Machine Learning with Applications 7 (2022) 100208

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Airbnb rental price modeling based on Latent Dirichlet Allocation and MESF-XGBoost composite model

Md Didarul Islam [a,*], Bin Li [b], Kazi Saiful Islam [c], Rakibul Ahasan [d], Md. Rimu Mia [c], Md Emdadul Haque [e]

[a] Department of Geography and Geoinformation Sciences, George Mason University, Fairfax, VA 22030, USA
[b] Department of Geography & Environmental Studies, Central Michigan University, Mount Pleasant, MI 48859, USA
[c] Urban & Rural Planning Discipline, Khulna University, Khulna 9208, Bangladesh
[d] Department of Geography, Texas A&M University, College Station, Texas 77843, USA
[e] Department of Geology and Mining, Barishal University, Barishal 8254, Bangladesh

ARTICLE INFO

Keywords:
Machine Learning
Latent Dirichlet Allocation
Eigenvector Spatial Filtering
XGBoost
Spatial Data Modeling

ABSTRACT

Airbnb price modeling is an important decision-making tool that determines the acceptability and profitability of the service. In this study, we demonstrated how proper descriptions of an Airbnb listing and location could influence determining the prices. We assumed the proper description of a listing property positively influences the renter's decision making; therefore, we applied a Latent Dirichlet Allocation (LDA) based topic model for generating synthetic variables from the textual description of property aiming to improve price prediction accuracy. Additionally, we applied a Moran Eigenvector Spatial Filtering based XGBoost (MESF-XGBoost) model to address the spatial dependence of location data and improve prediction accuracy. Our study at the San Jose County Airbnb dataset found that the number of bedrooms, accommodations, property types, and the total number of reviews positively influence the listing price, whereas the absence of a super host badge and cancellation policy negatively influence the price. The experiment demonstrates that incorporating synthetic variables from both LDA and MESF into the model specification improves the prediction accuracy. The experiment reveals that the XGBoost model with only non-spatial features is not strong enough to address spatial dependence; therefore, it cannot minimize spatial autocorrelation issues.

1. Introduction

Airbnb has expanded at an exponential rate since its inception in 2008 and is now bigger than the top five hotel brands put together (Hartmans, 2017). The industry is booming, with over 4 million hosts in more than 220 countries worldwide as of 2021 (Zervas, Proserpio, & Byers, 2021). The growing demand for Airbnb and lucrative business opportunities are attracting more homeowners to join the industry. In addition to the business community, academia has also paid due attention as it has become increasingly popular among tourists. An accurate valuation model of new host listing prices is desired by both owners and renters to trade-off between owner profit and customer satisfaction. Numerous studies have focused on Airbnb's benefits, problems (Guttentag, 2013; Lampinen & Cheshire, 2016; Meleo, Romolini, & De Marco, 2016), and legal issues (Edelman & Geradin, 2015; Lee, 2016), but only a few have investigated the aspects that influence Airbnb prices and accurate estimation (Dudás, Boros, Kovalcsik, & Kovalcsik, 2017). Researchers are continuously trying to make an accurate price evaluation model to aid the property owners (Sainaghi, 2020). Nevertheless, Airbnb pricing is still an ongoing research agenda as the number of scholarly works has been growing substantially in the past few years with the availability of more Airbnb listing data and exploitation of sophisticated machine learning, and natural language processing algorithms have the potentiality to make the estimation accuracy more realistic.

Several studies applied machine learning algorithms earlier for the accurate estimation of rental prices (Priambodo & Sihabuddin, 2020; Yang, 2021; Zhu, Li, & Xie, 2020). The major limitation of earlier studies was that they excluded consideration of Tobler's first law of geography – "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Earlier

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: mislam25@gmu.edu (M.D. Islam), li1b@cmich.edu (B. Li), saiful@urp.ku.ac.bd (K.S. Islam), rahasan@tamu.edu (R. Ahasan), rimumiandc86@gmail.com (M.R. Mia), emdad6344@gmail.com (M.E. Haque).

https://doi.org/10.1016/j.mlwa.2021.100208
Received 15 August 2021; Received in revised form 3 November 2021; Accepted 4 November 2021
Available online 23 November 2021

studies that incorporated spatial data in their models to examine pricing trends often lacked consideration of spatial dependence (Ahasan, 2019a, 2019b). It is established among the scientific community that spatial data has spatial dependence property, which reflects a situation where observation at one location or region tends to exhibit similar values from neighboring locations (LeSage, 2008). Therefore, excluding the spatial dependence effect from the machine learning-based model building process results in a model that lacks the capability to detect the spatial pattern and minimize spatial autocorrelation issues (Griffith, 1987). A common phenomenon in price modeling is that clusters of high and low-price neighborhoods can be observed across the city (Brueckner, Thisse, & Zenou, 1999). Gyódi and Nawaro (2021), in their study of the determinants of Airbnb prices in 10 major EU cities, found that listing in central city areas forms more expensive clusters while suburban areas tend to be characterized by lower values. Their study concluded that location is a significant factor in determining Airbnb listing prices. The luxury of spaciousness and higher accessibility of attractions in the city centers plays a crucial role in higher prices (Brueckner et al., 1999; Cai, Zhou, (Jenny) MA, & Scott, 2019).

The spatial variation of prices across the space cannot be addressed by pure data-driven machine learning-based models as they lack the capability to address spatial dependence. This aspect is often missing in the price modeling, not only in the pricing in the hospitality industry but most frequently in housing price modeling (Ahasan, 2019a). Therefore, this study proposes an eigenvector spatial filtering-based XGBoost model, which is capable of addressing spatial dependence by incorporating spatially filtered eigenvectors while building a price model. The role of spatially filtered eigenvectors as a proxy variable of spatial information of the study area would be addressing spatial dependence by interacting with other covariates. This study also utilizes the Latent Dirichlet Allocation (LDA) topic model to enrich the data from the textual description of the property. The main objective of this study is to develop a reliable Airbnb price prediction model by leveraging the power of machine learning and demonstrate the significance of incorporating spatial information into the model specification to improve prediction accuracy.

