Predictive models of hotel booking cancellation: a semi-automated analysis of the literature

Modelos preditivos de cancelamento de reservas de hotéis: uma análise semiautomática da literatura

Nuno António
ISCTE-IUL and Instituto de Telecomunicações, Av. das Forças Armadas, 1649-026 Lisboa, Portugal,
nuno_miguel_antonio@iscte-iul.pt

Ana de Almeida
ISCTE-IUL, CISUC and ISTAR, 1649-026 Lisboa, Portugal, ana.almeida@iscte-iul.pt

Luís Nunes
ISCTE-IUL, Instituto de Telecomunicações and ISTAR, 1649-026 Lisboa, Portugal, luis.nunes@iscte-iul.pt

Abstract

This study sought to combine data science tools and capabilities with human judgement and interpretation in order to demonstrate how semiautomatic analysis of the literature can contribute to identifying and synthesising research findings and topics about booking cancellation forecasting. The study also focused on recording in detail the analysis's full experimental procedure to encourage other researchers to conduct automated literature reviews in order to understand more fully the current tendencies in their field of study. The data were obtained through a keyword search in Scopus and Web of Science databases. The methodology presented not only diminishes human bias but also enhances data visualisation and text mining techniques' ability to facilitate abstraction, expedite analysis and improve literature reviews. The results show that, despite the importance of forecasting booking cancellations to understanding net demand and improving cancellation and overbooking policies, further research on this subject is needed.

Keywords: Data science, forecast, literature review, prediction, revenue management.

Resumo

Através da aplicação de ferramentas e recursos de *data science*, com apoio do julgamento e interpretação humanos, temos como objetivo demonstrar como a análise semiautomática da literatura pode contribuir para sintetizar a investigação existente para a previsão de cancelamento de reservas, incluindo a identificação dos tópicos abordados pela investigação. Para além disso, ao detalhar o procedimento experimental da análise, temos como objetivo encorajar outros autores a realizar análises automatizadas de literatura. Os dados utilizados foram obtidos a partir das bases de dados da Scopus e da Web of Science. A metodologia utilizada, além de atenuar o viés humano, demonstrou como as técnicas de visualização de dados e de *text mining* facilitam a abstração, fomentam a aceleração da análise e contribuem para a melhoria das revisões. Os resultados mostram que, embora a previsão de cancelamento de reservas seja de reconhecida importância para a compreensão da procura líquida e melhoria das políticas de cancelamento e *overbooking*, há ainda necessidade de mais investigação sobre o assunto.

Palavras-chave: *Data Science*, Gestão de receitas, previsão, revisão de literatura, processamento de linguagem natural.

1. Introduction

Revenue management (RM) is considered one of the most successful application areas of operations research (Talluri & Van Ryzin, 2005). The main goal is that of increasing revenue through demand management decisions, i.e. decisions based on the estimation of demand and its characteristics, which make use of price and capacity controls to "manage" demand (Talluri & Van Ryzin, 2005). RM is customarily applied in industries that have fixed capacity, a variable and uncertain demand, a perishable inventory, a high-cost fixed structure, and customers with different sensitivities to prices (Kimes & Wirtz, 2003). It is therefore not surprising the success of RM in travel- and tourism-related industries like airlines, hotels, cruises, rent-a-cars, or golf (Kimes & Wirtz, 2003; Talluri & Van Ryzin, 2005).

Theoretically, the problems addressed by RM are not new. Its novelty comes from the methods being employed in the decision-making process, mainly the sophistication of technology, and the detail and strong operational approaches put into making decisions. Therefore, the role of revenue management systems (RMS) in modern RM is comprehensible since it enables us to manage demand on a scale and level of complexity that would have been impossible otherwise (Talluri & Van Ryzin, 2005). One of the most critical tools in RMS is forecast performance. Without accurate forecast performance, an RMS's rate and availability recommendations would probably be highly unreliable (Talluri & Van Ryzin, 2005; Weatherford & Kimes, 2003). Estimation and forecasting, together with data collection, optimisation, and control, are the fundamental steps in RM. An RMS's forecasts focus mainly on measures such as room nights, arrivals, price sensitivity, and booking cancellations. In fact, an accurate booking cancellation forecast is of foremost importance to determine net demand, i.e. demand deducted of bookings predicted as likely to be cancelled (Lemke, Riedel, & Gabrys, 2013; Talluri & Van Ryzin, 2005).

With cancellations reaching 20% to 60% of all bookings received by hotels (Liu, 2004; Morales & Wang, 2010) it is understandable that hotels implement overbooking or restrictive cancellations policies to reduce cancellation rates (Chen, 2016; Talluri & Van Ryzin, 2005). However, such policies can have adverse effects in terms of reallocation costs, social reputation, or revenue decrease due to the discounts granted (Guo, Dong, & Ling, 2016; Noone & Lee, 2010). To reduce the negative impact of overbooking and restrictive cancellation policies, as well as the uncertainty in demand management decisions, cancellations and no-show forecasts are used as inputs in RMSs (Antonio, Almeida, & Nunes, 2017a; Morales & Wang, 2010; Talluri & Van Ryzin, 2005). However, despite the importance of booking cancellation forecast models (Chen, 2016), only a few works have tested or developed forecast cancellation models.

This work takes advantage of advanced data science methods to synthesise the current research findings on the development of forecast/prediction models for booking cancellation in travel- and tourism-related industries and to identify the main topics covered by booking cancellation research.

2. Literature review

2.1 Revenue management and forecasting models

The importance of forecasting in RM research is recognised in the latest literature reviews (Chiang, Chen, & Xu, 2007; Denizci Guillet & Mohammed, 2015; Ivanov & Zhechev, 2012). Demand forecasting is identified as one of the focuses of research in RM. However, forecasting is believed to be difficult, costly, and occasionally failing to produce satisfactory results. The literature recommends that future research should take advantage of technological and mathematical/scientific methods, including big data, to develop new and improved forecasting models (Chiang et al., 2007; Ivanov & Zhechev, 2012; McGuire, 2017; Pan & Yang, 2017).

Despite the importance of forecasting hotel demand, the literature on the topic is scarce and does not take advantage of technological and scientific methods. The existing literature shows three types of methods: time series, advance booking, and combined modelling (Lee, 2018). However, these methods are parametric and assume that data has a known distribution and that it is possible to estimate the needed parameters (Talluri & Van Ryzin, 2005). However, a parametric model's performance can suffer when the demand distribution differs from the parameters' assumptions. Thus, a hotel demand forecasting model's performance could be improved by nonparametric models, like neural networks or other advanced machine learning methods (McGuire, 2017).

Although the terms "forecasting" and "prediction" are considered synonyms and are employed interchangeably (Matsuo, 2003), formally they have different meanings and definitions. While forecasting aims to calculate or predict future events and is frequently associated with a time series, a prediction is also used to reconstruct and explain some past outcome (Lewis-Beck, 2005; Matsuo, 2003). For predicting booking cancellations in the context of machine learning, these goals are cast as two different problems. When the objective is solely to estimate the cancellation rate — a regression— then the problem should be considered a forecasting problem. When aiming at estimating the likelihood of a booking to be cancelled, and understanding the cancellation drivers, then the problem should be considered a classification problem. In this case, the booking cancellation predictions enable the estimation of the overall cancellation rate.

2.2 Automatic literature review

The analysis of prior relevant literature is essential to identify the areas covered within a body of research and uncover areas where further research is necessary. An effective literature review is a foundation for the advancement of knowledge (Webster & Watson, 2002). Nevertheless, the task of conducting a comprehensive literature review is becoming increasingly complex. The abundance of potentially relevant research, not only in the research field, but also in related and even non-related fields, makes the task evermore demanding (Delen & Crossland, 2008; Nunez-Mir, Iannone, Pijanowski, Kong, & Fei, 2016). This progressive difficulty in carrying out an adequate literature review causes some authors to defend the need to take advantage of technological advances to automate literature reviews. This automation would potentially enable faster and less resource-intensive literature reviews (Bragge, Relander, Sunikka, & Mannonen, 2007; Feng, Chiam, & Lo, 2017; Tsafnat et al., 2014). In fact, in a literature review conducted by Feng et al. (2017), the authors found 32 relevant studies that advocated the use of automated or semi-automated solutions to support systematic literature reviews. Systematic literature reviews assess and interpret existing literature pertinent to a specific research topic, subject, or phenomenon (Kitchenham & Charters, 2017).

As identified by Tsafnat et al. (2014), the automation of the literature review has the potential to help researchers in almost all systematic review tasks, namely, formulation of the review questions, finding previous systematic reviews, writing the protocol, devising the search strategy, searching, finding duplicates, scanning abstracts, obtaining full-text articles, scanning full-text articles, forward and backward citation searching, data extraction, data conversion and synthesis, literature re-checking, and lastly, writing up the review. Delen & Crossland (2008) state that the application of advanced methods that allow for the automation of the literature review could potentially lead to the following: enhancement of the retrieval of data, characterisation of the research based on metadata (journal, authors, organisations), reveal of new technical concepts or technical relationships, identification of the main topics and sub-topics of the research, identification of the relationship between topics and metadata, and provision of insights on research directions. The application and benefits of these advanced methods are witnessed by some examples of automated, or at least, semiautomated literature analysis. For instance, Moro et al. (2015) employed text mining (TM) to identify relevant terms and topics of business intelligence research applied to the banking industry. Nunez-Mir et al. (2016) used TM methods to demonstrate how automated content analysis could help synthesize knowledge from the enormous volume of ecology and evolutionary biology literature. Guerreiro et al. (2016) conducted TM analyses to study research in cause-related marketing. Park & Nagy (2018) employed TM to study thermal comfort and building control research. Haneem et al. (2017) utilised data analytics and TM in a literature review in master data management to show how they could assist the process of literature analysis.

