Journal of Intelligent & Fuzzy Systems 40 (2021) 1879–1891
DOI:10.3233/JIFS-189193
IOS Press

Fuzzy metatopics predicting prices of Airbnb accommodations

Manuel J. Sánchez-Franco[a,*], José A. Troyano[b] and Manuel Alonso-Dos-Santos[c,d]
[a]*Business Administration and Marketing, University of Sevilla, Spain*
[b]*Computer Languages and Systems, University of Sevilla, Spain*
[c]*Departamento de Administración, Universidad Católica de la Santísima Concepción, Concepción, Chile*
[d]*Department of Marketing and Market Research, University of Granada, Spain*

Abstract. The purpose of this study is to guide pricing policies of Airbnb accommodation rentals to reduce inefficient pricing strategies through a novel application of topic modelling and a fuzzy clustering. In particular, the method proposes the application of Structural Topic Modelling, which explains a set of observations from latent topics. The associations between topics by Fuzzy C-Means Clustering are analysed to obtain new, more compact representations of topics (i.e., metatopics). This research identifies 15-metatopics related to Airbnb accommodations based on location and connectivity, enjoyment of domestic and everyday services, and the possibility of more authentic local experiences, among others. The influence of key metatopics on the price of Airbnb accommodations is determined by applying Extreme Gradient Boosting (an efficient and scalable implementation of gradient boosting framework) and Shapley Additive Explanations values. To sum up, our research provides an explicit contribution of user-generated content to promote the development of mutually beneficial relationships between guests and hosts, and detects future lines of research and practical and conceptual implications of the findings.

Keywords: Sharing economy, Airbnb, price, fuzzy c-means clustering, structural topic model

1. Introduction

Individuals employ recommendation systems to reduce the risk of expectation failure by consulting freely-sharing accommodation scores and travellers' comments, and also to generate by themselves a *new content* on their subjectively experienced encounters (User-Generated Content, hereinafter, UGC; *e.g.* [1–4]). The research specifically focuses on Airbnb (classified under the sharing economy) that differs from traditional hotels in terms of booking systems, facilities, software platforms and design, and recommendations for guests. Since its launch in 2008, Airbnb has become one of the largest single tourism distribution platforms for short-term accommodation rentals [5]; *i.e.*, individuals grant each other temporary access to underutilised physical assets, usually for money (*e.g.*, [6]).

In this regard, previous research has detected inefficient pricing by Airbnb hosts due to the uniqueness of the rental services offered on Airbnb [5] and emotional drivers applied by non-professional hosts (*cf.* [7]). Furthermore, scant research focuses on guests' expectations, predictions, goals, and desires from the linguistic attributes of the online textual reviews generated by customers to holistically determine the sharing economy price [8]. To solve this gap, a product, feature-oriented approach is applied to identify relevant topics (a distribution of terms through a fixed vocabulary, [9, 10]) concerning the core and basic sharing services, as well as surrounding features (*e.g.,* neighbourhood amenities or tourist hotspots), extracted from Airbnb lodgings, that significantly affect the prices of accommodation listings. Although UGC is poorly structured and is overall focused on

*Corresponding author. Manuel J. Sánchez-Franco, Business Administration and Marketing, University of Sevilla, Spain. E-mail: majesus@us.es.

a single entity or aspect of hospitality services, or is multi-lingual, it is mainly based on authenticity, and tends to be even more empathetic and trustworthy than other social communications [11, 12].

The method, therefore, lies in improving the performance of guests' reviews on prediction tasks by identifying users' experience-related frequent terms and relevant topics, examining the underlying semantic structure and reducing the number of topics into meaningful fuzzy clusters (or metatopics) that make them easier to interpret, determining the significant clusters among a large quantity of text data that influence the price of Airbnb accommodations, and consequently providing an explicit contribution of UGC to promote the development of mutually beneficial relationships.

The paper is structured as follows. The second section presents the theoretical background derived from the prior literature. The third section elaborates the method and the results of the study that provide a new framework to comprehensively understand the influential drivers of Airbnb pricing. The last section concludes with implications, limitations and future research.

2. Theoretical framework

Airbnb is primarily a low-cost option where travellers (or here, guests) find entire apartments or (shared) rooms at a more competitive price than hotels coordinated through community-based services [5]. Likewise, the demand for Airbnb rentals is significantly elastic because of the hosts' fixed costs of rent and utilities, together with minimal labour costs and probably untaxed extra income (*cf.* [13]). In this regard, Airbnb accommodation prices become one of the key determinants that affects guests' selection of them and appropriately warrants their revenue [14–16].

Assuming that "an Airbnb accommodation listing is a bundle of elements that influence the quality of the overall product and provide consumers with value and satisfaction" [5], previous research provides valuable discussions about the presence or absence of key subjective dimensions or features and their contributions on model output. In particular:

- Site-specific features measured as the distance from the touristy hotspots (or also a major transportation hub), or hospitality features based on home-like lodging conditions (*e.g.,* household amenities and basic functionalities such as a homely feel, real beds, wireless Internet, a large space, and free parking, among others) are traditional factors that determine a customer's final choice (*cf.* [17–19]).
- Airbnb guests also focus on the emergence of a society that desires value for money (based on more transparent pricing; *cf.* [18,20–22]) or experiences described as authentic staying at an Airbnb lodging (*cf.* [23–26]), or novelty [3, 18, 19].
- Airbnb is precisely focused on postmodern tourists, in contrast to modern tourists, i.e. travellers who enjoy multiple experiences embracing different, sometimes contrasting life values, or daily interaction with the host and local people [18, 19, 21, 25, 26].

Although guests' motivations for selecting Airbnb accommodations have been researched by a handful of scholars, "the research to date related to 'pricing and Airbnb' does little to explain the variables that make up the price of a listing" [5]. Previous research mainly focuses on examining a few influential drivers in an isolated manner, without providing a broad perspective on the issue [27]. Moreover, "this body of research also suffers from numerous limitations, (. . .) and the studies reach somewhat incongruent conclusions" [23]. Accordingly, to partially solve this research gap, rational attributes and affective or social topics are extracted from customer reviews (guests' perspectives) and the full dynamics that most often influence prices of Airbnb accommodations are analysed in a comprehensive pricing analysis. To ignore topics from guests' narratives can yield inaccurate estimates of the prices, and the size of the pricing errors can affect the true conclusions about the research's implication for guests' welfare.

3. Materials and methods

Airbnb accommodation rental offers are here analysed using a Hedonic Pricing Analysis-based approach in the five most populated New York urban neighbourhoods (Bronx, Brooklyn, Manhattan, Queens and Staten Island). New York is indeed selected as a case study because it generates rich feelings that can affect guests' perceptions. New York is the most visited destination in the United States. According to NYC & Company, New York attracted

around 65.2 million tourists in 2018, and in particular, 13.5 million foreign tourists.