2. Literature review

In the hospitality industry, such as Airbnb, setting the right price (or rent) is crucial for an efficient business model (Lampinen & Cheshire, 2016), and knowing the elements that influence pricing is therefore quite valuable, as it may assist hosts in setting a reasonable price so that both the hosts and the renter profit from the sharing economy (Zhang, Chen, Han, & Yang, 2017). However, past studies, as well as the owners, had acknowledged that determining a reasonable Airbnb listing price is a challenging task as several factors, including physical attributes of the property, neighborhoods, time of the week, month, or year, and most importantly, location of the property influence the listing price (Wang & Nicolau, 2017). Driven from previous research on Airbnb pricing, it has found that different type of factors affects the price of Airbnb in different magnitude and direction (Sainaghi, 2020). Most often, Airbnb rental prices are determined by the kind of accommodation the host offers (Chen & Xie, 2017; Gibbs, Guttentag, Gretzel, Morton, & Goodwill, 2017; Wang & Nicolau, 2017). In addition to that, the number of bedrooms and bathrooms also has a favorable effect on the price of an Airbnb rental (Chen & Xie, 2017; Ert, Fleischer, & Magen, 2016; Gibbs et al., 2017; Wang & Nicolau, 2017).

Past studies reported the influencing variables of Airbnb rental prices, not only considering the physical attributes of the listing but also incorporating the social and demographic variables as well. Kakar, Voelz, Wu, and Franco (2016) investigated the influence of host information and discovered that Asian and Hispanic hosts charge a lower price on Airbnb in San Francisco than their white counterparts. However, studies did not find significant evidence of influence by the gender, marital status, or sexual orientation of the Airbnb host on the room pricing (Ert et al., 2016; Kakar et al., 2016). Teubner, Hawlitschek, and Dann (2017) used a large-scale dataset from 86 German cities and found that indices such as hosts' ranking scores and membership length are related to economic worth. A 'super host' label on Airbnb listing also boosts the demand and prices (Gibbs et al., 2017; Kakar et al., 2016; Wang & Nicolau, 2017). Similarly, professional hosts with two or more Airbnb listings can earn a greater rental rate for each home than nonprofessional hosts (Gibbs et al., 2017). Gutt and Herrmann (2015) looked at how star ratings and visibility impact listing prices on the Airbnb site and found that rating star visibility increases prices. Rental policies such as instantly bookable listings are less expensive, and refundable cancellation policies are associated with lower rental prices (Gibbs et al., 2017; Wang & Nicolau, 2017), which is contrasting to the findings from similar studies with hotels (Jun & Arendt, 2016; Latinopoulos, 2018). Several studies also revealed that location is the most crucial element in Airbnb rental price determination (Chica-Olmo, González-Morales, & Zafra-Gómez, 2020; Gyódi & Nawaro, 2021). The number of places of interest (POIs) in the surrounding areas such as tourist attractions, restaurants, or shopping centers also positively influences an Airbnb unit's listing price.

Statistical and Machine Learning models can estimate Airbnb rental prices by building a functional relationship between factors and Airbnb prices. The individual effect of the abovementioned factors on Airbnb prices is considered while making functional relationships in the algorithms. Some flexible algorithms exhibit superiority over others in data modeling as they have better optimization algorithms for minimizing the loss function, such as stochastic gradient descent in XGBoost and backpropagation in Artificial Neural Network (ANN). Therefore, choosing an efficient model is also crucial for the accurate estimation of rental prices. Zhu et al. (2020) applied several machine learning algorithms to build the price prediction model for the New York Airbnb market and found that the XGBoost and Random Forest (RF) are highly effective for price prediction. Priambodo and Sihabuddin (2020) applied an Extreme Learning Machin (ELM) Model to predict Airbnb base price in London, where they found ELM can outperform the XGBoost model. Yang (2021) applied neural networks and XGBoost for the Beijing Airbnb market, where they showed the XGBoost could outperform the neural network model. This study would apply a composite MESF-XGBoost model on the enriched dataset by text mining model LDA aiming to improve the accuracy further. Additionally, this study would apply RF and statistical regression models, which would serve as a benchmark model to compare the improvement of the prediction accuracy of the proposed model.

3. Materials and method

3.1. Study area and data preprocessing

The San Jose County Airbnb dataset listed between 2009 and 2020, scrapped in May 2020 from the Airbnb website, is used for experimental demonstration. The dataset consisted of a total of 7222 data points across the county. After removing the missing values and unnecessary variables (e.g., city name, market, etc., there was a total of 4426 data points with 35 variables (Fig. 1).

We converted several category variables (e.g., property types, cancellation policy) into binary variables and treated each binary variable as explanatory variables in the model specification. Then, we applied a stepwise regression to subset the most influential variables from the dataset. Finally, the selected variables (Table 1) from the stepwise regression method were used for model building.

The histogram of listing price distribution shows it is rightly skewed (Fig. 2a); therefore, we applied log transformation to make it less skewed. The histogram shows that the log-transformed values are normally distributed (Fig. 2b).

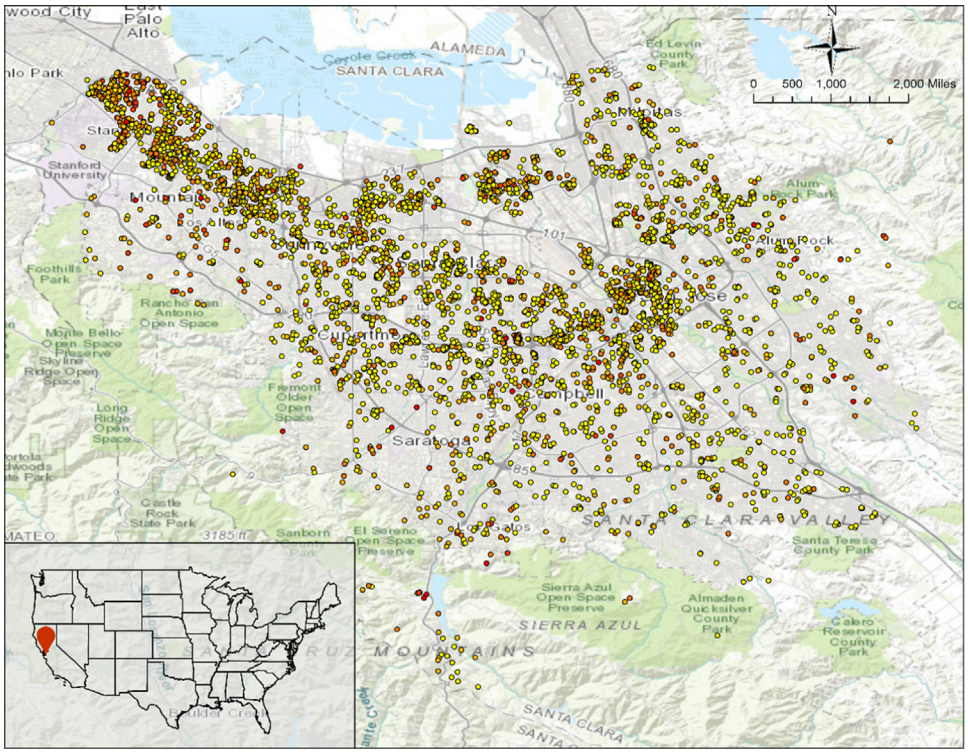

Fig. 1. Location of data points (Airbnb listings) in San Jose County, California.