All the above-cited examples of automated/semiautomated literature reviews employed TM. TM seeks to extract useful information from documents collections through the identification and exploration of patterns. While data mining (DM) assumes that data is stored in a structured format, TM data needs no structured format. Thus, TM data requires the application of preprocessing operations to identify and extract features representative of natural language documents (Welbers, Van Atteveldt, & Benoit, 2017). Due to the importance of natural language processing in TM, the latter draws on the advances of other computer science disciplines, like data science, to achieve its objectives. Data science is a fairly recent discipline that combines the use of mathematics, operational research, machine learning, natural language processing, statistics, data visualization, and other tools and capabilities from different disciplines; it is also complemented with a broad understanding of the problem domain to understand and analyze data (O'Neil & Schutt, 2013).

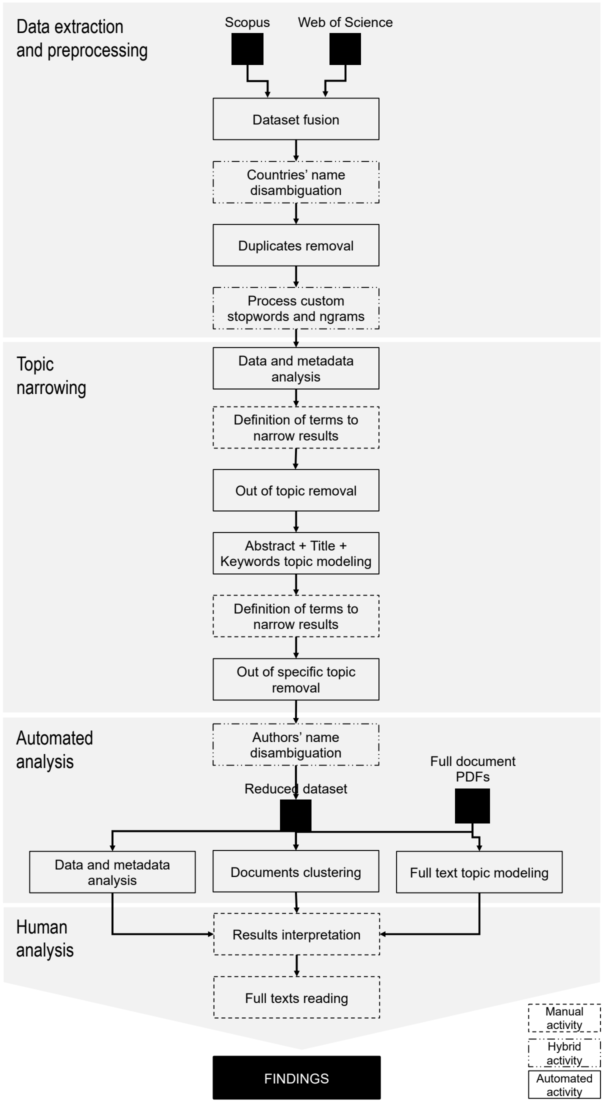
3 Methodology

Usually, authors consider two types of literature reviews: a mature topic that has an extended body of knowledge requiring analysis and synthesis; and a promising topic that could benefit from further theoretical groundwork (Webster & Watson, 2002). By focusing on the analysis of literature from a confined and recent topic, bookings cancellation forecasting/prediction in travel- and tourism-related industries, this work is an example of the latter type. In order to achieve the objective of this paper, data science tools and capabilities, especially TM, natural language processing, and data visualisation, are employed to conduct a semiautomated analysis of the existent literature.

Other works on the automation of literature review proposed somewhat similar processes to conduct literature analysis (Delen & Crossland, 2008; Feng et al., 2017; Haneem et al., 2017; Nunez-Mir et al., 2016; Tsafnat et al., 2014). However, those processes differ in the techniques or approaches employed. The present proposal is based on the process used by Haneem et al. (2017) adapted to the current topic; a diagram of this process is depicted in Figure 1. The new procedure is divided into four main steps splitting into diverse activities. Some are fully automated, some are manual, and others are hybrid activities, i.e. partially automated. The details are presented in the following sub-sections, jointly with some of the results to help explain the rationale behind the methodological choices.

All the steps of the experimental procedure presented in Figure 1 were conducted using R (R Core Team, 2016), which is a powerful statistical tool with numerous packages extending its capabilities and designed to facilitate data analysis. The R source code and datasets used in this work can be downloaded from https://osf.io/d2xr5/?view_only=957dba4a77724ebeb4f7963b8932e614.

Figure 1 - Experimental procedure workflow diagram


Data extraction and preprocessing
Scopus
Web of Science
Dataset fusion
Countries' name disambiguation
Duplicates removal
Process custom stopwords and ngrams
Topic narrowing
Data and metadata analysis
Definition of terms to narrow results
Out of topic removal
Abstract + Title + Keywords topic modeling
Definition of terms to narrow results
Out of specific topic removal
Automated analysis
Authors' name disambiguation
Reduced dataset
Full document PDFs
Data and metadata analysis
Documents clustering
Full text topic modeling
Human analysis
Results interpretation
Full texts reading
FINDINGS
Manual activity
Hybrid activity
Automated activity

Data visualisation is considered a powerful resource to ease data abstraction and improve reviews (Fabbri et al., 2013). Therefore, data and metadata analysis are involved in the elaboration of several data visualisations of the full-text corpus to aid the identification of patterns, namely, word clouds with top frequent unigrams and bigrams, heatmaps of frequent terms, and networks analysis of authors, countries, and keywords.

3.1 Data extraction and pre-processing

Quality literature analysis must cover relevant literature on the research topic and should not be confined to one specific research methodology, one set of journals, or one geographic region (Webster & Watson, 2002). The search strategy is an important component of an analysis of literature. The present approach is based on what Ali & Usman (2018) call an "automated search", that is, a search strategy that relies on electronic databases keyword searches. The number of databases used and their type are essential guidelines to guarantee the quality of the review (Ali & Usman, 2018). This work relies on two well-known databases: Scopus and Web of Science (WoS). These databases cover most sources of literature related to tourism and travel industries.

An adequate selection of keywords must be used in the correct construction of the search string (Ali & Usman, 2018; Delen & Crossland, 2008). Taking into consideration the problems that might arise from the differences among database search engines (Tsafnat et al., 2014), a simple query was executed on both databases, and the results were filtered to narrow the search to the present objective. The search string found simultaneous usage of the words "booking" and "cancellation" in the title or keywords of documents. In the case of the Scopus database, the words were also searched in the abstract; WoS does not have this functionality. Since sometimes the word "reservation" is employed instead of "booking" we also searched for "reservation". Variations in plural/singular nouns and in the UK and American English of the word "cancellation" were accounted for. The application of TM for multiple languages presents methodological difficulties. However, given that the most relevant research is published in English, the search was limited to English documents. The timespan was defined from 1990 to 2018, with May 4th 2018 being the final extraction date. Since each database has its own document classification categories, the type of documents selected in each of the databases was different. For Scopus, the chosen document types were article, article in press, book, book chapter, conference paper, or review. For WoS, the chosen documents were of types article, book, book chapter, proceedings paper, or review. The full search strings are shown in Figure 2 and Figure 3, respectively. Note that upon cancellation, the customer informs the service provider before the time of service provision, while in a "no-show" the customer does not inform the service provider and fails to check-in. Although "no-shows" are to be treated here as cancellations, the term "no-show" was not included in the search, thus avoiding the identification of works that solely address the problem of "no-shows" forecast (which is quite common for the airline industry). The Scopus search result was exported to a CSV (comma separated values) file using Scopus export functionality. The WoS search result was exported to a TSV (tab separated values) file. To assess the validity of each search's results, a randomly selected number of documents were checked to ascertain their inclusion of the search words. Next, the inclusion of known documents on the search topic results was also checked.