The dataset is obtained from the Inside Airbnb website (http://insideairbnb.com/ -a non-commercial, open source data tool on Airbnb).

3.1. Data collection

Initially the dataset contains:

- 1,106,639 reviews located exclusively in New York to avoid any risk of heterogeneity induced by a 'regional' effect that prevents relevant possible comparisons.
- 49,748 listings distributed in the following neighbourhoods: Bronx (1,001), Brooklyn (20,312), Manhattan (22,559), Queens (5,525), and Staten Island (351).
- Airbnb listings managed by 37,689 hosts.

In order to standardise the comparisons and to increase the comparability among listings, accommodations are included for only less than six guests and entire apartments during three consecutive years, from 2016 to 2018. Furthermore, a single language is analysed, English, to keep the language variable consistent across texts.

The dataset finally contains 40,572 reviews (and 9,710 listings). The numerical ratings average 94.47 (with $sd = 4.51$, and a minimum, median, and maximum of 20, 95, and 100, respectively). The price is taken at the accommodation level and does not include cleaning fees or additional charges for guests that are not included in the overall price. Removing outliers, the average price is \$157.59 (*sd*=\$58.34 with a minimum, median, and maximum of \$10, \$150, and \$320, respectively).

3.2. Data cleansing process

A cautious data cleansing process is carried out (to increase the quality of the metatopics), based on transforming free-form text into a structured form. It applies the following stages: it discards punctuation, capitalisation, digits, and extra whitespace, it recognises common abbreviations and acronyms, it removes a list of stopwords to filter out overly common terms without specific relevance for the research problem, and it tokenises and lemmatises the terms. To avoid tallying one term in various grammar contexts, only the stem of a term is retained. Terms shorter than a minimum of three characters are also omitted.

3.3. Extracting terms

Applying a topic modelling on all terms in a corpus is both computationally expensive and not very useful. The inclusion of redundant, irrelevant and noisy terms in the topic building process could also cause a poor predictive performance. A subset of terms is thus selected that minimises their redundancy and maximises their relevance, and a Bi-Normal Separation metric is applied (hereinafter, BNS; see [28]). Online reviews overall tend to be brief, with only a relatively small number of major topics standing out, and usually contain no terms that occur more than once per document.

3.4. Data mining

The relationships between extracted terms and documents are estimated by machine-learning algorithms based on text summarisation and the application of structural topic modelling (hereinafter, STM; *cf.* [29]).

3.4.1. Structural topic modelling: Model specification and selection

Topic modelling is nowadays a computer-assisted technique to address the costs and time associated with the growing amount of data and uncovers patterns of term co-occurrence across the corpus (defined as a mixture over terms where each term has a probability of belonging to a topic k; *cf.* [9] or [10], among others).

The approach focuses on STM -a generative model of term counts- and its implementation in the STM 1.3.3 R package [30]. STM, as an unsupervised method, allows the researchers to discover topics (inferred here from the guests' narratives) that can be correlated, and estimates their relationships to document metadata (*e.g.*, price) to take the context into account for a better understanding of the 'semantically interpretable themes' without forcing the metadata to be influential on the topics.

Managerial implications are prioritised, and extracting a small number of topics is proposed. Hosts prefer prediction models that not only provide technical insights but are also accessible. In this vein, different STM models between 30 and 50 topics are estimated. An initialisation is proposed based on the method of moments, "which is deterministic and globally consistent under reasonable conditions" ([30], p.11; *cf.* also [31]). Also, the models that have the lowest value for the bound are discarded [32]

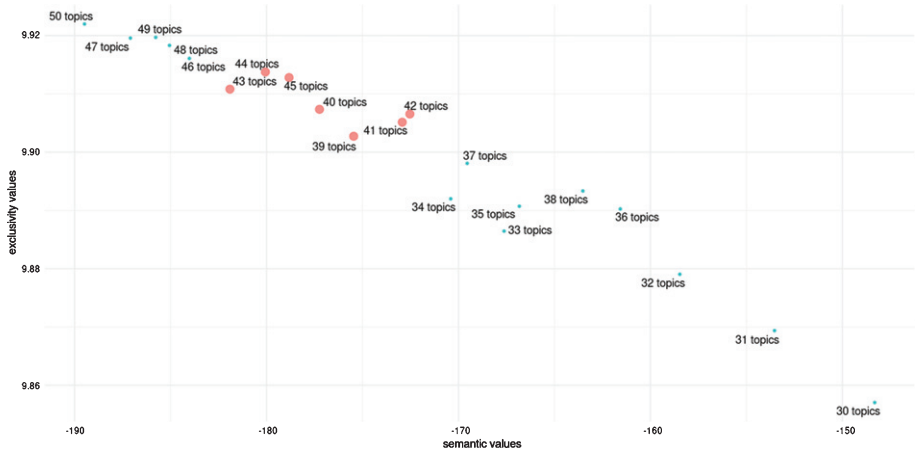

Fig. 1. Semantic coherence (x-axis) and exclusivity values (y-axis).

