Analysis of airbnb prices in European cities: Case of Paris

Nelly AGOSSOU, Jean-Baptiste GOMEZ & Ulrich SEGODO

Student in Master 2 EBDS and MAG3; Aix-Mairseille School of Economics (AMSE)

Mid Term

Octobre 2023

Abstract:

The objective of this study is to analyze the determinants of Airbnb prices in European cities. For this we have data on ten major European cities: Amsterdam, Athens, Berlin, Budapest, Barcelona, Paris, Lisbon, London, Rome and Vienna but we choose to work on data relative to Paris. We have carried out descriptive statistics and correlation tests on our data to identify the variables related to our variable of interest. We also made a geographical representation of apartments in the city of Paris for certain variables (attractiveness index, price and restaurant index). Then we estimated a linear regression model (Parametric model) et compare to non-parametric model that we estimated .

Key words: Airbnb prices, geographical representation, linear regression model, non-parametric model.

---

1. Contexte

Rental services in European cities have become increasingly popular in recent years. To face the ever-increasing demand and provide some good offers, these services typically include short-term rentals of apartments or homes, which are available to travellers and tourists looking for alternative accommodations to hotels.

Among all the rental services available in Europe, Airbnb is one of the most popular, with listings in almost every major city. The platform allows hosts to rent out their homes, apartments, or spare rooms to travellers for short-term stays. Airbnb also offers experiences and activities that are hosted by locals, allowing travellers to have an authentic and unique experience.

Airbnb has revolutionized the travel industry by providing an affordable and personalized lodging option to travellers around the world. As the platform has grown in popularity, the prices of Airbnb accommodations have become an important factor for both hosts and guests. In this context, many scientific works tried to analyse the prices of Airbnb accommodations and their determinants.

Using ordinary least squares and quantile regression analysis, Wang and Nicolau (2017) investigated the determinants of prices in different categories (host attributes, property attributes, online review ratings, amenities and services, and rental rules) and found that all these elements are good predictors of Airbnb prices. These results are confirmed by many other authors, such as Guttentag (2019), Toader and al. (2022) which found that Airbnb prices are determined by the location of the room, transportation accessibility, property type and rental duration.

Especially, Gyodi and Nawaro (2021) showed the spatial dependence in the determinants of Airbnb prices and used some innovative indices to measure the attractiveness of neighbourhoods. Using different spatial models, they found that Airbnb prices are dependent of the characteristics of the neighbourhood and highly linked to the size of the rooms and its type.

Furthermore, by examining the pricing strategies of Airbnb hosts, this analysis can help hosts optimize their rental income and improve their competitiveness in the market. Ultimately, a thorough analysis of Airbnb prices in European cities can contribute to a better understanding of the sharing economy and its impact on the travel industry.

2. Data and Methods

2.1. Data

To analyse the determinants of Airbnb prices in European cities, we use a database available on the Kaggle website. The database is entitled Airbnb prices in European cities and was provided Gyodi and Nawaro who used it to analyse the determinants of Airbnb prices using a spatial econometrics approach.

Concretely, our database is a collection of 20 databases giving information on Airbnb prices in 10 of the most popular European cities: Amsterdam, Athens, Barcelona, Berlin, Budapest, Lisbon, London, Paris, Roma and Vienna. For each of these cities, information on the characteristics of the rooms offered for rent and their host were collected from Airbnb listing site. This information is completed by some indices related to the neighbourhood collected by the authors of the database from TripAdvisor data.

Table1: Listing attributes (Kaggle)

Variables	Description	Type
realSum	The total price of the Airbnb listing	Numeric
room_type	The type of room being offered (e.g., private, shared, etc.)	Categorical
person_capacity	The maximum number of people that can stay in the room	Numeric
host_is_superhost	Whether the host is a superhost or not	Boolean
multi	Whether the listing is for multiple rooms or not	Boolean
biz	Whether the listing is for business purposes or not	Boolean
cleanliness_rating	The cleanliness rating of the listing	Numeric
guest_satisfaction_overall	The overall guest satisfaction rating of the listing	Numeric
bedrooms	The number of bedrooms in the listing	Numeric
dist	The distance from the city centre	Numeric
metro_dist	The distance from the nearest metro station	Numeric
lng	The longitude of the listing	Numeric
lat	The latitude of the listing	Numeric
Attr_index	Attraction index of the neighbourhood (Scale to 100)	Numeric
Rest_index	Restaurant Index (Scale to 100)	Numeric
Day	Weekends or Weekdays	Categorical

The table below shows the distribution of listings and the distribution of prices in Paris

Table 2: Some Descriptive statistics for dependant variable Price of room

City	Count	Mean	SD	Min	Max
Paris	6688	392.5314	330.9497	92.7393	16445.61

It is retained from the table above that is also a very large difference between the minimum and maximum value of the price. Also, the standard deviation is very larger . We will therefore use the price logarithm as a dependent variable for estimating our model.