Figure 2 – Scopus search string

```
TITLE-ABS-
KEY ( ( "booking"  OR  "bookings"  OR  "reservation"  OR  "reservations" )  AND  ( "cancellation"  OR  "cancel
lations"  OR  "cancelation"  OR  "cancelations" ) )  AND  ( DOCTYPE ( ar )  OR  DOCTYPE ( ip )  OR  DOCT
YPE ( bk )  OR  DOCTYPE ( ch )  OR  DOCTYPE ( cp )  OR  DOCTYPE ( re ) )  AND  PUBYEAR  >  1989  A
ND  ( LIMIT-TO ( LANGUAGE ,  "English " ) )
```

Figure 3 – WoS search string

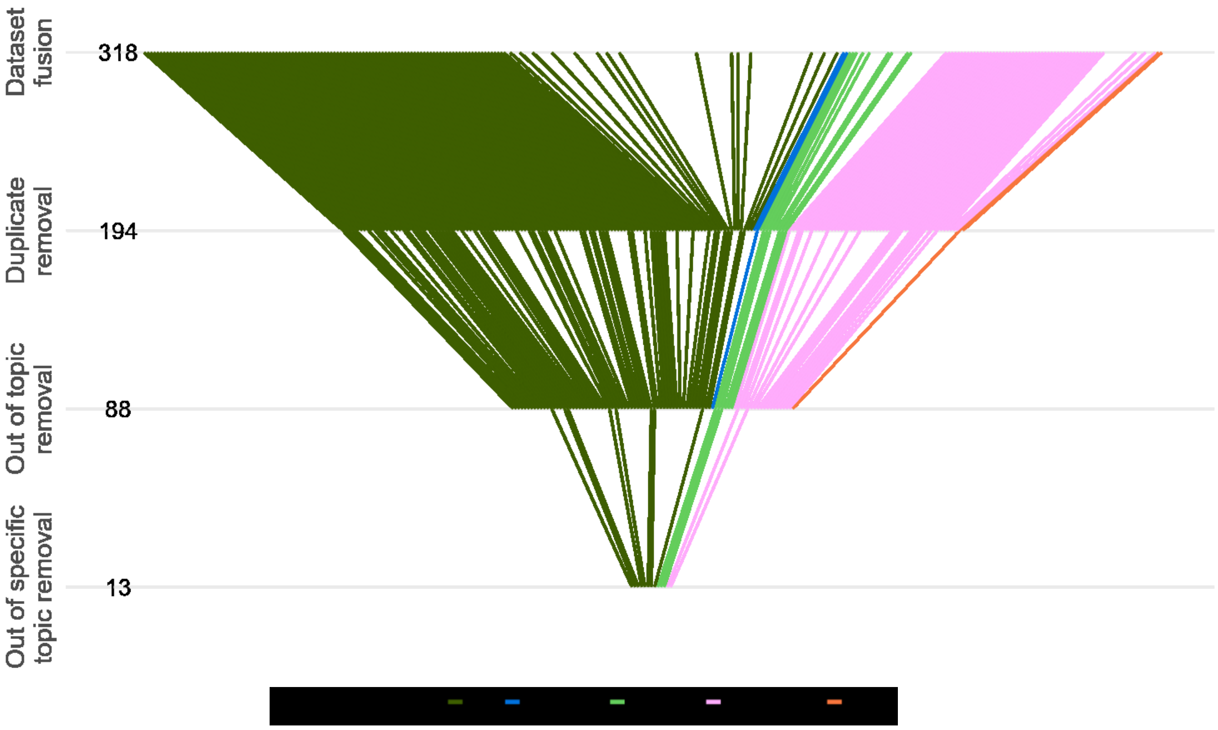
```
(((TS=(("booking" OR "bookings" OR "reservation" OR "reservations") AND ("cancellation" OR "cancellations"
OR "cancelation" OR "cancelations")) OR TI=(("booking" OR "bookings" OR "reservation" OR "reservations")
AND ("cancellation" OR "cancellations" OR "cancelation" OR
"cancelations")))))AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article OR Book OR Book
Chapter OR Proceedings Paper OR Review)

Timespan: 1990-2018. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, CCR-
EXPANDED, IC.
```

The fusion of multiple databases from different origins is always a challenge due to differences in the database structures, data formats, and data quality. For this reason, and according to the scope of this work, it was decided to create a single dataset based on the fields described in Figure 4 by fusion of the two results. This involved a normalisation process: the conversion of all text into lowercase, thus transforming all words into a uniform form (Welbers et al., 2017). All text preprocessing was performed using the "NLP" (Hornik, 2017) and "tm" (Feinerer & Hornik, 2017) R packages. An additional field was created to store the author's country. Its content was automatically extracted from the Affiliations field with further manual manipulation since some countries were written differently in some documents (e.g. USA, United States, or United States of America).

Figure 4– Dataset field matching diagram

As can be observed (Figure 5), the fusion of the two datasets resulted in a total of 318 documents (168 from Scopus and 150 from WoS) of which a substantial part were duplicates. The only common field identifier in both databases is the DOI, but not all documents have a DOI. Thus, the removal of duplicates was achieved after the titles of the documents were preprocessed; this was achieved by comparing the titles and then comparing the DOIs. Preprocessing is a process that tokenises full texts to smaller and specific features, including the normalisation of words, for improved analysis and enhanced computational performance. Preprocessing text involves the removal of punctuation, removal of numbers, removal of stopwords, and stemming. Stopwords are words that are common in a language, e.g. "the" or "a". Stemming normalises words with different morphological variations, such as verb conjugation suffixes or the plural form of a noun (Welbers et al., 2017). Title preprocessing allowed for the capture of duplicate documents that a simple comparison would not. For example, the title of one article in Scopus did not include the initial "The". One exception to the automatic identification of duplicates was found: two documents shared the same title and abstract but had different sources, DOIs, and authors. A manual verification showed that it was the same document, but was presented at two different conferences. The removal of duplicates reduced the dataset to 194 documents (Figure 5).

Figure 5 - Document selection funnel

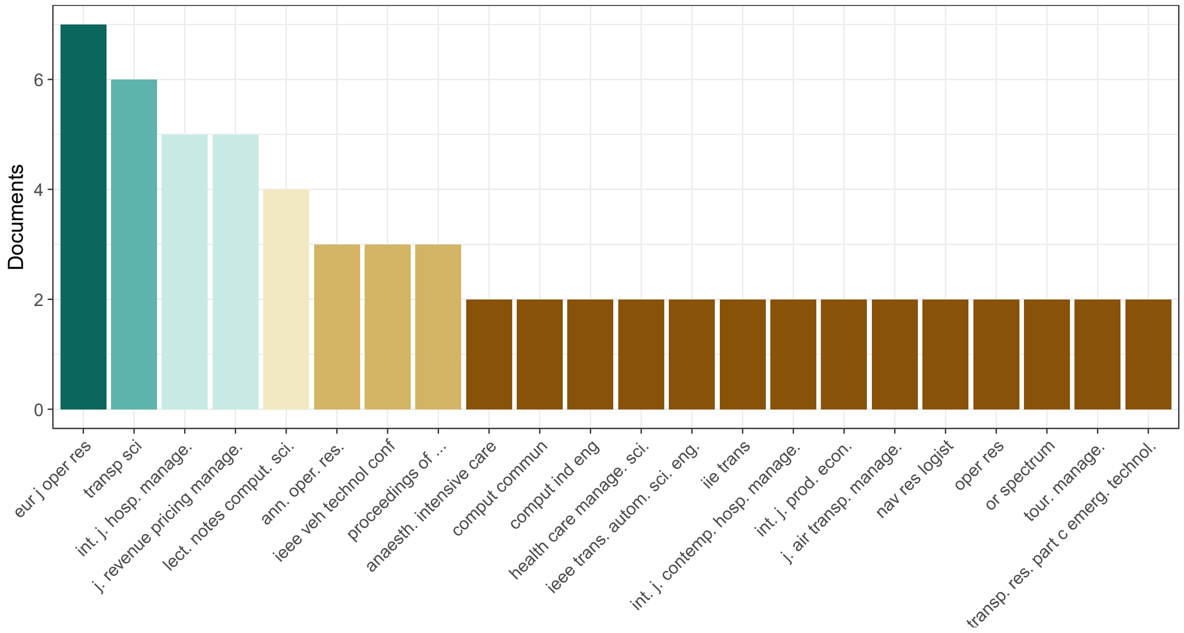


With the help of a document-term matrix (DTM) the frequency of common terms was verified. A DTM, or corpus, is a common form for representing a collection of documents. It assigns documents to the rows of a matrix and terms contained in the documents to columns. The cells of the matrix indicate the frequency of the terms in the documents, allowing for analysis with vector and matrix algebra (Welbers et al., 2017). This was necessary to identify terms that, although presenting a high frequency, were not relevant (e.g. "Elsevier bv", "all rights reserved", "et al", among others). On the other hand, terms that were relevant but composed of multiple words were converted to one-word terms (e.g. "revenue management" or "no-show"). In TM, terms can also be called "n-grams", where "n" indicates the number of words. A single word term is called a "unigram", a two-word term is called a "bigram", and so on (Welbers et al., 2017). The terms identified as not being relevant were simply removed from the title, author keywords, index keywords, and abstract, while the relevant bigrams were converted into unigrams (e.g. "revenue management" was converted into "revman", and "no-show" was converted into "noshow").

3.2 Out of topic removal

The 194 documents obtained from the previous step were composed by 128 (66%) articles, 53 (27%) conference papers, 9 (5%) book chapters, 2 (1%) articles in press, and 2 (1%) reviews. The documents came from 152 different sources. From those sources, only 22 had more than one document. The sources are depicted in Figure 6, and it is possible to verify that the documents came from different areas, such as operations research, hospitality management, medicine and health, and transportation and logistics management.

Figure 6 - Documents per source (sources with more than one document)



Source: Authors.

Although all previous works on the topic of bookings cancellation forecast/prediction could potentially be relevant for the objective of this research, this is not true for areas where business characteristics are not compatible with travel- and tourism-related industries. Therefore, it was decided that all documents unrelated to travel and tourism industries were to be excluded from the dataset.