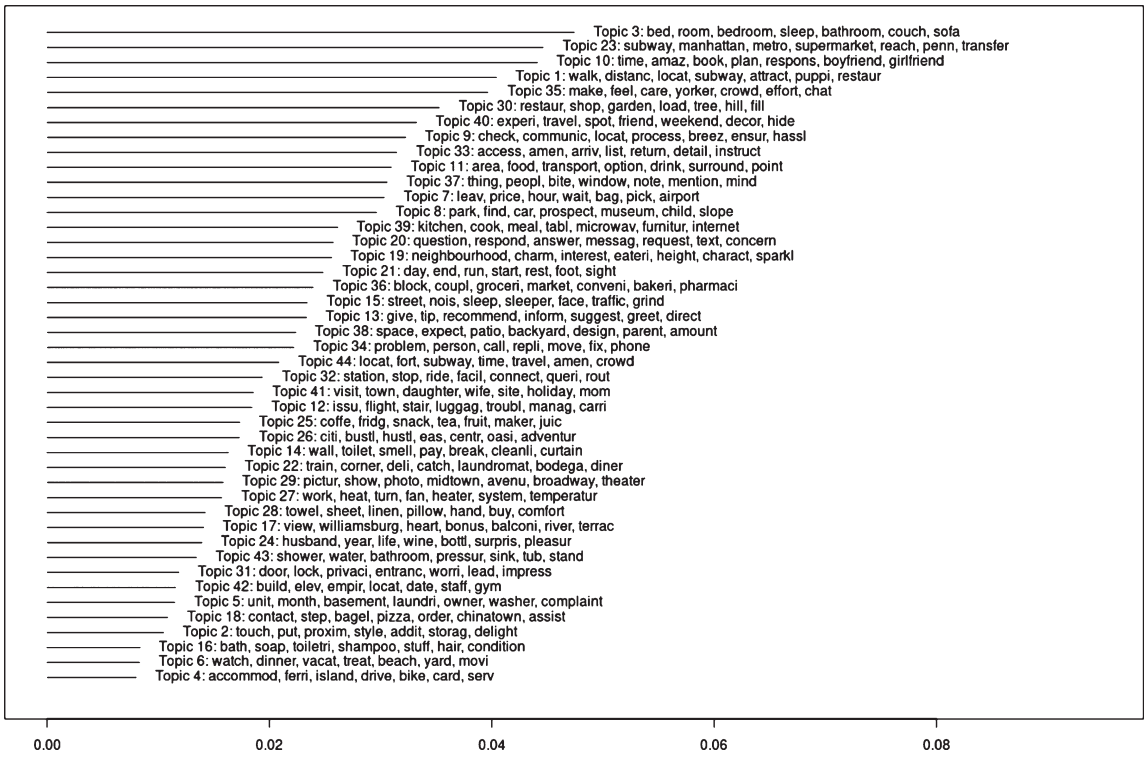

Fig. 2. Example of graphical display of estimated topic proportions ($\omega = 0.5$).

and, finally, the trade-off between semantic coherence and exclusivity is assessed (*i.e.,* internal consistency -cohesiveness- and differentiation, see [10, 32, 33]) (see Fig. 1). Accordingly, this process selects a subset of models prioritising diverse criteria (here, 39-45 topics). In particular, the research prioritises that extracted topics are capturing different conceptual aspects of Airbnb experiences (*cf.* discriminant validity). It fits into "recent research which has begun to show diversity in motivations for participating in the sharing economy" (*cf.* [34], p.188). In this regard, to identify intuitive meanings of topics, they are conceptualised with a list of the most representative customer terms - strongly connected to each topic. [35] propose using the Frequency-Exclusivity (FREX) statistic that combines term frequency and exclusivity to topics. It is here set to 0.5. Fig. 2 provides a summary of 44-topics as an illustrative

proposal, their mean prevalence and the most FREX terms. Topics 3 and 23 related to 'home benefits' (around 4.5% of the documents) and 'connectivity' (4.3%) show the most estimated topic proportions.

Following Quinn et al. [36], the coherent meaning of the metatopics is further evidence of the semantic validity of the topic model. One way to define topics is to categorise them by clusters, revealing their organisation to examine the semantic relationships within and across clusters of topics (hereinafter, metatopics) by applying (here) a fuzzy clustering analysis and by using of the theta matrix based on the posterior probability of a topic given a document as clustering inputs.

Finally, in order to select a definitive model based on its predictive validity, the candidate models' outputs with the highest capacity to predict price accommodations are assessed by an r2 metric from applying an extreme gradient boosting (hereinafter, XGBoost) and the candidate models that are based on a manageable number of metatopics.

3.5. Topic groupings: Fuzzy clustering analysis

A feature extraction technique is applied based on Fuzzy C-Means Clustering (hereinafter, FCM as an extension of the hard K-means algorithm to the fuzzy framework). FCM obtains new, more compact representations of documents than those initially provided by extracted topics. In particular, FCM allows the researcher to deal with lexical ambiguity because a single term (or a topic) could belong to numerous semantic categories. Despite the past research regarding fuzzy clustering conducted in the, scarce attention has been paid to its applications in tourism [37].

FCM was initially studied by Dunn [38], and it was generalised by Bezdek in [39–41]. FCM was inspired by the hard-clustering algorithm called K-means and was based on the centroid concept. In FCM, centroids are calculated with a weighted average (using membership degrees as coefficients) of a transformation of instances. Transformations consist in raising each instance using the parameter m as a fuzziness exponent or fuzzification degree; i.e., higher values of m cause lower degrees of membership and subsequently, higher fuzzy partitions. m is a real number greater than 1. In comparison to hard clustering or crisp clustering, each review is a mixture of topics and each topic is thus a member of distinct clusters with varying degrees of membership between 0 and 1. The features (i.e., dataset columns) are grouped instead of instances (i.e., dataset rows). The interest lies in distributing the information given by each feature in different clusters and keeping the centroids as representatives of extracted clusters acting as new features.

3.6. Predictive analysis

Although generalised regression models such as OLS regression or quantile regression are the most commonly adopted methods to analyse features affecting accommodation prices, XGBoost is applied here. This was initially proposed by Friedman [42], and it started as a research project by Tianqi Chen [43].

On the one hand, boosting is an ensemble technique that relies on the idea that a series of weak estimators (classifiers or regressors) can behave like a robust estimator. On the other hand, XGBoost is an efficient and scalable implementation of gradient boosting framework, or GBM [43, 44]. GBM is the most popular classifier, and it has the particularity of interpreting the boosting process in terms of the optimisation of a cost function, which allows the use of an adaptation of the gradient descent algorithm for guiding the training. XGBoost also includes a series of optimisations that make the training much faster than other implementations, along with regularisation techniques that help reduce overfitting.

4. Experimental results

The main goal is to obtain a new dataset with fewer features (metatopics) that gathers most of the information from the original dataset. The quality of the reductions is based on the performance of an XGBoost regressor trained with the new features, where the log(price) is the target variable of the regressor. All models are fitted in R (R-3.6.1). For FCM ppclust package version 0.1.3 is applied, and for XGBoost the xgboost package version 0.90.0.2 [45] and the caret package version 6.0–84 are used [46].

