Analyzing Key Influences on Airbnb Pricing in Lecce Municipality

Riccardo Russo

December 2023

1 Abstract

The emergence of the "sharing economy" in recent years has become more evident through the proliferation of online peer-to-peer platforms. Among these platforms, Airbnb stands out as a notable example. Functioning as an accommodation marketplace, Airbnb facilitates access to tourist lodging, embodying the principles of the sharing economy. The emerging business paradigm is distinguished by the peer-to-peer (P2P) economy, where direct interaction between providers and clients takes center stage, eliminating the need for intermediaries.

This study delves into the intricate spatial dynamics that influence Airbnb pricing in the municipality of Lecce, a small town with unique characteristics. Employing spatial analysis techniques, we examine the distribution of Airbnb listings. The analysis reveals that Lecce's Airbnb landscape is diverse, reflecting the town's cultural and geographical heterogeneity. The results guide the selection of an appropriate spatial specification for the model, leading to the estimation of a Spatial Autoregressive model. The study concludes with a reflection on the challenges posed by Lecce's unique characteristics and the limitations of linear spatial models in capturing nuanced pricing patterns. The findings contribute to the broader understanding of Airbnb pricing dynamics in smaller, tourist-centric towns, emphasizing the need for tailored analytical approaches in such contexts.

2 Introduction to the study area

According to the World Tourism Organization (WTO), Italy is renowned as one of the top tourist destinations globally, and within its diverse landscape, Lecce, situated in the enchanting region of Puglia, stands out as a captivating destination. Lecce, often referred to as the "Florence of the South," offers a unique blend of historical charm and modern allure, making it a significant attraction for travelers. With its rich cultural heritage, exquisite architecture, and delightful Mediterranean climate, Lecce has become a prominent destination, rivaling other popular spots like Rome or Florence. In recent years, Lecce has experienced a surge in tourist influx, attracting visitors from around the world. The city's historic center, characterized by ornate Baroque architecture, has earned it the title of a UNESCO World Heritage site, contributing to its appeal. The city's enduring popularity can be attributed also to its Mediterranean climate. The picturesque coastline, historical landmarks such as the Basilica di Santa Croce, Piazza del Duomo, and the Roman Amphitheatre, along with the vibrant atmosphere of its streets, make Lecce a compelling destination for those seeking a blend of history, culture, and relaxation.

In the context of Municipality of Lecce, this study delves into the intricate factors influencing the pricing of Airbnb. Understanding the determinants of Airbnb prices is crucial for both hosts seeking optimal returns and researchers aiming to contribute to the broader discourse on the evolving hospitality sector. The essence of this analysis lies in the utilization of spatial regression models to unravel the multifaceted aspects that contribute to Airbnb pricing in the unique setting of Lecce. Spatial regression techniques provide a nuanced understanding of the spatial relationships between accommodations and their proximities to various influential factors. Lecce, with its distinct cultural heritage and vibrant tourism scene, serves as an intriguing focal point for this study. The city's historical significance, coupled with the contemporary influence of the sharing economy, creates a rich tapestry for investigating the interplay between spatial dynamics and Airbnb pricing. The primary objective of this research is to identify, quantify, and analyze the key determinants that shape Airbnb pricing strategies in Lecce Municipality. By employing spatial regression models, we aim to go beyond

traditional pricing analyses and consider the spatial context, acknowledging the potential impact of geographic proximity to local attractions, amenities, and other relevant factors. Through this investigation, we seek to contribute empirical insights that not only enhance the understanding of Airbnb pricing mechanisms but also offer practical implications for hosts and stakeholders in the hospitality sector in Lecce.

As the sharing economy continues to evolve, studies of this nature become instrumental in adapting strategies to the unique characteristics of specific cities and regions. Van der Borg et al. [2017] present findings suggesting that Airbnb listings situated in close proximity to tourist attractions, lakes, and mountains experience higher occupancy rates. Perez-Sanchez et al. [2018] assert that the geographical placement of properties, particularly those in areas rich in dining options or close to the beach, positively influences the pricing of Airbnb accommodations. According to Deboosere et al. [2019], properties near public transportation tend to command higher prices, particularly when such transport facilitates convenient access to city centers. Beyond these investigations, we aim to illustrate the spatial autocorrelation among Airbnb prices as a potential influential factor.

3 Dataset and variables

Data Collection

The data for this study was sourced from the comprehensive dataset provided by Inside Airbnb, a platform that aggregates and makes available detailed information about Airbnb listings worldwide. Specifically, we focused on the region of Puglia, Italy, leveraging datasets tailored to this area.

Inside Airbnb's Puglia datasets, accessible through their data portal Airbnb [2023], offered a wealth of information crucial for our analysis. The following datasets were utilized in this study:

- `Puglia listings.csv.gz`: Detailed Listings data providing comprehensive information about individual Airbnb listings in Puglia. This dataset serves as the foundation for our analysis, offering insights into various aspects of each accommodation.
- `Puglia neighbourhoods.geojson`: A GeoJSON file of neighborhoods in Puglia. This file further enhances the spatial analysis, providing a geographic representation of neighborhoods for a more comprehensive understanding.

The availability of these diverse datasets enables us to perform a thorough investigation into the determinants of Airbnb pricing in Lecce, combining both quantitative and spatial analyses for a comprehensive perspective.

Data Preprocessing

Prior to conducting the analysis, the raw dataset underwent a series of preprocessing steps to ensure data quality and appropriateness for the intended investigation. The key preprocessing steps are outlined below:

- **Selection of the municipality:** From listing.csv and neighbourhoods.geojson we select only the rows relating to the bnbs in the municipality of Lecce and the geometry of the city, as the search is based on this area.
- **Selection of Relevant Variables:** The dataset was refined by selecting a subset of variables deemed essential for the analysis. The selected variables include `id`, `host_response_rate`, `host_has_profile_pic`, `host_identity_verified`, `room_type`, `host_listings_count`, `host_total_listings_count`, `accommodates`, `neighbourhood_cleansed` `bathrooms_text`, `bedrooms`, `beds`, `price`, `minimum_nights`, `maximum_nights`, `number_of_reviews`, `review_scores_rating`, `review_scores_location`, `review_scores_value`, `longitude`, `host_acceptance_rate` and `latitude`. Other columns such as host url, host name or description were not taken into account.
- **Price Conversion:** The `price` variable, originally formatted with currency symbols (i.e. $117.00), was converted to numeric format. This involved removing non-numeric characters.

- **Host Response Rates and Host Acceptance Rates Conversion:** The values were converted to a numeric percentage by removing percentage symbols. (i.e. values like 90% became 0.9).
- **Bathroom Text Standardization:** The `bathrooms_text` column, which contained varied entries like "Half-bath," "Private half-bath," and "Shared half-bath," was standardized. Entries representing half-baths were replaced with 0.5, and the column was converted to a numeric format. (i.e. 2 baths became 2).

These preprocessing operations aimed to create a refined and standardized dataset, ready for subsequent analyses. The standardized variables and cleaned dataset provide a solid foundation for exploring the determinants of Airbnb pricing in the context of Lecce.

3.1 Variables explained

1. **id:** A unique identifier assigned to each Airbnb listing. This variable is crucial for distinguishing and referencing individual listings in the analysis.
2. **host_response_rate:** The percentage of messages a host responds to. This metric reflects the host's responsiveness and engagement with potential guests.
3. **host_has_profile_pic:** A binary indicator (yes/no) signaling whether the host has a profile picture. This can contribute to the guest's perception of the host's legitimacy and professionalism.
4. **host_identity_verified:** A binary indicator indicating whether the host's identity has been verified by Airbnb. Verified hosts may be perceived as more trustworthy.
5. **room_type:** Categorization of the type of accommodation offered (e.g., entire home/apartment, private room, shared room). This variable captures the nature of the space provided by the host.
6. **host_listings_count:** The number of listings a host currently has. This variable provides insights into the scale of a host's operation.
7. **host_total_listings_count:** Similar to host_listings_count, providing information on the total number of listings a host has had over time.
8. **neighbourhood_cleansed:** The specific neighborhood or locality within Lecce where the Airbnb listing is situated. This variable captures the geographical location of the accommodation.
9. **accommodates:** The number of guests the listing can accommodate. This is a key factor influencing pricing, as larger accommodations may command higher prices.
10. **bathrooms_text:** Description of the number and type of bathrooms in the listing.
11. **bedrooms:** The number of bedrooms in the accommodation. This is a crucial factor influencing the capacity and perceived value of the listing.
12. **beds:** The number of beds available in the listing. Similar to bedrooms, this variable contributes to the overall capacity and pricing.
13. **price:** The nightly rate set by the host for the listing. This is the dependent variable in the regression analysis, representing what the study aims to understand.
14. **minimum_nights:** The minimum number of nights a guest must book to stay at the listing. Hosts often use this variable to manage booking durations.
15. **maximum_nights:** The maximum number of nights a guest can book for a stay. This variable, like minimum_nights, contributes to the host's booking policies.
16. **number_of_reviews:** The total number of reviews received by the listing. This variable reflects the level of guest engagement and satisfaction.

