = 1$, $w_i = \frac{1/d_i}{\sum_{i=1}^{n}(1/d_i)}$.

Unlike the nearest neighbors approach, in the IDW approach locations further away are assigned less weight in predicting the value at a location.

The results for the IDW method are presented in Figure 9. When $n = 1$ the results are consistent with what was found for Voronoi and NN. However, the predictions are different when neighborhoods start to grow. Although the interpolation is relatively smooth, there are certain regions where colors stand out. Despite this, there is a greater concentration of yellow in the center of the city.

4.4 Ensemble Method

We can calculate predictions using an ensemble where we combine the predictions obtained with each of the methods. Specifically, if M is the number of prediction methods considered,



(a) $n = 1$

(b) $n = 5$

(c) $n = 20$

Figure 9: Inverse Distance Weighting method

the predicted value z can be obtained as:

$$z = \sum_{i=1}^{M} z_i w_i$$

where z_i and w_i are the prediction value and weight corresponding to method i, with $i = 1, ..., M$. The weights can be chosen in different ways. For example, they can be proportional to the goodness of fit of each method and sum to 1.

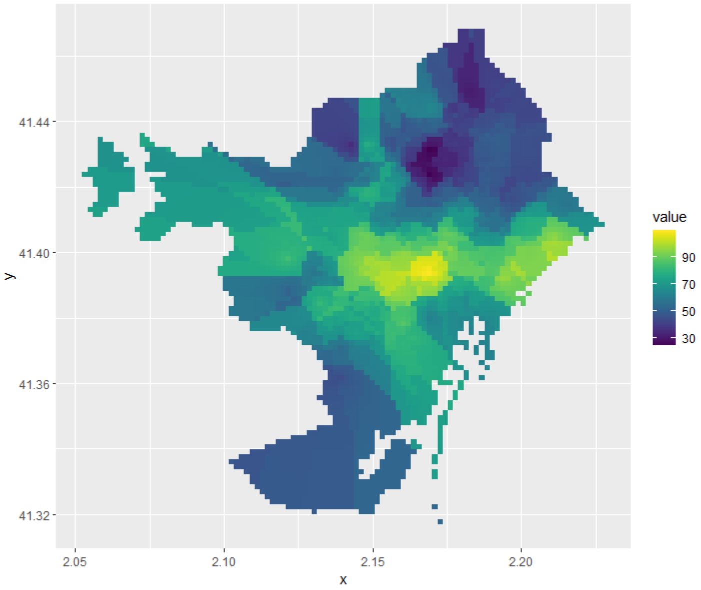


Figure 10: Ensemble method

4.5 Cross validation analysis

Cross validation is used to see how well the model works by sampling it a defined number of times (folds) to pull out sets of training and testing samples, and then comparing the model's predictions with the (unused) testing data. Each time through, the model is built out of that fold's training data. For each fold, a random selection is created by the sample function. The mean RMSE of all of the folds is the overall result, and provides an idea on how well the interpolation worked.

Here, we assess the performance of each of the following methods to predict the spatially continuous variable in the example above: (i) Voronoi; (ii) Nearest neighbors; (iii) IDW; and, (iv) Ensemble.

	Voronoi	Near Neigh	IDW	Ensemble
split 1	26.89	26.21	24.85	24.36
split 2	32.83	18.51	20.01	21.41
split 3	18.51	17.53	17.19	16.39
split 4	25.21	20.55	19.91	19.55
split 5	32.57	19.63	20.93	22.38
Average	**27.20**	**20.49**	**20.58**	**20.82**

Table 1: Cross validation analysis. RMSE obtained for each of the 5 splits

5 Conclusions

Based on the RMSE value, the Nearest Neighbors Interpolation Method with $n = 5$ yields much more accurate prediction values than the Voronoi or IDW Interpolation Methods. Areas with high Airbnb prices in 2022 are located around L'Eixample, El Poblenou and Les Corts, while areas with low prices are in the vicinity of Porta and Sant Andreu neighborhoods. The reasons behind these findings could be associated with the high demand for apartments by academics, students and foreign entrepreneurs and visitors in Barcelona; and the scarce offer of accommodation near the city center. However, for a deeper analysis of the determinants of Airbnb prices, it is necessary to consider a multivariate modeling of prices based on different covariates.