In particular, the coefficients of root mean squared error (RMSE) and r-squared (r2) scores are used to validate the models and evaluate their qualities. To ensure the effectiveness of the training process and to find the best-fitted model based on the r2 metric, the dataset is split into two subsets. 75% of the observations (train dataset) allow us to train the XGBoost model in order to find the best combination of parameters. Cross-validation helps with finding

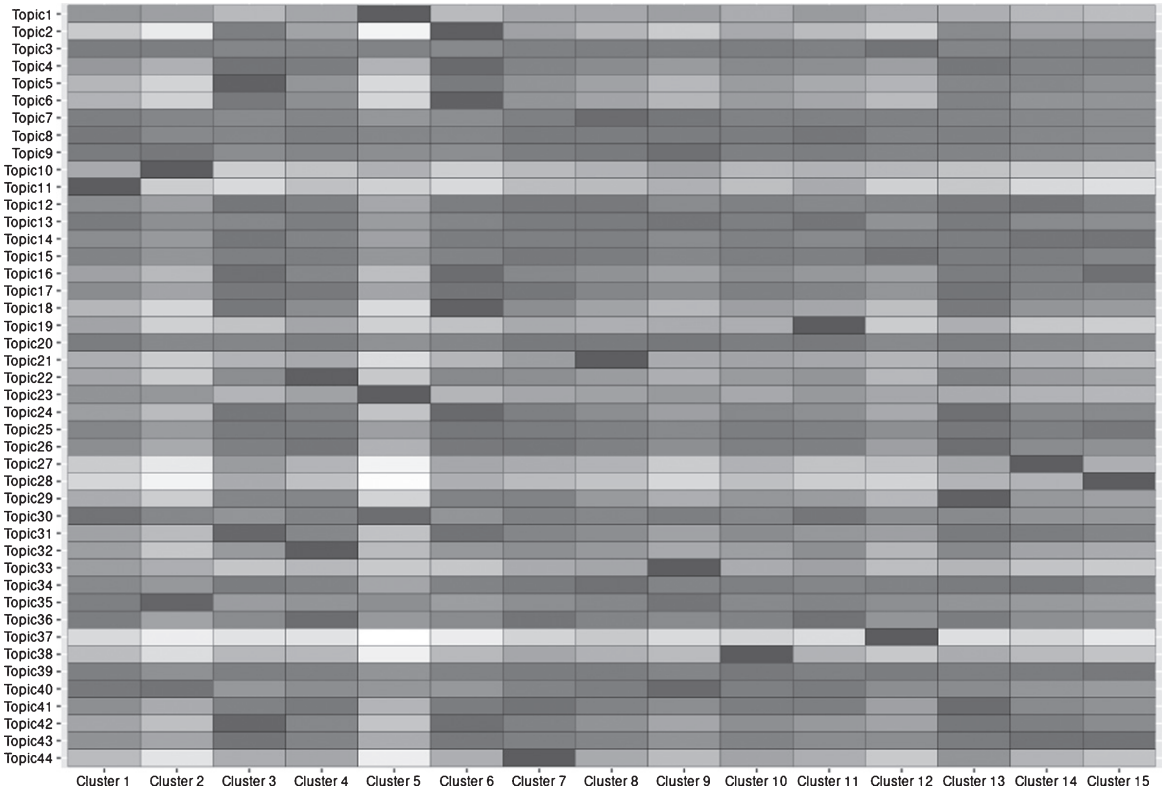

Fig. 3. Graphical display of fuzzy clustering analysis.

the optimal combination of parameters [44]. A 10-fold cross-validation is performed (50 rounds, and 100 tuning-iterations). The remaining 25% (validation dataset) enables us to validate the performance of the best-fitted model and to ensure that the model is generalisable [44].

Applying the tuning process of detecting the best parameter combination, the r2 value is 0.93 (and the RMSE is 0.1) with 44 topics, a learning rate (η) of 0.193, a subsample equal to 0.96 and a maximum depth of each tree of 10, i.e., a very high predictive capacity of the extracted topics. On the other hand, the best results based on the clustering of the topics are achieved for values of m close to 1.5, and employing 15 metatopics. This reaches an r2 of 0.79 (RMSE = 0.17) with a loss of only 15% with respect to the original dataset, and iterating 50 trees of 6 maximum depth, $\eta = 0.23$, and a subsample equal to 0.849, among other parameters. In the 15 metatopics proposed (as a manageable number of topics to facilitate the managerial implications of the research), some of them almost correspond to original topics (see Fig. 3), and other metatopics are algorithmically forced to be constructed by combinations of them.

Assuming that clustering is an unsupervised learning process, it is necessary to complement the external validation (see Section 3.1 above) with an internal validation of clustering with c number of clusters (5, 10, 15 and 20 features) and 9-values of m (from 1.1 to 1.9). Specifically, the partition coefficient pc is applied here [47]. pc measures the amount of overlapping between two fuzzy clusters (cf. [40]), and is usually used in the absence of supervised evaluations to determine optimal values of parameters. The pc metric here evolves for different values of parameter m (see Fig. 4), and r2 drops below 0.5 for values of m greater than 1.5. The pc-value evolution (<0.5) fits with the significant decrease in r2 for values of m greater than 1.5.

5. Interpretation of metatopics

The main goal when applying dimensionality reduction is to gain interpretability. To reduce from 44 features (topics) to 15 features (metatopics) allows more comprehensively handling (from the managers' perspectives) the recommendations, pre-

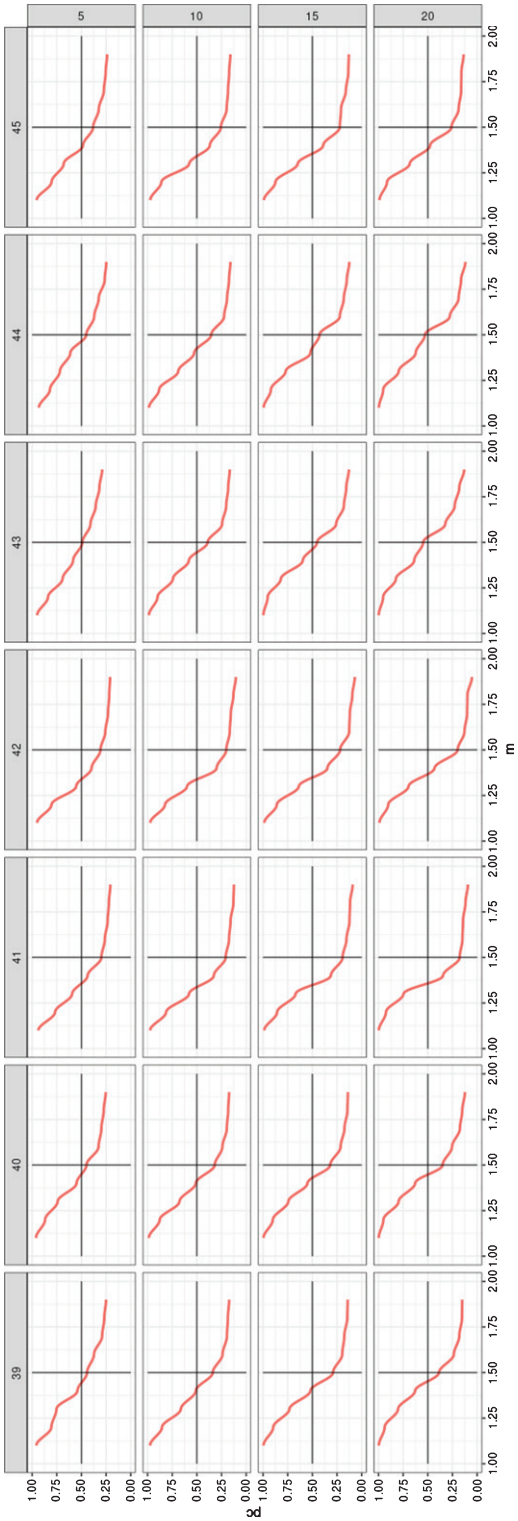

Fig. 4. Evolution of the pc metric for different values of parameter m, and number of features.