17. **review_scores_rating:** The overall rating score assigned to the listing by guests. It provides a quantitative measure of guest satisfaction.

18. **review_scores_location:** A specific aspect of the review scores, focusing on the perceived location of the listing.

19. **review_scores_value:** Another aspect of the review scores, emphasizing the perceived value of the accommodation.

20. **host_acceptance_rate:** The percentage of booking requests accepted by the host. This variable indicates the host's willingness to accept reservations.

21. **longitude and latitude:** Geographic coordinates of the Airbnb listing. These variables are essential for the spatial analysis, providing the precise location of each accommodation within Lecce.

Defining Territory and Spatial Grid

To enable a spatial analysis of Airbnb listings within Lecce, the territory was defined and segmented into a grid of hexagonal cells. The procedures involved in this spatial processing were executed using the `sf` package in R.

The perimeter of the Lecce municipality was delineated by extracting the boundary geometry of neighborhoods.

A hexagonal grid, covering the defined perimeter, was generated using the `st_make_grid` function. The grid was configured with a cell size of 0.007 in both longitude and latitude directions to ensure comprehensive coverage of the territory.

The intersection of the hexagonal grid with the defined neighborhood perimeter ensured the proper alignment of the hexagonal grid with the boundaries of neighborhoods within Lecce.

The R code utilized for these operations is presented below:

```
plot(neighborhoods_sf$geometry)
perimeter <- st_union(st_boundary(neighborhoods_sf$geometry))
perimeter <- st_cast(st_union(st_geometry(neighborhoods_sf)), "POLYGON")
grid <- st_make_grid(perimeter, cellsize = c(0.007, 0.007), what= "polygons", square=F)
result_grid_neighborhoods <- st_intersection(perimeter, grid)
plot(result_grid_neighborhoods, main = "Intersection of Perimeter and Grid")
```

The resulting hexagonal grid, aligned with the neighborhood boundaries, serves as the basis for spatially grouping Airbnb listings within each cell. This spatial structure enables a more nuanced exploration of the geographical distribution of listings and their relationships to various factors. The hole we see in the center is the municipality of Surbo, which, being a separate municipality, will not be used in the analysis despite its proximity.

Summarization

The Airbnb listings were transformed into spatial elements using longitude and latitude coordinates, resulting in a spatial dataframe (`listings_sf`), and the coordinate reference system (CRS) was set to match that of the hexagonal grid (WGS 84). In the end each bnb was associated with the corresponding hexagon based on whether its coordinates are within that polygon or not.

```
listings_sf <- st_as_sf(listings_df, coords = c("longitude", "latitude"),
                crs = st_crs(result_grid_neighborhoods))
result_grid_neighborhoods <- st_make_valid(result_grid_neighborhoods)
listings_sf$GRID_ID <- st_within(listings_sf, result_grid_neighborhoods)
```

The listings were then summarized within each hexagonal cell, considering the median values for the various numerical attributes and the mode for the categorical variables. The median is a robust statistic, meaning it is less affected by extreme values. If there are individual values within a cell that deviate greatly from the others, the average may be unstable while the median provides a more stable estimate. In the end we will have a single point for each cell which summarizes all the bnbs relating to it.

```
# R Code:
summary_data <- listings_sf %>%
  group_by(GRID_ID) %>%
  summarise(
    # Median values for various attributes
    host_response_rate = median(host_response_rate, na.rm = TRUE),
    host_acceptance_rate = median(host_acceptance_rate, na.rm = TRUE),
    host_listings_count = median(host_listings_count, na.rm = TRUE),
    ...
    price = median(price, na.rm = TRUE),
    host_has_profile_pic = names(table(host_has_profile_pic))
                           [which.max(table(host_has_profile_pic))],
    host_identity_verified = names(table(host_identity_verified))
                           [which.max(table(host_identity_verified))],
    room_type = names(table(room_type))[which.max(table(room_type))],
    bnb_density = n()
  )
```

We show the municipality of Lecce divided by hexagons where each point represents the bnb that summarizes that cell.

```
tm_polygons <- tm_shape(result_grid_neighborhoods) +
  tm_borders(col = "lightblue")

tm_points <- tm_shape(summary_data) +
  tm_bubbles(size = 0.05, col = "red")
```

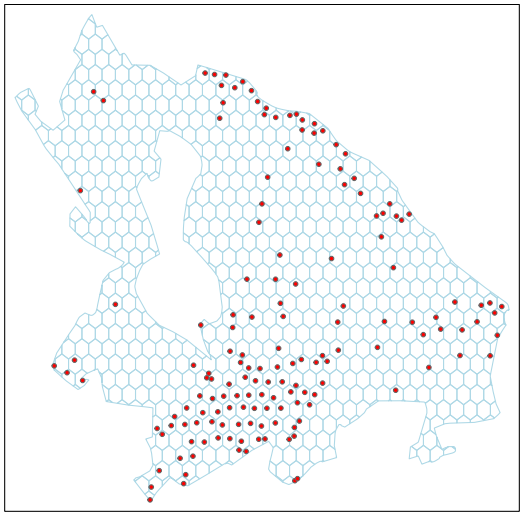

Figure 2: Lecce with points

These summarization steps set the stage for further analysis, providing a concise and representative overview of Airbnb listings within each hexagonal cell in Lecce.

Spatial Analysis: Room Type, Density, and Location

In this section, we present spatial visualizations that provide insights into key aspects of Airbnb listings in Lecce. Each map focuses on different factors, offering a comprehensive view of the geographical distribution of listings and their characteristics.

Room Type Distribution

The map illustrates the distribution of room types across Lecce's Airbnb listings. Hexagonal cells are shaded according to the predominant room type, allowing for an easy identification of areas with varying accommodation types.

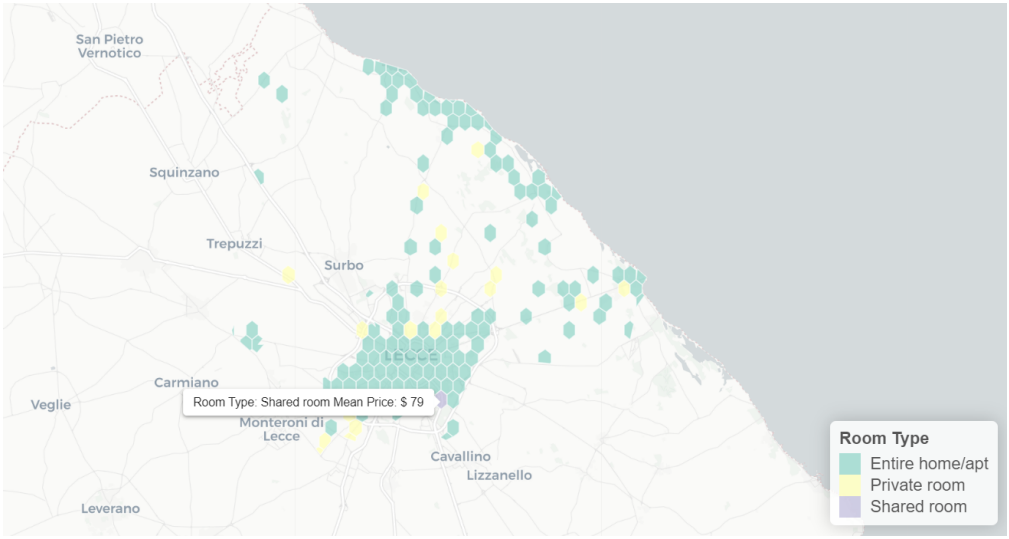

Figure 3: distribution of room types across Lecce's Airbnb

- Entire home/apt: 86.54%
- Private room: 12.82%
- Shared room: 0.64%

The majority of listings in Lecce are entire homes or apartments, followed by private rooms, and a small percentage of shared rooms. In principle, the type of room could greatly influence the price, in general the prices for full apartments are higher than for private rooms. As observed from the graph, the average price for entire homes is 105$ while for private ones it is 101$. Even less for shared (79$), but given the dominance of Entire home (especially in the city center where there are all of the entire apartments) we cannot be sure this result is significant.