Table 3: Some descriptive statistics for quantitative variables in data set

City	Person capacity	Cleanliness	Bedrooms	Dist_centre	Dist_metro
Paris	2.95	9.26	0.97	2.99	0.23
	(1.22)	(0.97)	(0.64)	(1.46)	(0.12)

This table presents some descriptive statistics on some quantitative variables. We can conclude that in the city of Paris, the capacity of the rooms is on average 3 people. Also, Airbnb are located on average 2.99 kilometers from downtown and 0.23 kilometers from the metro.

Table 4: Some descriptive statistics for qualitative variables by cities

	Room type			Host_superhost		Multi-sites listing		Listing on bus_sites	
	Home/apt	Private room	Shared room	False	True	False	True	False	True
Paris	0.76	0.23	0.01	0.86	0.14	0.86	0.14	0.75	0.25

From the table above, we notice that most of the rooms listing on Airbnb platforms in Paris are homes or apartments. In contrast, there is a very little proportion of offers of shared rooms in these big cities. The table also tells that most of the hosts aren’t considered superhosts.

2.2 Methodology

Our objective is to analyse the determinants of Airbnb prices in Paris. To do so, we are using the available data on Airbnb listing prices and attributes. Our analysis is done in two steps. Firstly, we estimate parametric model (Ordinary least squares regression), then we estimate nonparametric model and compare the result of all estimation.

Parametric Models (Ordinary least squares regression)

Our first step consists in a linear regression using Ordinary least squares to evaluate how the different variables of our study affects prices in Paris. This analysis uses the database for Paris city. The model estimated is the following:

$$Y= \alpha + \mu X + \varepsilon$$

Where:

- Y represents our vector of interest variable (the prices of Airbnb rooms) ;
- α is the intercept of the model ;
- X is a vector of the listing's attributes ;
- ε is error term.

The attributes used in this model are determined by a correlation test to choose the best variables that influence our interest's variable. Given the distribution of prices and the high values of standard deviations, we choose to use le logarithm of prices in our analysis.

Non-Parametric Model:

The model estimated is the following:

$$Y= m (X) + \varepsilon ;$$

Where:

- Y represents our vector of interest variable (the prices of Airbnb rooms) ;
- m is a kernel function ;
- X is a vector of the listing's attributes that have nonlinear relationship with the dependant variable ;
- ε is error term.

Using GAM model, we estimate assume that the function m is not the same for all explanatory variables and estimed the following model:

$$Y = m_1(x_1) + m_1(x_2) + \ldots\ldots\ldots\ldots + m_k(x_k) + \epsilon$$

3. Results

The following section presents the results of our analysis with R.

3.1. Some descriptive statistics

The graph bellow displays the relationship between Airbnb prices and the distance to the city-centre and the subway in our European cities.

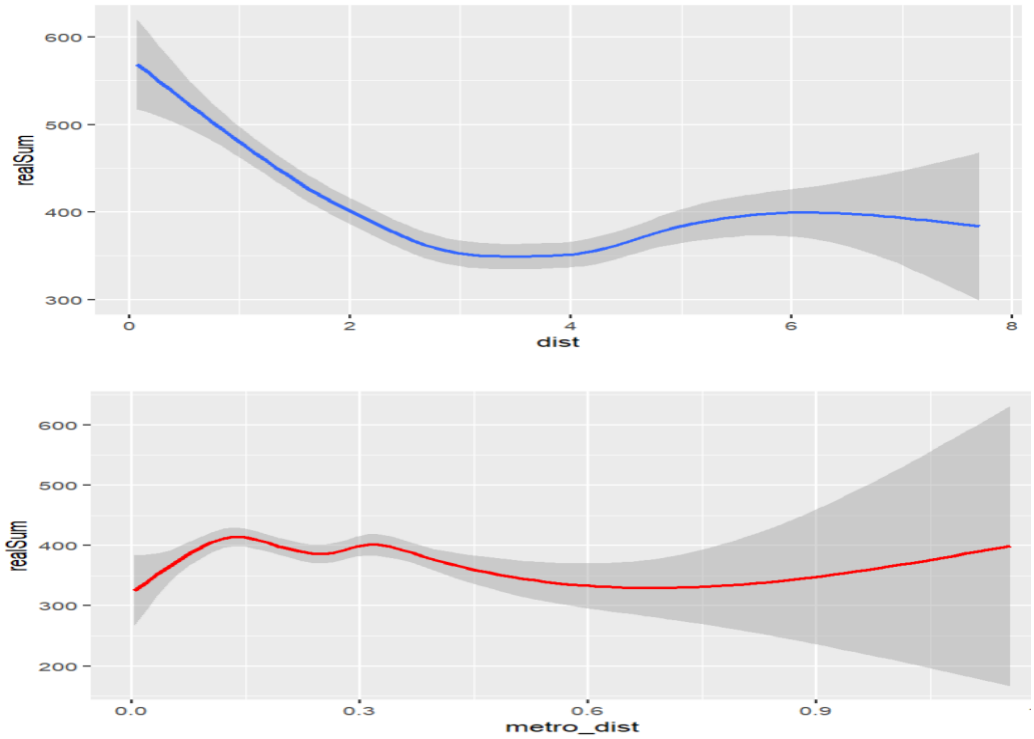


Figure 1: Evolution of Airbnb prices by distance to the city center

From this chart , we notice that in Paris, the lower the distance from the city center, the higher the price of the rooms, which is perfectly logical. Accommodations that are close to transportations and downtown tend to be more expensive. But the relationship into room prices and distance to the metro is not the same.