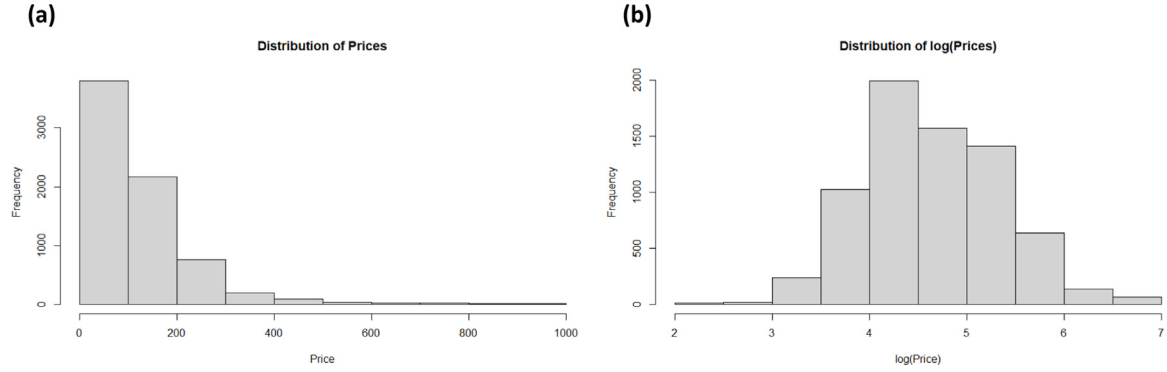

Fig. 2. Distribution of housing sales prices (a) without transformation (b) after log transformation.

Table 1
Airbnb data properties in San Jose County.

Variables	Description
Accommodates	Number of people who can stay
Bedrooms	Number of bedrooms
Number_of reviews	Total number of reviews
Host listing count	Total number of hosting in homes or shared rooms
Reviews_per_months	Total number of reviews in every month
Property type	If the property type is an apartment, house, or shared room.
Superhost	Whether the host is superhost based on certain criteria
Cancellation_policy	Whether the cancellation policy is flexible or moderate
Price	Price per night

3.2. Latent Dirichlet allocation (LDA)

LDA is a popular topic model in the natural language procession (NLP) domain that discovers topics from a collection of documents (Blei, Ng, & Jordan, 2003). It assumes a particular topic is a distribution of a set of words (N) in a document (m) and uncovers the most suitable topics (K) from a collection of documents (M). Initially, the LDA algorithm randomly assigns each word from a document to one of topics K. Later, it calculates the proportion of words (W) in each document M and represents as a probabilistic distribution of Dirichlet on latent topics. Specifically, its attempts to capture how many documents belong to topic K because of the word W. The graphical presentation of the LDA algorithm is shown in Fig. 3.

Here, α is a Dirichlet distribution parameter, θ is the expected topic proportion of a document m, Z represents the assignment of a word in a topic K.

The LDA model was applied to extract the probability distribution of words in the neighborhood overview description of each property. The following graph (Fig. 4) shows the probability distribution of words in the neighborhood overview feature in the dataset. It seems the LDA model generates two topics: the first topic is related to distance, indicating nearby streets, downtown, airport, restaurants, and spatial proximity, and the second topic is about neighborhood views such as parks, mountains, downtown, quiet places, etc. The probability of two topics in each neighborhood overview description of each property would be used as synthetic variables in model specification.

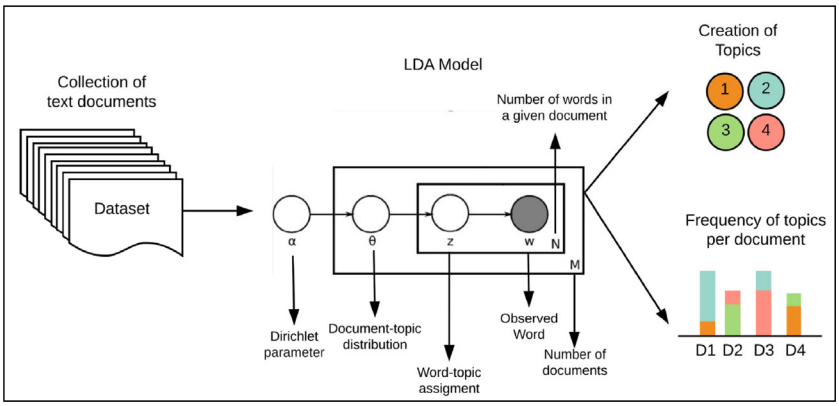

Fig. 3. Schematic of LDA algorithm by Buenano-Fernandez, Gonzalez, Gil, and Lujan-Mora (2020).

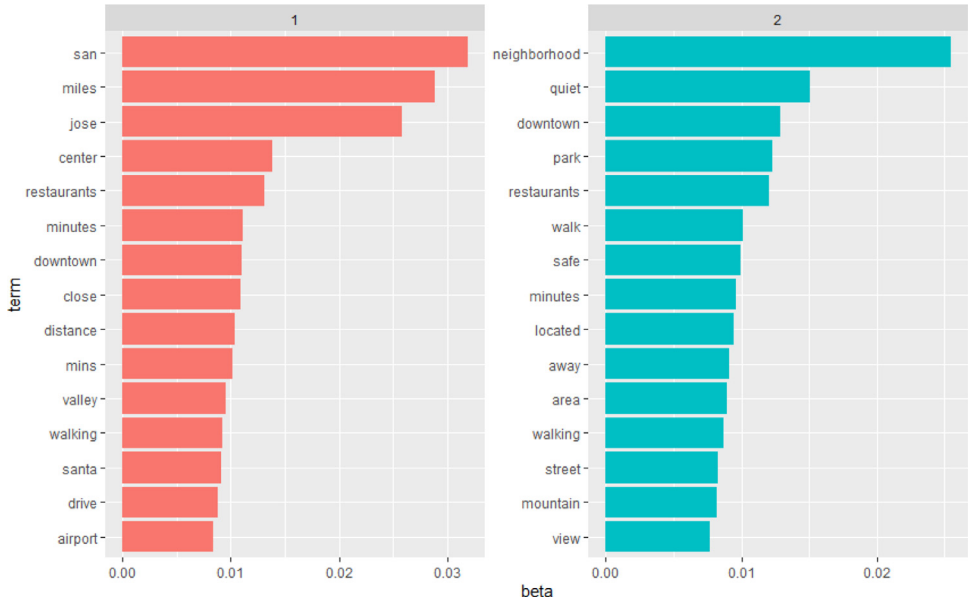

Fig. 4. LDA generated two topics from neighborhood overview description.