Density of Listings

The second map showcases the density of Airbnb listings within the hexagonal cells. It provides an overview of areas with higher concentrations of listings, aiding in the identification of cells with a higher availability of accommodations.

Figure 4: density of Airbnb listings within the hexagonal cells

As we can see, most bnb's are located in the city centre, where tourism is assumed to be greater than in neighbouring or on the outskirts of the city.

Location-Based of Listings

The third map focuses on the distribution of review scores for location. It also features markers for notable landmarks within Lecce, providing additional context to the spatial analysis.

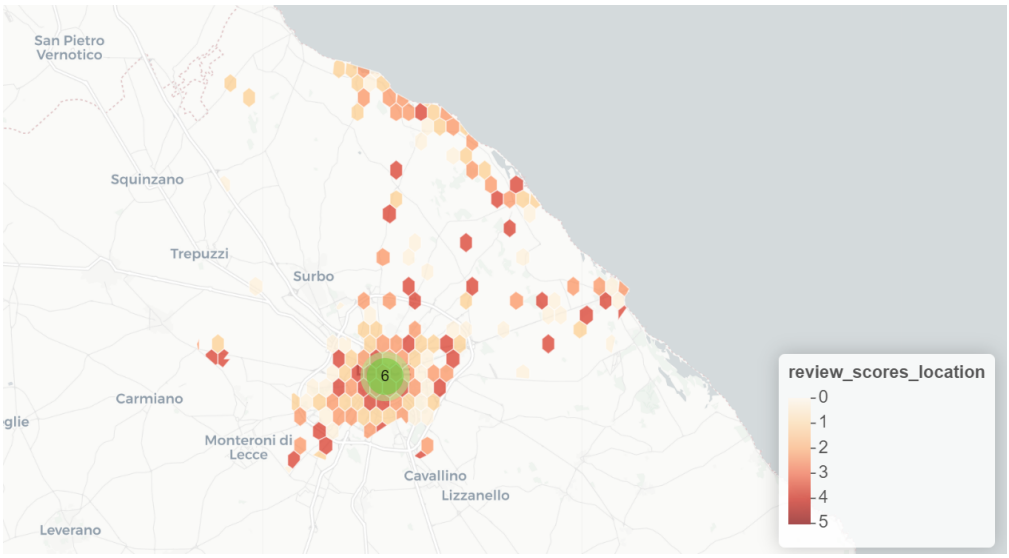

Figure 5: distribution of review scores for location

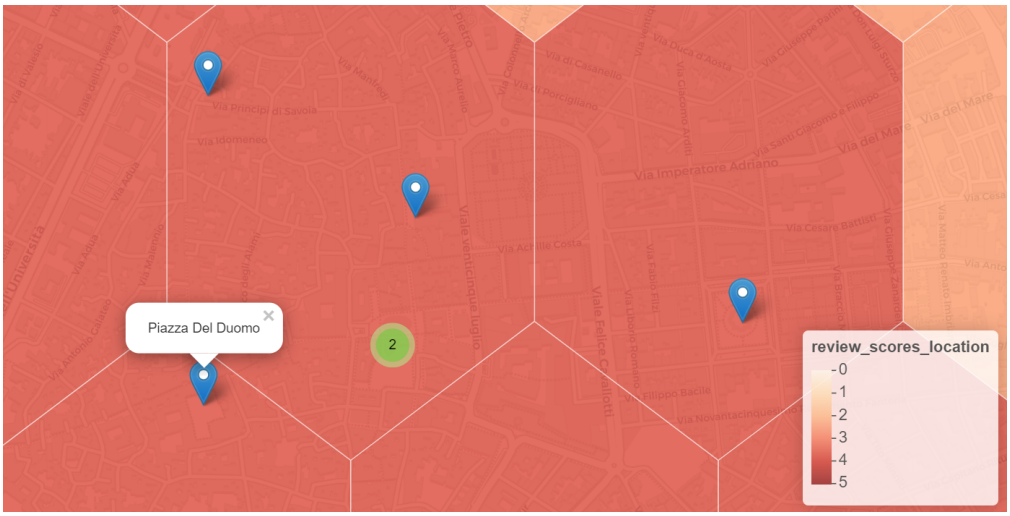

Figure 6: distribution of review scores for location

These two plots cluster the main tourist attractions. By clicking on the cluster we can see where they are located, their name and the score associated with their area. As we notice they are all located in the highest score areas as they constitute the essence of Lecce and contribute together with the historic centre to making it one of the best tourist destinations. We conclude that geographic proximity to different points of interest are relevant in the Airbnb's review scores for location.

These visualizations contribute to a spatial understanding of Airbnb listings in Lecce, helping to identify patterns, clusters, and trends within the city.

Creating Spatial Weights Matrices

The first crucial step in the analysis involves the creation of spatial weights matrices. This process begins by establishing the neighborhood relationships among spatial units. To statistically test whether the factors associated with location significantly affect apartment prices in Lecce, various spatial weights matrices were specified.

Defining Spatial Neighbours

Initially, each spatial unit is assigned a unique reference spatial coordinate. This is achieved by utilizing both polygons and points within the dataset.

```
gdf_polygons <- st_geometry(result_grid_neighborhoods)
gdf_points <- st_as_sf(summary_data, coords = c("longitude", "latitude"))
coords <- st_geometry(gdf_points)
st_crs(coords) <- st_crs("+proj=longlat +datum=WGS84")
```

k-Nearest Neighbours

The k-nearest neighbors criterion is applied to define neighborhood relationships among spatial units. The parameter k determines the number of neighbors for each spatial unit. Obviously, the higher the k, the richer the final graph.

```
# k=1
knn1 <- knn2nb(knearneigh(coords, k = 1))
plot(st_geometry(result_grid_neighborhoods), border = "lightblue")
plot(knn1, coords, add = TRUE, pch = 20, cex = 0.1, col = "red")

#k = 5
knn5 <- knn2nb(knearneigh(coords, k = 5))
plot(st_geometry(result_grid_neighborhoods), border = "lightblue")
plot(knn5, coords, add = TRUE, pch = 20, cex = 0.1, col = "red")
```

(a) k-Nearest Neighbours k=1

(b) k-Nearest Neighbours k=5

Figure 7: k-Nearest Neighbours comparison

Critical cut-off neighborhood

The critical cut-off neighborhood criterion considers two spatial units as neighbors if their distance is equal to or less than a fixed distance, representing a critical cut-off. The cut-off distance is determined based on the minimum threshold distance that ensures each spatial unit has at least one neighbor.

```
max_distance <- max(unlist(nbdists(knn1, coords)))
```

After having discovered that the minimum threshold is 4.11 km we plot two different graphs with an increasing cut-off distance. As the cut-off distance increases, the number of links among spatial units grows rapidly, visually represented in the plots.

Figure 8: critical cut-off neighborhood comparison

These neighborhood definitions lay the foundation for constructing spatial weights matrices, essential for subsequent spatial econometric analyses.

Defining Spatial Weights

Once the neighborhood relationships among the observations have been defined, the next step is to create the spatial weights matrix. This matrix encapsulates the spatial relationships between different spatial units.

Row-Standardized Spatial Weights Matrix

For each critical cut-off neighbors list previously created, row-standardized spatial weights matrices are generated. This ensures that the sum of weights for each spatial unit is equal to 1.

```
dnb4.listw <- nb2listw(dnb4,style="W",zero.policy=F)
dnb5.listw <- nb2listw(dnb5,style="W",zero.policy=F)
dnb6.listw <- nb2listw(dnb6,style="W",zero.policy=F)
```

Building Free-Form Spatial Weight Matrices

We can also construct free-form spatial weight matrices, using different distance functions among our coordinates. Here, we illustrate some examples:

```
distM <- st_distance(coords)
class(distM) <- "matrix" #distance matrix
```

$$W_1 = \frac{1}{1 + \text{distM}} \tag{1}$$

$$W_2 = \frac{1}{1 + \text{distM}^2} \tag{2}$$

$$W_3 = \frac{1}{1 + |\text{distM}|} \tag{3}$$

These free-form spatial weight matrices offer flexibility in capturing spatial relationships based on various distance functions. We concluded that the WM with the first function is the most appropriate spatial weight matrix to evaluate the spatial structure of Airbnb prices. The choice is dictated by the minimum p-value (about 0.03) compared to the other Moran's I test configurations which we will explain in the next section.