Two-term document matrices (TDM, i.e. a matrix where terms are assigned to rows and documents are assigned to columns.) were created using a preprocessed version (removal of numbers and stopwords, and stemmed) of all the abstracts in the dataset to identify the terms to be used in the search for documents strictly related to travel and tourism industries. One TDM for unigrams and another for bigrams allowed for the counting of the frequency of each term in the corpus. Using the "wordcloud" R package (Fellows, 2014), word clouds were elaborated to enable the analysis of the terms with a frequency equal to or above 10 (Figure 7 and Figure 8). Regarding unigrams, it is possible to verify that the most frequent terms are related to travel and tourism industries (e.g. "cancel" or "reserv") or are general terms (e.g. "model" or "system"). However, some terms like "patient" or "appoint" point to medicine and health industries. A similar pattern seems to emerge in bigrams, where the more frequent terms are related to the travel and tourism industries (e.g. "book limit" or "cancel noshow"), but others seem to be related to medicine and health (e.g. "miss appoint") or electronics (e.g. "papr reduct" or "multiplex ofdm")

Figure 7 – Unigram word cloud

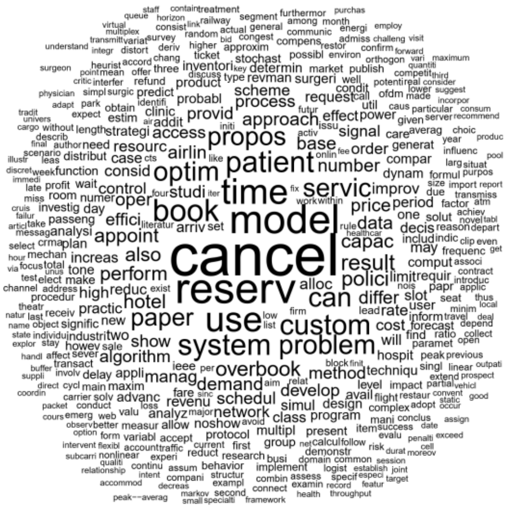
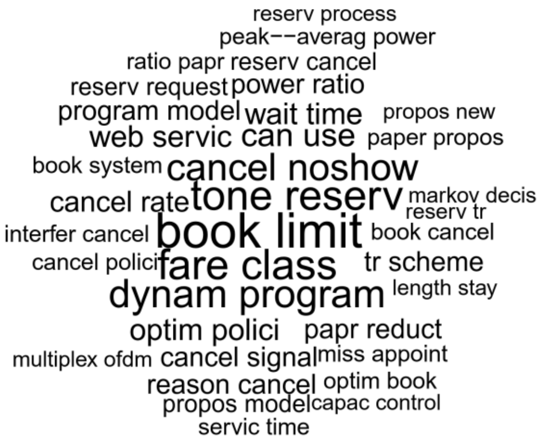


Figure 8 - Bigram word cloud

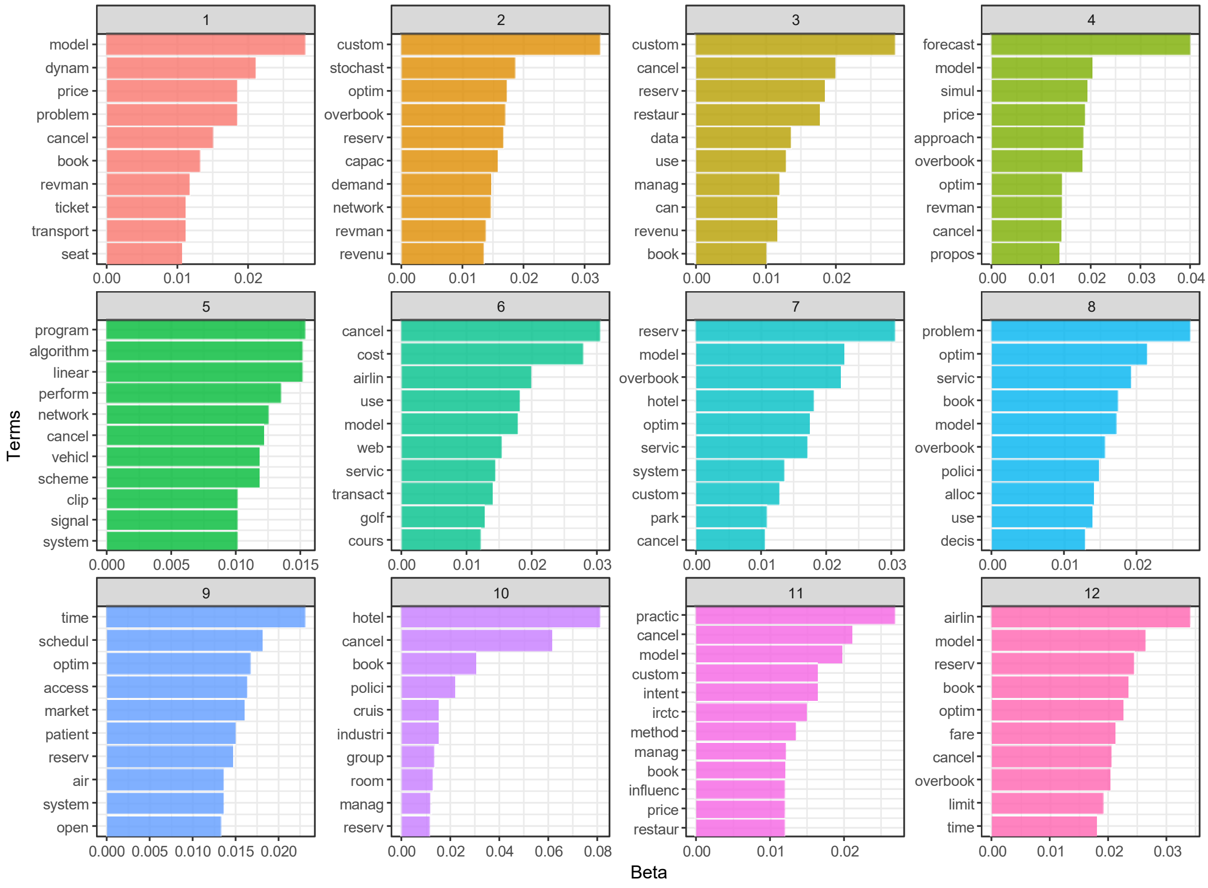


Considering the analysis of term frequency, the chosen filter terms to restrict the documents to be searched were as follows: tourism, hotel, airline, aviation, restaurant, golf, event, travel, transportation, railway, and "revman". A stemmed version of these terms was then looked for in preprocessed versions of the title, author keywords, index keywords, and abstract of the 194 documents in the dataset. Every document not containing one of the terms was excluded from the fused dataset, filtering the dataset to a total of 88 documents (Figure 5).

To understand which topics are covered by the resulting documents, we decided to employ the latent Dirichlet allocation (LDA) method (Blei, Ng, & Jordan, 2003) using the R package "topicmodels" (Grun & Hornik, 2011). LDA is the most popular and widely used method for topic modelling (Calheiros, Moro, & Rita, 2017), being a statistic model that groups text documents based on a classification given by computed measures representing the document's distance from a given topic and from the document to the known a priori (Arun, Suresh, Madhavan, & Murthy, 2010). Therefore, what should be the ideal number of topics was decided using the R package "ldatuning" (Nikita, 2016). This package uses four different methods to help decide the number of topics. Based on the results, the more adequate number of topics was determined to be 12. LDA was then applied to a corpus formed with the preprocessed versions of the title, author keywords, index keywords, and abstract of each document. Following an analysis of the top 10 terms identified in each of the 12 topics by the LDA beta, the Dirichlet prior on the per-topic word distribution (Figure 9) shows that at least topic 10 is related to both cancellations in hotels and cruises. Topic 4 seems to be on forecasting and modelling. Other topics, like 1 and 5, seem to relate more to methodological aspects.

