Pricing strategy on AirBnb based on the Spatial Influence of Touristic Attractions in New York

ABSTRACT

This study evaluates variables supporting pricing strategies for hosts interested in competing on a sharing economy icon dominating markets in many cities, AirBnb. Data created by Airbnb worldwide offers a rich repository for insights on the real estate realm by crossing multiple variables, including some with geographical characteristics that can support the decision making process of millions of property owners looking for obtaining profit out of their properties. New York City (NYC) is one of the most visited cities in the world and data published by the Airbnb website contains variables that support deciding whether Airbnb can be an revenue stream option for property owners. The methodology consisted of exploratory quantitative analysis (spatial point pattern, density maps of listings, evaluation of block proximity to attractions in price using buffer analysis) followed by spatial dependence models: spatial auto-correlation and spatial regression models (SAR), using socio-economic-demographic variables and distances from NYC attractions to properties. The key results indicate important moderate spatial auto-correlation of rental prices (Moran´s *I* of 56.85%) and the consequent increase of explanation of rental prices in the models of 21.3 percentage points with the incorporation of variables of distance to NYC attactions and of the auto-regressive spatial term (R^2 from 34.59% of OLS regression to 55.89% of the Spatial Autoregressive model).

Keywords: Airbnb; spatial regression; point pattern analysis; buffer analysis; consumer behavior

1. INTRODUCTION

Many people can find it difficult to decide whether to invest in an Airbnb vs. long term rental when they start in real estate investing. That can be a tough decision if not properly thought through. Despite the fact more users are joining the platform, many can lack experience on hot to price. Future AirBnb's hosts can obtain additional monetary value and extra sources of incomes by sharing idle resources. Airbnb is a platform that connects people who are willing to share a space with guests in return for payment and is growing in business and relevance as a disruption (GUTTENTAG; SMITH, 2017; MAKARAND, 2017; VARGAS-HERNÁNDEZ; RUEDA, 2019) and its relevance in the sharing economy (JUN, 2015; ZERVAS et al., 2017; HUEY-YEH et al., 2017). Also it has impacted many changes in traditional business models and traditional markets (BOECKMANN, 2013). Basically, the platform requires property owner to list their room, house, apartment on the website, and wait for reservations. Airbnb rentals are considered a short term rental strategy because the properties are usually rented up to a smaller period of time. The rapid growth of peer-to-peer platforms has arguably been enabled by two key factors: technology innovations and supply-side flexibility (ZERVAS et al., 2017). Also, Airbnb's platform is one of many that contains location-based services or 'Geo Services' and it can be used in innovative ways connecting people to nearest community services and linking consumers to local services (OXERA, 2013).

Transactions created on the platform and available online on Airbnb's own website, called Get the Data (AIRBNB, 2018), offer great opportunities for studies in the real estate and tourism industry. When crossed in models, it can offer answers to some business questions, including for those who want to enter their property in Airbnb and have doubts about profitability, risk and whether it is an attractive investment or not. Pricing has been an item of relevance in many studies (ZHANG et al., 2011; WANG and NICOLAU, 2016; ABDELAZIZ et al., 2017; ZHANG et al., 2017). AirBnb has developed digital tools to assist the platform users in pricing

but the decision is still for an individual to decide, decide how to compete with others and relate renting versus participating in a shared economy model.
Other variables have been studied to explain why people would use the platform presenting how relevant this platform has been to the academy and interest of the public in many studies (GIBBS et al., 2018; GUNTER, 2018; KENNEDY, 2018; LIANG et al., 2018; SO et al., 2018). Technology innovations have streamlined the process of market entry for suppliers, facilitated searchable listings for consumers, and kept transaction overheads low (ZERVAS et al., 2017), which motivates entrance to the platform however the on going pricing strategy is not well defined. In this article, we provide empirical evidence to this debate by studying the factors that can influence the price of the properties. It explores the relationship between Airbnb charged prices with the region as well and estimates the prices as a function of different conditions, property characteristics, approximation to touristic areas so that the decision making process for the platform can be facilitated.
Datasets with a limited number of variables can be mashed and explored with auxiliary datasets to create derivate variables with high value to be combined in business insight reports, support predictive models and cluster data to clear blurry and hidden information underneath data patterns. Data crossing or data wrangling creates even higher value when information gain is obtained from secondary datasets available publicly. This paper is an introductory work on how to obtain public data offered by Airbnb for a particular city, cross it with geographical datasources to create meaningful insights that could not be obtained without the support of geospatial analysis. The concepts of spatial statistcs can aid the decision making process (LEE; KANG, 2015; FRANCISCO et al., 2006). Our hypothesis is that some offerings with Airbnb serve as a substitute for traditional renting and it is differentiated by geographic region.

H_0 : Rental price of Airbnb properties are not spatial dependent
H_A : Rental price of Airbnb properties are spatial dependent

The consistent conclusions of the work are how much geointelligence has relevance in modelling regressions and improve accurancy. The concepts of geotechnology and geointelligence modelled in the pricing strategy for property owners can be considered an optimized methodology implemented in the analysis of joining Airbnb, notably for property owners competing in a distributed territory in an urban area.
Results achieved also enlights that the problem presents spatial autocorrelation, which indicates that rental prices can be explained by spatial dependence.

2. RESEARCH DEVELOPMENT

2.1 Project Objectives

As a result of the success of Airbnb, their current business scenario involves monumental numbers: according to official Airbnb numbers, in May 2018, the company had 4.85 global active listings offered in 191 countries, and a total value of US$ 38 billion. This paper about Airbnb is a humble piece of work that represents how Big Data inputs can generate useful and practical outcomes, as the rational to support the extra income to be charged by Airbnb owners, depending on its closeness to touristic attractions, as described further ahead in this article.
In general terms, the project objective is to explore a set of variables in the datasource to support the decision of which model to use a real estate property for, whether Airbnb could represent a profitable opportunity. With the available variables from New York city's listings, a geographical analitycs approach with a geospatial analysis was very important to visualize whether a particular set of listings located in an area in which a host is looking for opportunies had the movement of guests looking for properties hosts offering propertys, and then with

enriched information opt for more conservative forms of investment, such as regular renting. The process to create the analysis included spatial exploration through point pattern and polygon analysis. The average price per neighborhood was used as the target for the predictive model based on multiple regressors explaining the dependent variable, with the linear model technique.