4 Spatial Autocorrelation Analysis

To assess the presence of spatial autocorrelation in Airbnb pricing within the Lecce municipality, we conducted a Moran's I test. This statistical test examines whether there is a spatial pattern in the distribution of prices, indicating whether similar prices cluster together or are dispersed.

4.1 Moran's I Test Results

The Moran's I test under normality was applied to the Airbnb pricing data, using the row-standardized spatial weights matrix `listW1s`. The results of the test are as follows:

```
        Moran I test under normality

data:  merged_data$price
weights: listW1s

Moran I statistic standard deviate = 1.8681, p-value = 0.03088
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.0146204630     -0.0064516129      0.0001272441
```

Figure 9: Moran's I Test Results

```
        Monte-Carlo simulation of Moran I

data:  merged_data$price
weights: listW1s
number of simulations + 1: 1000

statistic = 0.01462, observed rank = 965, p-value = 0.035
alternative hypothesis: greater
```

Figure 10: Monte-Carlo simulation of Moran I

4.1.1 Interpretation

The Moran's I statistic standard deviate of 1.8681 corresponds to a positive spatial autocorrelation. The associated p-value of 0.03088 indicates statistical significance. Therefore, we reject the null hypothesis of no spatial autocorrelation, suggesting that there is a discernible spatial pattern in Airbnb pricing within the Lecce municipality. The positive Moran's I statistic (0.0146) implies that similar prices tend to be spatially clustered, indicating the presence of localized patterns in Airbnb pricing. This finding has important implications for understanding the geographic distribution of pricing dynamics in the Lecce area.

We also conducted a Monte-Carlo simulation of Moran I to further validate the results. The simulation was performed with 999 iterations. The Monte-Carlo simulation results align with the initial Moran's I test, confirming the presence of positive spatial autocorrelation in Airbnb pricing within the Lecce municipality.

4.1.2 Conclusion

The significant positive spatial autocorrelation identified in Airbnb pricing implies that neighboring accommodations tend to exhibit similar pricing patterns. This information contributes to a better understanding of the spatial dynamics that influence Airbnb pricing in Lecce. In the subsequent sections, we delve deeper into the exploration of factors contributing to this spatial autocorrelation.

4.2 Spatial Autocorrelation in OLS Residuals

To investigate the presence of spatial autocorrelation in the residuals of regression model, we performed a Moran's I test. This test serves as a diagnostic tool to assess whether the residuals exhibit a spatial pattern.

First let's add two variables that could be a good predictor for the price. The distance from the historic center of Lecce, using Piazza Sant'oronzo as a point as it can be considered the center of Lecce, and the distance from the beach (taking Frigole as the location).

4.2.1 Studentized Residuals

```
pso <- st_sfc(st_point(c(18.172443, 40.353178)), crs = 4326)
merged_data_sf <- st_as_sf(merged_data, coords = c("longitude", "latitude"), crs = 4326)
merged_data_sf$distance_from_pso <- st_distance(merged_data_sf, pso)
merged_data_sf$distance_from_pso <- as.numeric(merged_data_sf$distance_from_pso) / 1000
merged_data$distance_from_pso= merged_data_sf$distance_from_pso

frig <- st_sfc(st_point(c(18.253591,40.4303882)), crs = 4326)
merged_data_sf <- st_as_sf(merged_data, coords = c("longitude", "latitude"), crs = 4326)
merged_data_sf$distance_from_frig <- st_distance(merged_data_sf, frig)
merged_data_sf$distance_from_frig <- as.numeric(merged_data_sf$distance_from_frig) / 1000
merged_data$distance_from_frig= merged_data_sf$distance_from_frig

linear_model <- lm(price ~ host_response_rate+ host_listings_count+
                     accommodates +bedrooms+beds+number_of_reviews+
                     review_scores_rating+host_identity_verified+
                     room_type +review_scores_location + review_scores_value
                   + host_has_profile_pic + maximum_nights + minimum_nights
                   + bathrooms_text +distance_from_pso + distance_from_frig , merged_data)
```

We examined the studentized residuals to visually inspect any spatial dependence. The map below illustrates the distribution of studentized residuals across the Lecce municipality:

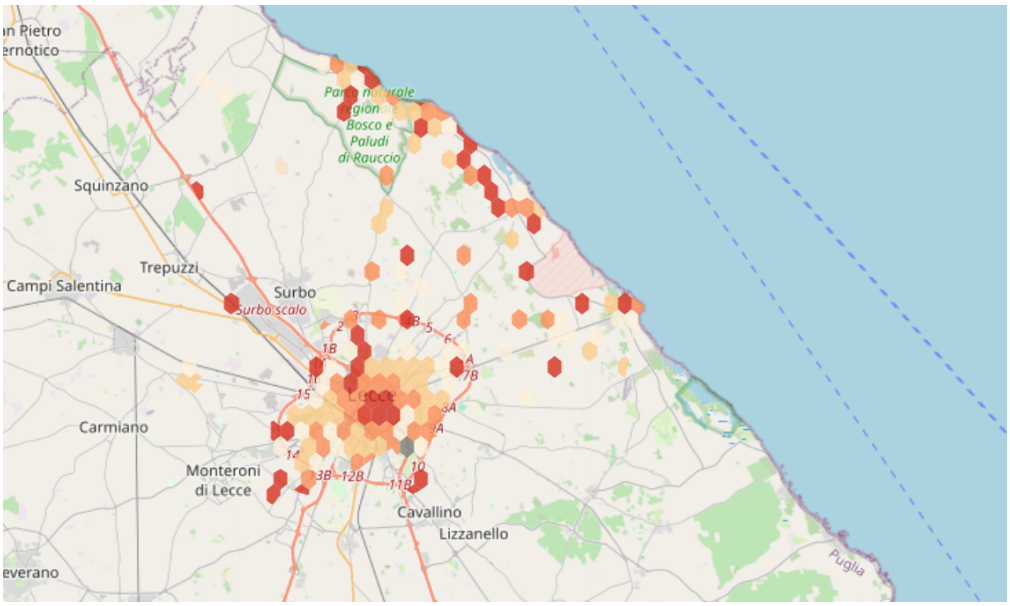

Figure 11: Spatial dependence in the residuals

The darker the cell, the larger the residual. The map suggests that there may be some spatial dependence in the residuals, with higher residuals concentrated in certain areas of the city. The darkest red cells are concentrated in the city center, suggesting that the residuals are highest in this area.

One possible explanation for this spatial dependence is that Airbnb rental prices are not evenly distributed across the city. For example, areas with a high concentration of Airbnb rentals may have higher prices due to increased demand (we will see whether in the city center, where we saw there is a greater density of bnb, the prices are higher). This could lead to higher residuals in these areas.

Another possible explanation for the spatial dependence is that there are other factors that are not included in the linear regression model or the relationship of our predictor with price is nonlinear, which means that a linear model would not be able to capture their true effect. This could lead to higher residuals, as the model would not be able to fit the data as well.

To further investigate the spatial dependence in the residuals, we need to use a spatial regression model. This would allow us to take into account the spatial distribution of the data and to obtain more accurate estimates of the model parameters. The overall distribution of red cells is somewhat clustered, suggesting that there may be some spatial autocorrelation in the residuals.

4.2.2 Moran's I Test on OLS Residuals

To formally test for spatial autocorrelation in the OLS residuals, we utilized the Moran's I test. We decided to apply a permutation bootstrap approach.

```
DATA PERMUTATION


Call:
boot(data = residuals(linear_model.lmx), statistic = MoraneI.boot,
    R = 999, sim = "permutation", listw = listW1s, n = length(listW1s$neighbours),
    S0 = Szero(listW1s))


Bootstrap Statistics :
       original       bias    std. error
t1* 0.001257802 -0.007771392  0.01102733
```

Figure 12: Bootstrap Results for Moran's I Test on OLS Residuals

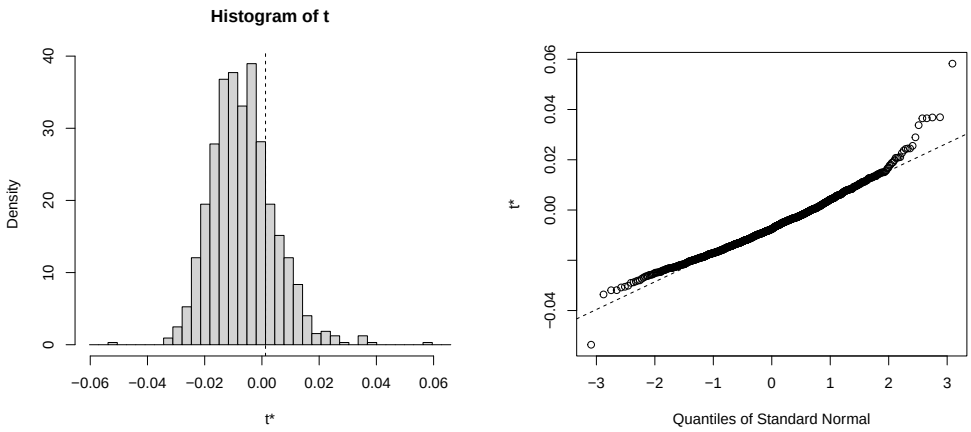

Figure 13: Bootstrap Results

Original Statistic ($t1^*$): The original Moran's I statistic calculated on the actual residuals from the linear regression model (0.00125).