3.3. Moran Eigenvector Spatial Filtering (MESF)

The Moran Eigenvector spatial filtering (MESF) is a popular method of addressing spatial autocorrelation issues in statistical regression models by expressing the spatial structure of a region at different spatial scales and incorporating them as a set of proxy variables of spatial information in the model specification (Griffith, 2000, 2013; Griffith & Chun, 2013; Griffith, Chun, & Li, 2019). The synthetic variables generated from the spatial filtering approach can address spatial dependence by interacting with systematic covariates (Wang, Kockelman, & Wang, 2013). The basic linear model of MESF is formulated as follows:

$$\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{E\gamma} + \boldsymbol{\varepsilon}, \quad \varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \tag{1}$$

Where $\mathbf{y}$ is an $N\times1$ vector of response variable values; $\boldsymbol{X}$ represents a set of covariates where β is the coefficient, $\boldsymbol{E}$ is an $N\times L$ eigenvector matrix, and γ is the coefficient of $\boldsymbol{E}$ eigenvectors; ε is an $N\times1$ vector of disturbances. Here, eigenvectors $\boldsymbol{E}$ portray distinct, selected orthogonal map patterns which describe spatial dependence. However, all spatial eigenvectors cannot capture spatial autocorrelation and filtering out the useful eigenvectors is essential to avoid producing noise in the data model. Murakami and Griffith (2015) outlined a Nystrom extension-based fast approximation of Moran eigenvectors and eigenvalues and found that 200 eigenpairs are sufficient to model spatial dependence and address spatial autocorrelation. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are estimated by Ordinary Least Squares (OLS) estimation-based stepwise methods. Later, Murakami and Griffith (2015) proposed Markov Chain Monte Carlo (MCMC) techniques to estimate parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ that made the estimation more unbiased and showed promising results over the initial method. The Random Effect-based ESF method can be expressed as follows:

$$\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{E\gamma} + \boldsymbol{\varepsilon}, \quad \gamma \sim N\left(\mathbf{0}, \sigma_\gamma^2\Delta\right), \quad \varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \tag{2}$$

where σ_γ^2 is a variance parameter, Δ is an $L\times L$ diagonal matrix whose diagonal elements are the selected set of eigenvalues λ_L corresponding to L eigenvectors in $\mathbf{E}$. The following graph shows the first eight (8) eigenvectors of the dataset (Fig. 5).

3.4. Extreme Gradient Boosting (XGBOost)

XGBoost is a popular and frequently used Classification and Regression Tree (CART) model designed to efficiently reduce computing time and scale up tree boosting algorithms (Chen & Guestrin, 2016). It is an implementation of gradient boosted decision trees and used for supervised learning where training data $\boldsymbol{xi}$ is used to predict a target variable $\boldsymbol{yi}$. The prediction score of each tree is summed up to get the final score, which is evaluated through N additive functions to predict

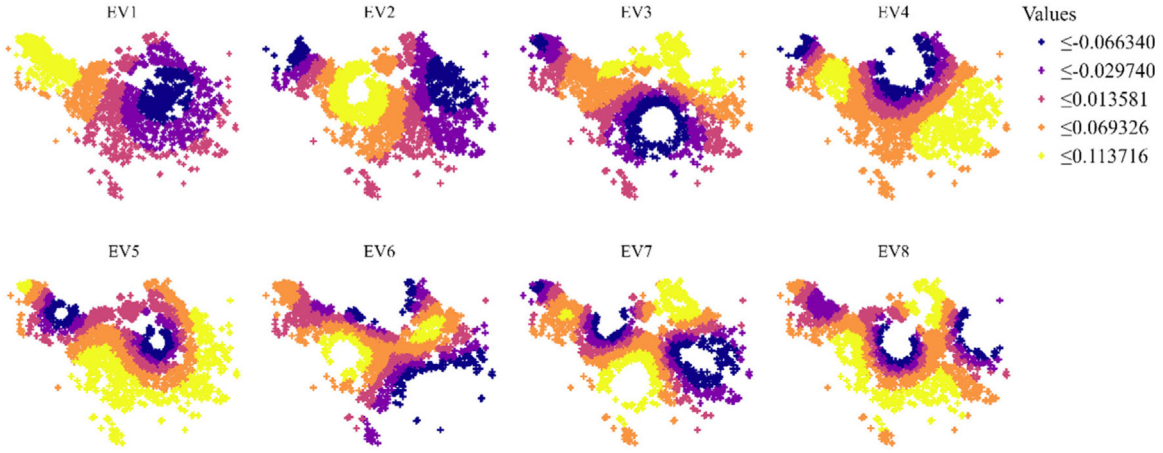

Fig. 5. First eight eigenvectors of the San Jose County location dataset.