2.2 Methodology

Figure 1 shows the macro-stages involved in developing the methodology. The sub-topics describe each stage in greater detail. Schrage (2014) brings to the fore the fact that companies still have difficulty analyzing data using data to drive decisions and Big Data comes to challenge this simple analysis, causing other quantitative analyzes to be rethought. This methodology serves as an example of good use of Big Data in complex business situations.

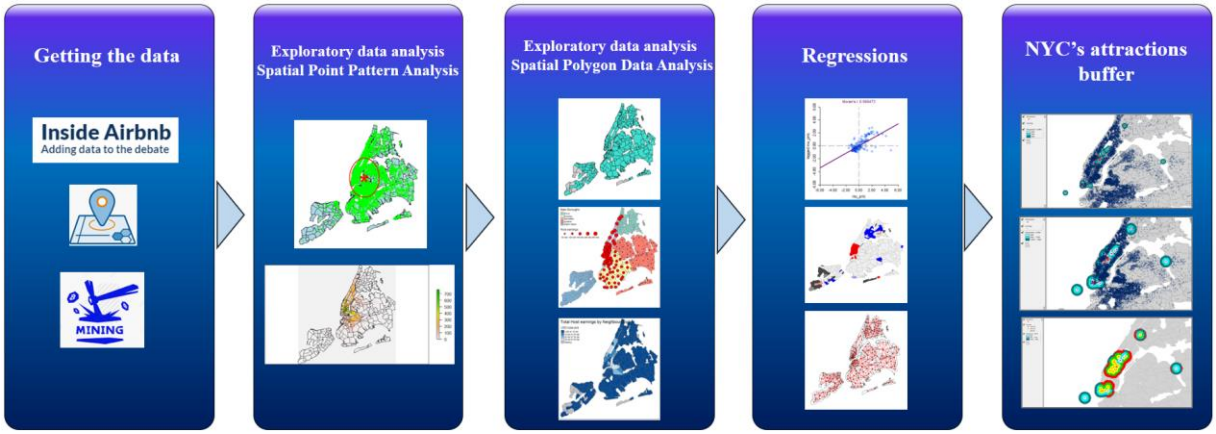


Figure 1. Stages through the different phases of the methodology
Source: The authors.

2.2.1 Getting the data

Airbnb's site, named Get the data, created incentives for data extraction. It is publicly offered to anyone and is composed of three different sets. One is the set of listings, the other is the review description given by guests and the third one, the geojson location of New York's neighborhoods. The last one was turned into spatial media, like shapefile, by internet websites which easily reads the geojson's content and returned set of files in shapefile format. The set of variables were numeric, categorical and geocoded as well.

The development of all the ideas and related solutions started with making of the database available so the process of analysis included normal exploratory data, at first, given it was publicly available and, after downloaded, it required the exploration from different angles, perspectives and techniques.

The goal of profitability required us to use the price, our monetary variable as the target variable in the regression models: linear regression, spatial regression and geographically weighted regression.

2.2.2 Enriching the variables with geographical data

The researchers nevertheless considered that the problem of understanding users' interest in booking a property involved not only the variables present in the original database but also the distribution of properties across the city's territory and its geography. So this work has presented special dedication so that all the properties retained a spatial geographic coherence

with the original data and, at the same time, allowed for the mathematical integration for the stage of the geospatial analysis computer model. Geospatial techniques can add reasoning to models and ease the making sense process by providing maps and other visual routines to the interfaces delivered (FRANCISCO, 2011).
The philosophy assumed in this work is that the geospatial frameworks could gain more popularity in explaining the phenomena in which location data is available.

2.2.3 Exploring other variables and enriching the model with text data

Exploring the sentiments mined off of the Airbnb's reviewer has proved its relevance (ABDELAZIZ, et al., 2017). One of the datasets loaded to the analysis contained the review comments each user had left after their experience in the property. It was considered that quantitative variables could be obtained by by mining data off of the reviews. By transforming and mashing its contents, the goal was to structure it numerically to plug it into the model by mining the language details running different lexicon algorithms through all its contents and transforming it into positive/negative scores or mining the sentiments, whether the review expressed joy or anger or any other feeling.
The lexicon used in our analysis, the NRC (from Saif Mohammad and Peter Turney) contains more than 6,000 words in English and such words receive scores either for positive or negative sentiments (in a binary fashion, that is, yes or no), as well as emotions classifications. The emotions classifications are anger, anticipation, disgust, fear, joy, sadness, surprise and trust.
When retrieving the sentiment, an average over the results was obtained and joined back into the listing's data source. In addition to common lexicons, the analysis included scraping words usually associated with the idea of location such as: close, far, near, nearby, station, central, minutes, waking, distance, etc. They were used to classify the importance of location. Words such as bathroom, kitchen, Wi-Fi, convenient, quiet, etc. were used for the amenities attribute. Moreover, words like spacious, chic, decorated, decoration, unique, etc were used to classify the listing as luxurious.

2.2.4 Data analytical approach

With the available variables from New York City's listings, the geospatial analysis was very important to visualize whether a particular set of listings represented an area in which guests would make bookings or hosts would offer property, and then opt for more conservative forms of investment, such as regular renting. The process to create the analysis included the spatial exploration through point pattern and polygon analysis. The average price per neighborhood was used as the target for the predictive model based on multiple regressors explaining the dependent variable, with the linear model regression. When the target is a continuous variable, the linear regression tries to understand whether changes in some variables will increase or decrease the value of y, as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots + \beta_p x_p + \varepsilon$$

The process to find which variables are important in the equation consisted of analyzing the statistical significance of these so called regressors (β_0, β_1, β_2, …, β_p) through the cutoff 0.05 P-value. Also in combination of the R^2 score, conclusions can drawn to whether predictor "y" was indeed explained by the chosen regressors. The higher the R^2, the better the model was considered, but in combination with the proper P-value. After a linear equation was considered valid, the following steps included the usage of spatial regression, which took longitude and latitude attributes into account in the dataset and then used to predict the "y" value. Also, there are more elegant spatial regression techniques, such as Autoregressive Spatial Regression.