Before running our linear regression, we check the correlation between our variables. The results are shown below.

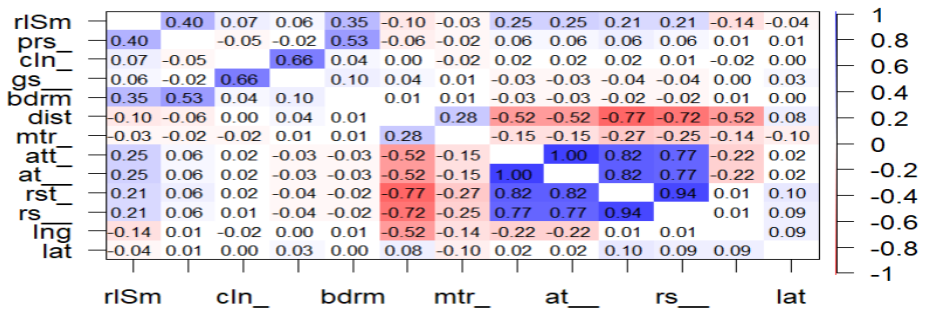


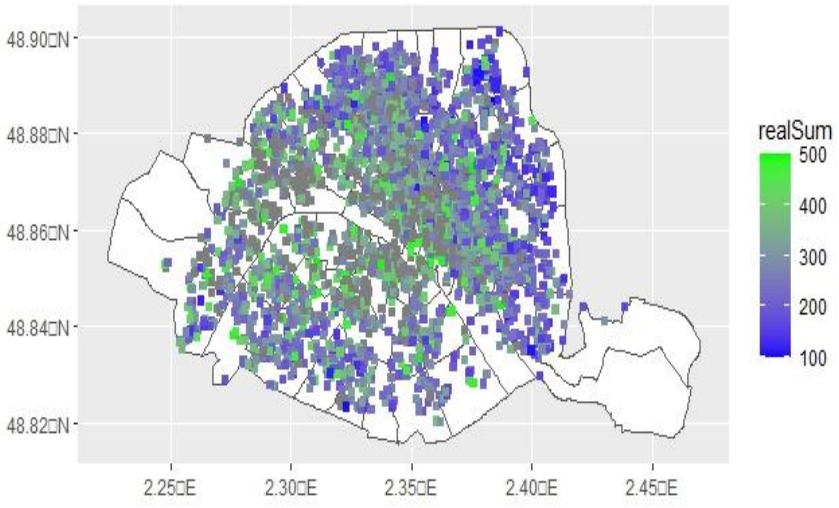
Figure 2: Correlation plot

The following graph shows the correlation matrix in numerical variables. It can be seen from this graph that there is a strong correlation relationship between cleanliness rating and the guests satisfaction. Anyways, we can see that the effects of our explanatory variables on the prices aren't very high.

Table 5: Result of Kolmogorov test for qualitative variables

variable	**p-value**
room_type	< 2.2e-16
multi	< 2.2e-16
biz	< 2.2e-16
day	< 2.2e-16

From the tables 5, we notice that all our qualitative variables are correlated with the dependent variable, which is Airbnb prices. Then, they will be used in our regression analysis.