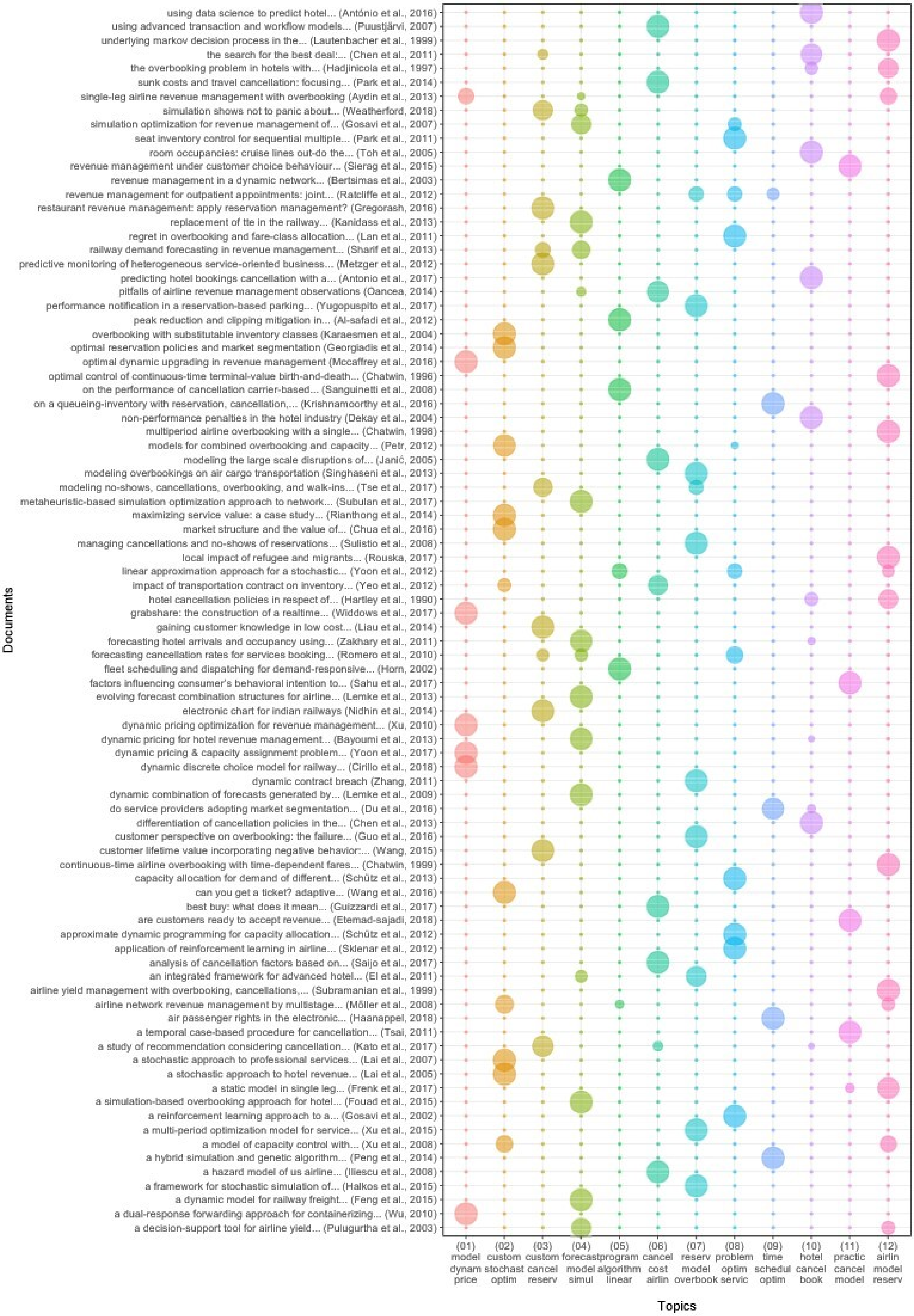
When looking at the probability of a document covering a topic, it is possible to identify which documents cover which topics by using the gamma distribution of LDA (Figure 10). An analysis of titles and topics revealed that some documents were not in the forecast/prediction modelling topic. For example, "Peak reduction and clipping mitigation in OFDM by augmented compressive sensing" (Al-Safadi & Al-Naffouri, 2012) is related to electronics, and "Local impact of refugee and migrants crisis on Greek tourism industry" (Krasteva, 2017) measures the impact of refugees on tourism (including cancellations, but not modelling them). Thus, it was decided to narrow the search even further. The new filtering was carried out through another automated search of terms. This time it was searched which documents had in the preprocessed title, author keywords, index keywords, or abstract a stemmed version of the terms "prediction" or "forecast". This final filtering resulted in a total of 13 documents (Figure 5).

Figure 9 - Top 10 terms per topic



Source: Authors.

Figure 10 - Documents probabilities per topic



Source: Authors.

3.3 Automated analysis

To continue with the determination of an adequate document dataset on booking cancellation forecast/prediction for travel and tourism industries, automatic disambiguation of author names was conducted. This enabled the identification of several authors whose name was written diversely in different documents (e.g. only with the first and last names or with the full name). The names were manually corrected for subsequent document analysis.

The PDF files for all 13 selected documents were manually downloaded from the documents' publisher websites or from scientific repositories. A new corpus was then created by including a preprocessed version of the full text of each document. This preprocessing included several normalisation processes, namely, case lowering, removal of numbers, removal of punctuation, removal of the non-informative terms previously identified, conversion of two-word terms to one-word terms (for the terms previously identified), and stemming.

The new corpus was used to classify documents by clusters and topics. Document clustering is perhaps the most commonly used method and is the determination of the number of clusters to be discovered (Kassambara, 2017). The R package "factoextra" (Kassambara & Mundt, 2017) was used to identify and analyse clusters. For determining the number k of expected clusters for the dataset, the "elbow" method and the "average silhouette" method were used. Despite using a weighted term frequency-inverse document frequency (tf-idf) DTM, results were k=1 for the "elbow" method and k=2 for the "average silhouette" method. This is probably associated with the small number of documents, which also explains the result for topic modelling (obtained with the "ldatunning" R package) that determined that the number of topics should be changed from 10 to 13.

3.4 Human analysis

Finally, to create a global view of the set of the 13 selected documents, the data and metadata analysis, document clustering, and full texts topic modelling results were interpreted and analysed together with an interpretation of the text resulting from manually reading the documents. The following sections are dedicated to the description of the output of this analysis.

4. Results and discussion

As previously discussed (Section 3.2), although the filtering of documents by terms related to travel and tourism industries was unable to capture all the documents from other areas, LDA topic modelling was revealed to be a good approach to understand the topics addressed by the 88 documents specific to booking cancellation. The two topics with more documents probabilistically assigned to them, topics 4 and 12, show "model" in the top three most relevant words. Documents associated with topic 4 consider forecasting of arrivals, occupation, and demand. Documents associated with topic 12 seem to be focused on airline models. Topic 7 also includes the word "model" in its top 3, and many documents associated with it consider issues related to overbooking. A reasonable number of documents are probabilistically assigned to topics 2, 3, 8, and 10. Topic 2 and topic 8 address different types of optimisation problems in service industries. Topic 3 seems to be related to the understanding of why customers cancel services. Topic 10 addresses different hotel cancellation-related problems. From the 13 documents finally filtered, 7 address the topics covered in topic 4 to some extent. That none of these documents address topics 2, 5, 6, and 9 is clearly because the terms in those topics mainly cover problems specific to airlines or other service industries; for example, some cover optimisation and scheduling problems.

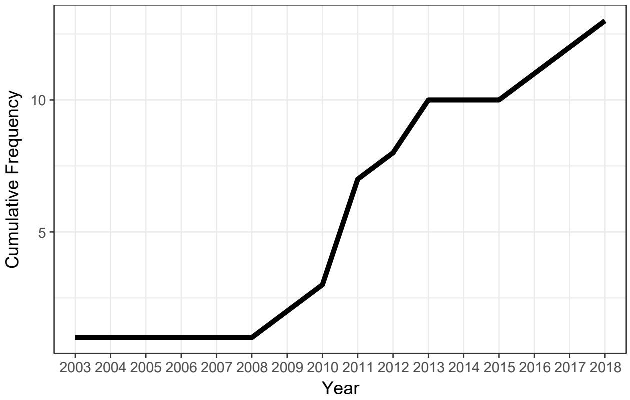
As can be seen in Figure 11 (and in Table 1), the first document on the topic dates from 2003 and only from 2009 onwards does the publication of documents continually increase.

Table 1 – Summary of the 13 final documents

Author (Year)	Type	Citations	Topics (Fig. 10)
Pulugurtha & Nambisan (2003)	Article	6	4, 12
Lemke, Riedel, & Gabrys (2009)	Conference paper	5	4
Morales & Wang (2010)	Article	21	8, 4, 3
Lan, Ball, & Karaesmen (2011)	Article	15	8
Tsai (2011)	Article	0	11
Gayar et al. (2011)	Article	24	7, 4
Zakhary, Atiya, El-Shishiny, & Gayar (2011)	Article	20	4
Metzger, Franklin, & Engel (2012)	Conference paper	23	3
Azadeh, Labib, & Savard (2013)	Article	1	4, 3
Lemke et al. (2013)	Article	4	4
Antonio, Almeida, & Nunes (2017c)	Book chapter	0	10
Antonio, Almeida, & Nunes (2017b)	Conference paper	0	10
Cirillo, Bastin, & Hetrakul (2018)	Article	0	1

Source: Authors.

Figure 11 - Documents published over the years for the final selection

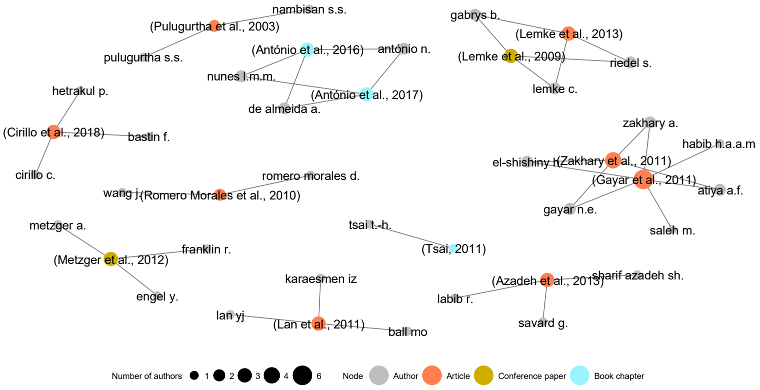


Source: Authors.

This helps to confirm Chen's claim that the number of papers on booking cancellation forecast/prediction is relatively low (Chen 2016), but it is important to remember that this is a relatively new subject of research.