The Spatial Autoregressive (SAR) model considers the effect of the dependent variable observed in neighboring units, that is, how much the school dropout rate in neighboring municipalities influences the school dropout rate in the municipality. The model is represented by the following equation:

$$y = \rho W y + X\beta + \varepsilon$$

where: y = vector with observations on the dependent variable; ρ = spatial autoregressive coefficient, which varies between -1 and 1; Wy = vector of spatial lags for the dependent variable; X = observation matrix of the independent variables (plus the constant); β = vector with the regression coefficients ($\beta_0, \beta_1, \beta_2, \ldots, \beta_p$); ε = variable that includes all residual factors and possible measurement errors (ALMEIDA, 2012).
By adding New York's touristic attractions into the model, it became possible to crosscheck if listings located closer to any given attraction offered advantage to hosts, when compared to others not as close as within a 219-meter-distance. The buffered distance obtained from 3 different distances, with 73/146/219 meters served as a basis for average price calculation and these averages were compared against properties not close enough from touristic places.

2.2.4.1 Spatial Point Pattern Analysis

Point pattern analysis is the evaluation of the pattern, or distribution, of a set of points on a surface. At a basic level for the data in this study, it simply consists of point coordinates describing the locations of the Airbnb's listings. These points are the lowest level elementary object and can lead to more complex structures that are also clear and useful, such as polygons. Originally, spatial point analysis was very much used in the classical fields of Ecology but since data is composed of points, we have expanded the methodology to this business case. To start out this part of the analysis, we have used centrography as a direction to obtain a visual summary of the pattern of the listings' spread and found the mean center as well as the standard deviational ellipse. This is analogue to the mean and standard deviation in descriptive statistics in non-geographical data. Measures the center of the spatial distribution and whether a distribution of features exhibits a directional trend (whether features are farther from a specified point in one direction than in another direction) (EBDON, 1987).

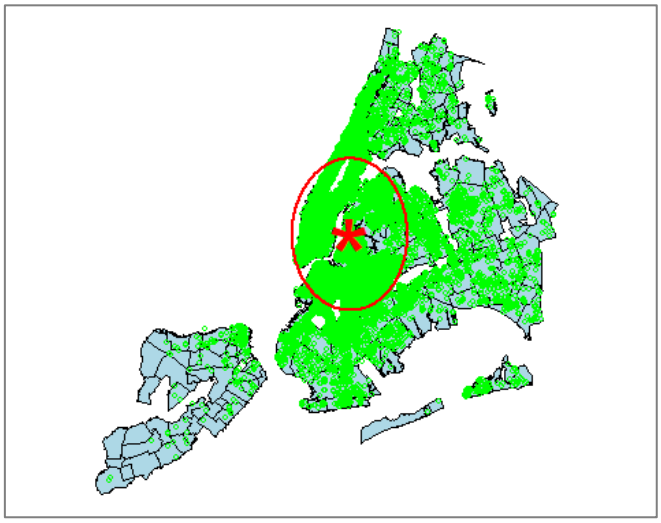


Figure 2. Mean center and Standard Deviational Ellipse
Source: The authors, using R 3.6.2 (package aspace).
Note: Map based on 37,934 listings

Although data is sparse and coverage is also high, due to location arrangement, most listings

were found in specific areas, so the mean center had to be somewhere among areas with more listings. We can see from Figure 2, that the mean center is somewhere between Manhattan and Brooklyn.
We have calculated the mean of all coordinates in the dataset. Therefore, guests have more options of places to stay in these two main neighborhoods.
Frequency is the most basic way of evaluating a spatial pattern and it is simply counting the number of points in our study area. We have processed 37,934 listings in the entire area.
To get density, we have counted the number of listings and divided it by the polygon area, which is, in this case, the specific neighborhood area. It was again noted that Brooklyn and Manhattan were very dense. Queens comes slightly in third place.

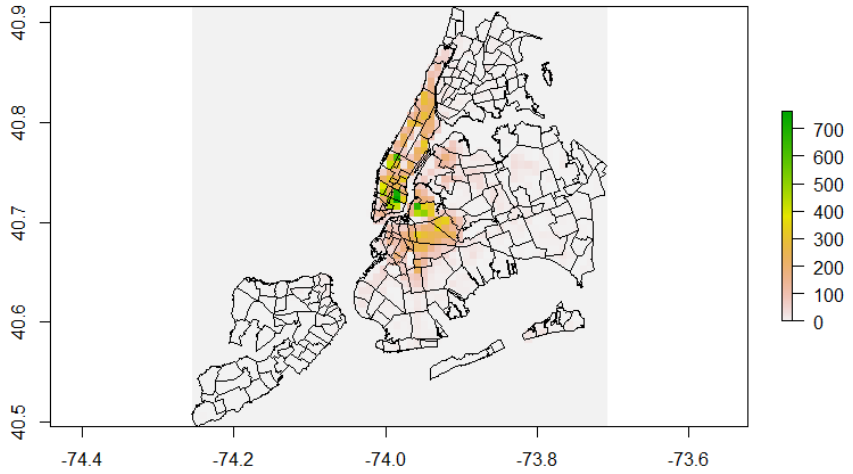


Figure 3. Density Map of Listings
Source: The authors, using R 3.6.2 (package GeoR).
Note: Map based on 37,934 listings

2.2.4.2 Spatial Polygon Data Analysis

In order to obtain better insights on spatial distribution in the map area, we have identified the company's coverage in the city and how largely spread was Airbnb in the area. People can book places everywhere, in all main boroughs and enjoy the numerous activities offered in the city.
The polygon analysis in this study was chosen as a technique to summarize the different Airbnb's listings or points into neighborhood level's measure to generalize analysis by an area as opposed to a particular data point, or property, in the NYC's space. It made it easier to detect patterns in the geospatial realm. Figure 4 tried to focus on the monetary earnings of hosts from some particular neighborhoods surrounding the Central Park area, were the ones with the highest earnings, excluding the neighborhoods in South Manhattan, where Wall Street and port area are.