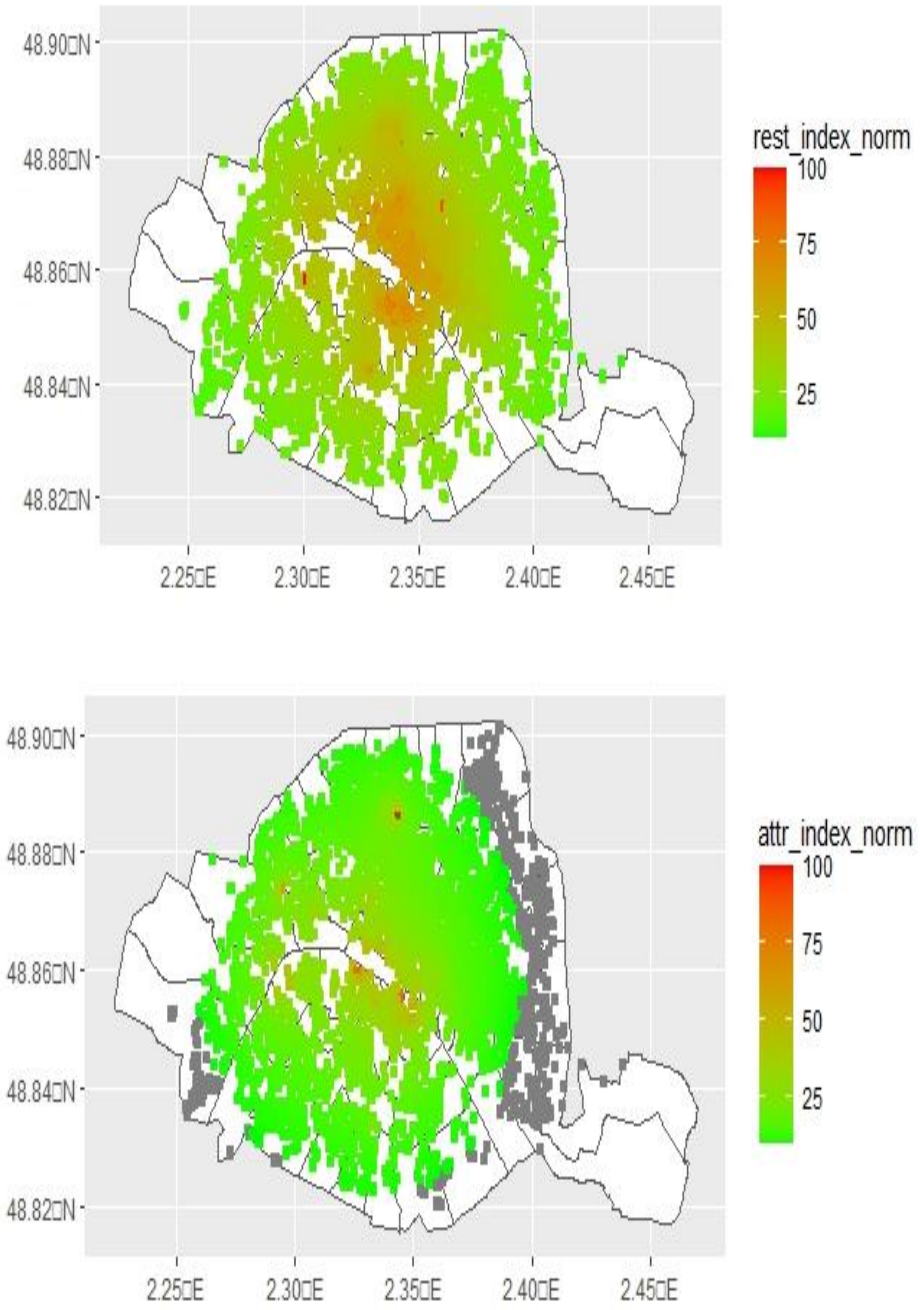


Figure 3: Geographical representation of the Airbnb in the city of Paris according to price, restaurant index and attractiveness index

The graphs above geographically represent Airbnb rooms according to price, restaurant index and attractiveness index in the city of Paris. It is noticeable that housing more at the ends of the city are those that cost the least while those in the center of the city cost more. It is also not surprising that both indices are higher in the center of the city, which is quite normal because economic activity is very concentrated in the center. This explains the fact that rooms are more expensive in this area. This makes this area very active so very popular and more expensive.

3.2. OLS

Table 7: Result of OLS estimation

	Paris	
variable	**Coeff**	**p- value**
Intercept	4.7412757	<2e-16
bedrooms	0.2195212	<2e-16
dist	0.0143316	<2e-16
metro_dist	0.0310706	3.59e-05
attr_index_norm	0.0237724	<2e-16
rest_index_norm	-	-
room_type_Private room	-0.2509490	<2e-16
room_type_Shared room	-1.0350032	<2e-16
multi	0.1085753	<2e-16
biz	0.2044214	<2e-16
day	-0.0178093	0.0344
host_is_superhostTrue	0.1137450	<2e-16
person_capacity	0.1273173	<2e-16
AIC	4685.346	-
BIC	4773.851	-
Adjusted R^2	0.5469	-

We observe that in the model, all variables are significative at 5%. We also notice that the number of rooms, distance from the city center, distance from the Metro, attractiveness index and restaurant index, multi, biz, person capacity positively influences the price of rooms. Thus, an increase of one unit in the number of rooms, increases the price of the Airbnb by 0.2%.Also the price of room decrease to 0.01% in weekdays compare to the weekends.

3.3 Non-Parametric Method: GAM

We will model spatial dependence nonparametrically using a Generalized Additive Model (GAM). To achieve this, we will visually depict the relationships between our dependent variable, named "realSum," and all explanatory variables.