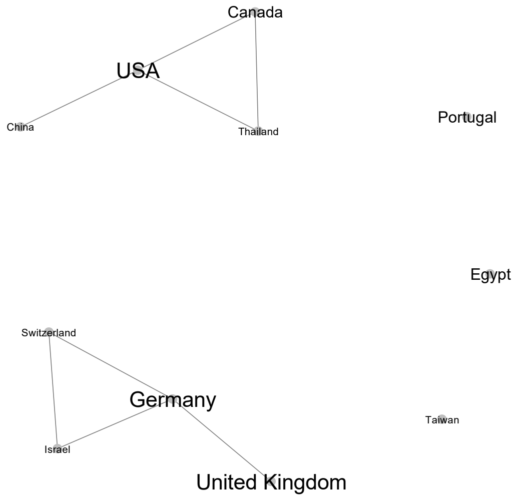
The network analysis shows that forecast and prediction is confined to a few groups of authors, with no collaborations in between (Figure 12). Only three groups show more than one publication on the topic. However, within some groups, there is collaboration between authors from different countries. As shown in Figure 13, only the groups of authors from Egypt, Portugal, and Taiwan do not show external collaborations. The countries with more collaborations are the USA, Germany, and the United Kingdom. Note that the size of the name of the country represents frequency.

Figure 12 - Authors' network



Source: Authors.

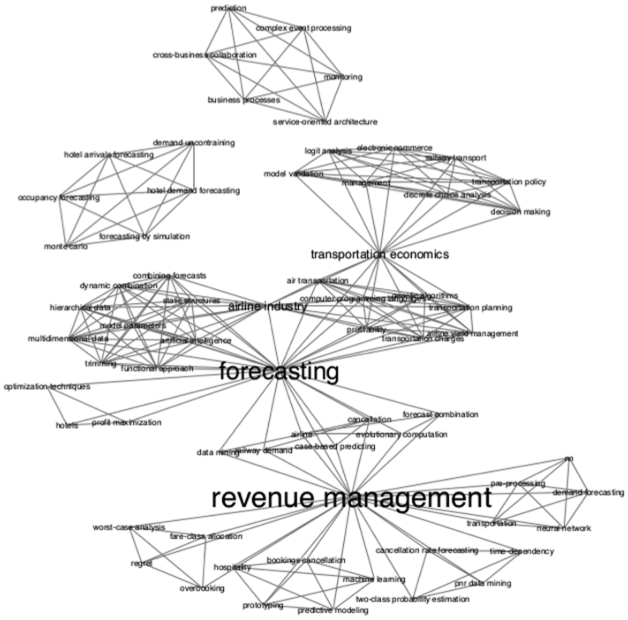
Figure 13 - Authors' countries network



Source: Authors.

The analysis of the keywords in the 13 selected documents shows that the two most frequent terms are "revenue management" and "forecasting", followed by "airline industry" and "transportation economics" (Figure 14). The use of other keywords is sparse. Like for the authors' network (Figure 12), the keywords network is composed of some isolated groups (Figure 14), suggesting that a study's topics are, to some extent, addressing different topics per group of authors.

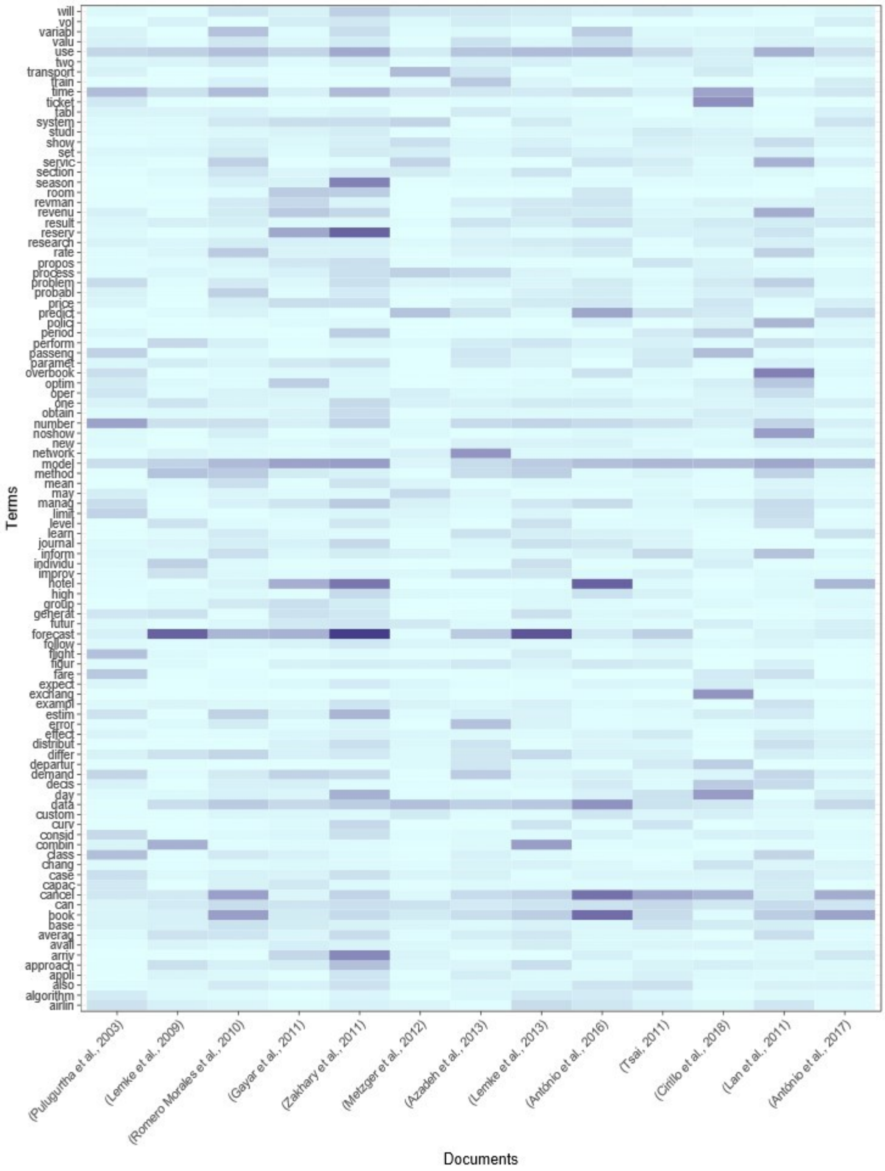
Figure 14 – Author and index keywords network



Source: Authors.