Bias: The bias represents the difference between the average value of the bootstrapped test statistics and the original statistic. In our results, the bias is -0.0076. A negative bias suggests that the original statistic may have been overestimated.

Standard Error: The standard error provides a measure of the variability of the bootstrapped test statistics. In your results, the standard error is 0.0113. A smaller standard error indicates less variability and greater precision in the estimation.

Conclusions: The bootstrap results indicate that the original Moran's I statistic is positive (0.00125), suggesting a positive spatial autocorrelation in the OLS residuals. However, the bias is negative, indicating a potential overestimation of the original statistic.

In summary, the bootstrap results give us additional information about the reliability and stability of the Moran's I estimate and suggest the presence of positive spatial autocorrelation in the OLS residuals.

4.3 Moran Scatterplot and Hat Values

The Moran scatterplot is a graphical representation used to explore the spatial autocorrelation between a variable of interest and its corresponding spatially lagged values. The four quadrants of the Moran scatterplot categorize regions into "High-High," "Low-Low," "High-Low," and "Low-High" based on their values of the variable of interest and its corresponding spatially lagged values. What lies in the high-high or low-low region are the points with positive spatial autocorrelation.

The Moran scatterplot for the variable of interest (*e.g., price*) is shown in Figure 14.

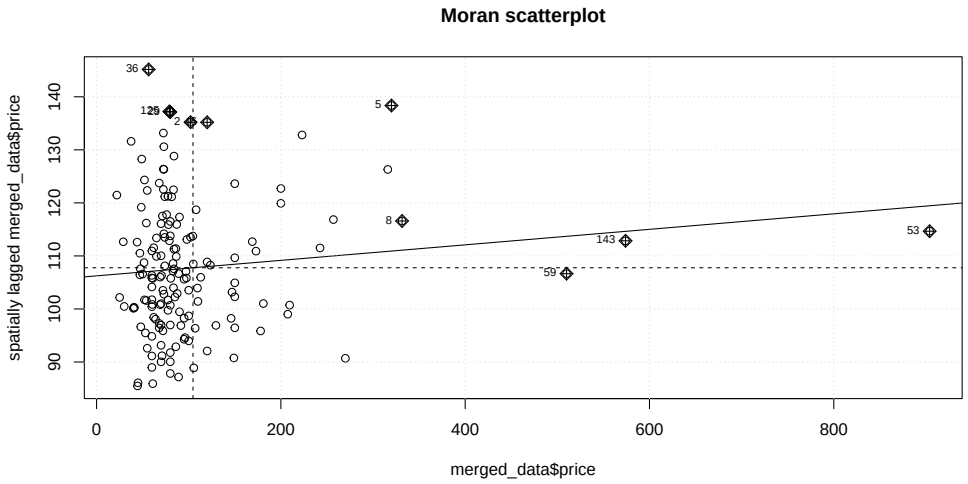

Figure 14: Moran Scatterplot

In Figure 14, the majority of the points fall in the High-High or Low-Low quadrants, indicating strong positive spatial autocorrelation. This means that regions with high variable of interest values tend to be surrounded by other regions with high variable of interest values, and regions with low variable of interest values tend to be surrounded by other regions with low variable of interest values.

The marked points on the plot represent influential observations that significantly impact the overall Moran's I statistic.

To further visualize the influence of points on the Moran scatterplot, the Hat Values are mapped onto a geographic representation. Figure 15 illustrates the distribution of Hat Values across different points.

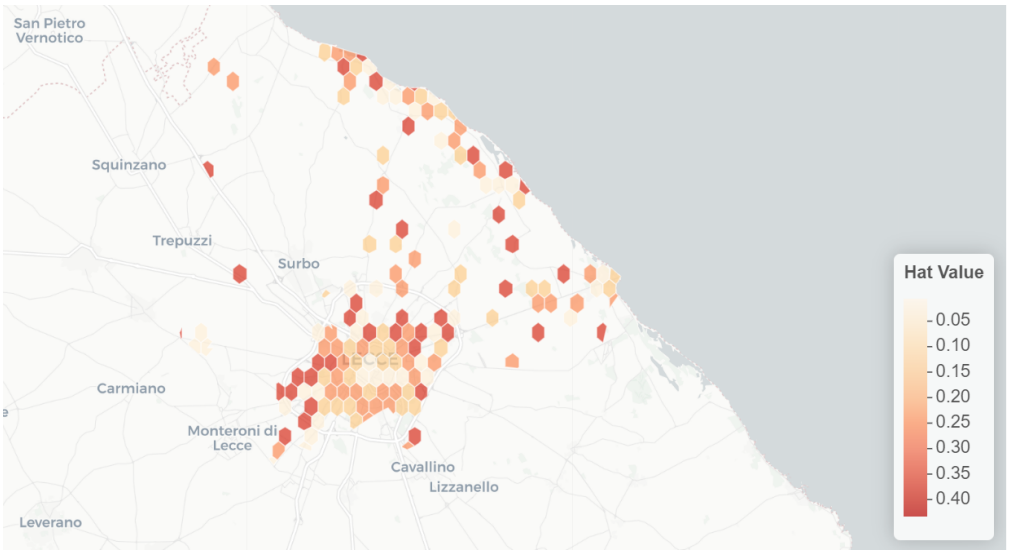

Figure 15: Map of Hat Values

In Figure 15, regions are colored according to their Hat Values, providing a spatial perspective on the influence of each point in the context of the Moran scatterplot. The map of hat values can be used to identify influential observations that may be driving the spatial autocorrelation pattern. Overall, the map of hat values suggests that the observations in the central and eastern parts of the study area have the greatest influence on the overall Moran's I statistic. This means that these regions are particularly important for understanding the spatial pattern of positive spatial autocorrelation in the dataset.

Another tool used to identify influential points is the Leverage-Residual Squared plot. It computes the product of leverage values and squared residuals for each observation. Points with higher values in this plot have a greater impact on the regression model.

Figure 16 displays the Leverage-Residual Squared Plot.

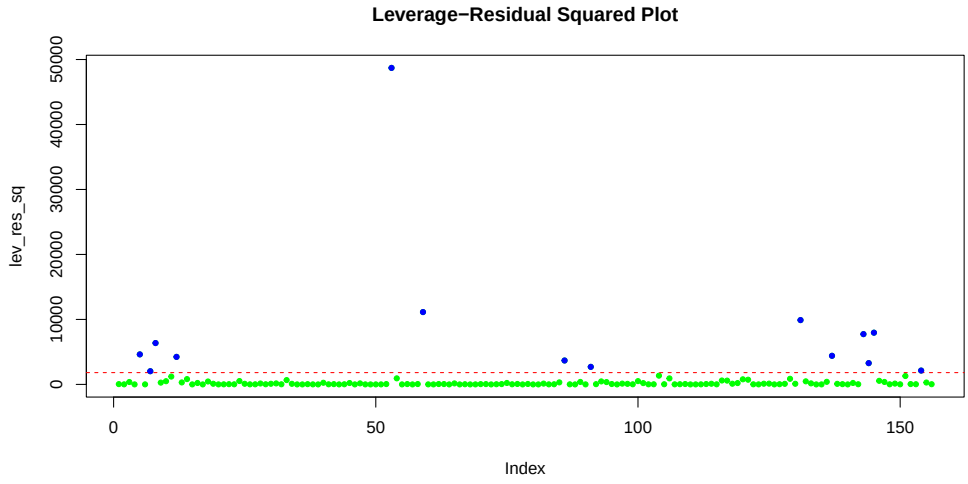

Figure 16: Leverage-Residual Squared Plot

In Figure 16, each point represents an observation, and its position on the plot is determined by the product of leverage and squared residual values. The red dashed line indicates the threshold used to identify influential points. Observations above this threshold are considered influential and are highlighted in blue.