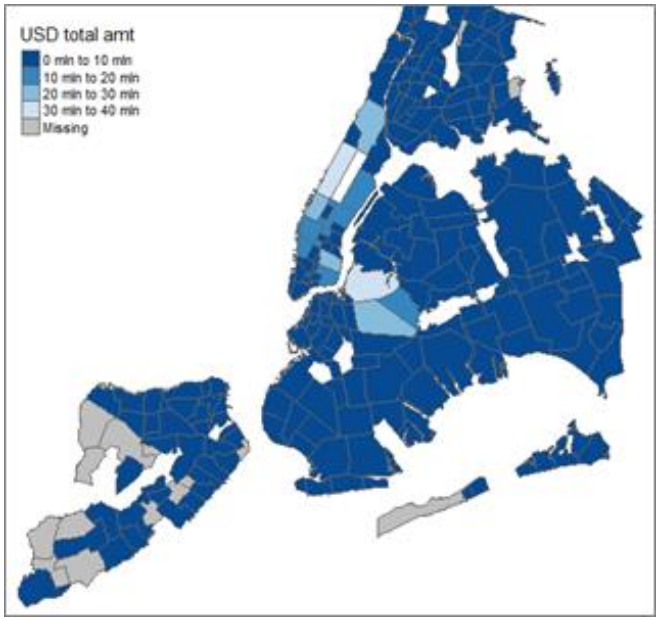


Figure 4. Map of Host earnings by neighbourhood
Source: The authors, using R 3.6.2 (package tmap).
Note: *n* = 233 polygons (neighbourhoods)

2.2.4.3 Local Spatial Autocorrelation Analysis

The exploratory analysis was around the geometric shapes isolated from one another. In predictive analytics, there are more elaborated techniques to obtain data pattern through the correlation of multiple different variables. This autocorrelation analysis in terms of space is a particular spatial technique which shows how much of the variation in a particular variable is explained by its spatial neighbors. Prior to identifying the correlation, a matrix had been created by identifying which neighborhoods are next to each other given a particular level, such as level 1 – "my neighbors are those crossing borders with me" or level 2 – "my neighbors' neighbors are also my neighbors" and so on. The one picked to identify the level of relationship in geography was level 1. Once picked a neighborhood's matrix, the next came to be the correlation of average price by neighborhood with how that previous matrix and not only see through the correlation graphic, the so called Moran's I metric as observed in Figure 5.

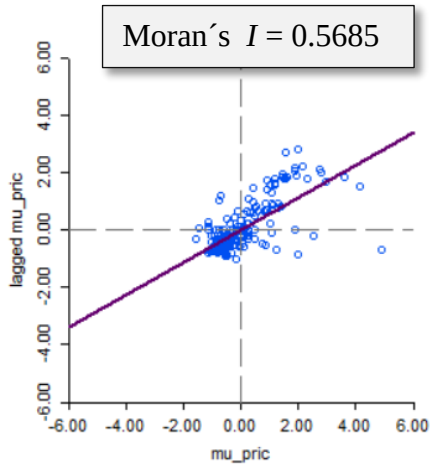


Figure 5. Moran´s *I* and Moran Scatterplot of Map of the Host earnings by neighbourhood
Source: The authors, using GeoDA 1.14.
Note: Using Weights Matrix (*W*) with Contiguity (adjacency), Queen, 1st order ; *n* = 233 polygons

Moran's *I* varies from -1 to 1 and it is an analogue to Pearson´s correlation, that is, the closer it is to 0, the lower the correlation. And the closer it is to 1, the higher the positive correlation. With a 0.5685, we have concluded the correlation is a moderate one, which could be used to further geospatial analysis. Still using the very same matrix, a clustering comparison for geography included the LISA relationship among areas and their neighborhood ones as in a technique to find out if there are clusters in a form of:

- High-High: High prices with high prices surrounding me
- Low-Low: Low prices with low prices surrounding me
- Low-High: Low prices with high prices surrounding me
- High-Low: High prices with low prices surrounding me

With presence of clusters in geography, there could be a form of data exploration with segmentation and identification of significant areas to the business to target particular marketing campaigns and incentives to current property owners to bring them into the Airbnb's platform. The cluster areas could be quantitatively used to derive dummy variables to a model or serve as a guideline for business decisions and tactical moves to the NYC area. Same technique could be used in any other city and geographically based business. As in a statistically significant analysis, a combination of the LISA Cluster Map, there was the LISA Significance Map.

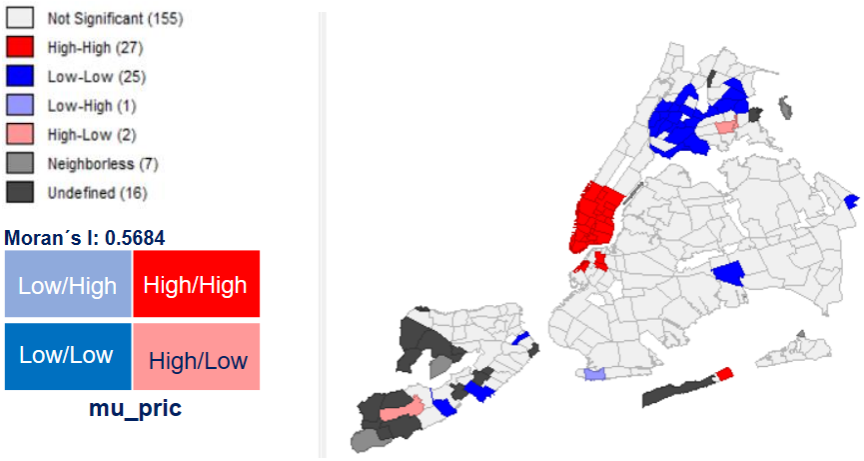


Figure 6. Local Moran´s *I* (LISA map) of the Host earnings by neighbourhood
Source: The authors, using GeoDA 1.14.
Note: Using Weights Matrix (*W*) with Contiguity (adjacency), Queen, 1st order ; *n* = 233 polygons

