THE INFLUENCE OF GEOGRAPHIC AND PSYCHIC DISTANCE ON ONLINE HOTEL RATINGS

Abstract

This study examines the relationship between distance measures and a Portuguese dataset consisting of 34,622 online hotel reviews extracted from Booking.com and TripAdvisor, written in Portuguese, Spanish, and English. Based on the country of origin of each review author, a geographic and a psychic distance measure is calculated for Portugal. Data and text mining analysis provides additional insights into online hotel ratings. We confirm that online travelers' evaluations are multifaceted constructs displaying varying patterns of rating behavior among the traveler base. By investigating the contemporary relevance of geographic and psychic distance, a key finding of this study is that travelers with less distance both in terms of psychic and geographic distance give a lower rating score than travelers with greater distance. The inclusion of psychic and geographic distance is advocated as a salient aspect for future researchers and for those practitioners who wish to enhance hotel product and service features.

KEYWORDS:

Hotels; Geographic distance; Psychic distance; Text mining; Online hotel ratings

INTRODUCTION

The increasing focus on customer experience by practitioners has led to the creation of a rich seam of research pertaining to online hotel ratings. Travelers' purchasing decisions are increasingly being influenced by online reviews (Cantallops and Salvi 2014; Kwok, Xie, and Richards 2017; Ring, Tkaczynski, and Dolnicar 2016; Tan, Lv, and Gursoy 2018). Indeed, from such investigations various questions arise, and are answered. We know from several contributions, that culture (Gao et al. 2018), language (Antonio et al. 2018a; Goethals 2016; Wu et al. 2017), travel experience (Lu and Stepchenkova 2015; Morosan and Bowen 2017), are among factors that have an influence on online hotel ratings. The continued growth of data-generating platforms have inspired new approaches to understanding the traveler experience. Such subjective rating information is now expressed and published in more than seventy diverse platforms including popular online booking websites such as Booking.com and TripAdvisor (Phillips et al. 2015).

Reviews from Booking.com and TripAdvisor possess two main types of ratings: quantitative (the overall review score) and qualitative (the textual component being the commentary). Although there are numerous studies on the subject of online reviews, most of them focus on the quantitative ratings of reviews to represent user opinion (Duan et al. 2016), but recent works are advocating the use of the textual component of reviews (Antonio et al. 2018a; Bjørkelund, Burnett, and Nørvag 2012; Duan et al. 2016; Han et al. 2016; Xiang et al. 2015; Xu and Li 2016). The rationale being the textual component has the potential to allow for better recognition of "guests' true feelings" (Han et al. 2016, 17).

The initial interest of online hotel ratings has been maintained by researchers advocating the merits of understanding evaluations posted on web and social media sites (Floyd et al. 2014; Kostyra et al. 2016; Schuckert, Liu, and Law 2015). Li, Xu, Tang, Wang, and Li, (2018) provide a thorough review of big data and online hotel ratings research. To date, researchers have mainly focused on using sentiment analysis which can automatically detect the valence of a piece of reviewer text, which can be positive, neutral or negative (Geetha, Singha, and Sinha 2017). Prior tourism related research has assessed the valence of reviews (Duverger 2013; Sparks and Browning 2011); volume of reviews (Xie, Zhang, and Zhang 2014); variance of reviews (Melian-Gonzalez, Bulchand-Gidumal, and Gonzalez Lopez-Valcarcel 2013). Additional insights can be derived from studying semantic relationships and meaning in online hotel ratings (Alaei, Becken, and Stantic 2017; Phillips et al. 2016; Xiang et al. 2015; Xu and Li 2016). The degree of positivity or negativity towards the main textual subject of online hotel ratings (semantic analysis) is currently a hotbed of research and development for academics and practitioners (Ge, Vazquez, and Gretzel 2018). The continual lower prices to travel to overseas locations together with a more favorable US dollar exchange rates have in part accelerated the international dimension of online hotel ratings. Online travelers' preferences have been investigated from many facets, but distance offers a fresh perspective. Indeed, in an online environment the concept of distance, needs to go beyond geographic distance (Deodhar, Subramani, and Zaheer 2017). Similar to Deodhar, Subramani and Zaheer (2017), this study incorporates a set of psychic measures, which are one of the most popular forms of distance (Safari, Thilenius, and Hadjikhani 2013). Psychic distance is liken as "the sum of factors" or the "differences" that goes beyond the objective criteria of geographic and cultural distance per se (see Yang, Liu, and Li 2019) , incorporating other factors such as business development, industrial development, and education differences too (Dow and Karunaratna 2006).

Using online hotel ratings, this study explores the relationship between distance measures. The authors generate a dataset consisting of reviews and associated ratings on Booking.com and TripAdvisor for Portuguese hotels in three different languages (Portuguese, Spanish and English). The country of origin of each review author is collected in order to derive a geographic and psychic distance measure between the author's country of origin and Portugal. A more technical aspect of this study is the sentiment analysis of reviews, where each review is in addition associated with a sentiment score based on a dictionary approach, where the ratio of terms with positive and negative sentiment is computed to derive an overall score. This technical approach of text mining analysis, provides additional insights into online hotel ratings.

Dissimilarities among travelers will influence their preferences with respect to hotel attributes (Banarjee and Chua 2016), which is rather pertinent for heterogeneous groups of travelers. Such a difference, leads onto social identity theory (Tajfel 1982), whereby networked-based communities may act not as an individual, but develop a social identity. Individuals perceive themselves and others as belonging to various social groups, which from the perspective of the hotel may result in different evaluative online hotel ratings statements, from guests based on their distance measures. For example, although culture is an important factor in decision-making, other factors such as online platforms may affect the decision-making process of travelers. Given the relatively nascent state of research, there is limited empirical work directly related to how the country of origin affects rating behavior. Furthering this path has led to some recent studies (Kim

2018; Gao et al. 2018) who look at the country of origin effects relating to online review ratings and culture and its effects relating to online review ratings respectively. Less well examined too, is the role of language in online hotel ratings (Schuckert, Liu, and Law 2015; Yong et al. 2017). Moreover, research is lacking on how distance influences travelers' ratings, together with the dangers of aggregating reviews written in multiple languages.

So why does this matter? Well, having an accurate understanding of salient online hotel ratings' relationships is essential for both strategic management and marketing theory and practice. The economic and societal impact of tourism across global markets is a priority for governments, private sector and societal oriented organizations. Hotels play a pivotal role in a country's tourism product. Travelers of varying distances may possess different expectations in areas unknown to those responsible for marketing strategies at the individual and destination level. So, by understanding such relationships may help advance a more effective connectivity among the online hotel ratings database, which is a key strategic resource.

The remainder of the paper is organized as follows. In the next section, we consider online hotel ratings, the relevant distance literature and present the research questions that describe the positioning of the study. The data and methods are outlined. Finally, we present the data analysis and results of our empirical analysis and round off with conclusion, theoretical and managerial implications.

ONLINE HOTEL RATINGS

Valence, variance, volume, and increasingly verbal features (4Vs) are the essential evaluation features of online hotel ratings (Maeyer 2012). Hoteliers believe that their performance will be hampered if they are unable to reliably monitor online hotel ratings, and results from recent academic studies support this (Duverger 2013; Phillips et al. 2016; Xie, Zhang, and Zhang 2014). A consequence of the popularity of online hotel ratings, is that reviews now constitute a new element of the marketing communication mix and have implications for both theory and practice. The decomposition of online reviews into their main elements of valence, variance and volume has been one way to obtain a better understanding into the relevance of each aspect of firm performance (see e.g. Floyd et al. 2014; Kostyra et al. 2016 for an overview).

Online hotel ratings not only captures online reviews, recommendations and opinions exchanged by consumers (Cantallops and Salvi 2014) but also form the bases on which consumers may revise their purchase decisions and ultimately change their buying behavior (Cantallops and Salvi 2014; Sparks and Browning 2011). Online hotel ratings create a resource whereby reviewers, review readers and managers can use either quantitative or qualitative techniques to consider outcomes in terms of consumer decision-making and business performance (Kwok, Xie, and Richards 2017). In fact, because of the diversity of opinion, independence, decentralization, and aggregation, users who post online reviews can be considered as Surowiecki (2005) calls a “crowd”. This being a diverse collection of independent individuals which are better at making certain decisions and predictions that its individual members or even, better than experts, which explains why, these days, consumers value more online hotel ratings than hotels’ official classifications or stars (Öğüt and Onur Taş 2012). A number of companies such as Olery.com (www.olery.com), ReviewPro (www.reviewpro.com) and Revinate (www.revinate.com) have sprung up to develop reputational management systems that show how improving guest satisfaction can translate and enhance revenues (Hensens 2015). Another firm with a strong presence is Brand Karma, which is an in-house market agency of Next story group (www.nextstory.com). This firm has the ability to filter western social media channels from Chinese social media channels.

To illustrate such benefits that can accrue, consider, for example Anderson (2012) who found that a 1% increase in a hotel's index score results in higher profitability in terms of 1.42% increase in Revenue per Available Room. But now, we accept that there is causality between review management, reputation and revenue development, but not as linear as presented by Anderson (2012). Recent studies have demonstrated that the effect of review management on revenues depends on the type of hotel, the destination, the customer structure, and the occupancy rate, for instance (Kim, Lim, and Brymer 2015; Phillips et al. 2015, 2016; Xie, Zhang, and Zhang 2014; Yang, Park, and Hu 2018).

Online hotel ratings is now a powerful resource, as the exchange of information by which the communicator (reviewer) transmits content (message) to several communicates (receivers), which can modify perceptions and behavior (Hernández-Ortega 2018). In terms of further opportunities, marketing managers ought to learn how to actively manage reviews, including negative reviews (Baka 2016; Cantallops and Salvi 2014). Yet, previously raised questions on what needs managing and measuring (Godes and Mayzlin 2004) have become more difficult to answer due to the increasing availability of data both to consumers and organizations.

Moreover, an area which has received scant scholarly attention is that of the influence of travelers' origin on online hotel ratings. In general, prior academic studies aggregate reviews from individual travelers of differing origins to compute an average rating. Such travelers may have consistently varying experiences. This practice raises particular issues for those hotels, where guests come from different nationalities (Wilson, Murphy, and Fierro 2012). Prior research of online hotel ratings aggregated data does reveal general trends, but as Mckercher (2008) notes aggregation camouflages significant changes that occur at sub-market levels. To illustrate this point consider (Pizam and Sussmann 1995) who espoused that travelers' perceptions in terms of satisfaction levels do vary according to country of origin. We know that when travelers select a tourism destination, they are influenced in part by both measurable and cognitive distances (Ankomah, Crompon, and Baker 1996; Massara and Severino 2013; Uchiyama and Kohsaka 2016; Zhang, Seo, and Lee 2013).

Understanding how distance and language influences online hotel ratings is important for several reasons. Dissimilarities among travelers will influence their preferences with respect to hotel attributes (Banarjee and Chua 2016), which is rather pertinent for heterogeneous groups of travelers. Having an accurate understanding of salient online hotel ratings' relationships is essential for both strategic management and marketing theory and practice. The economic and societal impact of tourism across global markets is a priority for governments, private sector and societal oriented organizations. Travelers of different origins will possess significantly different expectations. So, by understanding changes in online customer reviews beyond those written in

English, may help advance a more effective connectivity among the online hotel ratings database.

DISTANCE

The construct of distance can be disaggregated into multiple measures across the social sciences, economic, financial, political, administrative, cultural, as well as geographic (Berry, Guillén, and Zhou 2010). According to Johanson and Vahlne (1977), distance can be measured as an objective variable (e.g. geographic) and measured as a matter of decision-makers' perceptions (e.g. psychic distance). Both physical and perception distances are related but imperfectly correlated, and physical distance influences judgement and decision-making (Fujita et al. 2006).

Prior research within the business and management literature has considered cultural (Hofstede 1980; House et al. 2004), psychic (Beckerman 1956; Dow and Karunaratna 2006) and geographic (Choi and Contractor 2016; Mckercher 2008) distances as central to comprehending organizational performance. This study considers both objective and perceptive perspectives.