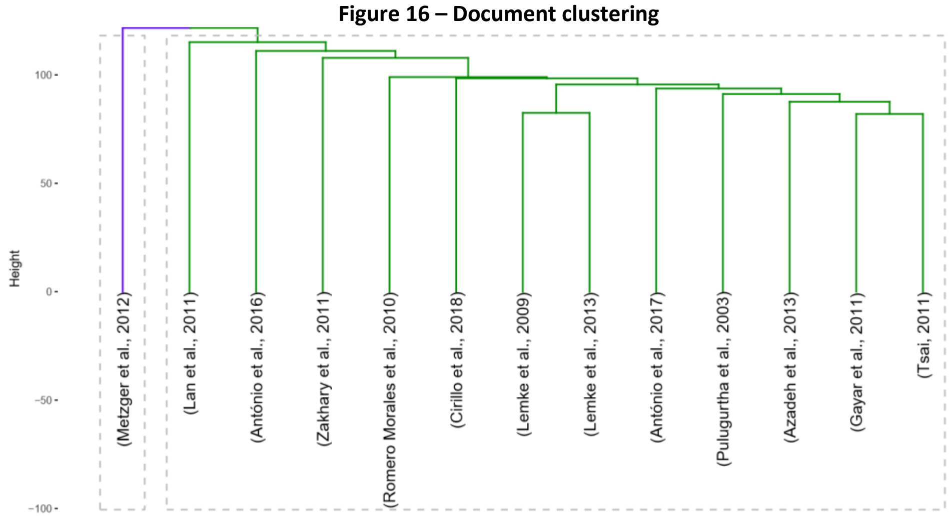
In fact, a heatmap showing the frequency of all stemmed terms in the corpus with a frequency equal to or above 100 (Figure 15) shows the topics each document deals with. For example, the term "forecast" presents high frequency in several documents; meanwhile, "predict" presenting a higher frequency is not so common (except for Lan et al. (2011)). This suggests that while some works address forecasting, others address prediction. Development or discussion of a model seems to be transversal to almost all documents since "model" is a highly frequent term in all documents except that by Metzger et al. (2012). The term "hotel" shows varying frequency, which might point to the document's specificity to the hospitality industry (Antonio et al., 2017b, 2016; Gayar et al., 2011; Zakhary et al., 2011). The suggestions arising from the analysis of the heatmap presented in Figure 15 are confirmed by reading the full texts of the selected documents. "Forecast" was not applied in the work by Lan et al. (2011). The authors intended to describe a model to optimise overbooking and fare class allocation instead of building a forecasting or prediction model. The works of Antonio et al. (2017b, 2017c), Gayar et al. (2011), and Zakhary et al. (2011) were the only ones focused on hotels while Lan et al. (2011), Lemke et al. (2009, 2013), Morales & Wang (2010), and Pulugurtha & Nambisan (2003) worked with airlines or employ airline data. Azadeh et al. (2013), Cirillo et al. (2018), and Tsai (2011) worked with railways. The only document that did not deal with industries related to travel and tourism is that by Metzger et al. (2012). This fact is also illustrated by the clustering analysis. The analysis of the two-cluster hypothesis shows a one-document cluster (Figure 16), which is precisely what Metzger et al. (2012) found related to logistics. From the 12 selected works identified for the travel and tourism industries, 8 are actually about modelling booking cancellation forecast/prediction. Lemke et al. (2009, 2013) employed real data from Lufthansa Systems AG to develop models to forecast cancellation rates as a way to improve demand forecast models. Morales & Wang (2010) employed hotel data in the development of a model to forecast cancellation rates while identifying variables relevant for a better understanding of cancellation drivers. Azadeh et al. (2013) developed a model to predict bookings and cancellations for a railway operator. Antonio et al. (2017a, 2017b) employed hotel data to develop models to predict each booking's cancellation probability, and, simultaneously, net demand. Tsai (2011) used railway data to develop a model to forecast the cancellation rate. Cirillo et al. (2018) developed a model to predict when railway customers would exchange or cancel their tickets. All of the previous works employed different methods: time series-based techniques (Lemke et al., 2009, 2013), economics-based techniques (Cirillo et al., 2018; Tsai, 2011), and more modern techniques usually applied in machine learning problems, like genetic algorithms, neural networks, or other advanced classification techniques (Antonio et al., 2017b, 2017c; Azadeh et al., 2013; Morales & Wang, 2010; Pulugurtha & Nambisan, 2003). Except the works by Antonio et al. (2017b, 2017c) and Morales & Wang (2010), who considered bookings cancellation estimation a classification problem, all other authors considered it a regression problem. The advantage of considering the problem a classification problem is that it is possible to calculate the cancellation rate (the regression measure) from the prediction of each booking´s likelihood to be cancelled(Antonio et al., 2017c). Although Morales & Wang (2010) argued that predicting the probability of a customer to cancel with high accuracy is difficult and that it would have no practical implications, Antonio et al. (2017b, 2017c) have shown the opposite. The authors showed that, in practice, the identification of bookings with a high likelihood of cancelling allows hotels to take measures to avoid the identified cancellations. Of the selected works, four considered cancellations but addressed other types of forecasting/prediction models. Pulugurtha & Nambisan (2003) developed a model to estimate the number of seats to allocate for each fare class. Gayar et al. (2011) developed a demand forecast model that generates demand scenarios to be used by an optimisation model. Zakhary et al. (2011) presented a model to forecast arrivals and occupancy. Lan et al. (2011) introduced a model for overbooking and fare class allocation. The remaining selected works related to only 5 of the 12 topics presented in Figure 10, namely, topics 3, 4, 8, 10, and 11. Thus, the abstracts of all documents shown in Figure 10 were read to understand if any of the documents was unduly filtered, but this turned out not to be the case. Overall, from the 12 documents related to travel or tourism industries, the documents from Antonio et al. (2017b, 2017c) and Morales & Wang (2010) show the most promising paths for research in booking cancellation prediction, that is, predicting each booking cancellation's likelihood and identifying cancellation drivers. Antonio et al. (2017b, 2017c) showed that by using advanced machine learning techniques, it is possible to reach an accuracy of 80% to 90% for the prediction of a booking's cancellation outcome. Moreover, Antonio et al. (2017c) and Morales & Wang (2010) showed that some of the features are of higher explanatory value than others (e.g. lead time, distribution channel, or agent).

Figure 15 - Terms frequency per document (frequency equal to or above 100)



Source: Authors.

Figure 16 – Document clustering



Source: Authors.

5. Conclusions

Data science methods were used to conduct a semiautomatic literature analysis of the research published about booking cancellation forecast/prediction for travel- and tourism-related industries. The objective was to synthesise research findings, including the understanding of the general research topics addressed by bookings cancellation research.

Although booking cancellation forecasting is of foremost importance to determine net demand in travel- and tourism-related industries, particularly in the hospitality industry, it was here shown that few works are addressing the development and testing of cancellation models. The existing works agreed on estimating cancellation rates as an important input for demand forecasting models. However, only two addressed the importance of predicting each booking's cancellation outcome (Antonio et al., 2017b, 2017c). Moreover, only three addressed the importance of forecast and prediction models in the identification of cancellation drivers (Antonio et al., 2017b, 2017c; Morales & Wang, 2010). As for the broader question about understanding the general topics addressed by research in booking cancellation, the LDA topic analysis revealed that the topics were mostly related to airline, hotel, and railway industries. The main topics addressed were the development of global demand forecast models, development of cancellation forecast/prediction models, development of overbooking models, development of fare-allocation models, and development of simulation and optimisation models.

This work makes several significant contributions. Methodologically, it confirms the importance of data visualisations as a tool for abstraction, time efficiency, and improvement of reviews. Examples are the identification of groups of collaborating authors and countries, terms to select as filters, topics addressed per documents, and others. Sharing the source code and the datasets employed, as well as detailing the steps of the experimental procedure together with the text mining techniques and tools, including the search strings applied in each activity, makes this work a replicable one and allows other authors to employ similar methods and techniques for literature reviews. It also confirms semiautomatic literature analysis as an excellent technique to overcome problems related to the availability of numerous sources of information, the need to conduct faster and less resource-intensive literature analysis and mitigate the effect of human biases.

Theoretically, the few works here identified highlight the fact that, despite the importance of booking cancellation forecast/prediction being recognised, there is still a need for further research on the subject. The present paper also shows that most of the existing works do not take advantage of advanced predictive methods, such as those linked to machine learning. Additionally, although most of the works employed real data, none used multiple data sources. All documents relied on a single data source (hotel property management systems data, airline bookings data, or railway booking data), and failed to use other data sources whose data could help explain cancellation drivers (e.g. weather, social reputation, competition prices, etc.). Furthermore, despite the value of prediction models to understand past behaviour, using prediction models to understand cancellation drivers is treated lightly. Understanding cancellation drivers can be important to improve overbooking and cancellation policies, which are two very important topics in RM research. Understanding cancellation drivers can help hoteliers improve cancellation policies. This would include more dynamic policies, that is, cancellation policies that could vary according to lead time, duration of stay, and other booking characteristics. By fostering adjustable and dynamic cancellation policies, hotels could also enhance net demand predictions, which in turn could reduce the necessity of overselling for compensating cancellations, thus lessening the risk of overbooking and incurring increased costs.

5.1 Limitations and future work

As with all analysis of literature, this work presents some limitations. The first stems from the research strategy used. Although an "automated search" allows for the quick identification of documents, it relies on the correct definition of search keywords and the electronic databases selected. Despite Scopus and WoS being two of the more well-known scientific databases, documents were not included in the analysis because they were not present in these databases or came from the so-called “grey literature”. Another limitation arises from the differences between electronic database formats and how information is presented for each publication; these issues make it challenging to automate certain activities, like the identification of authors, countries, or duplications. To avoid manual disambiguation or manual processing, automatic methods for performing such activities must be developed. Differences between the structure of electronic databases, such as the absence of specific fields in certain databases, are also hard to handle automatically. For example, WoS does not include a field with the references and thus does not allow for the inclusion of an automatic snowballing activity, that is, a process of analysis of the references of the selected documents. In fact, references analysis could potentially lead to the identification of other documents on the subject of investigation. To encourage automated literature analysis, electronic database providers should always include references in the search results. Finally, the last limitation is related to the keywords employed in the filtering of documents. Although document clustering found that one document was somehow “different” and lead to the understanding, it was not related to travel and tourism industries, the process to select keywords for the different filtering activities is still a time-consuming task. Future research should address the problem of defining appropriate keywords in a more automated way.

References

Ali, N. B., & Usman, M. (2018). Reliability of search in systematic reviews: Towards a quality assessment framework for the automated-search strategy. *Information and Software Technology*, *99*, 133–147. https://doi.org/10.1016/j.infsof.2018.02.002

Al-Safadi, E. B., & Al-Naffouri, T. Y. (2012). Peak reduction and clipping mitigation in OFDM by augmented compressive sensing. *IEEE Transactions on Signal Processing*, *60*(7), 3834–3839. https://doi.org/10.1109/TSP.2012.2193396

Antonio, N., Almeida, A., & Nunes, L. (2017a). Predicting hotel booking cancellation to decrease uncertainty and increase revenue. *Tourism & Management Studies*, *13*(2), 25–39. https://doi.org/10.18089/tms.2017.13203

Antonio, N., Almeida, A., & Nunes, L. (2017b). Predicting hotel bookings cancellation with a machine learning classification model. In *Proceedings from the 16th IEEE International Conference on Machine Learning and Applications* (1049–1054). Cancun, Mexico: IEEE. https://doi.org/10.1109/ICMLA.2017.00-11

Antonio, N., Almeida, A. de, & Nunes, L. (2017c). Using data science to predict hotel booking cancellations. In P. Vasant & K. M (Eds.), *Handbook of Research on Holistic Optimization Techniques in the Hospitality, Tourism, and Travel Industry* (141–167). Hershey, PA, USA: Business Science Reference.