As with other different statistical techniques, the significance level of a particular analysis is developed with aid of P-values. These metrics are calculated to obtain how accurate is the statistical hypothesis test. A hypothesis is an idea developed beyond the current assumption. Therefore, in this business case, the particular hypothesis was that geography had no influence in the business. This is how we have started this entire analysis. From that first assumption, the analyst intention was to prove that hypothesis was wrong with an alternative one, which yes indeed, the geography had some sort of influence in the business. These two opposed hypotheses can be validated through the P-value outcome plus a P-value cutoff number. The default cutoff in statistical tests is usually the 0.05, not exactly being a rule for each and every analysis, but rather a baseline to adopt and depending on the field of study, it could and it should be increased to 0.10, especially, for instance, when analyzing social behavior, in psychology, sociology or marketing. By adopting the default 0.05 one, the analysis had a map with gradient green colors to identify areas in which the P-value was low with really darker green colors and lighter green tones to areas in which the P-value was closer to the cutoff. Given this guideline and visually seeing its result on the map on Figure 7, the significant areas supporting the

hypothesis that average prices are influenced by geography are the darker areas, followed by medium darker ones and last by light green. The gray area is not statistically significant with all the P-value theory underneath it.

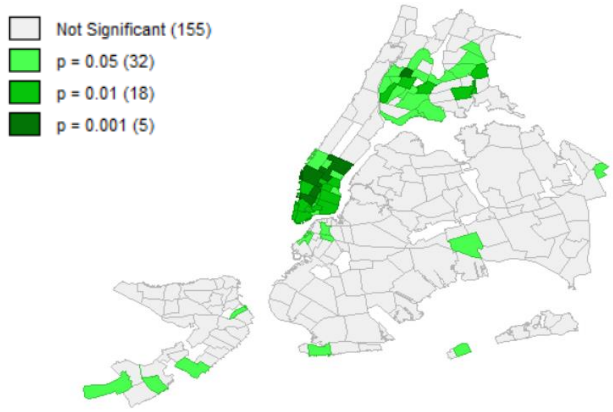


Figure 7. Local Significance Moran´s *I* of the Host earnings by neighbourhood
Source: The authors, using GeoDA 1.14.
Note: Using Weights Matrix (*W*) with Contiguity (adjacency), Queen, 1st order ; *n* = 233 polygons

Figure 8 illustrates how the neighborhood matrix was computed following the Queen, 1st oreder (contiguity weights matrix) so that by selecting a particular neighborhood, their neighbors would be highlighted in white. In GeoDa, as one selected an area, the ones surrounding it were depicted in the map. GeoDa is an interactive software that provides visually appealing maps that respond to selection as opposed to static maps designed in R. The queen matrix can be created in both softwares, but GeoDa gave this quite nice selecting feature and as a side note, their behavior while a neighborhood matrix was plotted for explanation.

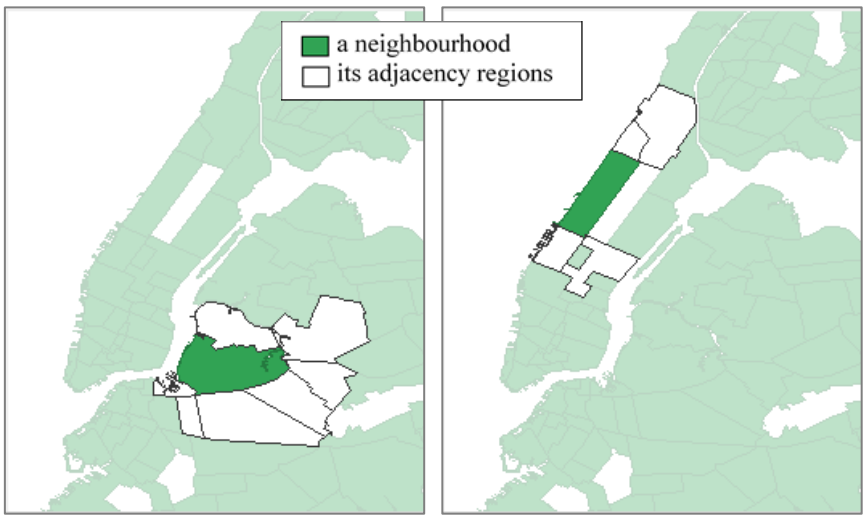


Figure 8. Examples of Neighborhoods using Contiguity (adjacency) Queen 1st order criterion
Source: The authors, using GeoDA 1.14.

2.2.8 Regression Models

Prior to modelling, the dataset was processed to summarize all numeric features by neighborhood in order to obtain a simpler and visual polygon regression and explain the target variable average price by neighborhood. All the listings have been grouped by the neighborhood

variable, so that it would obtain the mean and count scores.
We have started with a linear multivariable regression to discover which variables would better explain the mean price by neighborhood. All 22 variables were included in the regression. The very first regression model including all variables had issues with low significant variables and multicoliearity.
Severe multicollinearity is a major problem, because it increases the variance of the regression coefficients, making them unstable. The more variance they have, the more difficult it is to interpret the coefficients.
One way used to measure multicollinearity was through the variance inflation factor (VIF), which assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated. If the VIF is equal to 1 there is no multicollinearity among factors, but if the VIF is greater than 1, the predictors may be moderately correlated. A VIF between 5 and 10 indicates high correlation that may be problematic. And if the VIF goes above 10, states that the regression coefficients are poorly estimated due to multicollinearity.
Using a stepwise algorithm using P-value with a 0.05 alpha threashold and multicolinearity handling we have obtained a model explaining $R^2 = 34.59\%$ of price variation.