4.3.1 Mapping points with Noteworthy Influence

In addition to identifying influential points using the Leverage-Residual Squared Plot, it can be valuable to map points with noteworthy influence based on their quadrant in the Moran Scatterplot.

To achieve this, the following steps are taken:

1. The Moran Scatterplot is generated and influential points are identified.
2. Spatially lagged values of the variable of interest are obtained for the influential points.
3. Each influential point is assigned to the proper quadrant in the Moran Scatterplot based on the comparison of variable of interest and spatially lagged values.

Figure 17 illustrates the map of regions with noteworthy influence based on their quadrant in the Moran Scatterplot.

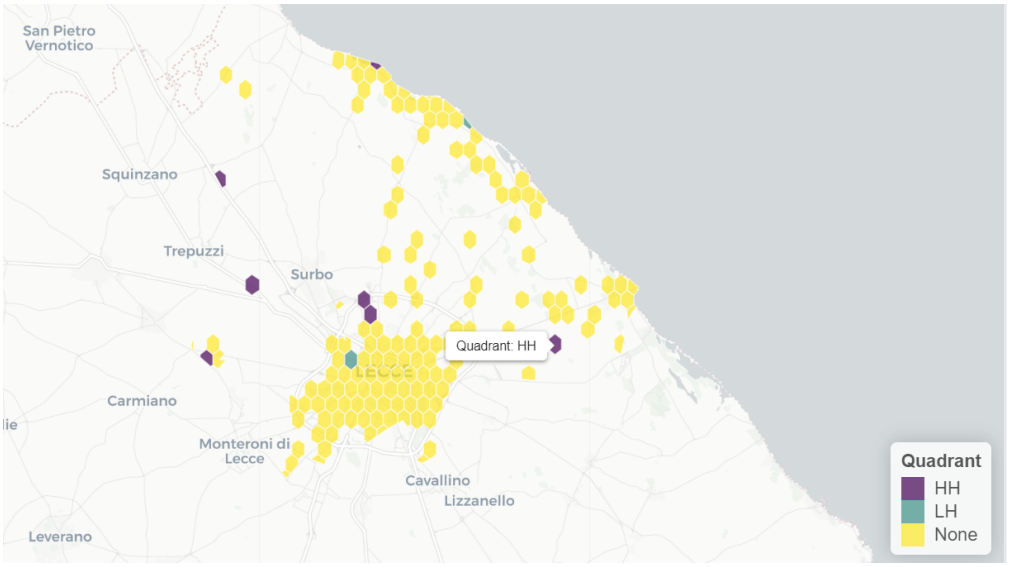

Figure 17: Map of points with Noteworthy Influence

In Figure 17, regions are color-coded based on their quadrant in the Moran Scatterplot. Each quadrant represents a specific spatial association pattern, such as High-High (HH), High-Low (HL), Low-Low (LL), and Low-High (LH).

4.4 Local Moran's I Analysis

The Moran Scatterplot provides a visual representation of local patterns of spatial association, but to assess the statistical significance of these patterns, we use the Local Moran's I index. Local Moran's I is a spatial statistic that measures the degree of spatial autocorrelation at a specific location, while taking into account the values of the variable at neighboring locations.

```
lmI <- localmoran(merged_data$price, listW1s)
```

The resulting `lmI` object contains the Local Moran's I index values for each spatial unit. Furthermore, a leaflet map can be created to illustrate the spatial distribution of Local Moran's I values:

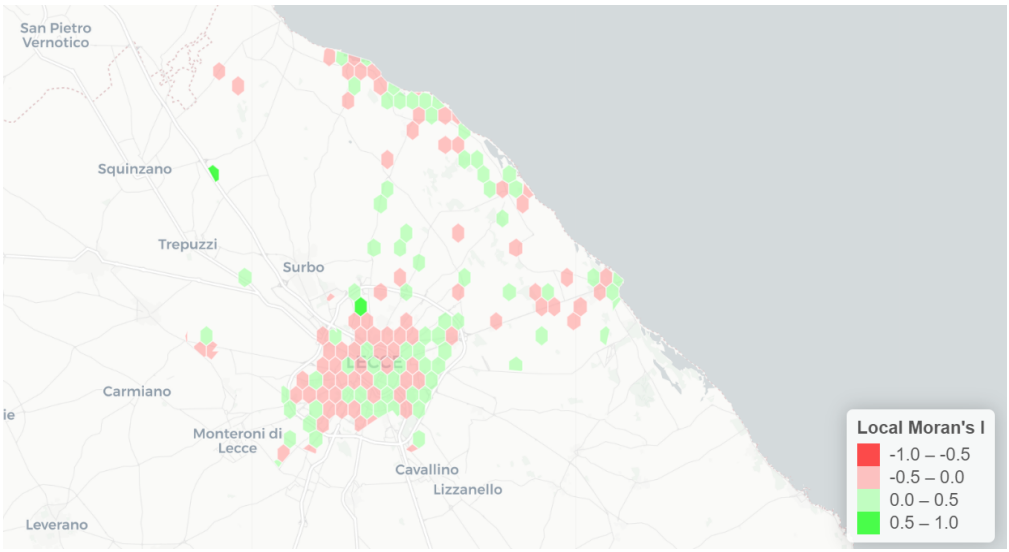

Figure 18: spatial distribution of Local Moran's I values

In the leaflet map, each spatial unit is color-coded based on its Local Moran's I value with high values (positive or negative) indicated by brighter colors and low values indicated by darker colors, providing insights into the local patterns of spatial autocorrelation.

4.4.1 Testing Local Moran's I Significance

The local Moran's I statistics obtained from the analysis can be subjected to hypothesis testing to determine the significance of the observed local spatial patterns. The leaflet map displays the adjusted p-values, allowing us to visually identify areas with significant local spatial patterns. The color classification provides an intuitive representation of the statistical significance of local Moran's I values across different spatial units.

The map reveals that there are several areas with statistically significant local spatial patterns in bnb prices. For example, near Squinzano and Surbo there are low adjusted p-values, indicating that the positive spatial autocorrelation in bnb prices in these areas is statistically significant. This means that high-priced bnbs in these areas are typically clustered together.

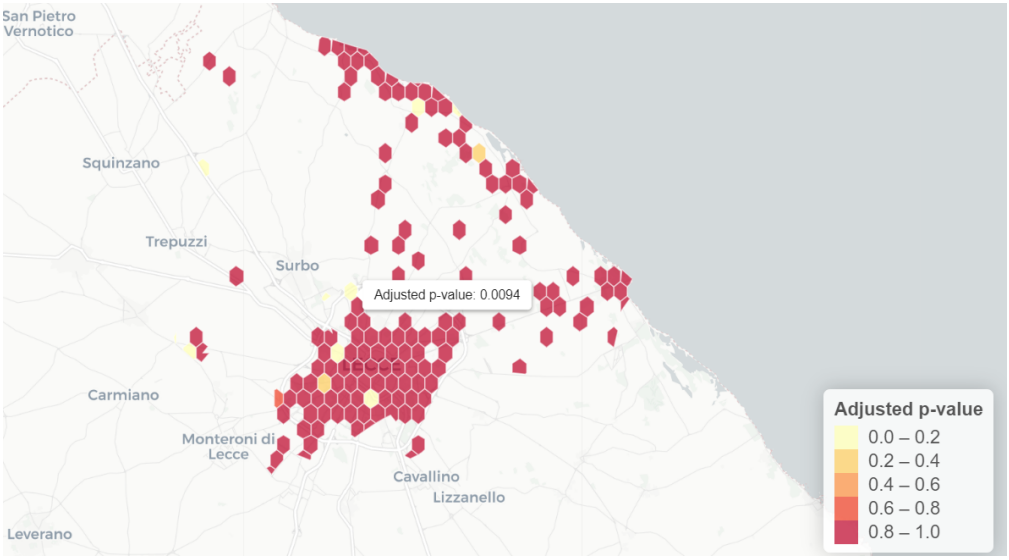

Figure 19: significance of local Moran's I values

5 Choosing the Proper Specification

This section presents the Lagrange multiplier (LM) test of spatial dependence, aiding in the selection of an appropriate spatial specification for the model.