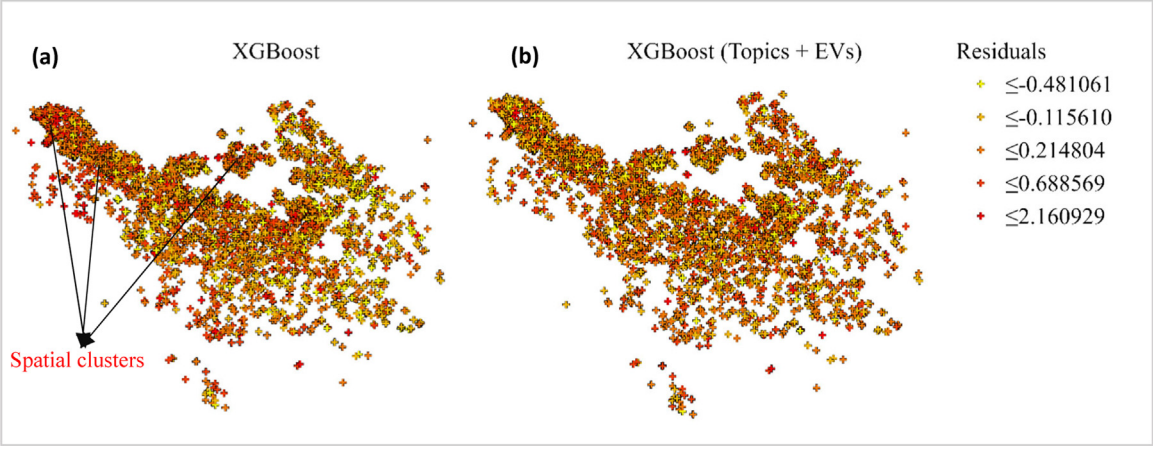

Fig. 6. Residuals distribution of (a) XGBoost with property variables and (b) XGBoost with topics and EVs.

the final output. It can be expressed as follows:

$$\hat{y}_i = \sum_{k=1}^{N} f_k(\boldsymbol{x}_i), f_k \in \mathcal{F} \tag{3}$$

Where N is the number of trees and $\mathcal{F}$ is the space of regression trees, and f is a function in the functional space $\mathcal{F}$.

3.5. Experimental LDA and MESF-based XGBoost composite model

This study applied LDA to extract the probability distribution of words for two distinct topics from the neighborhood description of each listed property and used them as explanatory variables in the model specification. Similarly, the MESF method was applied to extract approximated eigenvectors from spatial coordinates in order to use them as a set of proxy variables of spatial information in the model. Finally, the synthetic variables from LDA and MESF, along with selected covariates, were trained by the XGBoost algorithm for building the Airbnb listing price model. The LDA method was implemented by the 'topicmodels' package, the EVs were extracted by using the 'spmoran' package, and the XGBoost model was implemented by the 'XGBoost' package of R programming language. We applied the 'Caret'' package to implement K-fold cross-validation to validate the model's output.

3.6. Evaluation metrics

In this study, we used traditional regression assessment metrics for model validation. These metrics include R-square, RMSE, and MAE.

R-square measures the goodness of fit between the predicted and actual values. It can be expressed as follows:

$$R^2 = 1 - \frac{RSS}{TSS} \tag{4}$$

Here, RSS = sum of squares of residuals and TSS = total sum of squares.

Root Mean Square Error (RMSE) measures the standard deviation of the residuals. The residuals are the difference between observed values $\boldsymbol{y}_i$ and predicted values $\hat{\boldsymbol{y}}_i$.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^2} \tag{5}$$

Mean Absolute error (MAE) measures the average of absolute values of the residuals without considering their direction.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{y}_i - \boldsymbol{x}_i| \tag{6}$$

4. Analysis and results

The initial regression analysis shows that the selected variables can explain 54% variation of the response variable log_price (Table 2). All variables were found significant at a 95% confidence interval. The number of bedrooms and accommodates influences the listing price positively. If the property type is an apartment, the price will go up. However, if the super host badge is absent in a listing, it would negatively influence the price. The cancellation policy also negatively influences the price. That indicates if the listing offers flexible cancellation policy, those listing has a lower rental price.

