Università di Pisa
Sant'Anna Scuola Universitaria Superiore Pisa
Consiglio Nazionale delle Ricerche
Università degli Studi di Cagliari

Book of Short Papers

SIS 2021

SIS 2021
PISA, 21-25 GIUGNO 2021

SIS Società Italiana di Statistica

Editors: Cira Perna, Nicola Salvati and Francesco Schirripa Spagnolo

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891927361

A statistical model to identify the price determinations: the case of Airbnb.

Un modello statistico per identificare le determinanti del prezzo: il caso Airbnb.

Giulia Contu, Luca Frigau, Gian Paolo Zammarchi, Francesco Mola

Abstract We propose an indicator to estimate the effects of *transports*, *culture*, *crowd*, *managerial*, *accommodation* dimensions on price determination for Airbnb accommodations. The indicator is defined using a proportional odds model in two steps. In the first phase, we estimate for each accommodation the probability to belong to a specific category of the price moving from *very-low* to *very-high*. Then, we estimate the average concentration index to identify which is the price class more likely for each observation and which dimension can better explain the price. Then, we assign the price class to each observation using the median of the probabilities of the model's fitted values and identify the most significant dimension. Afterward, we aggregate the concentration index calculated for each observation for the district of *Trastevere* in Rome, with the aim to identify its most relevant dimensions. The results highlight a significant impact of the *managerial* and *accommodation* dimensions.

Abstract *Si propone un indicatore per stimare l'effetto delle dimensioni del* trasporto, cultura, affollamento, management *e* caratteristiche dell'alloggio *sulla determinazione del prezzo per gli alloggi Airbnb. L'indicatore è definito utilizzando un proportional odds model come spiegato nel Vector Generalized Additive Model. L'indicatore viene costruito in due fasi. Nella prima fase si stima la probabilità*

Giulia Contu
University of Cagliari, Department of Economics and Business Sciences, Viale Sant'Ignazio 17, 09123, Cagliari, e-mail: giulia.contu@unica.it

Luca Frigau
University of Cagliari, Department of Economics and Business Sciences, Viale Sant'Ignazio 17, 09123, Cagliari, e-mail: frigau@unica.it

Gian Paolo Zammarchi
University of Cagliari, Department of Economics and Business Sciences, Viale Sant'Ignazio 17, 09123, Cagliari, e-mail: gp.zammarchi@unica.it

Francesco Mola
University of Cagliari, Department of Economics and Business Sciences, Viale Sant'Ignazio 17, 09123, Cagliari, e-mail: mola@unica.it

di ogni alloggio di appartenere ad una delle cinque categorie di prezzo che vanno da molto basso *a* molto alto*. Successivamente, si stima l'indice di concentrazione medio per identificare la fascia di prezzo in cui ricade l'osservazione e quale dimensione possa spiegare meglio il prezzo. Poi, la fascia di prezzo viene assegnata a ciascuna osservazione utilizzando la mediana del vettore di probabilità e si identificano le dimensioni più significative. Successivamente, l'indice di concentrazione calcolato per ciascuna osservazione viene aggregato per il quartiere di Trastevere di Roma, con l'obiettivo di individuare la dimensione più rilevanti per il quartiere. I risultati hanno evidenziato un impatto significativo della dimensione* management *e* caratteristiche dell'alloggio*.*

Key words: proportional odds model, price determination, price dimensions, Airbnb accommodation, price indicator

1 Introduction

Airbnb is one of the most famous platforms where it is possible to book apartments, private and shared rooms. It has been founded in 2008 in San Francisco and has grown significantly over the years. It operates in more than 65,000 cities and 191 countries and sells millions of room nights to tourists and travelers all around the globe. Tourists choose Airbnb accommodation for different aspects, such as a wider range of listings, a favorable price-quality ratio, a lower price, the possibility of choosing between a private or a shared environment, the possibility of meeting new people, or living a more authentic experience [3].
Different researchers have previously investigated which aspects can impact on the Airbnb prices and they have proposed different models. Generally, these models include the price as the response variable, and features grouped in categories referable to site characteristics, reputation, convenience, personal, and amenities attributes as independent variables [2]. The methodologies included in these models are ordinary least squares (see for instance [4]), panel data analysis (see for instance [5]), quantile regressions (see for instance [6]), hedonic price models (see for instance [2]), price equations with spatial effects [7], pricing strategy model [8], machine learning (see for instance [9]).
In this work, we propose a statistical composite indicator of the price that estimates the impact of five different dimensions on Airbnb accommodations' prices. The indicator is defined using *proportional odds model* as explained in Vector Generalized Additive Model by [10]. In our analysis we considered the Airbnb accommodation located in the district of Rome called *Trastevere*.
Five sections, besides the introduction, complete this study. The second and third sections are related to the research design: data and methodology are described. The results are explained in the fourth section. Finally, the fifth section focuses on concluding remarks, limitations, and future developments.