5.1 Lagrange multiplier

We started from the general OLS estimation model then we computed LM tests to check for the existence of spatial structures in the model. The conventional Ordinary Least Squares (OLS) estimation method assumes independence among observations, as pointed out by Arbia [2006]. However, in the context of geographical data, this assumption may not hold true. The prices of two nearby apartments might exhibit a stronger relationship than those of two more distant apartments. This spatial proximity can lead to spatial dependence, where disturbances in the model show spatial autocorrelation, as emphasized by Anselin [1988] and reiterated by Pace and LeSage [2009]. Consequently, OLS estimates may become inefficient in the presence of spatial autocorrelation.

```
starting_model <- lm(price ~ host_response_rate+ host_listings_count+
                  accommodates +bedrooms+beds+number_of_reviews+
                  review_scores_rating+host_identity_verified+
                  room_type +review_scores_location + review_scores_value
                + host_has_profile_pic + maximum_nights + minimum_nights
                + bathrooms_text +distance_from_pso + distance_from_frig, merged_data)

lmtest_starting_model <- lm.LMtests(starting_model, listW1s,
                             test=c("all"))
summary(lmtest_starting_model)
```

The following points summarizes the results of the Lagrange multiplier (LM) diagnostics for spatial dependence on Ordinary Least Squares (OLS) in the context of the specified regression model:

- **LMerr (Lagrange Multiplier for Error Dependence):** The LMerr statistic is 0.0074847 and a p-value of 0.93106. The non-significant p-value suggests no evidence of error dependence.
- **LMlag (Lagrange Multiplier for Lag Dependence):** The LMlag statistic is 1.1305 and a p-value of 0.28767. This tests for spatial dependence in the residuals due to spatial autocorrelation. The p-value is not significant at the 0.05 level, indicating no strong evidence of lag dependence.
- **RLMerr (Robust Lagrange Multiplier for Error Dependence):** The RLMerr statistic is 2.9223 and a p-value of 0.08736. This is a robust version of LMerr that is less sensitive to outliers. Similar to LMerr, the p-value is not significant, suggesting no robust evidence of error dependence.
- **RLMlag (Robust Lagrange Multiplier for Lag Dependence):** The RLMlag statistic is 4.04541 and a p-value of 0.04429. This is a robust version of LMlag, testing for spatial autocorrelation in a robust manner. The p-value suggesting evidence of lag dependence.

5.2 Best specification

To address this spatial dependence, spatial models are employed. These models explicitly consider the spatial autocorrelation between data points, providing a more accurate representation of the underlying spatial relationships. In our study, we utilize a spatial model to elucidate the factors influencing BnB prices, leveraging data obtained from Airbnb and incorporating locational variables.

According to Elhorst [2010], in order to find the best specification under the ML estimation approach, we should:

1. Estimate an OLS model and then use the LM-test to verify if the SAR model or SEM are more proper.

2. If the OLS model is rejected in favor of only the SAR model, or only the SEM, or in favor of both models, then the SDM should be estimated.

3. Conduct a LRT to verify if SDM can be reduced to SAR (H0: $\theta = 0$); and conduct a LRT to verify if SDM can be reduced to SEM.

4. If both restrictions, H0: $\theta = 0$ and $\theta + \rho\beta = 0$, are rejected, then SDM best describes the data.

5. If only the SAR (or SEM) restriction cannot be rejected and the RLM-test points to the SAR (or SEM), then the SAR (or SEM) best describes the data; otherwise, the SDM is better.

6. If the procedure points to SEM, estimate an SDEM and verify if θ is statistically significant.

7. If the LM-test points to the OLS model, estimate an LDM model and then verify if θ is statistically significant.

After following these steps, and thus estimating SDM, we conduct LRT to check whether SDM can be reduced to SAR or SEM. Although LRT finds both models significant, we decide to use the SAR since the LM test had a lower pvalue in the RLMlag.

	Model	df	AIC	logLik	Test	L.Ratio	p-value
SDM	1	37	1829.2	-877.59	1		
SAR	2	20	1814.7	-887.35	2	19.521	0.29947

Figure 20: LRT (SDM-SAR)

5.3 SAR

Spatial Autoregressive (SAR) models are used to examine spatial dependencies in data, acknowledging that observations at one location may be influenced by neighboring locations. The key formula involves a spatial autoregressive coefficient (ρ), expressing how much the variable of interest at a given location depends on the same variable in nearby locations. The model is defined by:

$$Y = \rho WY + X\beta + \epsilon$$

Here, Y represents the dependent variable, ρ is the spatial autoregressive coefficient, W is the spatial weight matrix, X is the matrix of independent variables, β is the vector of coefficients, and ϵ is the error term. This formula captures the spatial interdependence crucial for understanding and predicting spatial phenomena. We can rewrite the SAR model in this way:

$$Y = (I - \rho W)^{-1}(X\beta) + \epsilon$$

And the matrix of partial derivatives of the expected values of y with respect to the kth explanatory variable is:

$$\begin{bmatrix} \frac{\partial E[y_1]}{\partial x_{1k}} & \frac{\partial E[y_1]}{\partial x_{2k}} & \cdots & \frac{\partial E[y_1]}{\partial x_{Nk}} \\ \frac{\partial E[y_2]}{\partial x_{1k}} & \frac{\partial E[y_2]}{\partial x_{2k}} & \cdots & \frac{\partial E[y_2]}{\partial x_{Nk}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial E[y_N]}{\partial x_{1k}} & \frac{\partial E[y_N]}{\partial x_{2k}} & \cdots & \frac{\partial E[y_N]}{\partial x_{Nk}} \end{bmatrix} = \tag{4}$$

.

$$= (I - \rho\mathbf{W})^{-1} \begin{pmatrix} \beta_k & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \beta_k & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \beta_k \end{pmatrix} \tag{5}$$

The partial derivatives matrix reveals insights into the implications of the SAR model. In this context, the matrix not only indicates the potential adjustment in the price of the specific BnB (referred to as the direct effect, observed in each diagonal element) but also highlights the impact on the prices of neighboring BnB (called as the indirect effect, residing in the off-diagonal elements of the partial derivatives matrix). This indirect effect can be interpreted as the consequences of a variation in an explanatory variable, extending its influence beyond the immediate BnB to affect the prices of other nearby BnB. Alternatively, it can be understood as the influence on the price of a specific apartment triggered by a modification in the explanatory variable of all other surrounding BnB (Pace and LeSage [2009]).

This is the summary of our SAR model:

```
Rho: 0.37665, LR test value: 1.0893, p-value: 0.29662
Asymptotic standard error: 0.29901
    z-value: 1.2597, p-value: 0.20779
Wald statistic: 1.5868, p-value: 0.20779

Log likelihood: -885.6495 for lag model
ML residual variance (sigma squared): 4983.4, (sigma: 70.593)
Number of observations: 156
Number of parameters estimated: 21
AIC: NA (not available for weighted model), (AIC for lm: 1812.4)
LM test for residual autocorrelation
test value: 0.65493, p-value: 0.41836
```

Figure 21: SAR summary

Rho (Spatial Autoregressive Coefficient): Rho is the spatial autoregressive coefficient, capturing the spatial dependence in the dependent variable. The estimated Rho is 0.3766. The LR test value is 1.0893, and the associated p-value is 0.29662, testing the null hypothesis that Rho is zero.

Wald Statistic for Rho: The Wald statistic for Rho is 1.5868 with a p-value of 0.2078, testing the null hypothesis that Rho is zero.

LM Test for Residual Autocorrelation: The LM test for residual autocorrelation has a test value of 0.65 with a p-value of 0.418, testing the null hypothesis that there is no residual autocorrelation.

Coefficients: We found few significant variables such as host_rate_rate, bathroom_text, accommodates and distance_from_frig.

Interpretation: The Rho coefficient (0.3766) suggests positive spatial autocorrelation in the dependent variable, indicating that the price in one location is positively correlated with the prices in neighboring locations. The LR test and Wald statistic for Rho provide mixed evidence regarding the significance of spatial lag dependence. The p-values (0.29662 and 0.2078, respectively) suggest that Rho is not statistically significant. The LM test for residual autocorrelation does not find significant evidence of autocorrelation in the residuals.

5.3.1 Simpler model

Let's try to simplify the model by eliminating the less significant variables such as minimum and maximum nights, room type and review scores location.

```
Rho: 0.39803, LR test value: 1.2719, p-value: 0.25941
Asymptotic standard error: 0.29243
    z-value: 1.3611, p-value: 0.17347
Wald statistic: 1.8527, p-value: 0.17347

Log likelihood: -885.9852 for lag model
ML residual variance (sigma squared): 5003.1, (sigma: 70.733)
Number of observations: 156
Number of parameters estimated: 16
AIC: NA (not available for weighted model), (AIC for lm: 1803.2)
LM test for residual autocorrelation
test value: 0.66721, p-value: 0.41403
```

Figure 22: Simplified SAR summary

The p-value of the LR test for the spatial lag coefficient decreased to 0.25941 in the simplified model. The p-value of the Wald Statistic decreased to 0.17347 in the simplified mode. Both models indicate the presence of spatial dependence (positive Rho), but the evidence is not so strong, it depends on the chosen significance level. Therefore, we cannot say with certainty that the price of each Airbnb accommodation is also infuenced by the prices of the Airbnb listings closest to it. This finding does not coincide with previous literature which examines the spatial autocorrelation structures in Airbnb's prices (López FA [2019]).