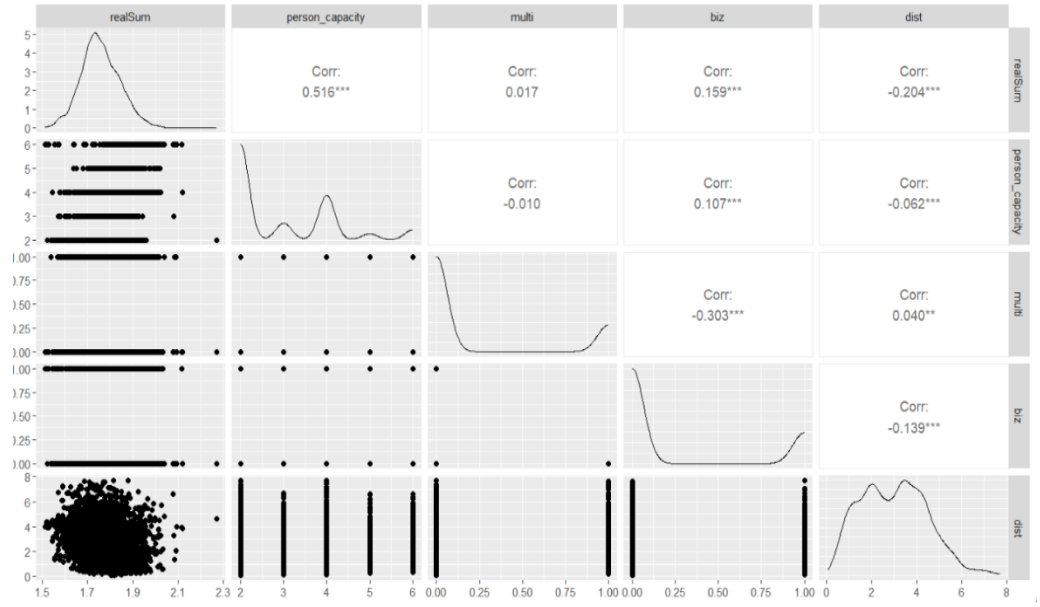


Figure 4: Variable distribution analysis and interaction visualization (linear relationship)

As observed in the preceding figure, there are predominantly linear relationships between these explanatory variables and "realSum." However, in contrast, as depicted in the subsequent figure, the remaining explanatory variables exhibit more nonlinear relationships with "realSum."

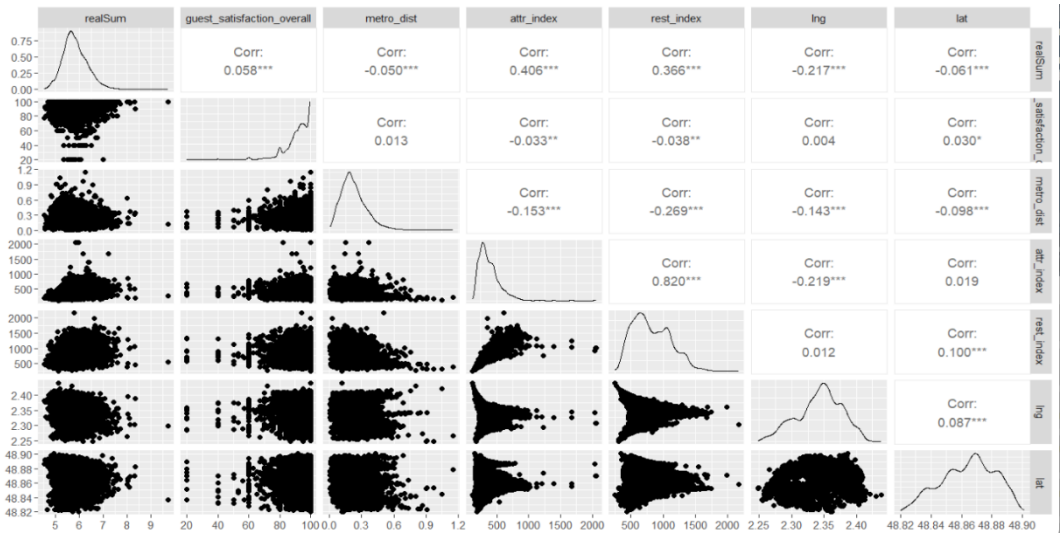


Figure 5: Variable distribution analysis and interaction visualization (nonlinear relationship)

Therefore, we will proceed with our GAM model using these variables. Initially, we will define a standard GAM model:

Table 8: Individual estimation GAM model

```
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.822399   0.005399    1078   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                               edf Ref.df      F  p-value
s(guest_satisfaction_overall) 5.932  6.910 17.254  < 2e-16 ***
s(metro_dist)                 7.374  8.374  4.773 5.48e-06 ***
s(attr_index)                 8.275  8.826 15.540  < 2e-16 ***
s(rest_index)                 2.293  3.053  3.102   0.0244 *
s(lng)                        8.015  8.748 11.165  < 2e-16 ***
s(lat)                        7.518  8.467  6.071  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =   0.25   Deviance explained = 25.4%
GCV = 0.19613  Scale est. = 0.19495   n = 6688
```