Geographic

Understanding, the influence of geographical distance is important for several reasons. The stimuli of geographic distance has been incorporated in prior empirical studies (Choi and Contractor 2016). Previous research defines geographic distance as the distance between two cities in kilometers (Brewer 2007). Blum and Goldfarb (2006) note how geographic distance influences the trade of digital goods sold over the Internet. Studies have had varying levels of success by assessing the distance between capital cities (Brock, Johnson, and Zhou 2011), major cities (Hutzschenreuter, Kleindienst, and Lange 2014), and geographic centers of countries (Ojala and Tyrväinen 2008).The geographical dispersal of travelers present opportunities for marketers to customize visitor packages. In the tourism literature, studies observe that travel demand decreases as distance from the origin market increases (Cai and Li 2009; Mckercher 2008; Mckercher and Lew 2003). Increasing the distance, adds time, costs and money, thus making the destination less attractive to the traveler (Prideaux 2000). The distance-decay model provides some theoretical foundations (Mckercher and Lew 2003), where the demand increases up to a certain distance and afterwards decreases exponentially. Nicolau and Más (2006) proposed that the effects of distance and prices are moderated by travelers' motivation. The digitized environment operating across different time zones can further reduce the efficacy of the communication effort. Geographical proximate destinations provide lower economic and social costs, together with a degree of environmental familiarity.

In short, both information networks and transportation costs may influence the impact of distance (Ghemawat 2001). Notwithstanding the improvements in transportation systems and digital technologies, travelers that are geographically distant may undergo differing experiences in their outbound trips. In fact, Ojala (2015) remark that modern air transport and communication have reduced the perceived distance, and have eased commercial interactions. Child, Ng and Wong (2002) allude to these as "distance-compressing factors". This study investigates how traveler distance between home and destination influences online hotel ratings by considering:
RQ1 How do varying levels of geographic distance influence travelers' online hotel ratings?

Psychic

In this section, we set out the positioning of our study in the broader destination image literature. We begin by briefly acknowledging the destination image literature, and highlight why we have framed our approach using psychic distance.

In examining how tourists view or have mental representations of a place, researchers generally consider destination image (Ryan and Cave 2005). Taking this perspective, destination image is commonly depicted as a concept formed by a traveler's interpretation of cognitive and perceptive evaluations and effective appraisals towards a destination (Crompton 1979; Hallmann, Zehrer, and Müller 2015). The topic has been the most popular tourism literature for more than four decades (Pike and Page 2014), and is considered to be a multidimensional construct. In the extant literature, destination image tourists' mental representation has been defined, operationalized and measured in a plethora of ways. Space precludes us from a detailed overview, but Kock, Josiassen, and Assaf (2016) provide a succinct overview of the destination image literature. Critically, two ways of depicting destination image include: the sum of beliefs, ideas, and impressions people have of an object, place, destination (Zhang et al. 2014). Another view relates to cognitive (beliefs or assessments), affective (positive or negative emotion) and conative (behavioural intention) (Choi, Hickerson, and Kerstetter 2018; Kim 2018).

On the credit side, although the considerable body of prior destination image research provide useful insights, they leave room for additional theorizing and empirical research. We shall outline our rationale.

First, as previously stated the topic of destination image has been one of the most popular topics of the tourism literature (Pike and Page 2014). However, the precise nature and scope of destination image remain vague (Hallmann, Zehrer, and Müller 2015; Lai and Li 2016). As Albert Einstein famously quoted ''We cannot solve problems by using the same kind of thinking we used when we created them''. Moreover, new approaches are required if organizations wish to prosper and survive new environments (Baden-Fuller and Stopford 1994; Markides 1998) . We wish to look outside this traditional destination image approach, indeed delve into another area.

Second, in a turbulent, chaotic and nonlinear tourism environment, strategies need to incorporate cultural and value differences (Phillips and Moutinho 2014). More specifically, Phillips and Moutinho (2014) lament about the methodological introspection of prior approaches in tourism and stress that new research methodologies are critically important in enhancing theory and practice. The implication being that new approaches may generate fresh knowledge and insights too.

Third, in addition, Ferrer-Rosell, Martin-Fuentes, and Marine-Roig (2019) findings revealed that the marketing promotion activities of higher-class hotels highlight their facilities, whereas lower-class hotels refer more to the destination. In this study, 4 and 5-star hotels made up more than 80% of our sample. This latter point reinforces that our unit of analysis is not the destination per se, but the hotel itself.

Finally, according to Mossberg and Kleppe (2005), destination image is an area for marketing practice, which can incorporate the sale of export products in the international arena. Psychic distance being the distance between the home country of the firm and export countries can be used as multidimensional concept and measured from the customer perspective at the individual level (Assarut and Srisuphaolarn 2018). So, the adoption of psychic measures can incorporate Mossberg and Kleppe (2005) views of destination image, and its legitimacy can be supported by our first three points. We replace the traditional unit of analysis of the firm with the traveler, and give attention to the international business management and marketing literatures and employ both psychic and distance measures. Kim (2018) considers post visit image rather than revisit, so that tourists are able to rate their experiences. The current study adopts the post visit approach and considers hotels and uses online hotel reviews to gauge perceptions.

Early psychic distance research commenced with Beckerman (1956), who coined the phrase by remarking on the special problem posed by its existence. The term was sporadically referred to in international trade flow research (Geraci and Prewo 1977; Linnemann 1966). During the 1970s prominence in the management-oriented literature was provided by the research at the University of Uppsala. In terms of measurement, the sum of the factors approach include differences in language, culture, political systems, level of education, and level of industrial development (Johanson and Wiedersheim-Paul 1975). International business researchers have since refined and added to the aforementioned list. The existing literature offers a wide range of studies, but developing and confirming a set of psychic scales that captures the characteristics that matter has posed a dilemma (Dow and Karunaratna 2006).

Psychic distance is not solely about nationality and cultural factors, but considers individuals and relationships of customers in an international online setting (Safari, Thilenius, and Hadjikhani 2013). The unit of analysis varies too, with some studies considering differences between countries and others between companies (Durand, Turkina, and Robson 2016). In the context of this study psychic distance is the gap or differences that a traveler might perceive between their origin (country) and the destination. In spite of a decade of online hotel ratings research, the field of tourism has not yet illuminated a comprehensive analysis of the fact that travelers with different origins may provide ratings, which are different on a number of distance dimensions beyond solely cultural studies (Assaf, Josiassen, and Agbola 2015; Bi and Lehto 2018; Martin, Jin, and Trang 2017; Qian, Law, and Wei 2018). Psychic distance goes beyond the objective criteria of geographic and cultural distance per se, as it incorporates business, industrial development, and education differences too. Dow and Karunaratna (2006) proposed and tested a range of potential psychic distance stimuli encompassing culture, language, religion, education and political systems. This school of thought concentrates upon more than one stimulus, such as culture, and demonstrates that the latter is only one indicator.

The concept of psychic distance is one of the most explored areas in the internationalization literature (Safari, Thilenius, and Hadjikhani 2013). Yet, conflicting findings on the issue of psychic distance indicates the need for further research (see Durand, Turkina and Robson 2016). This issue deserves attention in tourism too, as it prevents researchers and practitioners from making effective recommendations in deploying marketing strategies. The scarcity of available resources now makes it imperative that the salient drivers are identified (Durand, Turkina, and Robson 2016). Considering the importance of forming and maintaining effective customer relationships as drivers of competitiveness, innovation, customer satisfaction, and performance (Ulaga and Eggert 2006) in international settings (Zhang, Cavusgil, and Roath 2003), it is necessary to identify contingent factors that influence the effect of psychic distance on international travel (Durand, Turkina, and Robson 2016).

Another significant observation from the prior literature is the absence of studies applying psychic distance in online settings (Safari, Thilenius, and Hadjikhani 2013). In this study, we reflect and investigate this multifaceted concept by considering:

RQ2 How do varying levels of psychic distance influence travelers’ online hotel ratings?

DATA AND METHODS

DATASET

The study utilizes a unique dataset created by merging four different datasets: one with geographic distances between countries created by Mayer and Zignago (2011), a dataset with psychic distance between countries developed by Dow and Karunaratna (2006), a third one with ISO country codes (International Organization for Standardization 2017) with ISO 3166 two-digit country codes and their designations in English, Portuguese and Spanish, and a dataset of hotel online customer reviews. The latter was created using a custom-built web content extractor that retrieved a total of 39,425 hotel reviews published during the period 1st July 2015 to the 30th November 2016. The custom-built web extractor made use of a Firefox internet browser to automatically navigate through Booking.com and TripAdvisor reviews' web pages and process the content of those web pages. A process known as "web scraping" (Batrinca and Treleaven 2015; Braun, Kuljanin, and DeShon 2018). European law recognizes users can make copies of publicly available databases and use that data in research (Bosch 2017; Monkman, Kaiser, and Hyder 2018), but companies are making scraping increasingly difficult (Jennings and Yates 2009). Due to this difficulty we decided to extract data only from Booking.com and TripAdvisor as these are the two of the most popular platforms, and only in English, Spanish and Portuguese. Also, these three languages represent the main official languages of 70 per cent of Portugal's hotel guests (Instituto Nacional de Estatística 2016). This diversity in languages makes Portugal an ideal location to examine the influence of language. Difference in language is a stimuli that has received endorsement from numerous studies, from Beckerman (1956), Conway and Swift (2000) and Dow and Karunaratna (2006) to more recent works such as Avloniti and Filipppaios (2014), Cuypers, Ertug and Hennart (2015), and Antonio et al. (2018a).

One of the authors (responsible for the data collections) is actively involved in the Portuguese hotel sector and has access to many hotel contacts and sources of data. This enables the collection of both qualitative and quantitative data, which supports this study. Portugal is the setting for the destination with two, three, four and five star city and resort hotels providing the context of the study. The inclusion of city and resort hotels enable greater insights by category of hotel. Andriotis (2011) clusters destinations into three categories: urban, coastline and rural and so, in terms of hotel profiles, our study uses City (urban) hotels in Lisbon and Resort (coastline) hotels in the Algarve. Four city hotels and four resort hotels were initially selected and each hotel manager were asked to identify the top five hotels of their competitive set. This resulted in a total of 56 hotels being selected for online reviews retrieval, from two to five stars, as detailed in Table 1.

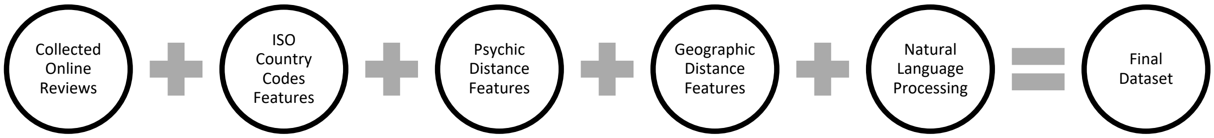
Table 1. Hotel summary

Hotel classification	**City**		**Resort**	
	Hotels	**Average rooms**	**Hotels**	**Average rooms**
Two stars	5	52	4	32
Three stars	6	65	6	77
Four stars	12	127	12	202
Five stars	6	224	5	116
Total	29	117	27	106

Dataset elaboration

To elaborate and analyze this dataset with respect to the two research questions, we employed the software package R because of its openness, statistical and visualization capabilities. As previously mentioned and illustrated in Figure 1, this study's dataset is a merger of four different datasets: hotel reviews, geographic distances, psychic distances and ISO country codes. The construction of the final dataset was based on the hotel reviews. From the 39,425 obtained reviews, 16 were removed because they were duplicated or in a language different from the ones chosen for this study. As the country of the traveler writing the review was not identifiable, another 3,877 reviews were removed. Most TripAdvisor reviews provide the user's identification and his/her location, but that is not the case in Booking.com reviews, where either location is not a mandatory field in the user profile or the user can ask to remain anonymous. Lastly, 448 reviews were removed because they were from countries that had less 20 reviews, or were from countries where there was no information on the geographic or psychic distances datasets, which was the case of 462 reviews from Serbia, Gibraltar, Georgia and Angola.

Figure 1. Dataset elaboration diagram


Collected Online Reviews
ISO Country Codes Features
Psychic Distance Features
Geographic Distance Features
Natural Language Processing
Final Dataset

An array of Data Science tools were employed to build this dataset, including Data Visualization, Natural Language Processing, Feature Engineering, Statistics and Machine Learning. Such tools enable the creation of new features, which were necessary because:

Booking.com and TripAdvisor use different rating scales in their quantitative components. Booking.com uses a continuous scale from 1 to 10 and TripAdvisor a discrete scale from 1 to 5. Besides this, there is a difference in the scales used. Yet, Booking.com scale actually has a minimum rating of 2.5, as highlighted by Mellinas, María-Dolores, and García (2015). Thus, it is necessary to normalize the quantitative ratings from both sources in order to study them.

There is a need for clearer analysis and interpretation of the impact of geographic and psychic distances in ratings. Thus, geographic and psychic distances, originally continuous variables, need to be converted into categorical features.

A summary of the features included in the final dataset is presented in Table 2.