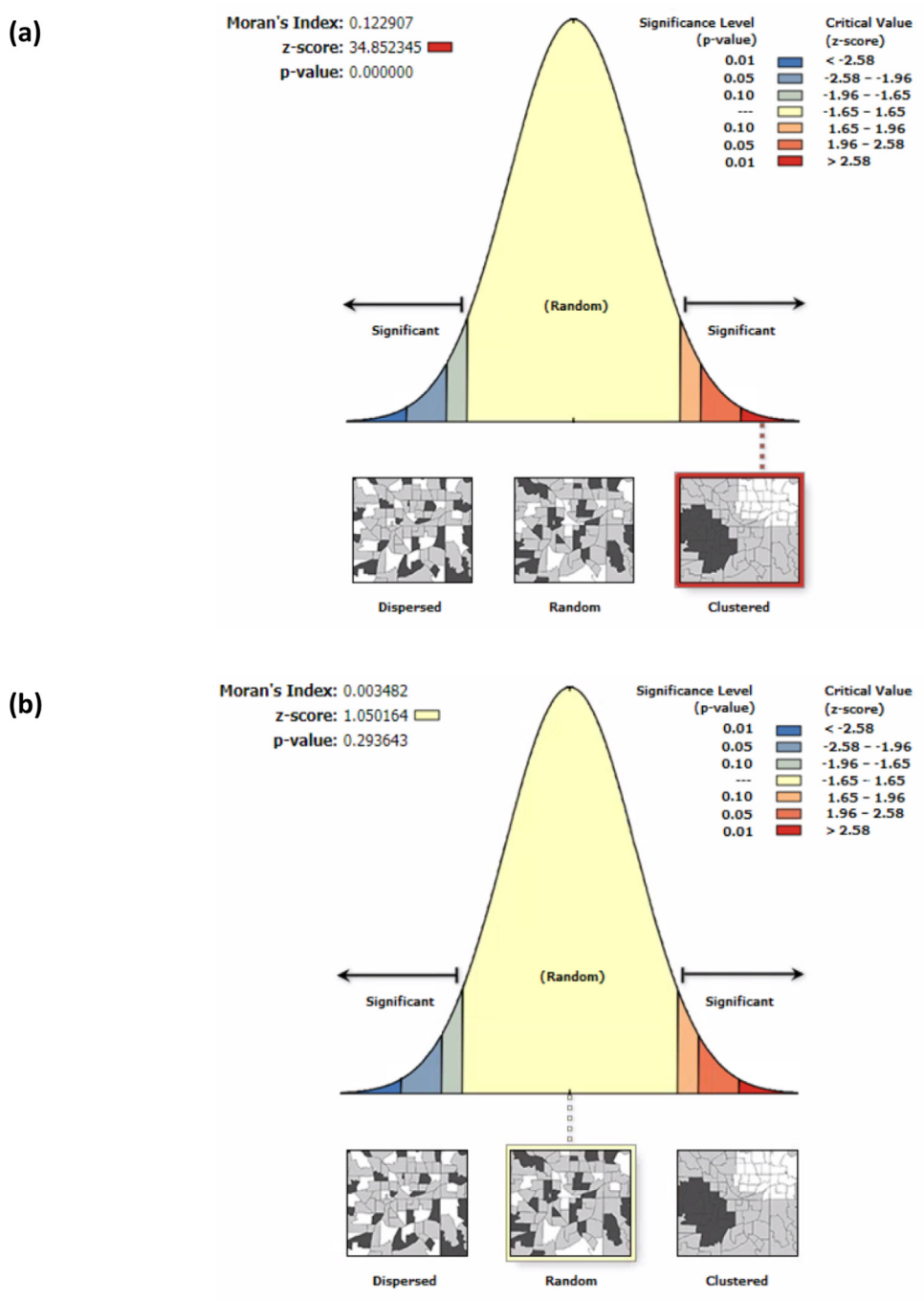

Fig. 7. Global Moran's I test for model (a) XGBoost with property variables and (b) XGBoost with topics and EVs.

The selected property variables were set as default for every model to understand how the synthetic variables from topic modeling and spatial eigenvectors improve the model's prediction accuracy. Then, the probability score of each topic and spatial eigenvectors are added in the process. Finally, a 10-fold cross-validation method was applied to obtain prediction diagnostics in order to compare prediction accuracy among the models (Table 3).

The cross-validation table shows that the initial regression model can achieve 54% of R-squared and 0.47 RMSE. With the same number of variables, the RF can achieve 64% of R-squared and 0.44 RMSE whereas the XGBoost model can achieve 0.67% of R-squared and 0.410 RSME. After adding two synthetic variables (topics) from the topic model, the RMSE decreased from 0.41 to 0.40, and R-squared increased from 65.7% to 67.4% in XGBoost. Finally, by adding synthetic variables (EVs) from MESF with topics variables and the selected property variables, the XGBoost model can achieve 70% of R-squared, which is 5% higher than the initial XGBoost model; the RMSE and MAE score reduced further. The diagnostic scores indicate a significant improvement in the prediction accuracy of the XGBoost model over the initial XGBoost model after adding synthetic variables.

The spatial distribution of residuals is visualized to understand how the addition of synthetic variables enables the XGBoost model to capture spatial dependence effects. Additionally, the global Moran's I test has been performed to detect any spatial autocorrelation in the residuals. The global Moran's I test measures spatial autocorrelation in a variable where +1 indicates a perfect spatial clustering of similar values, 0 indicates no autocorrelation, and −1 indicates perfect

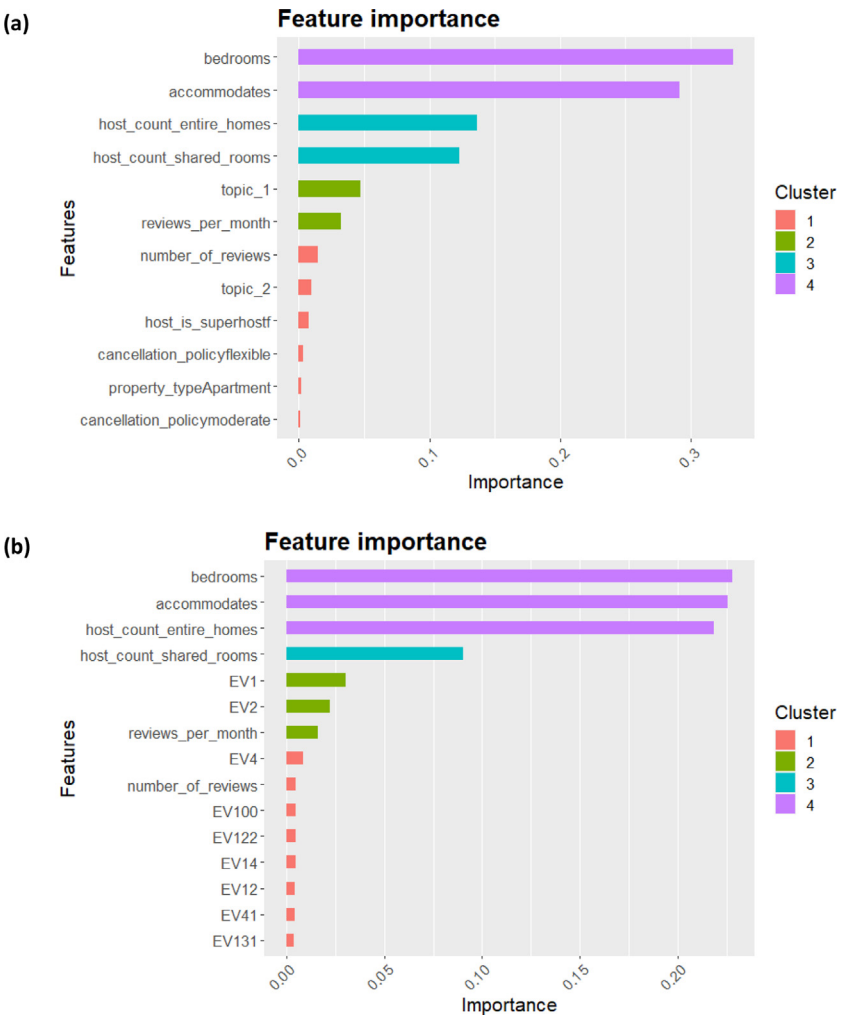

Fig. 8. Variable importance plot (a) XGBoost with topics (b) XGBoost with topics and EVs (top 15 variables).

clustering of dissimilar values or perfect dispersion. A spatial cluster is visible in the residuals of the XGBoost model with only property variables (Fig. 6a). The Moran's Index is 0.1229 in the residuals of the initial XGBoost, which indicates there is a less than 1% likelihood that the spatially clustered pattern of residuals could be the result of random chance (Fig. 7a). Therefore, the XGBoost model with only property variables is not strong enough to address spatial dependence and minimize spatial autocorrelation.

After incorporating, topics variables and spatial eigenvectors (EVs) into the model, the spatial clusters in the residuals almost disappeared (Fig. 6b). The Moran's Index of residuals decreased to 0.00348, which indicates the spatial distribution of residuals is randomly distributed (Fig. 7b). The global Moran's I test shows that incorporating synthetic variables from LDA and MESF into the XGBoost model enables it to address spatial dependence; consequently, its prediction accuracy improves significantly.