6 Impact Analysis

The following values show the estimated impact on the dependent variable due to changes in the corresponding independent variable. The effects are divided into direct effects (Direct impact), indirect impacts (impact through spatial relationships), and the total impact (sum of direct and indirect impacts).

	Direct	Indirect	Total
host_response_rate	-49.16950777	-32.23252044	-81.40202821
host_listings_count	-0.02766477	-0.01813533	-0.04580011
accommodates	-9.88764261	-6.48173344	-16.36937605
bedrooms	9.08990997	5.95878874	15.04869871
beds	4.11540420	2.69780716	6.81321136
number_of_reviews	-0.24758126	-0.16229913	-0.40988039
review_scores_rating	36.40804346	23.86688536	60.27492882
host_identity_verifiedTRUE	6.11424482	4.00812475	10.12236957
review_scores_value	-40.51105330	-26.55656753	-67.06762084
host_has_profile_picTRUE	15.07690636	9.88349720	24.96040356
bathrooms_text	119.46096382	78.31129765	197.77226147
distance_from_pso	1.85526691	1.21619946	3.07146637
distance_from_frig	3.06557613	2.00960412	5.07518025

Figure 23: Impacts

Notable variables with significant impacts include bathrooms_text, beds and bedrooms.

7 Conclusion

Concerning the control variables, our findings align with the anticipated directions established in prior literature. Specifically, we observed a positive correlation between the number of beds and bedrooms in each Airbnb listing, consistent with studies by Gibbs et al. [2018], Wang D [2017], and Lorde T [2019]. In contrast, the negative coefficient associated with accommodation suggests that there is a trend towards cost savings for larger group sizes, supporting the idea that more exclusive and private accommodation tends to command

higher prices. However, it is noteworthy that the majority of our results lack statistical significance. Reasons will follow regarding the non-significant results.

In the municipality of Lecce, the challenge of discerning significant spatial autocorrelation patterns in Airbnb prices is multifaceted and can be attributed to several factors unique to the local context. Unlike larger cities such as Barcelona, London, or Chicago, where distinct spatial patterns may emerge due to the enormous volume of accommodations and diverse neighborhoods, Lecce presents a scenario marked by its smaller scale and territorial dispersion. Lecce, with its population of around one hundred thousand inhabitants, stands out for its scattered territorial structure. This spatial distribution contributes to a nuanced and diverse landscape where the prices of Airbnb accommodations are influenced by various factors that might elude a linear model, despite being spatial. Unlike the assumed pattern in more centralized urban areas, Lecce's pricing dynamics are intricate, reflecting the municipality's distinctive geography and the variety of factors influencing accommodation costs. The municipality encapsulates diverse locales, such as San Cataldo on the sea or BnBs situated in the countryside, each offering unique experiences tied to Salento traditions. The small-town setting and the scattering of accommodations present challenges for capturing consistent patterns, especially given the varying characteristics of different areas within the municipality. Moreover, the temporal aspect plays a pivotal role in Lecce's Airbnb market dynamics. The city, particularly bustling during the summer, witnesses fluctuations in demand, with people often opting for accommodations closer to the sea (distance_from_frig and distance_from_pso encapsulate the distances from the center and the sea) rather than exclusively within the city center. This seasonal aspect further complicates the identification of consistent spatial autocorrelation patterns throughout the year. Furthermore, the price of bnbs, especially in summer, can change, seeing an increase for bnbs near the sea and a decrease in the city centre. So the results may change depending on the season. The visual representation of Airbnb prices in Lecce, as seen in the following graph, underscores the complexity of the pricing landscape. Unlike the conventional urban model, Lecce exhibits a pattern where prices in the city center are high, gradually decreasing in more suburban areas, only to rise again near the Adriatic. This intricate pattern reflects the diverse preferences of visitors who may prioritize proximity to the sea or embrace the allure of suburban tranquility over the conventional and chaotic city-center.

In summary, the absence of significant spatial autocorrelation results in Lecce's Airbnb pricing can be justified by the town's unique characteristics, including its smaller scale, territorial dispersion, diverse locales, and the seasonal nature of tourism. The spatial dynamics in Lecce deviate from the typical urban model, highlighting the need for nuanced modeling approaches that consider the intricate interplay of geographical, temporal, and cultural factors influencing Airbnb prices in this distinctive municipality.

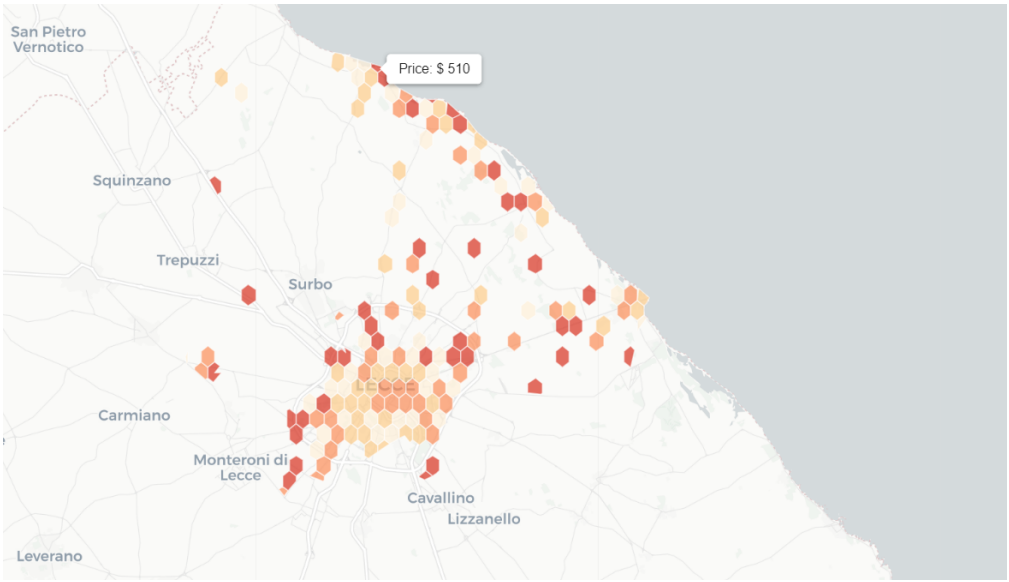

Figure 24: Price distribution

References

Inside Airbnb. Inside airbnb. *Inside Airbnb*, 2023.

L. Anselin. *Spatial econometrics: Methods and models*. Kluwer Academic Publishers, 1988.

G. Arbia. *Spatial econometrics: Statistical foundations and applications to regional convergence*. Springer Science & Business Media, 2006.

R. Deboosere, E. Claes, and R. Dekimpe. Airbnb in brussels: A tale of two cities. *Annals of Tourism Research*, 77:100–114, 2019.

J. P. Elhorst. *Spatial econometrics: From cross-sectional data to spatial panels*. Springer Science & Business Media, 2010.

Gibbs, Guttentag D, Gretzel, U Morton J, and Goodwill A. Pricing in the sharing economy: a hedonic pricing model applied to airbnb listings. *J Travel Tour Mark*, 35:46–56, 2018.

Weekes Q Lorde T, Jacob J. Price-setting behavior in a tourism sharing economy accommodation market: a hedonic price analysis of airbnb hosts in the caribbean. *Tourism Management Perspectives*, 30:251–261, 2019.

Mur J López FA, Minguez R. Ml versus iv estimates of spatial sur models: evidence from the case of airbnb in madrid urban area. *Annu Reg Sci*, 64:313–347, 2019.

R. K. Pace and J. P. LeSage. *Introduction to spatial econometrics*. CRC Press, 2009.

D. Perez-Sanchez, R. L. G. de Arce, and D. Iribarren. The impact of short-term rentals on housing affordability: Evidence from the spanish market. *Land Use Policy*, 76:161–170, 2018.

J. Van der Borg, P. Costa, and L. Van den Berg. Understanding the impact of airbnb as a new mode of accommodation: An analysis of recent trends in lisbon. *Tourism Management Perspectives*, 24:65–76, 2017.

Nicolau JL Wang D. Price determinants of sharing economy based accommodation rental: a study of listings from 33 cities on airbnb.com. *Int J Hosp Manag*, 62:120–131, 2017.