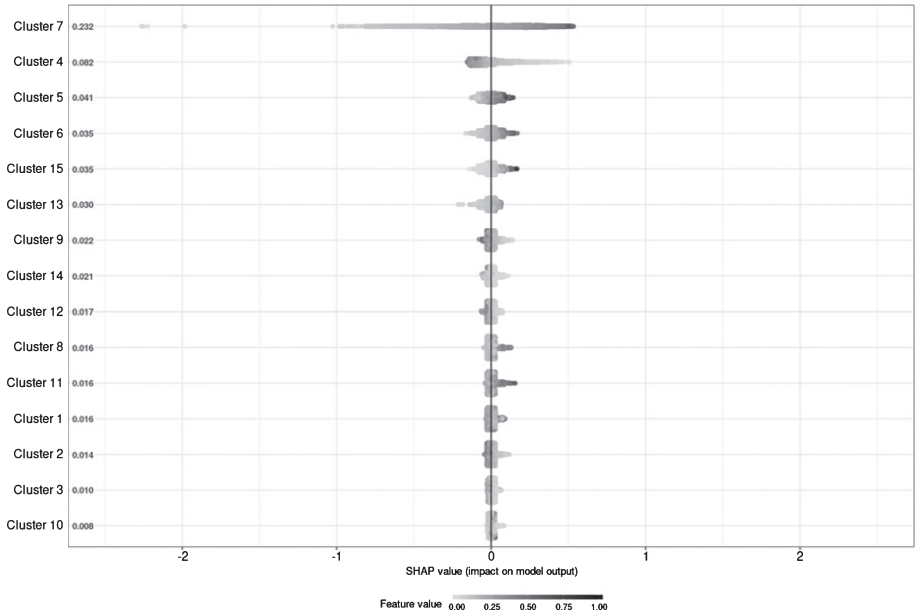

Fig. 5. Feature importance displayed by local SHAP values.

serving the broad range of motivations that draw guests to the Airbnb service. In particular, Fig. 5 shows the distribution of feature (metatopic) contributions (directionality and density) to the model output using Shapley Additive Explanations values (hereinafter, SHAP values) of each metatopic for every document. Figure 5 also represents the range and distribution of the impacts that each metatopic has on the model output applying one additive feature attribution method, SHAP. SHAP decomposes the prediction of an individual observation into components attributable to each feature [48, 49].

On the one hand, the highest influential metatopics in the model output are, in this order, the following: 7, 4, 5, 6 and 15. Metatopic 7 is related to location measured as distance from the touristy hotspots and transportation hub (cf. [5]). Guests preferentially (and unsurprisingly) pay more (less) for entire apartments in highly (lowly) rated locations. Guests highlight advantages related to closeness to tourist hotspots or transportation hubs, among others, and highly rated locations tend to be the most expensive ones. Metatopic 4 (based on the distance to neighbourhood amenities, e.g., short/long walking distances to local restaurants, supermarkets, groceries, or shops, among other amenities) shows the opposite effect — higher values (distances) lead to a lower prediction of pricing. In this regard, SHAP dependence plots help to identify features which negatively influence the model output (see Fig. 6). For instance, increasing x-values up to 0.02 results in a sharp decrease in predictions of Airbnb prices; the trend softens at middle values. A long (here, left) tail means that the accommodations' locations (metatopic 7) can be extremely influential for specific guests (Fig. 5). On the contrary, metatopic 5 (based on the transportation hub, or connectivity) affects many predictions by a small amount (high density). Although connectivity is less of a concern among all travellers, metatopic 5 could here be conceptualised as one of the growing trends influencing booking decisions. Growing values of metatopic 6 (e.g., authenticity) and metatopic 15 (e.g., home amenities) are associated with increasing values in the Airbnb price of accommodations. The density of metatopics 6 and 15 based on authentic character of staying related to 'touring like a local' or homely feel related to 'cooking and cleaning at home' (cf. [19, 23]) is high. Both metatopics therefore affect all the predictions by a small amount.

On the other hand, SHAP dependence plots reflect how single values of a metatopic can have different effects on the model output depending on the context of the other metatopics present in a document. For this purpose, the study also plots Fig. 7, and selects other metatopics for colouring to highlight interactions, to identify additional influential features on Airbnb prices, and to elaborate the results based on dispersion of the data points as a probable case of interaction with other main metatopics. For instance,