The variable importance plot (Fig. 8a) shows that XGBoost identifies bedrooms as the most important variable for model building. It can be seen that topic 1 and topic 2 also contribute to the model building that means including the words from those topics in the textual descriptions of the property in Airbnb would make a positive impression on the renters about the listing property and might increase the demand and profit. Fig. 8b shows that spatial eigenvectors (EVs) also dominate the model structure building, indicating that spatial information is also a crucial determinant and improves the model's accuracy.

The textual description of data carries valuable information which can be used to improve classification and prediction accuracy in machine learning models. Similarly, spatial data is influenced by surrounding geographical settings, which plays a key role in identifying and predicting a spatial phenomenon. Therefore, incorporating the spatial information in the form of spatial eigenvectors can address the spatial dependence and minimize spatial autocorrelation issues; consequently, it improves the prediction accuracy and helps to understand spatial phenomena better.

5. Conclusion

Spatial data tends to be spatially dependent where nearby locations influence a spatial phenomenon. In the pure data-driven machine learning domain, the importance of spatial dependence is largely neglected while modeling spatial data. However, addressing spatial dependence in machine learning models is crucial as it can improve the model's performance in real-world scenarios. This study shows that spatially filtered eigenvectors, which represent spatial information and address spatial dependence by interacting with other covariates, can improve the prediction performance of the machine learning modes. It confirms that geography plays a crucial role in determining Airbnb rental prices. The study also reveals that the synthetic topic variables by LDA also play an important role in further improving accuracy. By applying LDA, useful information can be extracted from the textual description in the data. In the analysis, we have seen that synthetic topic variables

Table 2
Linear regression results with only property variables.

log_price *Predictors*	*Estimates*	*CI*	*p*
(Intercept)	4.0987	4.0563–4.1411	<0.001
Accommodates	0.1304	0.1193–0.1415	<0.001
Bedrooms	0.1525	0.1261–0.1789	<0.001
Number_of_reviews	0.0010	0.0006–0.0014	<0.001
Calculated_host_listings_count_entire_homes	0.0097	0.0082–0.0111	<0.001
Calculated_host_listings_count_shared_rooms	−0.0297	−0.0326–−0.0269	<0.001
Reviews_per_month	−0.0492	−0.0632–−0.0352	<0.001
Property_typeApartment	0.1434	0.0999–0.1870	<0.001
Host_is_superhostf	−0.1035	−0.1345–−0.0724	<0.001
Cancellation_policyflexible	−0.1017	−0.1408–−0.0626	<0.001
Cancellation_policymoderate	−0.0503	−0.0855–−0.0152	0.005
Observations	3940		
R^2/R^2 adjusted	0.542/0.541		

Table 3
10-fold cross-validation result.

Model	RMSE	MAE	R-squared
LR	0.469	0.360	0.542
RF	0.448	0.316	0.641
XGBoost	0.410	0.305	0.657
XGBoost (topics)	0.400	0.297	0.674
XGBoost (topics+EVs)	0.381	0.281	0.704

from neighborhood overview data property improve the prediction accuracy that supports the assumption: a better description of the Airbnb rental unit usually make it a more attractive listing and subsequently, generates higher revenue. Finally, we can conclude that the synthetic variables from LDA and MESF can strengthen the machine learning models in detecting spatial patterns and improve predicting accuracy and performance.

This study would guide the homeowners in describing their property while listing it on the Airbnb platform. The findings can assist the homeowners in determining the accurate listing prices considering the location and surrounding neighborhoods. They could include details regarding the accommodations (how many people could live at once) along with the no. of bedrooms, whether or not they have a superhost badge, and the significant POIs around the listing, which we found will positively influence the price. On top of that, the homeowners could describe their cancellation policies explicitly that is also correlated with the listing price as well as the attraction to the renter. It can also encourage researchers to incorporate spatial information in machine learning algorithms to improve accuracy while modeling spatial data. The limitation of the study includes property description variables such as transit, summary, space, etc. were excluded while addressing missing values and keeping the data consistent. The study can be extended further by using spatial dictionaries in the LDA model to sense the location proximity, such as distance from nearby bus transit or parks, and generate useful spatial information from textual description.

CRediT authorship contribution statement

Md Didarul Islam: Conception and design of study, Writing – original draft, Literature Review (Revised Version). **Bin Li:** Conception and design of study. **Kazi Saiful Islam:** Conception and design of study. **Rakibul Ahasan:** Conception and design of study, Writing – original draft. **Md. Rimu Mia:** Literature Reviewe (Revised Version). **Md Emdadul Haque:** Conception and design of study, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

All authors approved the version of the manuscript to be published.

References

Ahasan, R. (2019a). *Graduate theses and dissertations*, *Transit proximity and affordable housing investments: Application of hedonic model in Des Moines, Iowa*. Article 17385, Available at: https://lib.dr.iastate.edu/etd/17385.

Ahasan, R. (2019b). Transportation accessibility, housing investments, and housing prices: Application of hedonic price model in Des Moines, Iowa. In *ACSP 2019 annual conference, Toronto, Canada*. http://dx.doi.org/10.31235/osf.io/ghdx4.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Brueckner, J. K., Thisse, J. F., & Zenou, Y. (1999). Why is central Paris rich and downtown detroit poor?: An amenity-based theory. *European Economic Review*, *43*(1), 91–107.

Buenano-Fernandez, D., Gonzalez, M., Gil, D., & Lujan-Mora, S. (2020). Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach. *IEEE Access*, *8*, 35318–35330. http://dx.doi.org/10.1109/access.2020.2974983.

Cai, Y., Zhou, Y., (Jenny) MA, J., & Scott, N. (2019). Price determinants of airbnb listings: Evidence from Hong Kong. *Tourism Analysis*, *24*(2), 227–242. http://dx.doi.org/10.3727/108354219x15525055915554.

Chen, T., & Guestrin, C. (2016). Xgboost. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. http://dx.doi.org/10.1145/2939672.2939785.

Chen, Y., & Xie, K. (2017). Consumer valuation of airbnb listings: A hedonic pricing approach. *International Journal of Contemporary Hospitality Management*, *29*(9), 2405–2424. http://dx.doi.org/10.1108/ijchm-10-2016-0606.

Chica-Olmo, J., González-Morales, J., & Zafra-Gómez, J. (2020). Effects of location on airbnb apartment pricing in Málaga. *Tourism Management*, *77*, Article 103981. http://dx.doi.org/10.1016/j.tourman.2019.103981.