Table 1. Linear regression (OLS) with Socio-economic-demographic variables

Variable	Estimate	P-value	VIF
(Intercept)	-40.2	0.141002	
Average of recency in days since last booking	0.2	0.0000000000865	1.1
Average of luxurious related words present in user's review	4.1	0.001305	1.0
Average of the afinn lexicon related words present in user's review	49.2	0.000357	1.1
Number of Entire home property type	0.2	0.0000000150	4.3
Number of Private Room property type	-0.1	0.00000349	4.1
$\boldsymbol{R^2 = 34.59\%}$			

Note: *n* = 233

The explanation of the model would be that the average price of a neighborhood is positively related to how recent the property has been booked on a factor of 0.2, followed by the presence of luxurious expressions detected in the users' reviews in a factor of 4.1. The highest factor was found in the afinn lexicon expressions that when show positive feeling out of the users' reviews score positive values and the model obtained a factor of 49.2 out of the lexicon parsing. The number of entire home type of property influenced the price on a factor of 0.2 and lastly one factor was negatively related to the price which was the number of private room type of properties present in the neighborhood by a factor of -0.1.

2.2.8.1 Including NYC attractions

We have performed feature engineering also with secondary data to support finding patterns in modelling behavior when adding external data to the Airbnb's business data. Combing the lessons learned so far through the visual observations and hypothesis validation, areas in Manhattan are more valued than in any other locations. One thing that is noticeably important in the city is how it attracts people by the variety of attractions it offers. As a first validation, it was checked if in buffered areas surrounding those touristic places, the average prices of the properties within a certain distance offered higher booking prices.
The NYC tourist attractions analysis was done in ArcGIS 10.2. The buffers meant to be a very nice visual tool to observe how far away some listings would have been affected by the designated distance. Buffers show the area that is within some distance of the input features.

The tool is popular because the concept of buffering is easy to understand and buffering plays an important role in many geoprocessing workflows involving proximity or distance analysis (FLATER, 2011).
For resolving how far the buffer would extend its reach from a tourist attraction, we have looked at a random block in NYC. Each block in the avenues intersected by streets was assumed to have 240 ft as shown in a sample corner observed in Google maps.



Figure 9. Distance between streets and buffered distance from NYC's attractions
Source: The authors, using ArcGIS ArcView 10.2.

It is visually noticed the high density of listing offers, as well as attractions in the Manhattan area, with one or two attractions further away. The idea was to identify if being closer to a hot spot touristic place could offer advantage over other areas and how much of an advantage that would be. The listings within the buffered areas were computed as average calculation per each internal buffered group and resulted in a clear difference on average prices, 54% higher within a block distance than properties not as close.

Table 2. Average price within the 73/146/219 meters buffer

Buffer Distance	# of listings	Average price
73 meters	51	**$254.00**
146 meters	153	$230.00
219 meters	294	$212.00
>219	37436	$138.00

There was some correlation found in this particular analysis and averages indicated some sort of price variation by closeness to certain areas.
After the buffer analysis, we have calculated for every listing in the database its distance to all 32 attractions, creating what we have called a matrix of distances. Other variables were derived from the listing's distance such as distance to the closest attraction, average of distances to all attractions, standard deviation to all atractions, number of attractions within 73, 146, 219, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000 and 10000 meters from the property.
The new set of variables resulted in a new model with improved results. The new variables were grouped by neighborhood and then used to run a regression model. Just as before we have

used a stepwise algorithm with a 0.05 alpha threshold and multicolinearity management through the VIF value, we have obtained a model explaining $R^2 = 50.43\%$ of price variation.

Table 3. Linear regression (OLS) with Socio-economic-demographic and Neighbourhood variables

Variable	Estimate	P-value	VIF
(Intercept)	-33.6	0.15577	
Average of recency in days since last booking	0.2	0.0000000000232	1.2
Average of luxurious related words present in user's review	2.9	0.00983	1.1
Average of the afinn lexicon related words present in user's review	47.1	0.0000872	1.1
Average of number of attractions within 2000 meters	3.1	0.03457	2.0
Average of number of attractions within 5000 meters	3.7	0.00000000278	2.1

$$\boldsymbol{R^2 = 50.43\%}$$

Note: *n* = 233

Feature engineering with distance variables were an indicator of how finding variables out of the business domain were able to help explaining the phenomenon and these new features were an indicator of how location and geography could influence the price variation of the properties. So that was also a indicator of running geo spatial analysis over the data to observe how it could aid improving the model. Next, we took the resulting variables from the previous regression and used them in a SAR model. Spatial regression methods capture spatial dependency when running regression analysis. It provides information on spatial relationships among the variables involved. The geography of New York, at the neighborhood level, was the basis for the distance matrix in the model. The queen-weighted matrix with 1-order of contiguity was calculated and used in the model. In the first model the *Average of number of attractions within 2,000 meters* did not show statistical significance so it had to be removed. The next regression showed all significant variables and resulted in a higher performing model, with $R^2 = 55.89\%$. The use of spatial statistics in the Airbnb data created a model with a higher predictive capability than from a multivariable regression. The autocorrelation model addressed the optimization of model results and resulted in a model containing the Rho (ρ) estimate along with the very same variables from the optimal linear regression, with slightly different estimates, though.

Table 4. Spatial Auto-Regressive (SAR) Model with all variables

Variable	Estimate	P-value
Rho ($\boldsymbol{\rho}$)	0.3	0.00000
(Intercept)	-46.6	0.03805
Average of recency in days since last booking	0.1	0.00000
Average of luxurious related words present in user's review	2.6	0.01181
Average of the afinn lexicon related words present in user's review	42.1	0.00015
Average of number of attractions within 5000 meters	2.9	0.00000

$$\boldsymbol{R^2 = 55.89\%}$$

Note: Using Weights Matrix (*W*) with Contiguity (adjacency), Queen, 1st order ; *n* = 233 polygons

3. RESULTS

Few characteristics have correlation to estimating the average price of properties in the areas in NYC. When pulling the first tier of data with features describing the property uniquely, the analysis explored the data patterns in the city. With feature engineering we have mined data off of other the user's reviews source and joining them to the master data, resulting in a set of 22 variables to run a regression model. At first it was found that not all variables had enough significance and there was very high multicolinearity. A stepwise algorithm was used to combine the variables that could better explain the price variance, which resulted in 34.59% of explanation. It was not a very efficient model. However it helped explain the phenomenon.