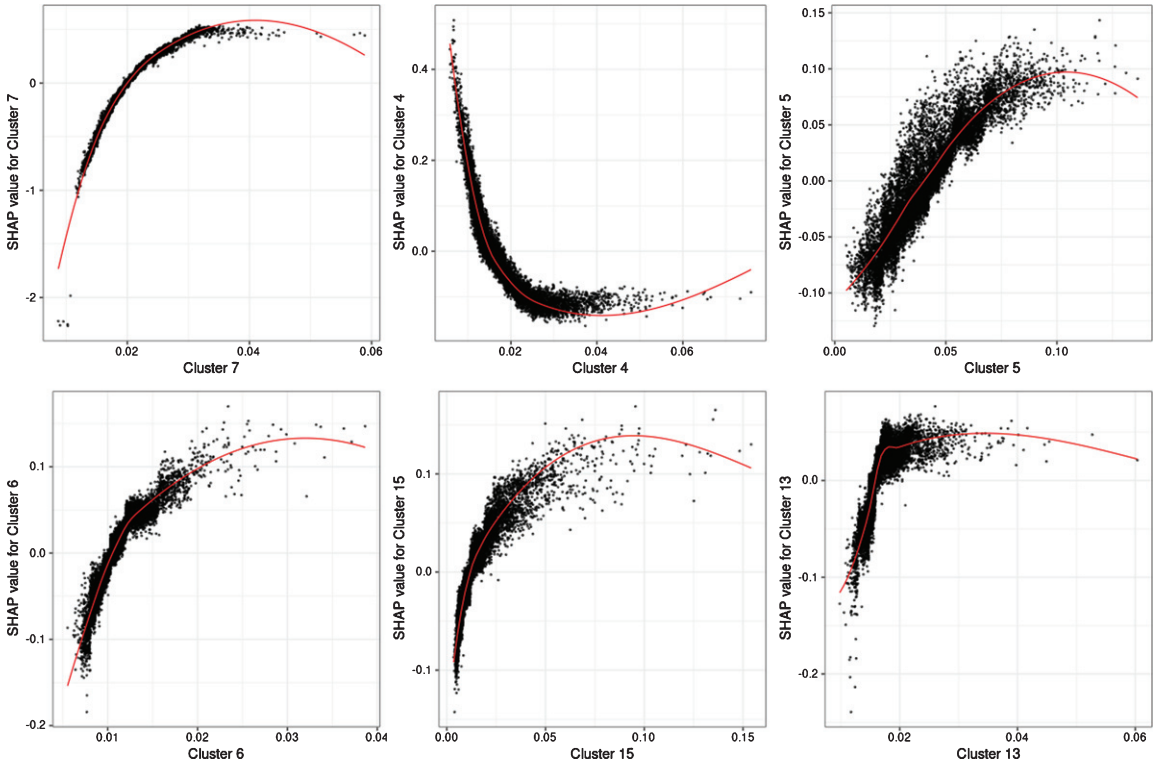

Fig. 6. SHAP dependence plots.

to plot just the interaction effect of metatopic 7 (highest influence on pricing) with other metatopics shows how the effect of location on predictions of pricing of Airbnb accommodations could vary.

- The less attractive the location (metatopic 7), the higher the influence of the values of metatopic 4 (more distant to neighbourhood amenities) on the prediction of the price. That is to say, smaller values of metatopic 4 push the prediction value down among specific traveller segments. Location is thus less concerned with the model output when it is accompanied by accommodations close to neighbourhood amenities. In this regard, Tussyadiah [50] proposes the crucial role of neighbourhood characteristics in addition to convenience in P2P accommodation evaluation to foster customer satisfaction.
- The greater the location value, the higher the influence of the connectivity (metatopic 5) on the prediction of the price. Indeed, a positive assessment of location is directly associated with high values in the price corresponding with, for instance, the presence of a subway line.
- Plotting the SHAP interaction value of location with housing space shows that guests could pay more for highly experiential housing-space based on comfort at home (metatopic 6) in a disadvantageous location (metatopic 7) despite higher costs and transportation expenses, and hosts could increase their price conferred by valuable authenticity. Previous studies suggest that staying at an Airbnb property (not located close to tourist hotspots) could offer 'real' experiences, and become an influential driver in using peer-to-peer accommodation (cf. [17, 24, 27, 50]). In other words, specific traveller segments could thus find it valuable to lodge outside of a tourist neighbourhood and enjoy the amenities of residential areas (cf. [23]).
- The lower the perception of the location (metatopic 7), the greater the influence of household amenities (metatopic 15) on the price of Airbnb accommodations. Distinct traveller segments could prefer paying more for higher housing benefits (access to functional amenities such as a full kitchen, a washing machine, and a dryer; cf. [17, 27]) in an advantageous location (metatopic 7). For instance, as Tussyadiah [50]

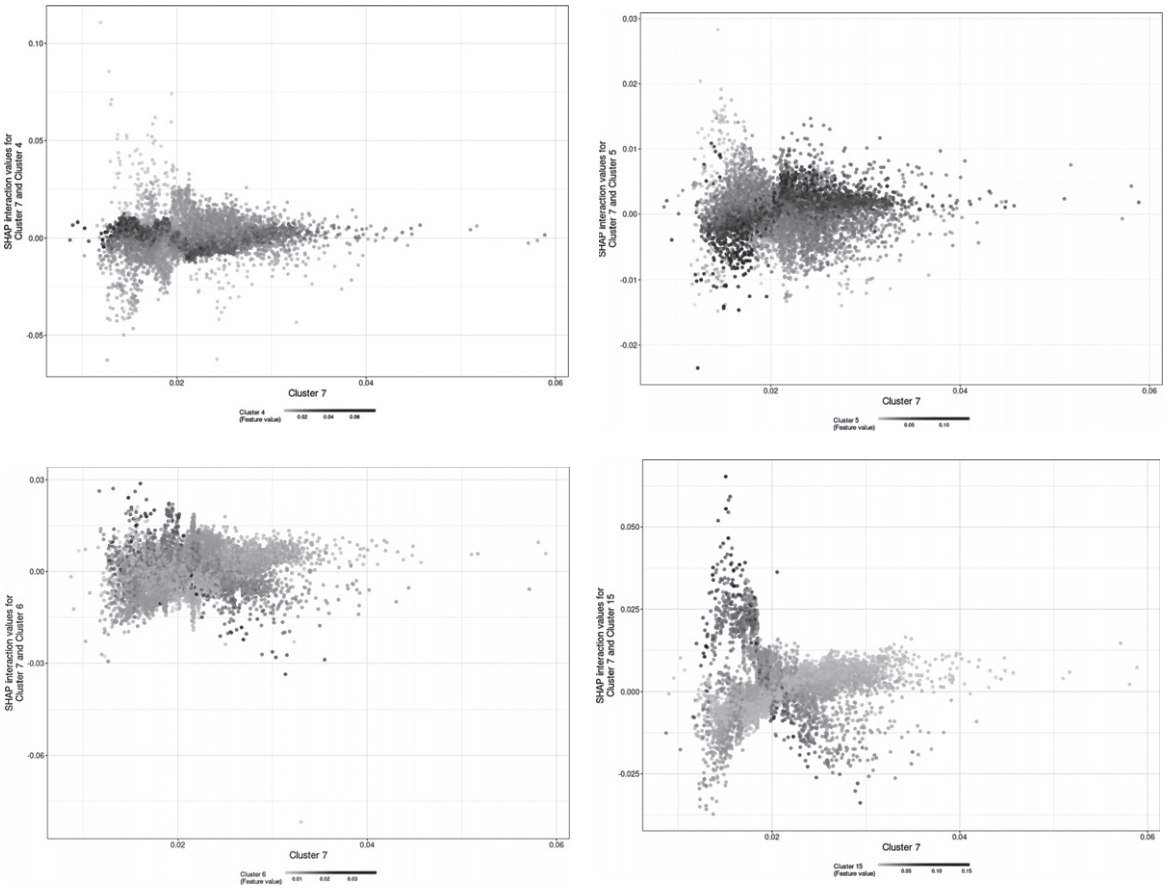

Fig. 7. Illustrating SHAP dependence plots based on interacting effects.

concludes, guests staying at entire apartments in less valuable locations highlight features related to the apartment reflecting the local authenticity and interacting with locals (metatopic 6) and valuable amenities (metatopic 15). Tussyadiah [51] conceptualises real authenticity as a social (enjoyable) benefit derived from interacting with hosts and local people.