Table 2. Final dataset features summary

Feature	Origin	Description
CEPII_dist	CEPII dataset	Geographic distance (in kilometers) from Portugal and review user country based on most important cities/agglomerations (of population)
HotelID	Reviews dataset	Hotel ID
HotelStars	Reviews dataset	Hotel official classification (2 to 5 stars)
HotelType	Reviews dataset	Type of hotel (City or Resort)
GeoDistanceFactor	CEPII dataset	Categorical version of "CEPII_dist"
Language	Reviews dataset	Language of the review (English, Spanish or Portuguese)
PD_PD_DK	Psychic distance dataset	Psychic distance from Portugal and the review user country (numeric)
PsychicDistanceFactor	Psychic distance dataset	Categorical version of "PD_PD_DK" (PT, Near or Far)
RevID	Reviews dataset	Review unique ID
RevRating	Reviews dataset	Review overall rating (normalized in a 1 to 5 scale)
RevSentences	Reviews dataset	Number of sentences in the textual component of the review
RevSentimentStrength	Reviews dataset	Sentiment analysis polarity value, calculated from the textual component of the review
RevTotalWords	Reviews dataset	Number of words in the textual component of the review
RevUserCountyISOCode	Reviews dataset	ISO country code based on the location mentioned on the review
Source	Reviews dataset	Website were review were extracted from (Booking.com or TripAdvisor)

Some of the features in Table 2 were engineered, namely:

- *GeoDistanceFactor*: geographic distance was transformed and resulted in a three-valued categorical feature: PT (Portugal), Near and Far. We considered values from 0 to 114.9 km as "PT", values from 115 to 4,999.9 km as "Near" (this includes most European countries) and from 5,000 km upwards as “Far”. The process of transforming continuous features to categorial features is called discretization. This process is usually done for allowing a feature to be employed by machine learning algorithms who do not work with continuous features, to speed processing, or to increase interpretability (Dougherty, Kohavi, and Sahami 1995; Kotsiantis and Kanellopoulos 2006). Discretization methods are usually divided into two groups: unsupervised and supervised. Unsupervised methods, such as equal interval binning or equal frequency binning, do not make use of class membership information in the discretization process. Conversely, supervised methods make use of class membership information to establish the discretization limits. Since supervised discretization methods only produce slightly better performance results than unsupervised methods (Dougherty, Kohavi, and Sahami 1995) and our objective was not to build a predictive model, we decided to employ an unsupervised approach that would guarantee what Kotsiantis and Kanellopoulos (2006) designate as the compromise between information quality (homogenous intervals) and statistical quality (sufficient sample size to ensure generalization).

- *PsychicDistanceFactor*: psychic distance was also transformed to a categorical feature. As for geographic distance, psychic distance was divided into three named values using a similar distance criterion: PT (Portugal), Near and Far. We considered a null (zero) distance as “PT”. From 0.1 to 1.49, which cover most Latin countries and other countries that Portuguese feel as “familiar” like Brazil, as “Near” (in terms of religion, language, and even in historic background). All other countries with a psychic distance above 1.5 were considered “Far”.

- *RevRating*: due to the aforementioned differences in the quantitative rating scales used by Booking.com and TripAdvisor, ratings were normalized so that the quantitative overall rating of reviews could be analyzed together. We opted to normalize ratings according to the TripAdvisor scale, that is from 1 to 5. Since Booking.com only allows a minimum rating of 2.5, we employed *binning*, a common technique used to convert numeric variables to discrete (Abbott 2014; Dougherty, Kohavi, and Sahami 1995). We divided the amplitude of the Booking.com scale (7.5 = 10 – 2.5) by 5 to obtain each bin amplitude, which resulted in the following bins classification intervals: [2.5, 4.0[, [4.0, 5.5[, [5.5, 7.0[, [7.0, 8.5[and [8.5, 10], respectively represented by 1 to 5.
- *RevSentences*: this feature is a by-product of the sentiment analysis of the review textual component. By recording the number of sentences, we can explore the existence of a possible relationship with the opinion or quantitative rating of the review.
- *RevSentimentStrength*: a numeric feature that reflects the polarity of the opinion (also known as sentiment analysis) based on the textual component of review. In the case of Booking.com, since it has two textual components, one for positive aspects and one for negative aspects, we concatenated both texts. Sentiment analysis, or opinion mining, is the computational study of people's opinions toward entities, individuals, events, topics, and their attributes. Sentiment analysis allows for the quantification of opinions according to their polarity (positive, negative, or neutral) (Liu and Zhang 2012). By assigning each review with a polarity value based on the textual component, it is possible to compare how users rate hotels in the textual component of reviews against what they rate in the quantitative component, therefore, obtaining two ratings for the same review. Prior to the execution of sentiment analysis, text preprocessing was performed. As

recognized by Han et al. (2016), text preprocessing is an arduous and time-consuming task, because it requires going back and forth while creating a document-term matrix (a document-term matrix is a common form for representing a collection of documents, where documents are assigned to rows, words are assigned to columns, and each cell populated with the frequency of the word in the document). This is even more difficult when it must be applied to three different languages. Text preprocessing consisted of the following steps:

- Transform all text to lowercase.
- Normalize related entities - transform words of similar meaning that appear in different formats in different languages to a consistent form. For example, “wi-fi” and “wi fi” were converted to “wifi”.
- Per language - perform stemming of common hospitality words like “rooms”, “restaurants”, and others that could be meaningful for data interpretation.
- Per language - normalize different spellings of the same words or expressions that could be written differently or could be misspelled. For example, in English, transform “didn’t” and “didnt” to “did not.”
- Per language - standardize domain-specific terms. For example, in English, “staff” is a common word used to describe hotel staff, but in Portuguese, numerous words like “equipa” (team), “pessoal” (personnel), “funcionários” (employees), or “colaboradores” (collaborators) are used. Other examples related to guest origin also had to be taken into consideration. Brazilian Portuguese has some differences from the European Portuguese language, and because Brazil is an important market in Portugal, terms from Brazilian Portuguese like “café da

manhã," "ônibus," or "metrô" had to be transformed to national equivalents, respectively, "pequeno-almoço," "autocarro," and "metro" (in English, "breakfast" "bus" and "metro").

- Removal of punctuation, numbers, and stop words (e.g. "a", "as", "at", "by", etc.).

After text preprocessing we then performed sentiment analysis to calculate the review sentiment strength. We adopted a dictionary-based approach, also known as a lexicon-based approach. Dictionaries are a collection of opinion words with a polarity classification (Ravi and Ravi 2015). Selection of dictionaries is an important methodological consideration (Han et al. 2016), with one essential aspect being its adequacy to the domain of the text, in this case, hospitality. Since we did not find hospitality dictionaries in any of the languages of this study, the criteria to choose dictionaries was based on relatively easy transformation, completeness (dictionaries had to have an extensive range of words), and openness (should be of general domain and broad). Based on these criteria, SentiLex-PT 02 sentiment lexicon (Silva, Carvalho, and Sarmento 2012) was chosen for Portuguese. For Spanish, the choice was the ElhPolar dictionary (Saralegi and San Vincente 2013). For English, the choice rested on the well-known Opinion Lexicon from Hu and Liu (2004). Sentiment strength was calculated by sentence, counting positive and negative words and then applying the same formula as used in Bjørkelund et al. (2012),

$$sentence\ sentiment\ strength = \frac{\sum positive\ words}{\sum positive\ words + \sum negative\ words}$$

which results in a value between 0 to 1, where 0 is perfectly negative and 1 is perfectly positive. Each review overall sentiment strength was calculated as the average of the reviews' sentences sentiment strength.

- *RevTotalWords*: as for *RevSentences* this feature is a byproduct of the sentiment analysis of the textual component the review. We kept a record of the number of words in the textual component to explore any links with the other features.
- *RevUserCountyISOCode*: from the location of the user of the review we extracted the name of the country and assigned its ISO 3166 two-digit country code.

The frequency and distribution of the resulting 34,622 observations in the final dataset, can be seen, per each categorical feature, per review source, in Table 3.

Table 3. Review frequency and distribution by source

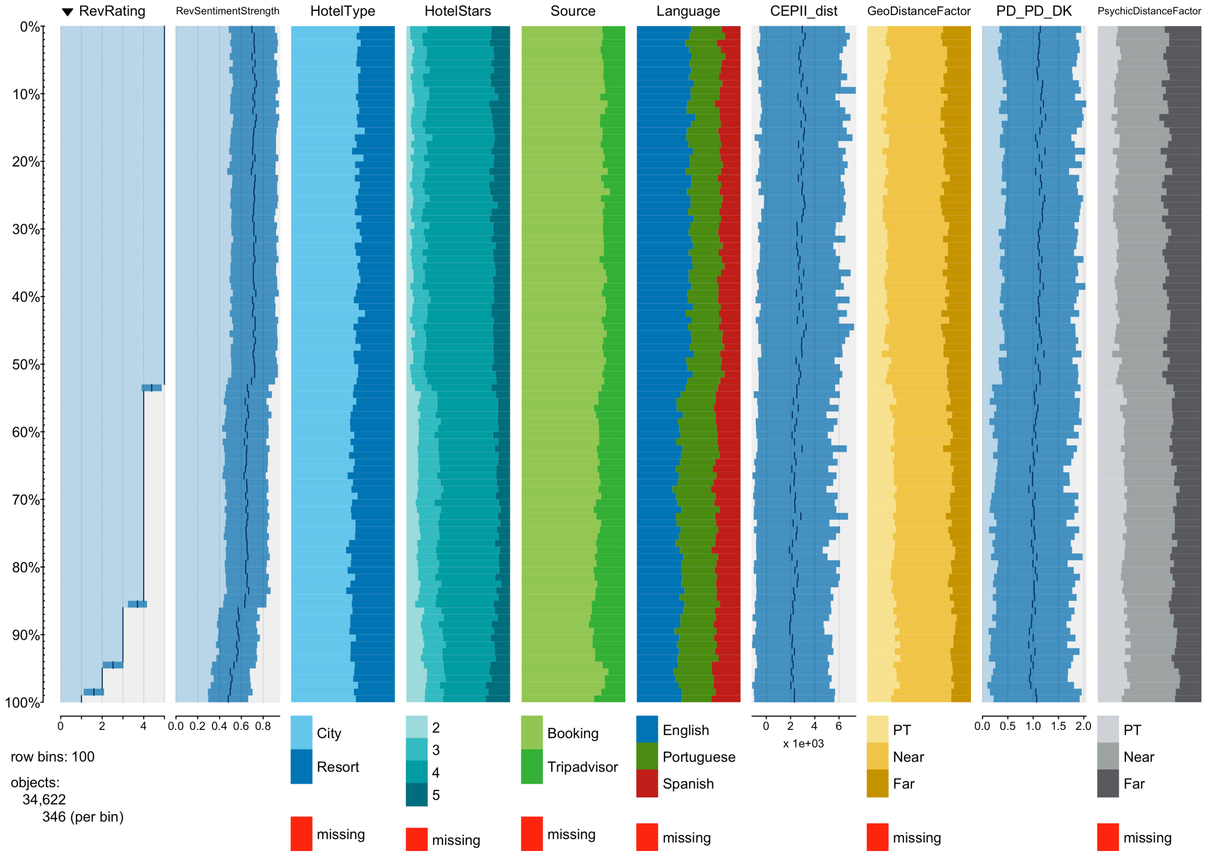
	Booking		**TripAdvisor**		**Total**	
	N	%	N	%	N	%
Hotel classification	**26,337**	**76.1**	**8,285**	**23.9**	**34,622**	**100.0**
Two stars	2,823	10.7	288	3.5	3,111	9.0
Three stars	4,635	17.6	835	10.1	5,470	15.8
Four stars	15,227	57.8	5,504	66.4	20,731	59.9
Five stars	3,652	13.9	1,658	20.0	5,310	15.3
Hotel type	**26,337**	**76.1**	**8,285**	**23.9**	**34,622**	**100.0**
City	17,925	68.1	3,404	41.1	21,329	61.6
Resort	8,412	31.9	4,881	58.9	13,293	38.4
Language	**26,337**	**76.1**	**8,285**	**23.9**	**34,622**	**100.0**
English	10,204	38.7	5,837	70.5	16,041	46.3
Portuguese	9,591	36.4	1,526	18.4	11,117	32.1
Spanish	6,542	24.8	922	11.1	7,464	21.6
Geographic distance	**26,337**	**76.1**	**8,285**	**23.9**	**34,622**	**100.0**
PT	6,355	24.1	1,197	14.4	7,552	21.8
Near	13,461	51.1	5,841	70.5	19,302	55.8
Far	6,521	24.8	1,247	15.1	7,768	22.4
Psychic distance	**26,337**	**76.1**	**8,285**	**23.9**	**34,622**	**100.0**
PT	6,355	24.1	1,197	14.4	7,552	21.8
Near	13,885	52.7	2,611	31.5	16,496	47.6
Far	6,097	23.1	4,477	54.0	10,574	30.5

DATA ANALYSIS AND RESULTS

Using a table plot, built with “tabplot”, an R package for visualization of large multivariate datasets (Tennekes and de Jonge 2017), we started by analyzing the distribution and looked for patterns in the dataset. This powerful visualization, as illustrated in Figure 2, shows each feature in a separated column and in each row, each bin aggregates a predefined number of observations of the dataset, in this case 100. Numeric features are represented in the form of bar charts and categorical, in the form of stacked bar charts.