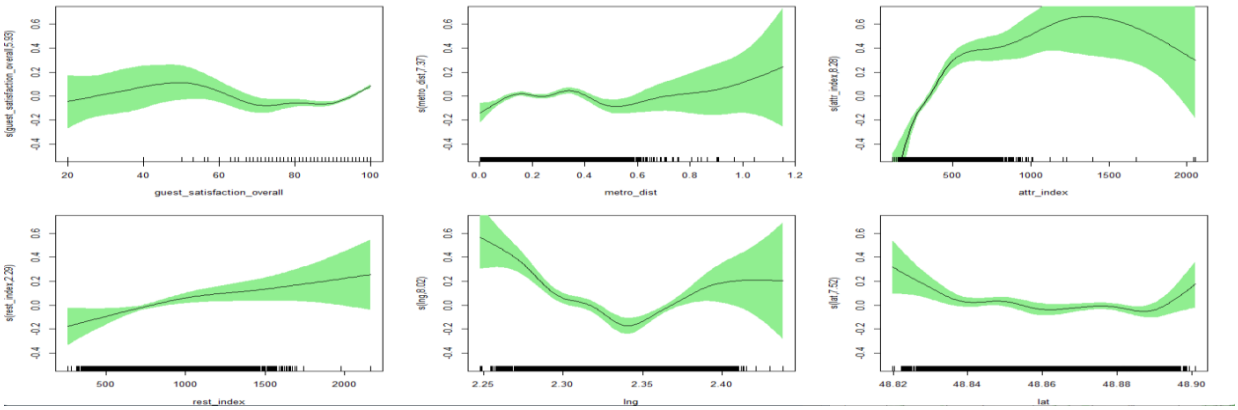


Figure 6: Graphic representation of individual estimation of GAM model

As illustrated in these graphs, there are notably significant nonlinear associations between these explanatory variables and "realSum." The p-values for each explanatory variable indicate that "guest_satisfaction_overall," "metro_dist," "attr_index," "lng," and "lat" exhibit highly nonlinear relationships.

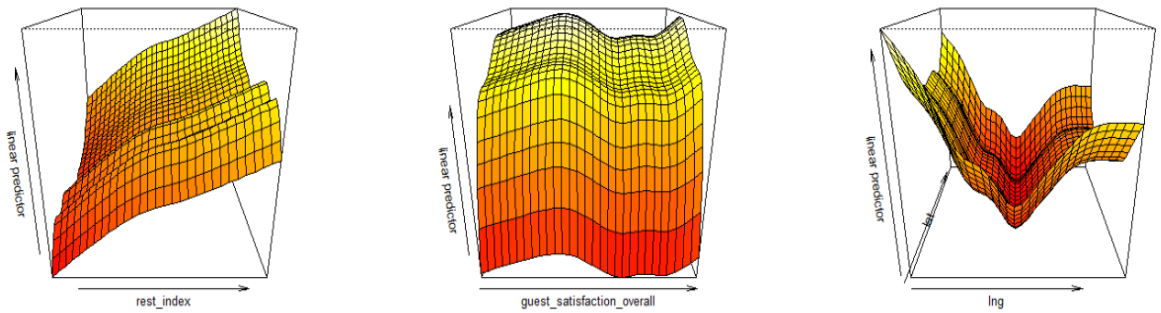


Figure 7: 3D representation of individual estimation of GAM model

Now, we should consider a more flexible model by combining the variables "guest_satisfaction_overall," "attr_index," "lng," and "lat." The results reveal the following:

Table 9: Estimation of GAM Model to have flexible model

```
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.822399   0.005358    1087   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                                                edf Ref.df      F  p-value
s(rest_index)                                 4.916  6.217  2.976  0.00599 **
s(metro_dist)                                 7.455  8.425  5.189 2.58e-06 ***
s(guest_satisfaction_overall,attr_index,lng,lat) 79.498 95.943 11.293  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.261   Deviance explained = 27.1%
GCV = 0.19468  Scale est. = 0.19197    n = 6688
```