Dudás, G., Boros, L., Kovalcsik, T., & Kovalcsik, B. (2017). The visualization of the spatiality of airbnb in budapest using 3-band raster representation. *Geographia Technica*, *12*(1), 23–30. http://dx.doi.org/10.21163/gt_2017.121.03.

Edelman, B., & Geradin, D. (2015). Efficiencies and regulatory shortcuts: How should we regulate companies like airbnb and uber? *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.2658603.

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in airbnb. *Tourism Management*, *55*, 62–73. http://dx.doi.org/10.1016/j.tourman.2016.01.013.

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J., & Goodwill, A. (2017). Pricing in the sharing economy: A hedonic pricing model applied to airbnb listings. *Journal of Travel & Tourism Marketing*, *35*(1), 46–56. http://dx.doi.org/10.1080/10548408.2017.1308292.

Griffith, D. A. (1987). *Spatial autocorrelation. A primer*. Washington, DC: Association of American Geographers.

Griffith, D. (2000). Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses. *Linear Algebra and its Applications*, *321*(1–3), 95–112. http://dx.doi.org/10.1016/s0024-3795(00)00031-8.

Griffith, D. A. (2013). *Spatial autocorrelation and spatial filtering: Gaining understanding through theory and scientific visualization*. Germany, Heidelberg: Springer Berlin.

Griffith, D., & Chun, Y. (2013). Spatial autocorrelation and spatial filtering. *Handbook of Regional Science*, *147*, 7–1507. http://dx.doi.org/10.1007/978-3-642-23430-9_72.

Griffith, D., Chun, Y., & Li, B. (2019). *Spatial regression analysis using eigenvector spatial filtering* (1st ed.). London: Academic Press.

Gutt, D., & Herrmann, P. (2015). Sharing means caring? Hosts' price reaction to rating visibility. In *ECIS 2015 research-in-progress papers, paper 54*. Retrieved from http://aisel.aisnet.org/ecis2015_rip/54/.

Guttentag, D. (2013). Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism*, *18*(12), 1192–1217. http://dx.doi.org/10.1080/13683500.2013.827159.

Gyódi, K., & Nawaro, Ł. (2021). Determinants of airbnb prices in European cities: A spatial econometrics approach. *Tourism Management*, *86*, Article 104319. http://dx.doi.org/10.1016/j.tourman.2021.104319.

Hartmans, A. (2017). Airbnb now has more listings worldwide than the top five hotel brands combined. *Business Insider*, Retrieved from: https://www.businessinsider.com/airbnb-total-worldwide-listings-2017-8.

Jun, J., & Arendt, S. (2016). Understanding healthy eating behaviors at casual dining restaurants using the extended theory of planned behavior. *International Journal of Hospitality Management*, *53*, 106–115. http://dx.doi.org/10.1016/j.ijhm.2015.12.002.

Kakar, V., Voelz, J., Wu, J., & Franco, J. (2016). The visible host: Does race guide airbnb rental rates in San Francisco? *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.2967902.

Lampinen, A., & Cheshire, C. (2016). Hosting via airbnb. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. http://dx.doi.org/10.1145/2858036.2858092.

Latinopoulos, D. (2018). Using a spatial hedonic analysis to evaluate the effect of sea view on hotel prices. *Tourism Management*, *65*, 87–99. http://dx.doi.org/10.1016/j.tourman.2017.09.019.

Lee, D. (2016). How airbnb short-term rentals exacerbate los angeles's affordable housing crisis: Analysis and policy recommendations. *Harvard Law & Policy Review*, *10*(229).

LeSage, J. (2008). An introduction to spatial econometrics. *Revue D'économie Industrielle*, *123*, 19–44. http://dx.doi.org/10.4000/rei.3887.

Meleo, L., Romolini, A., & De Marco, M. (2016). *Lecture notes in business information processing, The sharing economy revolution and peer-to-peer online platforms. The case of airbnb* (pp. 561–570). http://dx.doi.org/10.1007/978-3-319-32689-4_43.

Murakami, D., & Griffith, D. A. (2015). Random effects specifications in eigenvector spatial filtering: A simulation study. *Journal of Geographical Systems*, *17*(4), 311–331. http://dx.doi.org/10.1007/s10109-015-0213-7.

Priambodo, F., & Sihabuddin, A. (2020). An extreme learning machine model approach on airbnb base price prediction. *International Journal of Advanced Computer Science and Applications*, *11*(11), http://dx.doi.org/10.14569/ijacsa.2020.0111123.

Sainaghi, R. (2020). Determinants of price and revenue for peer-to-peer hosts. The state of the art. *International Journal of Contemporary Hospitality Management*, *33*(2), 557–586. http://dx.doi.org/10.1108/ijchm-08-2020-0884.

Teubner, T., Hawlitschek, F., & Dann, D. (2017). Price determinants on airbnb: How reputation pays off in the sharing economy. *Journal of Self-Governance and Management Economics*, *5*(4), 53–80.

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, *46*(234), http://dx.doi.org/10.2307/143141.

Wang, Y., Kockelman, K., & Wang, X. (2013). Understanding spatial filtering for analysis of land use-transport data. *Journal of Transport Geography*, *31*, 123–131. http://dx.doi.org/10.1016/j.jtrangeo.2013.06.001.

Wang, D., & Nicolau, J. (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. http://dx.doi.org/10.1016/j.ijhm.2016.12.007.

Yang, S. (2021). Learning-based airbnb price prediction model. In *2021 2nd International conference on E-commerce and internet technology*. ECIT, http://dx.doi.org/10.1109/ecit52743.2021.00068.

Zervas, G., Proserpio, D., & Byers, J. W. (2021). A first look at online reputation on airbnb, where every stay is above average. *Marketing Letters*, *32*, 1–16. http://dx.doi.org/10.1007/s11002-020-09546-4.

Zhang, Z., Chen, R., Han, L., & Yang, L. (2017). Key factors affecting the price of airbnb listings: A geographically weighted approach. *Sustainability*, *9*(9), 1635. http://dx.doi.org/10.3390/su9091635.

Zhu, A., Li, R., & Xie, Z. (2020). Machine learning prediction of new york airbnb prices. In *2020 Third International Conference on Artificial Intelligence for Industries*. AI4I, http://dx.doi.org/10.1109/ai4i49448.2020.00007.