This powerful visualization reveals in a glance patterns in data which indicate potential areas of interest. More than 50% of reviews have a *RevRating* of 5 and an average *RevSentimentStrength* above 0.7, which means the data is not normally distributed and is highly skewed. Figure 2 also shows that the sentiment strength (*RevSentimentStrength*) of the textual component of reviews is in line with the behavior of the review ratings (*RevRating*), because as one decreases the other decreases as well, but it also shows a similar pattern with geographic distance (*CEPII_dist*) and psychic distance (*PD_PD_DK*). This could indicate that less distant users, both in geographic and psychic distance, give lower ratings than more distant users. This visualization also illustrates that lower ratings (*RevRating* and *RevSentimentStrength*) occur more often in hotels of lower classification (2 and 3 stars in *HotelStars*) and when there is a lower number of reviews in English (*Language*).

Figure 2. Visualization of the full final dataset



Another interesting visualization that illustrates the skewness and the spread (degree of dispersion) of both review ratings (*RevRating* and *RevSentimentStrength*) by geographic and psychic distance factors, per hotel type and hotel star rating, is the set of boxplots presented in Figure 3. These boxplots show that although there are some similarities in the distribution of the quantitative ratings (*RevRating*) between geographic and psychic distances, this does not apply to the qualitative ratings (*RevSentimentStrength*). Qualitative rating, i.e. the sentiment strength of the textual component of reviews, does not follow the same patterns in terms of geographic and psychic distances, as the quantitative review ratings. This figure also shows that the distribution of both ratings differs by hotel type and hotels star ratings. These similarities and differences are detailed in Table 4 and Table 5, where frequency of reviews, as well as the mean and standard deviation for each combination by hotel type and hotel stars ratings, respectively per geographic and psychic distance, are shown.

Figure 3. Distribution of ratings by psychic and geographic distances, per hotel type and hotel stars

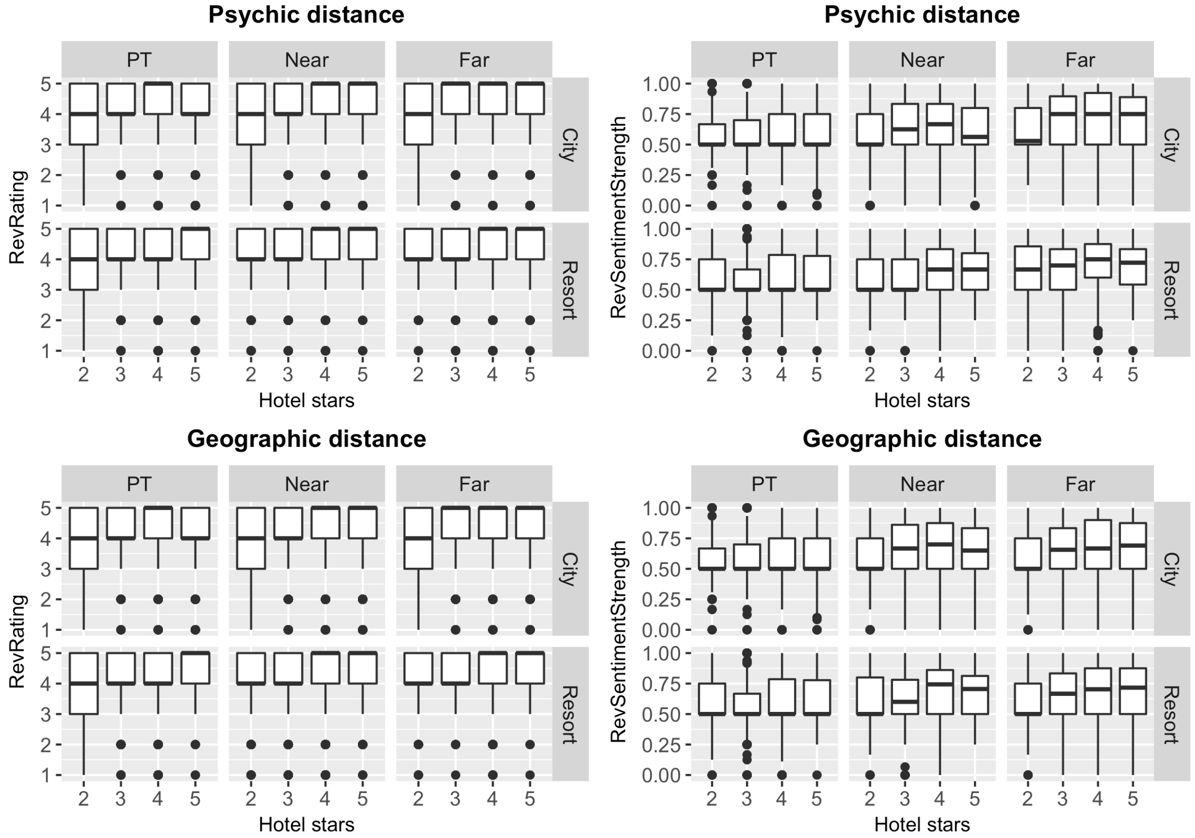


Table 4. Ratings statistics by geographic distance, hotel type and hotel stars

				RevRating		RevSentimentStrength	
Distance	Hotel Type	Hotel Stars	Frequency	Mean	Standard Deviation	Mean	Standard Deviation
PT	City	2	327	3.77676	1.03424	0.58779	0.20250
PT	City	3	646	4.10217	0.92112	0.58610	0.19350
PT	City	4	1,687	4.38234	0.81154	0.60024	0.21000
PT	City	5	956	4.12762	1.07539	0.58181	0.19650
PT	Resort	2	401	3.95012	0.95787	0.59943	0.20842
PT	Resort	3	850	4.02353	0.86656	0.58493	0.19822
PT	Resort	4	2,384	4.16233	0.94591	0.63436	0.21760
PT	Resort	5	301	4.60797	0.69700	0.64219	0.19470
Near	City	2	821	3.75761	1.09920	0.61570	0.21353
Near	City	3	1,846	4.27898	0.89240	0.68870	0.20930
Near	City	4	6,244	4.47213	0.77649	0.69620	0.21070
Near	City	5	1,989	4.15535	1.09485	0.66597	0.20894
Near	Resort	2	694	4.09366	0.93891	0.64361	0.21189
Near	Resort	3	895	4.09832	0.98838	0.64535	0.20346
Near	Resort	4	5,991	4.37423	0.85130	0.71107	0.19728
Near	Resort	5	822	4.57421	0.72788	0.68993	0.18074
Far	City	2	634	3.91956	1.03714	0.60112	0.22229
Far	City	3	1,177	4.34240	0.83555	0.66761	0.22239
Far	City	4	3,834	4.57303	0.73701	0.68872	0.22109
Far	City	5	1,168	4.48031	0.86617	0.69568	0.20951
Far	Resort	2	234	4.09829	0.98647	0.61607	0.22186
Far	Resort	3	56	4.12500	0.99201	0.66436	0.23211
Far	Resort	4	591	4.35871	0.92237	0.69947	0.21150
Far	Resort	5	74	4.43243	0.87712	0.67974	0.22309

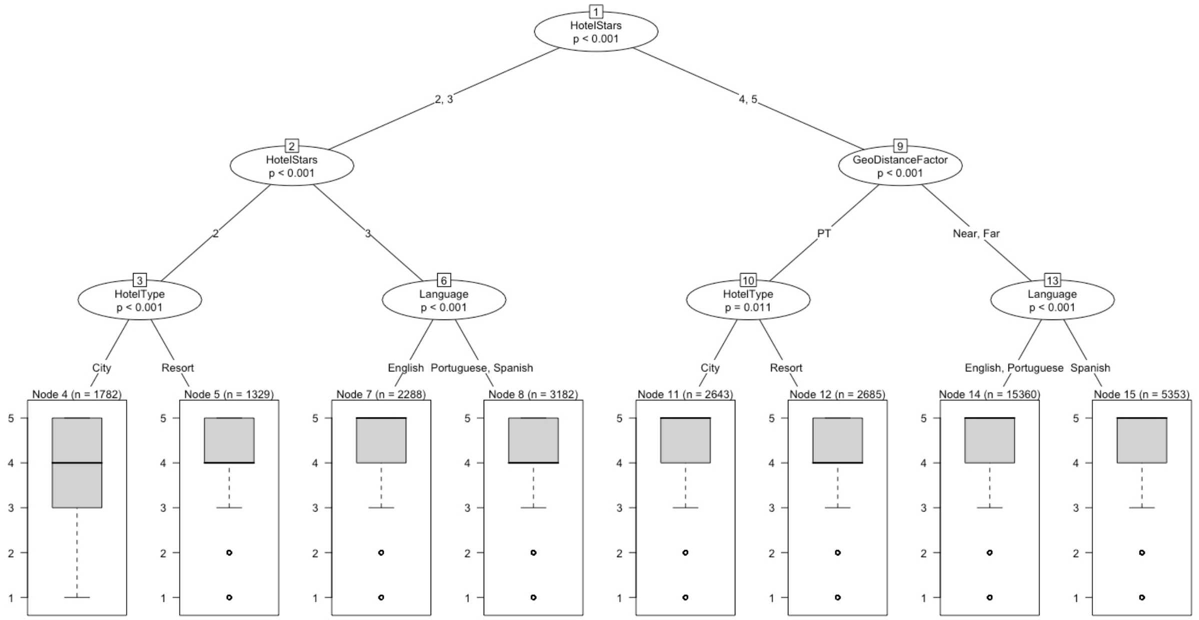
Table 5. Ratings statistics by psychic distance, hotel type and hotel stars

				RevRating		RevSentimentStrength	
Distance	Hotel Type	Hotel Stars	Frequency	Mean	Standard Deviation	Mean	Standard Deviation
PT	City	2	327	3.77676	1.03424	0.58779	0.20250
PT	City	3	646	4.10217	0.92112	0.58610	0.19350
PT	City	4	1,687	4.38234	0.81154	0.60024	0.21000
PT	City	5	956	4.12762	1.07539	0.58181	0.19650
PT	Resort	2	401	3.95012	0.95787	0.59943	0.20842
PT	Resort	3	850	4.02353	0.86656	0.58493	0.19822
PT	Resort	4	2,384	4.16233	0.94591	0.63436	0.21760
PT	Resort	5	301	4.60797	0.69700	0.64219	0.19470
Near	City	2	1,144	3.83392	1.06829	0.60188	0.21678
Near	City	3	1,962	4.26300	0.88856	0.66044	0.21691

				RevRating		RevSentimentStrength	
Distance	Hotel Type	Hotel Stars	Frequency	Mean	Standard Deviation	Mean	Standard Deviation
Near	City	4	6,949	4.48726	0.77695	0.67964	0.21490
Near	City	5	1,759	4.18533	1.06596	0.64555	0.20803
Near	Resort	2	748	4.06551	0.94550	0.62401	0.21129
Near	Resort	3	625	4.06400	1.00754	0.62273	0.20542
Near	Resort	4	2,936	4.33481	0.89555	0.67658	0.20885
Near	Resort	5	373	4.61394	0.67284	0.66799	0.18538
Far	City	2	311	3.80707	1.10194	0.63682	0.21797
Far	City	3	1,061	4.37889	0.83309	0.71756	0.20557
Far	City	4	3,129	4.56216	0.72940	0.72382	0.21123
Far	City	5	1,398	4.38913	0.96700	0.71649	0.20493
Far	Resort	2	180	4.21667	0.96460	0.68926	0.22107
Far	Resort	3	326	4.16871	0.94742	0.69197	0.19713
Far	Resort	4	3,646	4.40346	0.82511	0.73696	0.18566
Far	Resort	5	523	4.52581	0.78589	0.70414	0.18253

Analysis conducted with CTree, a conditional decision tree (Hothorn, Hornik, and Zeileis 2006) implemented with the R package “partykit” (Hothorn and Zeileis 2015) with the top three nodes predicting the value of RevRatings as depicted in Figure 4, shows that geographic distance is an important predictor of the quantitative review rating among four and five star hotels. This seems to be confirmation that some form of relationship exists between the geographic distance and review ratings. Figure 4 only shows three levels due to space constrains. CTree is a non-parametric class of regressions trees that embeds tree-structured regression models to the well-defined theory of conditional inference techniques. As CTree deals with overfitting and variable selection problems by inducing a recursive fitting procedure and application of appropriate statistical tests, on both variable selection and stopping, it is a good tool to explore the predictive importance of features in a determined outcome.