The features able to explain the variation in the prices were how recent the property has been booked on a factor of 0.2. Users might look at the last date of a review and make conclusions that the last user's experience is just as recent that they could also experience the same thing. The next event was the presence of luxurious expressions detected in the users' reviews in a factor of 4.1. Off of the text mining technique, we have obtained information that turned out to be important to model the pricing explanation. A few words that came out to be expressed many times in the reviews were words we have classified as luxury involved in the experience. Reviews containing words such as *boutiques*, *decorated*, *spacious*, *intimate*, *loft*, *chic*, *gorgeous*, *unique*, *style*, *decor*, *charms*, *cozy*, *visual*, *sophisticated*, *artsy* and *elegant.* The highest factor was found in the afinn lexicon expressions that when show positive feeling out of the users' reviews score positive values and the model obtained a factor of 49.2 out of the lexicon parsing. The neighborhood with more Entire Homes than Private Rooms or Shared Rooms also increases the chances of higher prices by a factor of 0.2. The number of Private Room properties was the last variable with significance that negatively influence the price by -0.1. This regression could be a baseline for property owners to base their decision on the characteristics of the property, whether it is a Private Room or Entire Home, however they would only be able to understand more about joining Airbnb if they ran an experiment in the platform to obtain data for the other variables such as the recency of reviews and the reviews content with positive words and the description of an experience of a luxury stay.

The linear regression did not present a satisfactory result which left uncertainties to the capacity of predicting prices and creating a solid ecosystem for owners to base off from to influence the decision making processes so the extent of the methodology included more variables with the touristic attraction. There was a total of 32 attractions mapped out: Bethesda foutain, Brooklyn Bridge, Carnegie hall, Central Park, Charging Bull, Chinatown, Columbus circle, Empire State Building, Grand central terminal, Guggenheim museum, Liberty island, Lincoln center, Macys 5th ave, Madame Tussauds, Madison square garden, Metropolitan, Museum of Modern Art, Natural museum, New York Mets, New York Stock Exchange, Ny public library, Radio City music hall, Rockfeller center, Saint paul's cathedral, Saks 5th, Times square, Titanic Memorial Park, Top of the rock, Trmp tower, Wall street, World Trade Center and Yankee Stadium. The distance to these places created other possibilities for analysis resulting in a model with 50% of explanation to the price variation. The new model no longer would need the property types and substituted them by the average of the number of attrations withing 2000 and 5000 meters respectively. That added a new perspective to property owners that being within 5 kilometers from touristic places would influence the property prices. So a decision could be made based on property location and it had higher influence over prices.

The last model also explored location moving forward and 55.89% of price variation would be explained by the neighbourhood next to the neighborhood where the property was. So Airbnb property's prices would need to follow prices from the neighborhoods surrounding them. That would also add more material for property owners deciding whether to join the platform by

anticipating their offering prices by looking at average prices of closer neighbourhoods.
The distributed geographical features added to the model resulted in better predicting results. The hypothesis H_0 was rejected and, so, the spatial distribution does help explain the pricing variable in the Airbnb properties.

4. CONCLUSIONS

First this study adds insights to the sharing economy literature as pricing strategies is a challenge for individuals to participate and compete without the support of a marketing department companies can count on such as hotels. Regular platform users need a set of tools to support the decision process in the lack of experience on how to be a player in this new raising economy model.
The introduction of the concepts and possibilities of geotechnology and geointelligence into the pricing strategy for property owners can be considered an optimized methodology implemented in the business practice, notably for property owners competing in a distributed territory in an urban area. The findings add complementary insights to the pricing strategy literature.
Most property owners could be looking for other options for their investments. One option for investing could be Airbnb as opposed to the conservative rental model with long-lasting contract arrangements. The company has been growing as a platform for booking properties to travelers and short-term renting using Airbnb's interface. Not only this new model offers potential earnings when properly used, it also publishes their data for the data-driven community to obtain insights on the transactions performed in this platform.
To those investors, the use of this study not only in New York, but potentially also in any other cities in the world, the host-investors can expand the analysis if it is worth going with Airbnb or not. The incentives of trying this new platform out is based on how transactions are performed, but, most importantly, where they are performed and what kind of results could be obtained, depending on where the property is, whether it is close to a touristic attraction or not, or next to the city center or surrounded by high-end neighborhoods.
So, picking up a metric, whether it is the earnings, the number of reviews, how recent the last booking was, a combination of the three and many others comparing them in relation to where the property is located, the result could create incentives to go with Airbnb or detract the interest and incentive property owners to stick to the conservative model to potentialize earnings. Geography plays a big role in New York when people are about to decide where to stay for the night when travelling around.
The geointelligence had relevance in modelling regressions and improved accurancy of explanation. The concepts of geotechnology and geointelligence modelled in the pricing strategy for property owners were an optimized methodology implemented in the analysis of joining Airbnb, notably for property owners competing in a distributed territory in an urban area. The geographic variables increased the predictability of the model by 62%, from the original model with R^2 of 34.59% to the final one with R^2 of 55.89%. The results achieved also enlights that the problem presents spatial autocorrelation, which indicates that rental prices can be explained by spatial dependence.
Other factors, such as cleanliness', the presence of Wi-Fi, nice décor, furniture, host welcoming are important but those are attributes that can be arranged easier than the location of a property. In fact, a property could have all amenities and luxurious furniture, a nice host with high availability and great customer service but if their house is located far away from popular spots, chances of bookings are lower than those located in a nicer popular area.
Taking this geographical model and scaling to other places can result in similar results. Other

variables would have to be considered as well, like violence, criminality, presence of public transportation and others, as different cities differ in so many aspects from NYC, but geography could be again impactful to the strategy for hosts. Location matters and closeness to tourist attractions matters a lot in New York. In addition, some neighborhoods tend to behave in price just similarly as their close neighborhoods. So, if an area next to my house has higher booking procis, I tend to higher up my booking price as well. Same behavior was detected with lower prices. Higher prices were spotted in North Manhattan area, around Central Park and North Brooklyn. Bronx was a neighborhood where price averages were lower and so were neighborhoods in the surroundings. Queens was very insignificant in discriminant analysis and kept neutral in most exploratory techniques. Staten Island acted just like Bronx.
And last, New York attractions show big influences on prices so the closer your property is to Times Square, for example, the higher you could charge guests.