2 Data

The study has been realized using a dataset composed of six groups of variables: i.e. *transports*, *culture*, *crowd*, *managerial*, *accommodation* (Table 1). Specifically, *transports* data consider the minimum distance from the accommodations to the closest subway stop, the number of bus stops in a range of 200 meters, and the distance from the accommodations to the city center (i.e. Pantheon). The *crowd* dimension measures the number of other Airbnb accommodations and hotels in a range of 50 and 500 meters. The data about the *culture* include the number of monuments in a range of 500 and 2000 meters. The *managerial* dimension identifies the aspects that can be directly chosen by the hosts as the provided services and the additional fees to be applied. The *accommodation* data include the listing types offered on the Airbnb platform, the number of bedrooms and bathrooms. Finally, the accommodation published nightly rate (*price*) has been categorized into five classes according to their quantiles and labeled as *very-low*, *low*, *medium*, *high*, *very-high*. The data have different origins: those related to *transport*, *crowd*, and *culture* have been downloaded from the website *Open Data Roma Capitale* [1]; the others have been provided by the company *Airdna*.
In this study, we take into account only the year 2016 because the data we had available covered only that year.

Table 1 Variables

Label	Group	Description
price	instrumental	Classes of Published Nightly Rate (euros)
distCenter	transports	Distance from city center (meters)
minDistSubway	transports	Minimum distance from the closest subway stop (meters)
distBus200	transports	Number of bus stops in a range of 200 meters
airbnb_close50	crowd	Number of airbnb in a range of 50 meters
airbnb_close500	crowd	Number of airbnb in a range of 500 meters
hotels_close50	crowd	Number of hotels in a range of 50 meters
hotels_close500	crowd	Number of hotels in a range of 500 meters
monuments_close500	culture	Number of monuments in a range of 500 meters
monuments_close2000	culture	Number of monuments in a range of 2000 meters
Bedrooms	accommodation	Number of bedrooms in a vacation rental listing
Bathrooms	accommodation	Number of bathrooms in a vacation rental listing
Privateroom	accommodation	Dummy variable: private room type (binary variable)
Entirehome	accommodation	Dummy variable: entire home type (binary variable)
Cancellationpolicy	managerial	Cancellation policy for the vacation rental listing (binary variable)
ResponseTimemin	managerial	Average time in minutes a host responds to (minutes)
MinimumStay	managerial	The default minimum night stay required by host
BusinessReady	managerial	Host who provides business facilities (binary variable)
Superhost	managerial	High quality experienced host (binary variable)
NumberofPhotos	managerial	Number of photos in a vacation rental listing

3 Methodology

To define the indicator, we have used the *proportional odds model*, also called *cumulative logit model*, as explained in the Vector Generalized Additive Model by [10]. The model is used when the Y is ordinal and it is defined through:

$$\text{logit } P(Y \leq j|\mathbf{x}) = \eta_j(\mathbf{x}), \tag{1}$$

subject to the following constraint

$$\eta_j(\mathbf{x}) = \beta^*_{(j)1} + \mathbf{x}^{\mathrm{T}}_{[-1]}\beta^*_{[-(1:M)]}, \qquad j = 1,\ldots,M \tag{2}$$

where j identifies the level of Y and moves form 1 to M; $\mathbf{x}_{[-1]}$ is the $\mathbf{x}$ with the first element deleted; $*$ denotes the regression coefficient that are to be estimated [10, p. 11]. We fit a model for each of the five dimensions, in which we use the corresponding features to estimate the price class. In other words, the model allows estimating the probability that a specific accommodation belongs to one of the five classes of the price by using the feature of a single dimension. We assume that the higher concentration of probabilities of the model's fitted values in a single price class the better capability of that dimension in explaining the price variable. In order to measure the probability concentration for each observation i we used the complementary of the normalized Gini index

$$\rho = 1 - \frac{1 - \sum_{j=1}^{5} f_{ij}^2}{4/5} \tag{3}$$

where j identified the five price classes.
We define two tools to discover which group of variables is more important in explaining the price. The first tool computes the average of the ρ_i as

$$\bar{\rho} = n^{-1} \sum_{i=1}^{n} \rho_i \tag{4}$$

that assumes a value in the range $[0,1]$ and provides directly the information about the importance of the single group in the definition of the price class. The second tool assigns a price class to each observation by using the median of the probabilities of the model's fitted values, and then operates an aggregation of the ρ_i by the price classes estimated by the model, where the median corresponds to the estimation of the price class for the whole zone.