Figure 4. Conditional inference tree by top predictors of RevRating



We also applied a set of filter-based techniques to evaluate how each feature of the dataset was relevant in terms of the prediction of the *RevRating*. The objective of this test is to understand if geographic and psychic distances have predictive power over the quantitative rating of the review, which could indicate the importance of these features (see Table 6). The tests we applied, with the help of Microsoft Azure Machine Learning, were: Pearson correlation, Mutual information, Kendall correlation, Chi squared and Spearman correlation.

Since our dataset is not normally distributed, to compare means of review ratings by geographic and psychic distances, per hotel type and per hotels star ratings, we chose to employ the Kruskal-Wallis (Kruskal and Wallis 1952), which is considered to be the non-parametric equivalent of the one-way ANOVA. With the Kruskal-Wallis results presenting values below the defined threshold (that we defined as 0.05) this indicates that rating mean values differ in each category of analyzed features. In these instances, it is necessary to conduct a posthoc analysis by the categories of each feature in the study. For this posthoc analysis we employed the R package "pgirmess" (Giraudoux 2016).

Table 6. Filter-based feature selection Results

Method	Pearson		Mutual Information		Kendall		Chi Squared		Spearman		Rank Mean	Rank Median
	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score		
GeoDistanceFactor	7	0.09712	7	0.00619	4	0.09845	7	428.58	4	0.10880	5.8	7
HotelID	11	0.03552	2	0.01583	11	0.02066	2	1084.97	11	0.02611	7.4	11
HotelStars	3	0.12674	3	0.01446	3	0.12019	3	994.31	3	0.13315	3	3
HotelType	10	0.03778	11	0.00150	10	0.04742	11	103.47	10	0.04992	10.4	10
Language	8	0.08724	9	0.00550	6	0.08845	9	380.28	7	0.09809	7.8	8
PsychicDistanceFactor	5	0.09850	8	0.00615	5	0.09729	8	426.01	5	0.10802	6.2	5
RevSentences	4	0.10089	6	0.00788	7	0.22339	6	534.19	6	0.09877	5.8	6
RevTotalWords	6	0.09789	5	0.00871	9	0.06326	5	621.48	8	0.08031	6.6	6
RevUserCountryISOCode	2	0.14515	4	0.01409	2	0.12709	4	980.57	2	0.15366	2.8	2
RevSentimentStrength	1	0.28980	1	0.03928	1	0.22339	1	2556.78	1	0.26567	1	1
Source	9	0.05230	10	0.00259	8	0.06475	10	179.79	9	0.06817	9.2	9

The result of Kruskal-Wallis test is used to evaluate if the means of *RevRating* and *RevSentimentStrength* differ by each of the features in the scope of the study (geographic and psychic distances, per hotel types and hotel star ratings) and is presented in Table 7. Since p-values for all categorical features presented values below 0.05, means do differ by categories for each feature. In other words, with respect to the two research questions, *RevRating* and *RevSentimentStrength* distributions differ by hotel type, hotel star ratings, geographic distance and psychic distance, which mean that users from different geographic and psychic distances rate hotels differently according to the hotel type and hotel stars.

Table 7. Kruskal-Wallis test results

	RevRating		RevSentimentStrength	
	p-value	KW statistic	p-value	KW statistic
HotelType	1.5604663-20	86.28169	9.370848e-06	19.63557
HotelStars	6.035435e-226	1,043.67400	3.187812e-86	399.26901
GeographicDistanceFactor	2.999200-92	421.47897	9.319089e-199	911.96474
PsychicDistanceFactor	3.220689e-92	421.33647	7.205815e-316	1,451.28400

As the Kruskall-Wallis test revealed that there were differences in the mean values of the categories in this study, a posthoc analysis was performed to determine which categories possess different means. This analysis is achieved by pairwise comparison for each combination of categories. Results of this test are presented in Table 8 and Table 9. In this test when the observed differences are higher than the critical value considered as significant (we opted for 0.05), we identify a difference between the categories.

Table 8. Geographic distance Kruskal-Wallis pairwise comparison

			RevRating				RevSentimentStrength			
Distance pair	Hotel Type	Measure	2 stars	3 stars	4 stars	5 stars	2 stars	3 stars	4 stars	5 stars
PT - Near	City	O.Dif.	1.632286	218.58573	377.5145	52.6228	68.86610	512.30538	1470.3708	484.4130
		C.Dif.	80.55358	115.92541	223.1080	111.8762	80.55358	115.92541	223.1080	111.8762
		Difference	FALSE	**TRUE**	**TRUE**	FALSE	FALSE	**TRUE**	**TRUE**	**TRUE**
	Resort	O.Dif.	60.99697	72.71656	586.47342	11.34392	78.58512	172.36841	987.3956	96.15210
		C.Dif.	57.63328	59.62671	150.0522	55.7540	57.63328	59.62671	150.0522	55.7540
		Difference	**TRUE**	**TRUE**	**TRUE**	FALSE	**TRUE**	**TRUE**	**TRUE**	**TRUE**
PT – Far	City	O.Dif.	73.861811	277.98615	838.4594	381.7184	26.51697	414.76299	1355.4141	643.1468
		C.Dif.	83.86921	124.17255	237.5557	123.9847	83.86921	124.17255	237.5557	123.9847
		Difference	FALSE	**TRUE**	**TRUE**	**TRUE**	FALSE	**TRUE**	**TRUE**	**TRUE**
	Resort	O.Dif.	70.44365	92.32849	637.21634	55.57726	37.75505	208.77441	810.1646	82.09789
		C.Dif.	75.58314	171.76107	284.7406	107.3796	75.58314	171.76107	284.7406	107.3796
		Difference	FALSE	FALSE	**TRUE**	FALSE	FALSE	**TRUE**	**TRUE**	FALSE
Near - Far	City	O.Dif.	72.229525	59.40042	456.9449	329.0956	42.34913	97.54239	114.9567	158.7338
		C.Dif.	65.12904	94.59142	166.8284	104.7948	65.12904	94.59142	166.8284	104.7948
		Difference	**TRUE**	FALSE	**TRUE**	**TRUE**	FALSE	**TRUE**	FALSE	**TRUE**
	Resort	O.Dif.	9.44668	19.61193	50.74292	44.23335	40.83006	36.40601	177.2310	14.05420
		C.Dif.	69.45516	171.49396	267.1704	100.4402	69.45516	171.49396	267.1704	100.4402
		Difference	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table 9. Psychic distance Kruskal-Wallis pairwise comparison

			RevRating				RevSentimentStrength			
Distance pairs	Hotel Type	Measure	2 stars	3 stars	4 stars	5 stars	2 stars	3 stars	4 stars	5 stars
PT - Near	City	O.Dif.	35.35794	194.2310	455.4339	75.29413	33.41208	378.776	1220.0371	367.2979
		C.Dif.	77.24650	115.03362	220.6876	114.2259	77.24650	115.03362	220.6876	114.2259
		Diference	FALSE	**TRUE**	**TRUE**	FALSE	FALSE	**TRUE**	**TRUE**	**TRUE**
	Resort	O.Dif.	49.20347	56.04866	510.5273	3.148166	45.59563	105.8142	523.9990	51.98803
		C.Dif.	56.86633	65.60103	170.8352	64.12061	56.86633	65.60103	170.8352	64.12061
		Difference	FALSE	FALSE	**TRUE**	FALSE	FALSE	**TRUE**	**TRUE**	FALSE
PT – Far	City	O.Dif.	24.81997	329.5172	764.3681	299.04971	112.94986	651.021	1885.3835	764.3891
		C.Dif.	97.56985	126.55507	245.5971	119.3059	97.56985	126.55507	245.5971	119.3059
		Difference	FALSE	**TRUE**	**TRUE**	**TRUE**	**TRUE**	**TRUE**	**TRUE**	**TRUE**
	Resort	O.Dif.	122.28617	108.04080	655.8555	23.447711	162.59565	306.2184	1331.8248	125.66106
		C.Dif.	82.43237	81.10535	163.2112	59.87353	82.43237	81.10535	163.2112	59.87353
		Difference	**TRUE**	**TRUE**	**TRUE**	FALSE	**TRUE**	**TRUE**	**TRUE**	**TRUE**
Near - Far	City	O.Dif.	10.53797	135.2862	308.9342	223.75559	79.53778	272.245	665.3104	397.0913
		C.Dif.	78.77668	96.63815	175.0506	101.8573	78.77668	96.63815	175.0506	101.8573
		Difference	FALSE	**TRUE**	**TRUE**	**TRUE**	**TRUE**	**TRUE**	**TRUE**	**TRUE**
	Resort	O.Dif.	73.08269	51.99214	145.3282	20.299545	117.00001	200.4042	807.8258	73.67303
		C.Dif.	76.27902	85.05612	153.6545	56.08595	76.27902	85.05612	153.6545	56.08595
		Difference	FALSE	FALSE	FALSE	FALSE	**TRUE**	**TRUE**	**TRUE**	**TRUE**

Unlike the quantitative review rating, the textual component allows users to fully expose their opinions. In other words, although a user gives a hotel a top rating (5 in TripAdvisor quantitative scale), in the text the user can express views differently (e.g. "Excellent hotel. Staff very helpful. Very good breakfast. The only downside is that it is slightly away from the center"), which is a sentence that is not fully positive. Therefore, as expected, the results presented in Table 8 show that there is a difference between the results of the quantitative rating (*RevRating*) and the sentiment of the textual component (*RevSentimentStrength*) (Antonio et al. 2018b). Nevertheless, results show that from the 48 combinations of categories, results only differ in seven of the combinations. This illustrates the correlation between review ratings and sentiment polarity of the textual component of reviews (Antonio et al. 2018a). This correlation is also illustrated in Table 9, with only nine of the combinations out of the 48 presenting different results.

Table 8 also shows that out of the 16 combinations of hotel type and hotel stars for each geographic category, 12 present different distributions for "PT – Near" distance category, nine for "PT – Far" and five for "Near – Far". The results suggest users from Portugal tend to have a very different opinion to users from a "Near" distance, but not so different to users from "Far". Opinions of users from "Near" do not differ much from users from "Far". However, for psychic distance, as presented in Table 9, results differ even more between combinations. From the 16 combinations of hotel type and hotel stars for each psychic category, eight present different distributions for "PT – Near", 14 for "PT – Far" and 11 for "Near – Far". The observations illustrate that the further away a user is in terms of psychic distance, the less similarity there is in ratings, independently of the hotel type and star rating.

CONCLUSION

This study reinforces the importance of both psychic and geographic distance as an influence of hotel online reviews, and provides theoretical and managerial guidance. We seek to draw upon a multi-disciplinary social science approach by incorporating strands of prior literature from international management, international marketing and tourism. Tourism is an integral aspect of contemporary society and is an area of interest across the social sciences (Holden 2004). This research considers some important gaps in the literature by strengthening understanding of the online hotel ratings during times of significant demand shifts (Wong, Fong, and Law 2016). In this article, we specifically ask, in terms of hotels, if distance is a factor, then to what extent does language of the review, hotel location and hotel star rating matter?

Based on the influence of preferences with respect to hotel attributes, the results reveal dissimilarities among travelers based on geographic and psychic distances. This is in agreement with prior research (Banarjee and Chua 2016). For hoteliers with a significant number of foreign guests this is worthy of further investigations, and is rather pertinent for heterogeneous groups of travelers. Social identity theory (Tajfel 1982), refers to social identity within communities. Hotels need to identify social groups within their customer database. With the prevalence of digital transformation, even the smallest hotel has to act. The marketing processes need to keep abreast of external shifts, customer expectations and from the employee perspective. To remain competitive hotels cannot allow the likes and dislikes of guests and communities to remain unknown. This study provides a platform that illuminates why aggregated evaluative online hotel ratings statements, from guests need to be disaggregated for effective marketing decision-making. With the advance and growing importance of personalization to the hotel sector (Buhalis and Amaranggana 2015), identifying patterns in terms of geographic and psychic distance will provide fresh knowledge.

The application of big data represents a new era for data exploration and utilization, which can be a driver for innovative processes in customer facing practices. Tourism is not an exception and the results from this study can provide opportunities to enrich marketing processes by delving into influence of traveler distance and language. For this to be effective hoteliers will need to demonstrate higher levels of analytical, interpretive and strategic knowledge (Phillips and Moutinho 2014). As research into the opportunities available through the use of big data in hotels remain nascent, this study provides theoretical and empirical evidence of fresh insights that can be obtained.