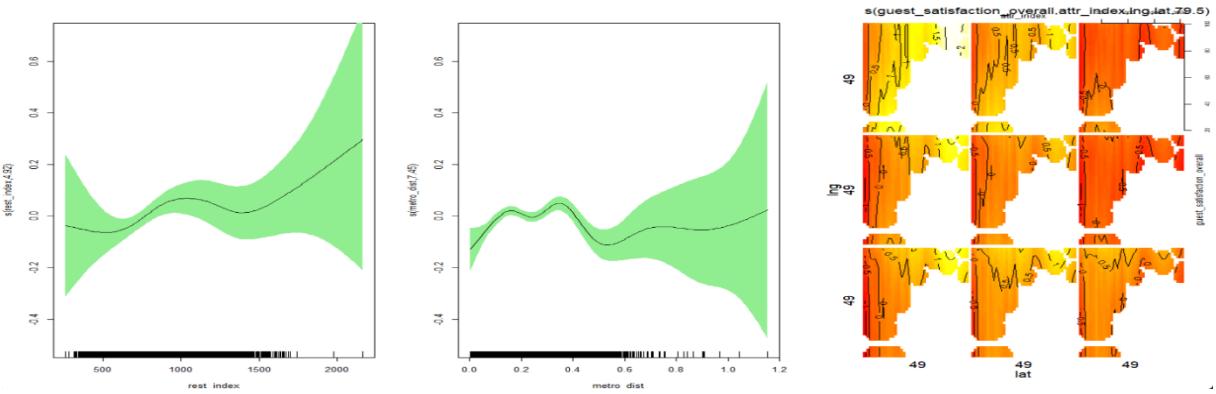


Figure 8: Graphic representation of flexible GAM model

To gain a deeper understanding of the relationships involving the "guest_satisfaction_overall," "attr_index," "lng," and "lat" variables, we can visualize these interactions using a 3D plot:

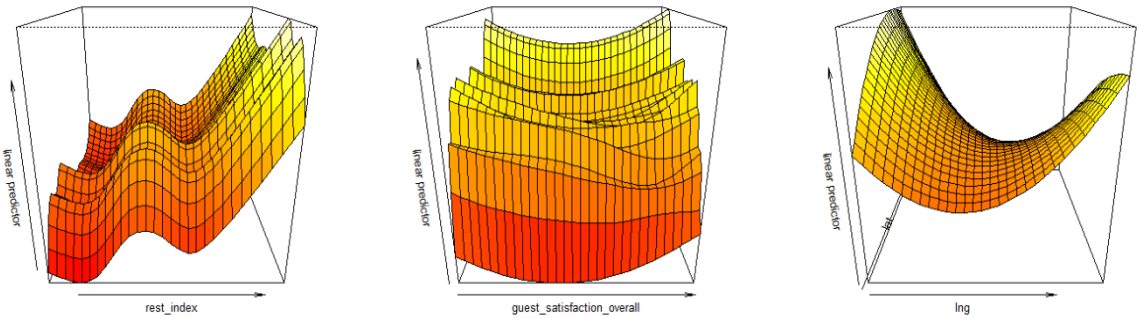


Figure 9: 3D representation of flexible GAM model

To conclude, we will include all explanatory variables with nonlinear relationships to create a more flexible model with basic splines. The outcome is as follows:

Table 10: Basic Splines estimation of more flexible GAM model

```
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.822399   0.005369    1085   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                                                        edf Ref.df      F p-value
s(rest_index)                                         4.797  6.117  2.446  0.0224 *
s(guest_satisfaction_overall,metro_dist,attr_index,lng,lat) 61.004 63.475 17.921  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.258   Deviance explained = 26.5%
-REML = 4180.5  Scale est. = 0.19277   n = 6688
```

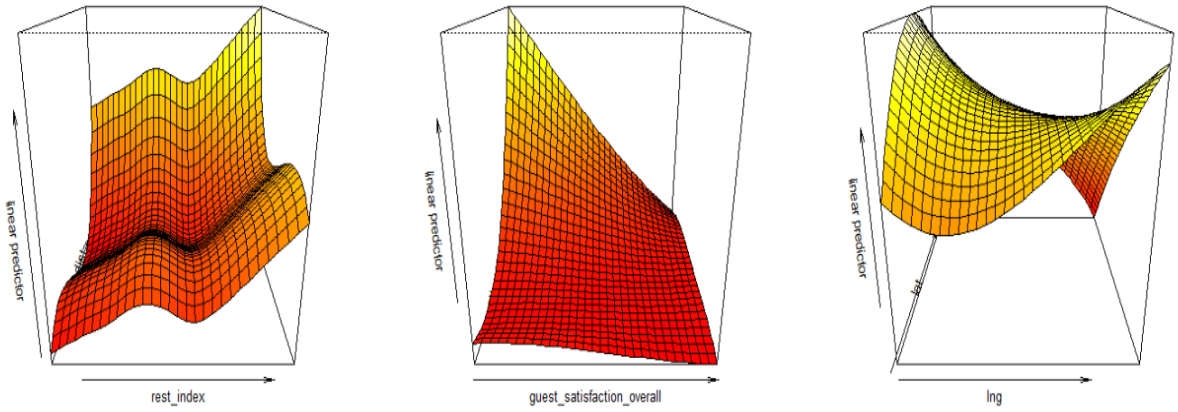


Figure 10: Graphic representation of Basic Splines estimation of more flexible GAM model

4. Comparison of OLS and GAM ,Alternative Method

4.1 Comparison of Models

In conclusion, as we make our GAM model more flexible, we observe a change in the relationships between each variable and the linear predictor. The models become increasingly efficient, as indicated by the increasing R-squared values. Furthermore, the significance of nonlinear relationships between explanatory variables and the dependent variable "realSum" becomes more pronounced.