BIBLIOGRAPHY

ABDELAZIZ, L., REED, M., TYLER, M., & YUQING, Z. Impact of reviews on price: Evidence from sentimental analysis of Airbnb reviews in Boston. Kentucky: Western Kentucky University, 2017.

ALMEIDA, Eduardo. Econometria Espacial Aplicada. São Paulo: Editora Alínea, 2012.

BOECKMANN M. The shared economy: It is time to start caring about sharing; value creating factors in the shared economy. University of Twente, Netherlands, June 27, 2013. [Online]. Available: http://thelovettcenter.com/
wpcontent/uploads/2014/11/bockmann-shared-economy.pdf

DUBIN, R. A. Predicting House Prices Using Multiple Listings Data. Journal of Real Estate Finance and Economics, 17(1), 35-59, 1998.

EBDON, D. 1987. Statistics in geography. 2nd edition. New York, NY Basil Blackwell Ltd. 232 p.

FLATER, D. Understanding Geodesic Buffering. Esri Geoprocessing Development Team. 2011. Developer's Corner. [Online]. Available:
https://www.esri.com/news/arcuser/0111/files/geodesic.pdf

FRANCISCO, E. R., ARANHA, F., ZAMBALDI, F., GOLDSZMIDT, R. Electricity Consumption as a Predictor of Household Income: a Spatial Statistics Approach. Advances in Geoinformatics. pg. 267-282. GEOINFO 2006. Campos de Jordão, Brazil, 2006.

FRANCISCO, E. R. RAE-eletrônica: Exploração do Acervo à Luz da Bibliometria, Geoanálise e Redes Sociais. ERA (Impresso), v. 51, p. 280-306, 2011.

GIBBS C., GUTTENTAG D., GRETZEL U., MORTON J., & GOODWILL A. Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings. Journal of Travel & Tourism Marketing, 35, 46-56. GOBIERNO DE ESPAÑA. MINISTERIO DE ENERGÍA, T. Y. Encuesta de movimientos turísticos en fronteras (Frontur), May, 2018. [Online]. Available: http://estadisticas.tourspain.es/eses/estadisticas/frontur/paginas/default.aspx

GUNTER, U. What makes an Airbnb host a superhost? Empirical evidences for San Francisco and the Bay Area. Tourism Management,66, 26-37, 2018.

GUTTENTAG, D. A., & SMITH, S. J. Assesing Airbnb as a disruptive innovation relative to hotels: Substitution and comparative performance expectations. International Journal of Hospitality Management, 64, 1-10, 2017.

HUEY-YEH LIN; MEI-HSIANG WANG; MIN-JHEN WU. A Study of Airbnb Use Behavior in the Sharing Economy. International Journal of Organizational Innovation, [s. l.], v. 10, n. 1, p. 38–47, 2017.

JUN, L., MORENO, A., & ZHANG, D. J. Agent Behavior in the Sharing Economy: Evidence from Airbnb. Michigan: Michigan Ross, 2015.
KENNEDY, R. D., DOUGLAS, O., STEHOUWER, L., & DAWSON, J. The availability of smoking-permitted accomodations from Airbnb in 12 Canadian cities. Tobacco Control, 27(1), 112-116, 2018.
LIANG, L. J., CHOI, H. C., & JOPPE, M. Exploring the relationship between satisfaction, trust and switching intention, repurchase intention in the context of Airbnb. International Journal of Hospitality Management, 41-48, 2018.
MAKARAND, A. M., SUESS, C., & XINRAN, L. The accommodation experiencescape: a comparative assessment of hotels and Airbnb. International Journal of Contemporary Hospitality Management, 29, 2377-2404, 2017.
SCHRAGE, M. How is Big Data transforming your 80/20 analytics? Harvard Business Review Insight Center Report From Data to Action, 2014. [Online]. Available: https://www.sas.com/pt_br/offers/sem/hbr-from-data-to-action.html?utm_source=Google&utm_medium=Search&utm_campaign=Big_Data&gclid=EAIaIQobChMIia2U24ek5AIVgRx9Ch2lKwmWEAMYASAAEgKY3PD_BwE
OXERA. What is the economic impact of Geo services? 2013.
SO, K. F., OH, H., & MIN, S. Motivations and constraints of Airbnb consumers: Findings from a mixedmethods approach. Tourism Management, 224-236, 2018.
VARGAS-HERNÁNDEZ, J. G.; RUEDA GÓMEZ, L. V. Consumption Characteristics of the Millennial Generation and Strategies Implemented to Satisfy Them: Airbnb Case. ANWESH: International Journal of Management & Information Technology, [s. l.], v. 4, n. 2, p. 1–6, 2019.
WANG, D., & NICOLAU, J. L. Price determinants of sharing eocnomy based accommodation rental: A study of listing from 33 cities on Airbnb.com. International Journal of Hospitality Management, 120-131, 2016.
ZERVAS, G., PROSERPIO, D., BYERS, J. W. The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. Journal of Marketing Research (JMR), [s. l.], v. 54, n. 5, p. 687–705, 2017. DOI 10.1509/jmr.15.0204.
ZHANG, H., ZHANG, J., LU, S., CHENG, S., & ZHANG, J. Modeling hotel room price with geographically weighted regression. International Journal of Hospitality Management, 30, 1036-1043, 2011.
ZHANG, Z., CHEN, R. C., HAN, L. D., & YANG, L. Key factors affecting the price of Airbnb listings: a geographically weighted approach. Sustainablity, 9, 1635, 2017.