THEORETICAL AND MANAGERIAL IMPLICATIONS

The present study extends extant research in three important ways. First, this paper provides new insights into how geographic and psychic distance influence online hotel ratings. Even though the concept of distance occupies a central role in business and management literature, tourism research to-date has not delved into the influence of the origin of travelers in their online rating behavior. In general, prior academic studies aggregate reviews from individual travelers of differing origins to compute sentiment scores of simple average ratings, which are also written in different languages. Aggregated hotel ratings may not provide the full story, as guest opinions may be buried and lost.

The choice of concept of distance needs to go beyond geographic distance (Deodhar, Subramani, and Zaheer 2017), and this study deploys a set of psychic measures, which are one of the most popular forms of distance (Safari, Thilenius, and Hadjikhani 2013). Online travelers' preferences have been investigated from many facets, but a paucity of prior studies focus upon traditional distance metrics. Gao et al. (2018) analyzed the relationship between online ratings (quantitative review ratings) and power distance (a metric distance different from this study). Our study distinguishes from Gao et al. (2018), by using two distance measures (geographic and psychic) and incorporating sentiment polarity of the qualitative component of reviews**.**

Second, the results of Kruskal-Wallis test illustrate the difference in the means of *Rerating* and *RevSentimentStrength* for each of the features in the scope of the study (geographic and psychic distances, per hotel types and hotel star rating). By using original data of 34,622 online customer reviews written in English, Portuguese and Spanish, we also confirm that online customer reviews are multifaceted constructs. By investigating the contemporary relevance of distance, whether psychic or geographic, our results reveal that both types of distance matter differently to hotels in terms of language, location and star rating. Travelers' rating patterns by language vary across hotel profiles too. We observe that less distance users both in terms of geographic and psychic distance give lower scores than more distance users. Figure 2 illustrates that lower ratings (*RevRating* and *RevSentimentStrength*) occur more often in hotels of lower classification (2 and 3 stars in *HotelStars*) and when there is a lower number of reviews in English (*Language*).

Third, in light of these gaps and concerns spotted in the literature, this study provides theoretical and managerial guidance for future research. Travelers of different origins may possess significantly different expectations. So, by understanding changes in online customer reviews beyond those written in English, may help advance a more effective connectivity among the traveler base. With the already high English penetration rates on the Internet, future growth will come from non-English languages. English is the most common language on the Internet (25.4%), but Chinese is already 19.3% ("Top Ten Internet Languages in the World - Internet Statistics (2017)" 2019). So, further research could develop and validate the influence of distances and alternative languages in differing tourism contexts. The study provides a platform for further exploration of geographic and psychic distance. The findings of this study provide a starting point to design a more focused investigation. We analyzed hotels in a Portuguese setting and suggest future work analyzing hotels in other countries. Each country may impact both geographic and psychic distance differently. This would strengthen the generalization of our results.

Moreover, as tourism has matured, increasing numbers of academics and practitioners' attention is drawn to developing creative ways for firms to enhance distinctiveness in their offer. To enhance the service offer, managers need to closely monitor customer voice (Phillips et al. 2016). This study contributes to contemporary research on online hotel ratings by incorporating distance and language. In fact, after a near decade of eWOW research, there have not been a comprehensive analysis of how differences of origin influence online hotel ratings. Distance influences not only travelers' assessment and choice of destination, but also the activities selected during their stay. This illustrates the potential for distance to be used as a segmentation variable (Nyaupane and Graefe 2008). The relationship between language and the Internet is not unimportant, but rather neglected in hospitality and tourism research (Schuckert, Liu, and Law

2015; Yong et al. 2017). So what causes this? Well distance compression may be making countries less distinct over time, and easier for travelers from greater distances. The promotional material received by such travelers may be making the hotel and its location more attractive. We content that this holistic distance results in travelers being more discerning as they possess improved availability of information together with increased knowledge of hotel experience. The uncertainty in greater psychic distance appears to make travelers less critical of the hotel experience. In this instance, the hotel may appear more attractive and the associated network relationships determine the impact (Ojala 2015).

In terms of practical implications, hotels need to listen to travelers, as they form an invaluable resource and are part of the brand strategy. However, it is critical to have effective processes in place to make the necessary operational and service improvements. This will enhance the level of the traveler experience. Tracking what travelers are saying requires the firm to develop a sound management of online customer reviews. By tapping into rich bespoke datasets firms can ascertain their strengths and weaknesses and make better quality decisions. Hotels should understand the impact of geographic and psychic distance when evaluating the customer journey. The advent of technology makes it possible to design bespoke customer journey strategies for differing customer segments beyond traditional demographics such as purpose of visit and age.

The results of the study suggest that local travelers tend to be more critical than travelers from a distance, which suggests that incentives could be made available to local travelers. In these instances, understanding the motivation of the trip and behavioral approaches of key segments will create a platform for better managing the salient information flows between the traveler and hotel. Our results highlight that when travelers perceive a gap and this is beyond an acceptable level, this will lead to dissatisfaction, which should be avoided.

Collectively this demonstrates how “the sum of factors” or the “differences” interact in the formation of asymmetric distance perceptions. The individual experience of travelers are influences by psychic distance and this will impact hotel marketing strategies and promotional activity, product development, and pricing strategies. Future research may delve into the moderating effects of distance and hotel online reviews, such as the size of the host country in GDP terms, attractiveness of host country, historical events, hotel entry modes.

LIMITATIONS

As with any other study, this research possesses limitations. Our research employs a sampling frame of reviews and associated ratings on Booking.com and TripAdvisor for Portuguese hotels in three different languages (Portuguese, Spanish and English). So, our findings may not be generalizable to other hotel markets and languages. However, the development of any new seam of research needs to be repositioned in terms of an over focus on the theoretical aspects, such as the rigor-relevance debate. Prior research in social sciences, have argued for a slight shift in abstract philosophical debate around research epistemologies. The consequence is very significant illustrating a lack of what (Ven 2007) outlined as engaged scholarship which has both rigor and relevance. With more than three quarters of traveler purchasers visiting TripAdvisor prior to making a booking, its influence and significance makes the platform useful for academic research.

Another limitation, as identified by Antonio et al. (2018a), relates to the small number of users that do not write reviews in the official language of their country. For example, such users will tend to write reviews in English rather than in their mother tongue. Therefore, the analysis of ratings or sentiment polarity based on the language of the review could not reflect the cultural background of all users communicated their review.

Another difficulty relates to the difficulty of performing text analysis across multiple languages, it was decided to use only reviews in English, Spanish, and Portuguese. Although reviews in these languages represent 70% of Portugal's tourists official languages, they are not representative of all tourists. Therefore, future research could explore the analysis of reviews in other languages.

We also recognize that as every language has a different degree of expressive power (Ravi and Ravi 2015), it is possible that some differences in the sentiment strength exist due to the differences in the dictionaries employed per language. Future research should explore the analysis of sentiment with dictionary-free approaches or with domain-specific dictionaries.

REFERENCES

Abbott, Dean. 2014. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. Indianapolis, IN, USA: Wiley.

Alaei, Ali Reza, Susanne Becken, and Bela Stantic. 2017. “Sentiment Analysis in Tourism: Capitalizing on Big Data.” Journal of Travel Research, December, 0047287517747753. doi:10.1177/0047287517747753.

Andriotis, Konstantinos. 2011. “A Comparative Study of Visitors to Urban, Coastal and Rural Areas. Evidence from the Island of Crete.” European Journal of Tourism Research 4: 93–108.

Ankomah, Paul K., John L. Crompon, and Dwayne Baker. 1996. “Influence of Cognitive Distance in Vacation Choice.” Annals of Tourism Research 23 (1): 138–50. doi:10.1016/0160-7383(95)00054-2.

Antonio, Nuno, Ana de Almeida, Luis Nunes, Fernando Batista, and Ricardo Ribeiro. 2018a. “Hotel Online Reviews: Different Languages, Different Opinions.” Information Technology & Tourism 18 (1–4): 157–85. doi:10.1007/s40558-018-0107-x.

———. 2018b. “Hotel Online Reviews: Creating a Multi-Source Aggregated Index.” International Journal of Contemporary Hospitality Management 30 (12): 3574–91. doi:10.1108/IJCHM-05-2017-0302.

Assaf, A. George, Alexander Josiassen, and Frank W. Agbola. 2015. “Attracting International Hotels: Locational Factors That Matter Most.” Tourism Management 47 (April): 329–40. doi:10.1016/j.tourman.2014.10.005.

Assarut, Nuttapol, and Patnaree Srisuphaolarn. 2018. “Applying Psychic Distance to Services Internationalization: A Case Study of Thai Caregivers and Japanese Elderly.” Journal of Asia-Pacific Business 19 (4): 228–45. doi:10.1080/10599231.2018.1525247.

Avloniti, Anthi, and Fragkiskkos Filipppaios. 2014. “Unbundling the Differences between Psychic and Cultural Distance: An Empirical Examination of the Existing Measures.” International Business Review 23 (3): 660–74. doi:10.1016/j.ibusrev.2013.11.007.

Baden-Fuller, Charles, and John M. Stopford. 1994. Rejuvenating the Mature Business : The Competitive Challenge. (Revised ed.). Boston, Massachusetts: Harvard Business School Press.

Baka, Vasiliki. 2016. “The Becoming of User-Generated Reviews: Looking at the Past to Understand the Future of Managing Reputation in the Travel Sector.” Tourism Management 53 (April): 148–62. doi:10.1016/j.tourman.2015.09.004.

Banarjee, Sneshasish, and Alton Y. K. Chua. 2016. “In Search of Patterns among Travellers’ Hotel Ratings in TripAdvisor.” Tourism Management 53 (April): 125–31. doi:10.1016/j.tourman.2015.09.020.

Batrinca, Bogdan, and Philip C. Treleaven. 2015. “Social Media Analytics: A Survey of Techniques, Tools and Platforms.” AI & SOCIETY 30 (1): 89–116. doi:10.1007/s00146-014-0549-4.

Beckerman, Wilfred. 1956. “Distance and the Pattern of Intra-European Trade.” The Review of Economics and Statistics 38 (February): 31–40.

Berry, Heather, Mauro F. Guillén, and Nan Zhou. 2010. “An Institutional Approach to Cross-National Distance.” Journal of International Business Studies 41 (9): 1460–80. doi:10.1057/jibs.2010.28.

Bi, Juan, and Xinran Y. Lehto. 2018. “Impact of Cultural Distance on International Destination Choices: The Case of Chinese Outbound Travelers.” International Journal of Tourism Research 20 (1): 50–59. doi:10.1002/jtr.2152.

Bjørkelund, Eivind, Thomas H. Burnett, and Kjetil Nørvag. 2012. “A Study of Opinion Mining and Visualization of Hotel Reviews.” In Proceedings of the 14th International Conference on Information Integration and Web-Based Applications & Services, 229–238. New York, NY: ACM. http://dl.acm.org/citation.cfm?id=2428773.

Blum, Bernardo S., and Avi Goldfarb. 2006. “Does the Internet Defy the Law of Gravity?” Journal of International Economics 70 (2): 384–405. doi:10.1016/j.jinteco.2005.10.002.

Bosch, Olav ten. 2017. “An Introduction to Web Scraping, IT and Legal Aspects.” European Comission. https://circabc.europa.eu/webdav/CircaBC/ESTAT/ESTP/Library/2017%20ESTP%20PROGRAMME/45.%20Automated%20collection%20of%20online%20prices_%20sources%2C%20tools%20and%20methodological%20aspects%2C%2023%20%E2%80%93%2026%20October%202017%20-%20Organiser_%20EXPERTISE%20FRANCE/20170919%20ESTP%20Prices%20-%206%20-%20Introduction%20IT%20and%20Legal.pdf.

Braun, Michael T., Goran Kuljanin, and Richard P. DeShon. 2018. “Special Considerations for the Acquisition and Wrangling of Big Data.” Organizational Research Methods 21 (3): 633–59. doi:10.1177/1094428117690235.

Brewer, Paul A. 2007. “Operationalizing Psychic Distance: A Revised Approach.” Journal of International Marketing 15 (1): 44–66. doi:10.1509/jimk.15.1.044.

Brock, Jurgen Kai-Uwe, Jeffrey E. Johnson, and Josephine Yu Zhou. 2011. “Does Distance Matter for Internationally-Oriented Small Firms?” Industrial Marketing Management 40 (3): 384–94. doi:10.1016/j.indmarman.2010.08.007.

Buhalis, Dimitrios, and Aditya Amaranggana. 2015. “Smart Tourism Destinations Enhancing Tourism Experience through Personalisation of Services.” In Information and Communication Technologies in Tourism 2015, 377–89. Springer, Cham. doi:10.1007/978-3-319-14343-9_28.

Cai, Liping A., and Mimi Li. 2009. “Distance-Segmented Rural Tourists.” Journal of Travel & Tourism Marketing 26 (8): 751–61. doi:10.1080/10548400903356137.