6. Conclusion

This research has contributed to the literature on the sharing economy, offering insights for theoretical and managerial implications on the effect of guests' reviews on the price of accommodations, and deriving predictions and future recommendations. It has been consistent with the proposal that the guests' narratives offer valuable information to subsequent guests, affecting their demand and, subsequently, the price recommended. As far as the authors know, this research has represented a novel application of topic modelling (here STM) and fuzzy clustering to price analysis (from the demand side) in the sharing economy.

Based on an FCM cluster analysis on Airbnb reviews in five large metropolitan areas in New York, 15-metatopics have been extracted; among others, location, distance to neighbourhood amenities, connectivity, authenticity, apartment amenities and touristy hotspots. These main metatopics are comparable with those identified in previous studies (cf. the theoretical framework section) and described by guests who are not only looking for real experiences of staying at a locally charming accommodation, but also visiting, e.g., The Rockefeller Centre. The fuzzy algorithms have been "the most suitable as they are able to capture the 'undefined' tourists' behaviour, preferences, emotions, or other feelings" [37]. Postmodern reviews might belong to more than one cluster. Although a common practice in previous research has been to assign each instance to a cluster

-adopting a defuzzification procedure-, Airbnb travellers' reviews contain a wide variety of nouns and verbs (among other core part-of-speech categories) related to social experiences that need to be analysed in this fuzzy-mode.

According to the metatopics extracted, location has been the most relevant proxy of the appeal of Airbnb accommodations. As Guttentag et al. [23] concluded, this could be unexpected "because Airbnb accommodations tend to be scattered in residential neighbourhoods rather than clustered like hotels in a downtown tourism core". Although location has greatly impacted pricing strategy, sharing apartments in a disadvantageous location for tourism has also provided (among travellers' segments) a distinct and convincing Airbnb value proposition (e.g., access to authentic residential amenities or functional home amenities). Hosts should thus offer accurate descriptions with comments related to experiential housing-space, based on comfort at home, or the influence of household amenities that reflect real authenticity.

7. Implications

The research findings have provided diverse implications. On the one hand, traveller-generated comments expressed in natural language allow them to share their (latent) opinions and authentic local interactions related to sharing hospitality services. This study has not been restricted to quantitative variables, and consequently identified hospitality topics (and metatopics) that are subtle yet difficult to diagnose, and which may damage the Airbnb host reputation if left unaddressed. On the other hand, it has demonstrated the relevance of applying different methodologies to summarise and interpret essential cues hidden in a huge volume of data, and to explain travellers' decisions and pricing strategies. In particular, the topics (and metatopics) and their predictive contributions to pricing have been identified by applying STM, FCM and XGBoost regressors. The application of text analytics has provided a summarised structure of UGC, by clustering topics into metatopics. Moreover, by using techniques for NLP or FCM as unsupervised algorithms, the analysis has confirmed and addressed the results of the previous literature about Airbnb accommodation features. Finally, the results have indicated an increasing predictive capacity when metatopics have been integrated with XGBoost.

8. Limitations and future research

Several limitations and future research need to be acknowledged. Although one price point in time is here considered and does not capture seasonal evolution, topic modelling represents a dynamic that potentially evolves over time intervals such as seasons of the year. Topics rise and fall in popularity over time. Future research should therefore (1) detect bursty topics related to subjectively experienced encounters, and consequently (2) track the different contribution of UGC (during seasonal evolution) to pricing policies and mutually beneficial relationships. Moreover, online textual reviews are influenced by culture-based response styles or additional demographic features of guests (gender, age or income, among others). SHAP values should thus be analysed not only based on the impact of the pricing recommendation but also on additional metadata that describe the database's structure. For instance, future studies should analyse travellers' gender-based differences because males and females cognitively structure hospitality experiences using different criteria. In particular, inconsistent findings from previous studies precisely foster future research to assess the influence of the biological male-versus-female dichotomy as a moderating factor in the relationship between the most salient gender-based preferences on Airbnb experiences and pricing policies. Likewise, future research should analyse the infrequent terms in the long tail of the distribution. And an improvement of algorithms is also necessary to more easily develop topic models and select the optimal number of topics and metatopics. Finally, given the limitation that only one destination (New York city) is examined, it is necessary to develop future studies for other destinations.

References

[1] D.M. Boyd and N.B. Ellison, Social network sites: Definition, history, and scholarship, *Journal of Computer-Mediated Communication* **13**(1) (2007), 210–230.

[2] J.A. Chevalier and D. Mayzlin, The effect of word of mouth on sales: Online book reviews, *Journal of Marketing Research* **43**(3) (2006), 345–354.

[3] Z. Mao and J. Lyu, Why travelers use Airbnb again? An integrative approach to understanding travelers' repurchase intention, *International Journal of Contemporary Hospitality Management* **29**(9) (2017), 2464–2482.

[4] W.W. Moe and D.A. Schweidel, Online product opinions: Incidence, evaluation, and evolution, *Marketing Science* **31**(3) (2011), 372–386.

[5] C. Gibbs, D. Guttentag, U. Gretzel, J. Morton and A. Goodwill, Pricing in the sharing economy: A hedonic pricing model applied to Airbnb listings, *Journal of Travel & Tourism Marketing* **35**(1) (2018), 46–56.

[6] A. Sundararajan, The sharing economy: The end of employment and the rise of crowd-based capitalism, The MIT Press, Cambridge, MA, 2016.

[7] T. Ikkala and A. Lampinen, Defining the Price of Hospitality: Networked Hospitality Exchange via Airbnb, in: Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, New York, USA, 2014, pp. 173–176.

[8] Y. Zhao, X. Xu and M. Wang, Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews, *International Journal of Hospitality Management* **76** (2019), 111–121.

[9] D.M. Blei, Surveying a suite of algorithms that offer a solution to managing large document archives, *Communication of the ACM* **55**(4) (2012), 77–84.

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

[11] U. Gretzel and K.H. Yoo, Use and impact of online travel reviews, in: Information and communication technologies in tourism, P. O'Connor, W. Hopken, and U. Gretzel, eds., 2008, Springer Vienna, New York, 2008, pp. 35–46.

[12] B.A. Sparks, K.K.F. So and G.L. Bradley, Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern, *Tourism Management* **53** (2016), 74–85.

[13] O. Jeroen, Airbnb: The future of networked hospitality businesses, *Journal of Tourism Futures* **2**(1) (2016), 22–42.

[14] D.A. Guttentag and S.L.J. Smith, Assessing Airbnb as a disruptive innovation relative to hotels: Substitution and comparative performance expectations, *International Journal of Hospitality Management* **64** (2017), 1–10.

[15] A. Lampinen and C. Cheshire, Hosting via Airbnb: Motivations and Financial Assurances in Monetized Network Hospitality, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, 2016, pp. 1669–1680.