4 Results

We have applied the model on the data related to the 1802 Airbnb accommodations located in the district of *Trastevere* in Rome. Firstly, we have fitted the models and then estimated the probabilities for each observation to belong to the five price classes. We have estimated that the number of Airbnb accommodation in the class *very-low* is equal to 98, in the class *low* is equal to 257, in the class *median* s equal to 424, in the class *high* s equal to 540 and, finally, in the class *very-high* s equal to 483. The main results are illustrated in Table 2. It emerges that the price is mainly determined by *accommodation* (58.76%) and *managerial* (40.92%), whilst the other three dimensions are less significant in explaining the price in this district. Then we can define the most representative price class for the district of *Trastevere*. The results show that the accommodations are rented at a *high* price.

Table 2 Price classes and dimensions relevance (ρ) in Trastevere.

Price class	*transports*	*crowd*	*culture*	*accommodation*	*managerial*	Tot
very-low	0.00	0.00	0.00	0.41	220.86	221.27
low	0.00	0.00	0.00	309.23	414.55	723.78
median	3.50	3.88	1.44	176.01	87.92	272.75
high	0.00	0.00	0.00	287.40	388.84	676.24
very-high	0.00	0.00	0.00	876.58	36.72	913.30
Tot	3.50	3.88	1.44	1649.63	1148.89	
Tot in %	0.12	0.14	0.05	58.76	40.92	

5 Conclusions

We have defined a price indicator composited by five dimensions using a proportional odds model. The indicator allows evaluating the relevance of each dimension on the price and the dominant price class range in the district. We focused on *Trastevere*, one of the most famous and tourist neighborhoods of Rome. The results have highlighted the relevance of dimensions *managerial* and *accommodation*. Less impact has been recorded for the other groups of variables.
We recognize some relevant aspects in the definition of the indicator. Firstly, we use the average concentration index to define the relevance of each dimension. The index allows attributing weights able to measure the impact of each group of variables has on the price. The second innovative aspect is related to the estimation level. We offer a double point of view of the price determination. To one side, we define a model able to evaluate the indicator for each accommodation offering the possibility to support the single host in the price determination. On the other side, we aggregate the results in terms of the geographical area to better explain the price and its determinants taking into account the geographical proximity.

We identify some limitations in this study. Firstly, we have analyzed the impact of the determinants' price only for one district. Taking into account more districts it can be interesting to comprehend if weights attribute at the dimensions can assume different values concerning the characteristics of the specific area of the city. Secondly, the price indicator has been estimated for a geographical area. However, we believe that inserting the time variation can be interesting to evaluate possible differences in terms of dimensions' impact. We are not sure for instance if the proximity of the subway and bus stops have the same impact in the different annual seasons.

References

1. https://dati.comune.roma.it/
2. Faye B (2021) Methodological discussion of airbnb's hedonic study: A review of the problems and some proposals tested on bordeaux city data. Annals ofTourism Research 86:103079
3. Skalska T (2017) Sharing economy in the tourism market: Opportunities and threats. KNUV (4 (54)):248–260
4. Voltes-Dorta A, Sánchez-Medina A (2020) Drivers of airbnb prices according to property/room type, season and location: A regression approach. Journal of Hospitality and Tourism Management 45:266—275
5. Falk M, Larpin B, Scaglione M (2019) The role of specific attributes in determining prices of airbnb listings in rural and urban locations. International Journal of Hospitality Management 83:132—140
6. Wang D, Nicolau JL (2017) Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb. com. International Journal of Hospitality Management 62:120–131
7. Tang LR, Kim J, Wang X (2019) Estimating spatial effects on peer-to-peer accommodation prices: Towards an innovative hedonic model approach. International Journal of Hospitality Management 81:43—53
8. Ye P, Qian J, Chen J, Wu Ch, Zhou Y, De Mars S, Yang F, Zhang L (2018) Customized regression model for airbnb dynamic pricing. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 932—940
9. Kalehbasti PR, Nikolenko L, Rezaei H (2019) Airbnb price prediction using machine learning and sentiment analysis. arXiv preprint arXiv:190712665
10. Yee TW (2015) Vector generalized linear and additive models: with an implementation in R. Springer