Cantallops, Antonio S., and Fabiana Salvi. 2014. “New Consumer Behavior: A Review of Research on EWOM and Hotels.” International Journal of Hospitality Management 36: 41–51. doi:10.1016/j.ijhm.2013.08.007.

Child, John, Sek Hong Ng, and Christine Wong. 2002. “Psychic Distance and Internationalization: Evidence from Hong Kong Firms.” International Studies of Management & Organization 32 (1): 36–56.

Choi, Jeongho, and Farok J. Contractor. 2016. “Choosing an Appropriate Alliance Governance Mode: The Role of Institutional, Cultural and Geographical Distance in International Research &amp; Development (R&amp;D) Collaborations.” Journal of International Business Studies 47 (2): 210–32. doi:10.1057/jibs.2015.28.

Choi, Youngjoon, Benjamin Hickerson, and Deborah Kerstetter. 2018. “Understanding the Sources of Online Travel Information.” Journal of Travel Research 57 (1): 116–28. doi:10.1177/0047287516683833.

Conway, Tony, and Jonathan S. Swift. 2000. “International Relationship Marketing - The Importance of Psychic Distance.” European Journal of Marketing 34 (11/12): 1391–1414. doi:10.1108/03090560010348641.

Crompton, John L. 1979. “An Assessment of the Image of Mexico as a Vacation Destination and the Influence of Geographical Location upon That Image.” Journal of Travel Research 17 (4): 18–23. doi:10.1177/004728757901700404.

Cuypers, Ilya R. P., Gokhan Ertug, and Jean-François Hennart. 2015. “The Effects of Linguistic Distance and Lingua Franca Proficiency on the Stake Taken by Acquirers in Cross-Border Acquisitions.” Journal of International Business Studies 46 (4): 429–42. doi:10.1057/jibs.2014.71.

Deodhar, Swanand J., Mani Subramani, and Akbar Zaheer. 2017. “Geography of Online Network Ties: A Predictive Modelling Approach.” Decision Support Systems 99 (July): 9–17. doi:10.1016/j.dss.2017.05.010.

Dougherty, James, Ron Kohavi, and Mehran Sahami. 1995. “Supervised and Unsupervised Discretization of Continuous Features.” In Proceedings of the 12th International Conference on Machine Learning, 194–202. San Francisco, CA, USA. http://robotics.stanford.edu/users/sahami/papers-dir/disc.pdf.

Dow, Douglas, and Amal Karunaratna. 2006. “Developing a Multidimensional Instrument to Measure Psychic Distance Stimuli.” Journal of International Business Studies 37 (5): 578–602.

Duan, Wenjing, Yang Yu, Qing Cao, and Stuart Levy. 2016. “Exploring the Impact of Social Media on Hotel Service Performance: A Sentimental Analysis Approach.” Cornell Hospitality Quarterly 57 (3): 282–96. doi:10.1177/1938965515620483.

Durand, Aurélia, Ekaterina Turkina, and Matthew Robson. 2016. “Psychic Disistance and Country Image in Exporter–Importer Relationships.” Journal of International Marketing 24 (3): 31–57. doi:10.1509/jim.15.0056.

Duverger, Philippe. 2013. “Curvilinear Effects of User-Generated Content on Hotels’ Market Share: A Dynamic Panel-Data Analysis.” Journal of Travel Research 52 (4): 465–78. doi:10.1177/0047287513478498.

Ferrer-Rosell, Berta, Eva Martin-Fuentes, and Estela Marine-Roig. 2019. “Do Hotels Talk on Facebook about Themselves or about Their Destinations?” In Information and

Communication Technologies in Tourism 2019, edited by Juho Pesonen and Julia Neidhardt, 344–56. Springer International Publishing.

Floyd, Kristopher, Ryan Freling, Saad Alhoqail, Hyun Young Cho, and Traci Freling. 2014. "How Online Product Reviews Affect Retail Sales: A Meta-Analysis." Journal of Retailing 90 (2): 217–32. doi:10.1016/j.jretai.2014.04.004.

Fujita, Kentaro, Yaacov Trope, Nira Liberman, and Maya Levin-Sagi. 2006. "Construal Levels and Self-Control." Journal of Personality and Social Psychology 90 (3): 351–67. doi:10.1037/0022-3514.90.3.351.

Gao, Baojun, Xiangge Li, Shan Liu, and Debin Fang. 2018. "How Power Distance Affects Online Hotel Ratings: The Positive Moderating Roles of Hotel Chain and Reviewers' Travel Experience." Tourism Management 65 (April): 176–86. doi:10.1016/j.tourman.2017.10.007.

Ge, Jing, Marisol Vazquez, and Ulrike Gretzel. 2018. "Sentiment Analysis: A Review." In Advances in Social Media for Travel, Tourism and Hospitality: New Perspectives, Practice and Cases, edited by Marianna Sigala and Ulrike Gretzel, 243–61. New York, NY: Routledge.

Geetha, M., Pratap Singha, and Sumedha Sinha. 2017. "Relationship between Customer Sentiment and Online Customer Ratings for Hotels - An Empirical Analysis." Tourism Management 61 (August): 43–54. doi:10.1016/j.tourman.2016.12.022.

Geraci, Vincent J., and Wilfried Prewo. 1977. "Bilateral Trade Flows and Transport Costs." The Review of Economics and Statistics 59 (1): 67–74. doi:10.2307/1924905.

Ghemawat, Pankaj. 2001. "Distance Still Matters. The Hard Reality of Global Expansion." Harvard Business Review 79 (8): 137–40, 142–47, 162.

Giraudoux, Patrick. 2016. Pgirmess: Data Analysis in Ecology (version 1.6.5). R package. https://cran.r-project.org/package=pgirmess.

Godes, David, and Dina Mayzlin. 2004. "Using Online Conversations to Study Word-of-Mouth Communication." Marketing Science 23 (4): 545–60. doi:10.1287/mksc.1040.0071.

Goethals, Patrick. 2016. "Multilingualism and International Tourism: A Content- and Discourse-Based Approach to Language-Related Judgments in Web 2.0 Hotel Reviews." Language and Intercultural Communication 16 (2): 235–53. doi:10.1080/14708477.2015.1103249.

Hallmann, Kirstin, Anita Zehrer, and Sabine Müller. 2015. "Perceived Destination Image: An Image Model for a Winter Sports Destination and Its Effect on Intention to Revisit." Journal of Travel Research 54 (1): 94–106. doi:10.1177/0047287513513161.

Han, Hyun J., Shawn Mankad, Nagesh Gavirneni, and Rohit Verma. 2016. "What Guests Really Think of Your Hotel: Text Analytics of Online Customer Reviews." Cornell Hospitality Report 16 (2): 3–17.

Hensens, Wouter. 2015. "The Future of Hotel Rating." Journal of Tourism Futures 1 (1): 69–73. doi:10.1108/JTF-12-2014-0023.

Hernández-Ortega, Blanca. 2018. “Don’t Believe Strangers: Online Consumer Reviews and the Role of Social Psychological Distance.” Information & Management 55 (1): 31–50. doi:10.1016/j.im.2017.03.007.

Hofstede, Geert. 1980. “Motivation, Leadership, and Organization: Do American Theories Apply Abroad?” Organizational Dynamics 9 (1): 42–63. doi:10.1016/0090-2616(80)90013-3.

Holden, Andrew. 2004. Tourism Studies and the Social Sciences. 1st ed. Routledge. doi:10.4324/9780203502396.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. “Unbiased Recursive Partitioning: A Conditional Inference Framework.” Journal of Computational and Graphical Statistics 15 (3): 651–74.

Hothorn, Torsten, and Achim Zeileis. 2015. “Partykit: A Modular Toolkit for Recursive Partytioning in R.” Journal of Machine Learning Research 16: 3905–9.

House, Robert J., Paul J. Hanges, Mansour Javidan, Peter W. Dorfman, and Vipin Gupta, eds. 2004. Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies. 1 edition. Thousand Oaks, Calif: SAGE Publications, Inc.

Hu, Minqing, and Bing Liu. 2004. “Mining and Summarizing Customer Reviews.” In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, edited by W Kim and R Kohavi, 168–177. New York, NY: ACM. http://dl.acm.org/citation.cfm?id=1014073.

Hutzschenreuter, Thomas, Kleindienst, and Sandra Lange. 2014. “Added Psychic Distance Stimuli and MNE Performance: Performance Effects of Added Cultural, Governance, Geographic, and Economic Distance in MNEs’ International Expansion.” Journal of International Management 20 (1): 38–54. doi:10.1016/j.intman.2013.02.003.

Instituto Nacional de Estatística. 2016. “Tourism Statistics-2015.” Estatísticas Do Turismo. July. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=265858123&PUBLICACOEStema=55581&PUBLICACOESmodo=2.

International Organization for Standardization. 2017. “ISO Country Codes.” Country Codes. February 13. https://www.iso.org/obp/ui/#search.

Jennings, Frank, and John Yates. 2009. “Scrapping over Data: Are the Data Scrapers’ Days Numbered?” Journal of Intellectual Property Law & Practice 4 (2): 120–129. doi:0.1093/jiplp/jpn232.

Johanson, Jan, and Jan-Erik Vahlne. 1977. “The Internationalization Process of the Firm-A Model of KnowledgedDevelopment and Increasing Foreign Market Commitments.” Journal of International Business Studies 8 (1): 23–32.

Johanson, Jan, and Finn Wiedersheim-Paul. 1975. “The Internationalization of the Firm - Four Swedish Cases.” Journal of Management Studies 12 (3): 305–23.

Kim, Jong-Hyeong. 2018. “The Impact of Memorable Tourism Experiences on Loyalty Behaviors: The Mediating Effects of Destination Image and Satisfaction.” Journal of Travel Research 57 (7): 856–70. doi:10.1177/0047287517721369.

Kim, Woo Gon, Hyunjung Lim, and Robert A. Brymer. 2015. “The Effectiveness of Managing Social Media on Hotel Performance.” International Journal of Hospitality Management 44 (January): 165–71. doi:10.1016/j.ijhm.2014.10.014.

Kock, Florian, Alexander Josiassen, and A. George Assaf. 2016. “Advancing Destination Image: The Destination Content Model.” Annals of Tourism Research 61 (November): 28–44. doi:10.1016/j.annals.2016.07.003.

Kostyra, Daniel S., Jochen Reiner, Martin Natter, and Daniel Klapper. 2016. “Decomposing the Effects of Online Customer Reviews on Brand, Price, and Product Attributes.” International Journal of Research in Marketing 33 (1): 11–26. doi:10.1016/j.ijresmar.2014.12.004.

Kotsiantis, Sotiris, and Dimitris Kanellopoulos. 2006. “Discretization Techniques: A Recent Survey.” In , 32:47–58. 1.

Kruskal, William H., and W. Allen Wallis. 1952. “Use of Ranks in One-Criterion Variance Analysis.” Journal of the American Statistical Association 47 (260): 583. doi:10.2307/2280779.

Kwok, Linchi, Karen L. Xie, and Tori Richards. 2017. “Thematic Framework of Online Review Research: A Systematic Analysis of Contemporary Literature on Seven Major Hospitality and Tourism Journals.” International Journal of Contemporary Hospitality Management 29 (1): 307–54. doi:10.1108/IJCHM-11-2015-0664.

Lai, Kun, and Xiang (Robert) Li. 2016. “Tourism Destination Image: Conceptual Problems and Definitional Solutions.” Journal of Travel Research 55 (8): 1065–80. doi:10.1177/0047287515619693.

Linnemann, Hanz. 1966. An Econometric Study of International Trade Flows. North Holland, Amsterdam.

Liu, Bing, and Lei Zhang. 2012. “A Survey of Opinion Mining and Sentiment Analysis.” In Mining Text Data, edited by C. C. Aggarwal and C. X. Zhai, 415–463. Springer. http://link.springer.com/chapter/10.1007/978-1-4614-3223-4_13.

Liu, Yong, Thorsten Teichert, Matti Rossi, Hongxiu Li, and Feng Hu. 2017. “Big Data for Big Insights: Investigating Language-Specific Drivers of Hotel Satisfaction with 412,784 User-Generated Reviews.” Tourism Management 59: 554–63. doi:10.1016/j.tourman.2016.08.012.

Lu, Weilin, and Svetlana Stepchenkova. 2015. “User-Generated Content as a Research Mode in Tourism and Hospitality Applications: Topics, Methods, and Software.” Journal of Hospitality Marketing & Management 24 (2): 119–54. doi:10.1080/19368623.2014.907758.

Maeyer, Peter De. 2012. “Impact of Online Consumer Reviews on Sales and Price Strategies: A Review and Directions for Future Research.” Journal of Product & Brand Management 21 (2): 132–39. doi:10.1108/10610421211215599.