It's also worth noting that the GAM model effectively manages interactions between explanatory variables. Through the inclusion of interactive smooth terms, it models how the relationships between explanatory variables change with respect to each other. As the p-values become increasingly significant, it is evident that all selected explanatory variables hold significant predictive power.

Moreover, it's important to mention that we can specify the smooth terms according to our assumptions about the relationships between the variables and fit them using various methods such as basis splines, regression splines, or P-splines. However, in this case, we have used only basic splines, as regression and P-splines are designed for univariate splines (1D).

4.2 Alternative Model : SAR

In their analysis, Gyodi and Nawaro (2021) found that there is a spatial dependence in the determinants of Airbnb prices. This is somehow intuitive, since Airbnb hosts tend to check the prices in their area to set their own prices. Moreover, the plateform provides some tools that suggest rates for their rooms, among which there are prices (Hill, 2015). Thus, a spatial model will be interesting to analyse more accurately the determinants of Airbnb prices.

Is this framework, the Manski model (Manski, 1993) is known as the most general model in spatial econometrics.

$$Y = \rho WY + \alpha \iota n + X\beta + WX\theta + u \text{ and } u = \lambda Wu + \varepsilon$$

Where:

- Y is a vector of the N observations for the dependant variable ;
- W is a spatial weight matrix of dimension N*N ;
- X is a matrix of explanatory variables of dimension N*K ;
- β is the vector of coefficients that measure the effects of the independent variables on the dependent variable ;
- θ is a vector to measure the effects of lagged explanatory variables ;
- u is the error term ;

- ε is the disturbance term.

Although this model is complete, it's complex and difficult to implement. To analyse the spatial dependence of Airbnb prices, we will just run a Spatial autoregression analysis in which we consider that prices are affected by lagged prices.

$$Y = \rho WY + \alpha\imath n + X\beta + \varepsilon$$

The analysis will be previewed by a Moran test to check if there is a spatial dependence in the Airbnb prices distribution.

- **Moran's test**

The Moran's test is a commonly used spatial autocorrelation test in spatial data analysis. It is used to determine whether there is spatial clustering or spatial dispersion in the values of a variable across a spatial dataset.

The test is based on the concept of spatial autocorrelation, which refers to the tendency of spatially proximate values to be similar or related to each other. The Moran's I test calculates a measure of spatial autocorrelation, which ranges from -1 to 1. A value of 0 indicates no spatial autocorrelation, while a value close to -1 indicates negative spatial autocorrelation (i.e., values that are dissimilar tend to be close together) and a value close to 1 indicates positive spatial autocorrelation (i.e., values that are similar tend to be close together).

The test is a hypothesis test to check if there is spatial dependence in the distribution of the variable with two hypotheses:

H_0 :: No special dependence

H_1 :: There's spatial dependence.

If the p-value is less than 5%, H_0 is validated at 5%.

- **RESULT**

Table 11: Result of Moran's test

Moran test			
Moran I statistic	**Expectation**	**Variance**	**p-value**
0.5261737115	-0.0001495439	0.0002359035	< 2.2e-16

Table 12: Result of SAR estimation

	Paris	
variable	**Coeff**	**p- value**
Intercept	2.57404111	< 2.2e-16
bedrooms	0.15117233	< 2.2e-16
dist	0.02722950	1.799e-14
metro_dist	0.04907828	0.07388
attr_index_norm	0.00987859	< 2.2e-16
rest_index_norm	0.00395429	2.354e-14
room_type_Private room	-0.15596817	< 2.2e-16
room_type_Shared room	-0.63547628	< 2.2e-16
multi	0.06986535	2.220e-16
biz	0.14048711	< 2.2e-16
day	0.03345683	6.029e-05
host_is_superhostTrue	0.06858076	2.198e-13
person_capacity	0.09236108	< 2.2e-16
AIC	2057.4,	-
Log Likelihood	-1013.697	-
Adjusted R^2	-	-

5. Discussion and Conclusion

During the last years, Airbnb accommodations have become an excellent solution to travellers looking for rooms for short stays. To provide some insights about the determinants of the prices of this rooms and helps people make better choices, we use a dataset entitled Airbnb prices in European cities and available on the Kaggle website to study the determinants of the prices through an OLS regression ,GAM and a spatial analysis. Our regression model shows that Airbnb prices are different to one another, according to the basic prices applied in the area. We also notice that factors like the number of bedrooms, the capacity, the attractiveness and the distance have an influence on the prices. However, the nature of these effects isn't the same for all cities. While for example the distance to the downtown has a positive impact in Paris.

Unfortunately, our limited knowledge in the domain of spatial econometrics and spatial analysis weren't sufficient to succeed in the complete spatial analysis. Thus, our work didn't completely reach its objective.