[16] D. Wang and J.L. Nicolau, Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com, *International Journal of Hospitality Management* **62** (2017), 120–131.

[17] D. Guttentag, Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector, *Current Issues in Tourism* **18**(12) (2015), 1192–1217.

[18] D. Guttentag, Why tourists choose Airbnb: A motivation-based segmentation study underpinned by innovation concepts (2016). http://hdl.handle.net/10012/10684

[19] A.-G. Johnson and B. Neuhofer, Airbnb – an exploration of value co-creation experiences in Jamaica, *International Journal of Contemporary Hospitality Management* **29**(9) (2017), 2361–2376.

[20] S. Satama, Consumer adoption of access-based consumption services-Case AirBnB, Aalto University School of Business, 2014. https://aaltodoc.aalto.fi/handle/123456789/13723

[21] I.P. Tussyadiah and J. Pesonen, Drivers and barriers of peer-to-peer accommodation stay – an exploratory study with American and Finnish travellers, *Current Issues in Tourism* **21**(6) (2016), 703–720.

[22] S. Yang and S. Ahn, Impact of motivation in the sharing economy and perceived security in attitude and loyalty toward Airbnb, *Advanced Science and Technology Letters* **129** (2016), 180–184.

[23] D. Guttentag, S. Smith, L. Potwarka and M. Havitz, Why tourists choose Airbnb: A motivation-based segmentation study, *Journal of Travel Research* **57**(3) (2018), 342–359.

[24] L.J. Liang, Understanding repurchase intention of Airbnb consumers: perceived authenticity, EWoM and price sensitivity, University of Guelph, Guelph, 2015.

[25] M.A. Mody, C. Suess and X. Lehto, The accommodation experiencescape: A comparative assessment of hotels and Airbnb, *International Journal of Contemporary Hospitality Management* **29**(9) (2017), 2377–2404.

[26] K.Y. Poon and W.-J. Huang, Past experience, traveler personality and tripographics on intention to use Airbnb, *International Journal of Contemporary Hospitality Management* **29**(9) (2017), 2425–2443.

[27] K.K.F. So, H. Oh and S. Min, Motivations and constraints of Airbnb consumers: Findings from a mixed-methods approach, *Tourism Management* **67** (2018) 224–236

[28] G. Forman, BNS Feature Scaling: An Improved Representation over Tf-idf for SVM Text Classification, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM), ACM, New York, NY, USA, 2008, pp. 263–270.

[29] M.E. Roberts, B.M. Stewart, D. Tingley and E.M. Airoldi, The structural topic model and applied social science, in: Advances in neural information processing systems workshop on topic models: computation, application, and evaluation, Harrahs and Harveys, Lake Tahoe, 2013, pp. 1–20.

[30] M.E. Roberts, B.M. Stewart and D. Tingley, stm: R package for structural topic models, *Journal of Statistical Software* **91**(2) (2019).

[31] M.E. Roberts, B.M. Stewartand and D. Tingley, Navigating the local modes of big data: The case of topic models, in: Computational social science: Discovery and prediction, R.M. Alvarez, Ed., Cambridge University Press, New York, 2016, pp. 51–97.

[32] M.E. Roberts, B.M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S.K. Gadarian, B. Albertson and D.G. Rand, Structural topic models for open-ended survey responses, *American Journal of Political Science* **58**(4) (2014), 1064–1082.

[33] J. Gerring, Social science methodology: A unified framework, Cambridge University Press, Cambridge, 2001.

[34] C. Lutz and G. Newlands, Consumer segmentation within the sharing economy: The case of Airbnb, *Journal of Business Research* **88** (2018), 187–196.

[35] J. Bischof and E.M. Airoldi, Summarizing topical content with word frequency and exclusivity, in: Proceedings of the 29th International Conference on Machine Learning (ICML-12), J. Langford, and J. Pineau, Eds., Omnipress, New York, NY, 2012, pp. 201–208.

[36] K.M. Quinn, B.L. Monroe, M. Colaresi, M.H. Crespin and D.R. Radev, How to analyze political attention with minimal assumptions and costs, *American Journal of Political Science* **54**(1) (2010), 209–228.

[37] P. D'Urso, M. Disegna, R. Massari and L. Osti, Fuzzy segmentation of postmodern tourists, *Tourism Management* **55** (2016), 297–308.

[38] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *Journal of Cybernetics* **3**(3) (1973), 32–57.

[39] J.C. Bezdek, Cluster validity with fuzzy sets, *Journal of Cybernetics* **3**(3) (1974), 58–73.

[40] J.C. Bezdek, Pattern recognition with fuzzy objective function algorithms, Kluwer Academic Publishers, Norwell, USA, 1981.

[41] J.C. Bezdek, R. Ehrlich and W. Full, FCM: The fuzzy c-means clustering algorithm, *Computers & Geosciences* **10**(2) (1984), 191–203.

[42] J.H. Friedman, Greedy function approximation: A gradient boosting machine, *The Annals of Statistics* **29**(5) (2001), 1189–1232.

[43] T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2016, pp. 785–794.

[44] F. Climent, A. Momparler and P. Carmona, Anticipating bank distress in the Eurozone: An Extreme Gradient Boosting approach, *Journal of Business Research* **101** (2019), 885–896.

[45] T. Chen, T. He, M. Benesty, V. Khotilovich and Y. Tang, XGBoost: eXtreme Gradient Boosting (R package ver-sion 0.90.0.2), (2019). https://https//cran.r-project.org/web/packages/xgboost/xgboost.pdf

[46] M. Kuhn, Caret: classification and regression training (R package version 6.0–84), (2019). https://cran.r-project.org/web/packages/caret/caret.pdf

[47] X.L. Xie and G. Beni, A validity measure for fuzzy clustering, *IEEE Transactions on Pattern Analysis & Machine Intelligence* **13**(8) (1991), 841–847.

[48] S.M. Lundberg and S.-I. Lee, A Unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems 30 (NIPS 2017), I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 4765–4774.

[49] S.M. Lundberg, G.G. Erion and S.-I. Lee, Consistent individualized feature attribution for tree ensembles, arXiv preprint, 2018. http://arxiv.org/abs/1802.03888

[50] I.P. Tussyadiah, Factors of satisfaction and intention to use peer-to-peer accommodation, *International Journal of Hospitality Management* **55** (2016), 70–80.

[51] I.P. Tussyadiah, An Exploratory study on drivers and deterrents of collaborative consumption in travel, in: Information & Communication Technologies in Tourism 2015, I. Tussyadiah, and A. Inversini, Eds., Springer International Publishing, Switzerland, 2015, pp. 817–830.