Markides, Constantinos. 1998. “Strategic Innovation in Established Companies.” MIT Sloan Management Review. https://sloanreview.mit.edu/article/strategic-innovation-in-established-companies/.

Martin, Brett A. S., Hyun Seung Jin, and Nhu Vi Trang. 2017. “The Entitled Tourist: The Influence of Psychological Entitlement and Cultural Distance on Tourist Judgments in a Hotel Context.” Journal of Travel & Tourism Marketing 34 (1): 99–112. doi:10.1080/10548408.2015.1130112.

Massara, Francesco, and Fabio Severino. 2013. “Psychological Distance in the Heritage Experience.” Annals of Tourism Research 42 (July): 108–29. doi:10.1016/j.annals.2013.01.005.

Mayer, Thierry, and Soledad Zignago. 2011. “Notes on CEPII’s Distances Measures: The GeoDist Database.” CEPII. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1994531.

Mckercher, Bob. 2008. “Segment Transformation in Urban Tourism.” Tourism Management 29 (6): 1215–25. doi:10.1016/j.tourman.2008.03.005.

Mckercher, Bob, and Alan A. Lew. 2003. “Distance Decay and the Impact of Effective Tourism Exclusion Zones on International Travel Flows.” Journal of Travel Research 42 (2): 159–65. doi:10.1177/0047287503254812.

Melian-Gonzalez, S., J. Bulchand-Gidumal, and B. Gonzalez Lopez-Valcarcel. 2013. “Online Customer Reviews of Hotels: As Participation Increases, Better Evaluation Is Obtained.” Cornell Hospitality Quarterly 54 (3): 274–83. doi:10.1177/1938965513481498.

Mellinas, Juan Pedro, Soledad-María Martínez María-Dolores, and Juan Jesús Bernal García. 2015. “Booking.Com: The Unexpected Scoring System.” Tourism Management 49: 72–74. doi:10.1016/j.tourman.2014.08.019.

Monkman, Graham George, Michel Kaiser, and Kieran Hyder. 2018. “The Ethics of Using Social Media in Fisheries Research.” Reviews in Fisheries Science & Aquaculture 26 (2): 235–42. doi:10.1080/23308249.2017.1389854.

Morosan, Cristian, and John T. Bowen. 2017. “Analytic Perspectives on Online Purchasing in Hotels: A Review of Literature and Research Directions.” International Journal of Contemporary Hospitality Management 30 (1): 557–80. doi:10.1108/IJCHM-10-2016-0566.

Mossberg, Lena, and Ingeborg Astrid Kleppe. 2005. “Country and Destination Image – Different or Similar Image Concepts?” The Service Industries Journal 25 (4): 493–503. doi:10.1080/02642060500092147.

Nicolau, Juan L., and Francisco J. Más. 2006. “The Influence of Distance and Prices on the Choice of Tourist Destinations: The Moderating Role of Motivations.” Tourism Management 27 (5): 982–96. doi:10.1016/j.tourman.2005.09.009.

Nyaupane, Gyan P., and Alan R. Graefe. 2008. “Travel Distance: A Tool for Nature-based Tourism Market Segmentation.” Journal of Travel & Tourism Marketing 25 (3–4): 355–66. doi:10.1080/10548400802508457.

Öğüt, Hulisi, and Bedri Kamil O. Onur Taş. 2012. “The Influence of Internet Customer Reviews on the Online Sales and Prices in Hotel Industry.” The Service Industries Journal 32 (2): 197–214. doi:10.1080/02642069.2010.529436.

Ojala, Arto. 2015. “Geographic, Cultural, and Psychic Distance to Foreign Markets in the Context of Small and New Ventures.” International Business Review 24 (5): 825–35. doi:10.1016/j.ibusrev.2015.02.007.

Ojala, Arto, and Pasi Tyrväinen. 2008. “Market Entry Decisions of US Small and Medium-sized Software Firms.” Management Decision 46 (2): 187–200. doi:10.1108/00251740810854113.

Phillips, Paul, Stuart Barnes, Krystin Zigan, and Roland Schegg. 2016. “Understanding the Impact of Online Reviews on Hotel Performance: An Empirical Analysis.” Journal of Travel Research, April, 0047287516636481. doi:10.1177/0047287516636481.

Phillips, Paul, and Luiz Moutinho. 2014. “Critical Review of Strategic Planning Research in Hospitality and Tourism.” Annals of Tourism Research 48 (September): 96–120. doi:10.1016/j.annals.2014.05.013.

Phillips, Paul, Krystin Zigan, Maria Manuela Santos Silva, and Roland Schegg. 2015. “The Interactive Effects of Online Reviews on the Determinants of Swiss Hotel Performance: A Neural Network Analysis.” Tourism Management 50 (October): 130–41. doi:10.1016/j.tourman.2015.01.028.

Pike, Steven D., and Stephen Page. 2014. “Destination Marketing Organizations and Destination Marketing : A Narrative Analysis of the Literature.” Tourism Management 41 (April): 202–27.

Pizam, Abraham, and Silvia Sussmann. 1995. “Does Nationality Affect Tourist Behavior?” Annals of Tourism Research 22 (4): 901–17. doi:10.1016/0160-7383(95)00023-5.

Prideaux, Bruce. 2000. “The Role of the Transport System in Destination Development.” Tourism Management 21 (1): 53–63. doi:10.1016/S0261-5177(99)00079-5.

Qian, Jianwei, Rob Law, and Jiewen Wei. 2018. “Effect of Cultural Distance on Tourism: A Study of Pleasure Visitors in Hong Kong.” Journal of Quality Assurance in Hospitality & Tourism 19 (2): 269–84. doi:10.1080/1528008X.2017.1410079.

Ravi, Kumar, and Vadlamani Ravi. 2015. “A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications.” Knowledge-Based Systems 89 (November): 14–46. doi:10.1016/j.knosys.2015.06.015.

Ring, Amata, Aaron Tkaczynski, and Sara Dolnicar. 2016. “Word-of-Mouth Segments: Online, Offline, Visual or Verbal?” Journal of Travel Research 55 (4): 481–92. doi:10.1177/0047287514563165.

Ryan, Chris, and Jenny Cave. 2005. “Structuring Destination Image: A Qualitative Approach.” Journal of Travel Research 44 (2): 143–50. doi:10.1177/0047287505278991.

Safari, Aswo, Peter Thilenius, and Amjad Hadjikhani. 2013. “The Impact of Psychic Distance on Consumers’ Behavior in International Online Purchasing.” Journal of International Consumer Marketing 25 (4): 234–49. doi:10.1080/08961530.2013.803899.

Saralegi, Xabier, and Iñaki San Vincente. 2013. “Elhuyar at TASS 2013.” In Proceedings of “XXIX Congreso de La Sociedad Española de Procesamiento de Lenguaje Natural,” edited by A. D. Esteban, I.A. Loinaz, and J. V. Román, 143–50. Madrid, Spain: El Congreso Español de Informática.

Schuckert, Markus, Xianwei Liu, and Rob Law. 2015. “A Segmentation of Online Reviews by Language Groups: How English and Non-English Speakers Rate Hotels Differently.” International Journal of Hospitality Management 48 (July): 143–49. doi:10.1016/j.ijhm.2014.12.007.

Silva, Mário J., Paula Carvalho, and Luís Sarmento. 2012. “Building a Sentiment Lexicon for Social Judgement Mining.” In Computational Processing of the Portuguese Language, edited by Helena Caseli, Aline Villavicencio, António Teixeira, and Fernando Perdigão, 218–28. Lecture Notes in Computer Science 7243. New York, NY: Springer Berlin Heidelberg. doi:10.1007/978-3-642-28885-2_25.

Sparks, Beverley A., and Victoria Browning. 2011. “The Impact of Online Reviews on Hotel Booking Intentions and Perception of Trust.” Tourism Management 32 (6): 1310–23. doi:10.1016/j.tourman.2010.12.011.

Surowiecki, James. 2005. The Wisdom of Crowds. Reprint edition. New York, NY: Anchor.

Tajfel, H. 1982. “Social Psychology of Intergroup Relations.” Annual Review of Psychology 33 (1): 1–39. doi:10.1146/annurev.ps.33.020182.000245.

Tan, Huimin, Xingyang Lv, and Dogan Gursoy. 2018. “Evaluation Nudge: Effect of Evaluation Mode of Online Customer Reviews on Consumers’ Preferences.” Tourism Management 65 (April): 29–40. doi:10.1016/j.tourman.2017.09.011.

Tennekes, Martjn, and Edwin de Jonge. 2017. Tabplot: Tableplot, a Visualization of Large Datasets (version 1.3-1). R package. https://CRAN.R-project.org/package=tabplot.

“Top Ten Internet Languages in the World - Internet Statistics (2017).” 2019. Internet World Stats. Accessed April 30. https://www.internetworldstats.com/stats7.htm?utm_source=lasindias.info/blog.

Uchiyama, Yuta, and Ryo Kohsaka. 2016. “Cognitive Value of Tourism Resources and Their Relationship with Accessibility: A Case of Noto Region, Japan.” Tourism Management Perspectives 19 (July): 61–68. doi:10.1016/j.tmp.2016.03.006.

Ulaga, Wolfgang, and Andreas Eggert. 2006. “Relationship Value and Relationship Quality: Broadening the Nomological Network of Business-to-business Relationships.” European Journal of Marketing 40 (3/4): 311–27. doi:10.1108/03090560610648075.

Ven, Andrew H. Van de. 2007. Engaged Scholarship: A Guide for Organizational and Social Research. 1 edition. Oxford ; New York: Oxford University Press, USA.

Wilson, Alan, Hilary Murphy, and Jesus Cambra Fierro. 2012. “Hospitality and Travel: The Nature and Implications of User-Generated Content.” Cornell Hospitality Quarterly 53 (3): 220–28. doi:10.1177/1938965512449317.

Wong, IpKin Anthony, Lawrence Hoc Nang Fong, and Rob Law. 2016. “A Longitudinal Multilevel Model of Tourist Outbound Travel Behavior and the Dual-Cycle Model.” Journal of Travel Research 55 (7): 957–70. doi:10.1177/0047287515601239.

Wu, Laurie, Han Shen, Alei Fan, and Anna S. Mattila. 2017. “The Impact of Language Style on Consumers′ Reactions to Online Reviews.” Tourism Management 59 (April): 590–96. doi:10.1016/j.tourman.2016.09.006.

Xiang, Zheng, Zvi Schwartz, John H. Jr. Gerdes, and Muzaffer Uysal. 2015. “What Can Big Data and Text Analytics Tell Us about Hotel Guest Experience and Satisfaction?” International Journal of Hospitality Management 44 (January): 120–30. doi:10.1016/j.ijhm.2014.10.013.

Xie, Karen L., Zili Zhang, and Ziqiong Zhang. 2014. “The Business Value of Online Consumer Reviews and Management Response to Hotel Performance.” International Journal of Hospitality Management 43 (October): 1–12. doi:10.1016/j.ijhm.2014.07.007.

Xu, Xun, and Yibai Li. 2016. “The Antecedents of Customer Satisfaction and Dissatisfaction toward Various Types of Hotels: A Text Mining Approach.” International Journal of Hospitality Management 55 (May): 57–69. doi:10.1016/j.ijhm.2016.03.003.

Yang, Yang, Hongbo Liu, and Xiang (Robert) Li. 2019. “The World Is Ffatter? Examining the Relationship between Cultural Distance and International Tourist Flows.” Journal of Travel Research 58 (2): 224–40. doi:10.1177/0047287517748780.

Yang, Yang, Sangwon Park, and Xingbao Hu. 2018. “Electronic Word of Mouth and Hotel Performance: A Meta-Analysis.” Tourism Management 67 (August): 248–60. doi:10.1016/j.tourman.2018.01.015.

Zhang, C., S. T. Cavusgil, and A. S. Roath. 2003. “Manufacturer Governance of Foreign Distributor Relationships: Do Relational Norms Enhance Competitiveness in the Export Market?” Journal of International Business Studies 34 (6): 550–66. doi:10.1057/palgrave.jibs.8400051.

Zhang, Hongmei, Xiaoxiao Fu, Liping A. Cai, and Lin Lu. 2014. “Destination Image and Tourist Loyalty: A Meta-Analysis.” Tourism Management 40 (February): 213–23. doi:10.1016/j.tourman.2013.06.006.

Zhang, Jun, Sangyun Seo, and Hoonyoung Lee. 2013. “The Impact of Psychological Distance on Chinese Customers When Selecting an International Healthcare Service Country.” Tourism Management 35 (April): 32–40. doi:10.1016/j.tourman.2012.05.007.