Arun, R., Suresh, V., Madhavan, C. E. V., & Murthy, M. N. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In *Advances in Knowledge Discovery and Data Mining* (391–402). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-13657-3_43

Azadeh, S. S., Labib, R., & Savard, G. (2013). Railway demand forecasting in revenue management using neural networks. *International Journal of Revenue Management*, *7*(1), 18. https://doi.org/10.1504/IJRM.2013.053358

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

Bragge, J., Relander, S., Sunikka, A., & Mannonen, P. (2007). Enriching literature reviews with computer-assisted research mining. Case: profiling group support systems research (243a–243a). IEEE. https://doi.org/10.1109/HICSS.2007.209

Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management*, *0*(0), 1–19. https://doi.org/10.1080/19368623.2017.1310075

Chen, C.-C. (2016). Cancellation policies in the hotel, airline and restaurant industries. *Journal of Revenue and Pricing Management*, *15*(3–4), 270–275. https://doi.org/10.1057/rpm.2016.9

Chiang, W.-C., Chen, J. C., & Xu, X. (2007). An overview of research on revenue management: current issues and future research. *International Journal of Revenue Management*, *1*(1), 97–128.

Cirillo, C., Bastin, F., & Hetrakul, P. (2018). Dynamic discrete choice model for railway ticket cancellation and exchange decisions. *Transportation Research Part E: Logistics and Transportation Review*, *110*, 137–146. https://doi.org/10.1016/j.tre.2017.12.004

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, *34*(3), 1707–1720. https://doi.org/10.1016/j.eswa.2007.01.035

Denizci Guillet, B., & Mohammed, I. (2015). Revenue management research in hospitality and tourism: A critical review of current literature and suggestions for future research. *International Journal of Contemporary Hospitality Management*, *27*(4), 526–560. https://doi.org/10.1108/IJCHM-06-2014-0295

Fabbri, S., Hernandes, E., Di Thommazo, A., Belgamo, A., Zamboni, A., & Silva, C. (2013). Using information visualization and text mining to facilitate the conduction of systematic literature reviews. In J. Cordeiro, L. A. Maciaszek, & J. Filipe (Eds.), *Enterprise Information Systems* (Vol. 141, 243–256). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40654-6_15

Feinerer, I., & Hornik, K. (2017). tm: Text mining package (Version 0.7-3). Retrieved from https://CRAN.R-project.org/package=tm

Fellows, I. (2014). wordcloud: Word clouds (Version 2.5). Retrieved from https://CRAN.R-project.org/package=wordcloud

Feng, L., Chiam, Y. K., & Lo, S. K. (2017). Text-mining techniques and tools for systematic literature reviews: A systematic literature review. In *2017 24th Asia-Pacific Software Engineering Conference (APSEC)* (41–50). https://doi.org/10.1109/APSEC.2017.10

Gayar, N. F. E., Saleh, M., Atiya, A., El-Shishiny, H., Zakhary, A. A. Y. F., & Habib, H. A. A. M. (2011). An integrated framework for advanced hotel revenue management. *International Journal of Contemporary Hospitality Management*, *23*(1), 84–98. https://doi.org/10.1108/09596111111101689

Grun, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(11), 1–30. https://doi.org/10.18637/jss.v040.i13

Guerreiro, J., Rita, P., & Trigueiros, D. (2016). A text mining-based review of cause-related marketing literature. *Journal of Business Ethics*, *139*(1), 111–128. https://doi.org/10.1007/s10551-015-2622-4

Guo, X., Dong, Y., & Ling, L. (2016). Customer perspective on overbooking: The failure of customers to enjoy their reserved services, accidental or intended? *Journal of Air Transport Management*, *53*, 65–72. https://doi.org/10.1016/j.jairtraman.2016.01.001

Haneem, F., Kama, N., Ali, R., & Selamat, A. (2017). Applying data analytics approach in systematic literature review: Master data management case study. In *Frontiers in Artificial Intelligence and Applications* (Vol. 297, 705–715). Kitakyushu, Japan.

Hornik, K. (2017). NLP: Natural language processing Infrastructure (Version 0.1.11). Retrieved from https://CRAN.R-project.org/package=NLP

Ivanov, S., & Zhechev, V. (2012). Hotel revenue management–A critical literature review. *Turizam: Znanstveno-Strucnicasopis*, *60*(2), 175–197.

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning*. STHDA.

Kassambara, A., & Mundt, F. (2017). factoextra: Extract and visualize the results of multivariate data analyses (Version 1.0.5). Retrieved from https://CRAN.R-project.org/package=factoextra

Kimes, S. E., & Wirtz, J. (2003). Has revenue management become acceptable? Findings from an international study on the perceived fairness of rate fences. *Journal of Service Research*, *6*(2), 125–135.

Kitchenham, B. A., & Charters, S. (2017). *Guidelines for performing Systematic Literature Reviews in Software Engineering (version 2.3)* (EBSE Technical Report No. EBSE-2007-01). Durham, UK: Keele University.

Krasteva, R. (2017). Local impact of refugee and migrants crisis on Greek tourism industry. *Economic Studies Journal*, (4), 182–195.

Lan, Y., Ball, M. O., & Karaesmen, I. Z. (2011). Regret in overbooking and fare-class allocation for single leg. *Manufacturing & Service Operations Management*, *13*(2), 194–208. https://doi.org/10.1287/msom.1100.0316

Lee, M. (2018). Modeling and forecasting hotel room demand based on advance booking information. *Tourism Management*, *66*, 62–71. https://doi.org/10.1016/j.tourman.2017.11.004

Lemke, C., Riedel, S., & Gabrys, B. (2009). Dynamic combination of forecasts generated by diversification procedures applied to forecasting of airline cancellations. In *IEEE Symposium on Computational Intelligence for Financial Engineering, 2009. CIFEr '09* (85–91).

Lemke, C., Riedel, S., & Gabrys, B. (2013). Evolving forecast combination structures for airline revenue management. *Journal of Revenue and Pricing Management*, *12*(3), 221–234. https://doi.org/10.1057/rpm.2012.30

Lewis-Beck, M. S. (2005). Election forecasting: Principles and practice. *The British Journal of Politics & International Relations*, *7*(2), 145–164.

Liu, P. H. (2004). Hotel demand/cancellation analysis and estimation of unconstrained demand using statistical methods. In I. Yeoman & U. McMahon-Beattie (Eds.), *Revenue management and pricing: Case studies and applications* (91–108). Cengage Learning EMEA.

Matsuo, Y. (2003). Prediction, forecasting, and chance Discovery. In Y. Ohsawa & P. McBurney (Eds.), *Chance discovery*. Berlin, Heidelberg: Springer.

McGuire, K. A. (2017). *The analytic hospitality executive: implementing data analytics in hotels and casinos*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Metzger, A., Franklin, R., & Engel, Y. (2012). Predictive monitoring of heterogeneous service-oriented business networks: the transport and logistics case (313–322). IEEE. https://doi.org/10.1109/SRII.2012.42

Morales, D. R., & Wang, J. (2010). Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research*, *202*(2), 554–562.

Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*, *42*(3), 1314–1324. https://doi.org/10.1016/j.eswa.2014.09.024

Nikita, M. (2016). ldatunning: Tuning of the latent Dirichlet allocation model parameters (Version 0.2.0). Retrieved from https://cran.r-project.org/web/packages/ldatuning/ldatuning.pdf

Noone, B. M., & Lee, C. H. (2011). Hotel overbooking: The effect of overcompensation on customers' reactions to denied service. *Journal of Hospitality & Tourism Research*, *35*(3), 334–357. https://doi.org/10.1177/1096348010382238

Nunez-Mir, G. C., Iannone, B. V., Pijanowski, B. C., Kong, N., & Fei, S. (2016). Automated content analysis: addressing the big literature challenge in ecology and evolution. *Methods in Ecology and Evolution*, *7*(11), 1262–1272. https://doi.org/10.1111/2041-210X.12602

O'Neil, C., & Schutt, R. (2013). *Doing data science*. Sebastopol, CA, USA: O'Reilly Media.

Pan, B., & Yang, Y. (2017). Monitoring and forecasting tourist activities with big data. In M. Uysal, Z. Schwartz, & E. Sirakaya-Turk (Eds.), *Management science in hospitality and tourism: Theory, practice, and applications* (43–62). Apple Academic Press. Retrieved from http://www.crcnetbase.com/doi/pdfplus/10.1201/b19937-1

Park, J. Y., & Nagy, Z. (2018). Comprehensive analysis of the relationship between thermal comfort and building control research — A data-driven literature review. *Renewable and Sustainable Energy Reviews*, *82*, 2664–2679. https://doi.org/10.1016/j.rser.2017.09.102

Pulugurtha, S. S., & Nambisan, S. S. (2003). A decision-support tool for airline yield management using genetic algorithms. *Computer-Aided Civil and Infrastructure Engineering*, *18*(3), 214–223. https://doi.org/10.1111/1467-8667.00311

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Talluri, K. T., & Van Ryzin, G. (2005). *The theory and practice of revenue management*. New York, NY: Springer.

Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., & Coiera, E. (2014). Systematic review automation technologies. *Systematic Reviews*, *3*, 74. https://doi.org/10.1186/2046-4053-3-74

Tsai, T.-H. (2011). A temporal case-based procedure for cancellation forecasting: a case study. *Current Politics and Economics of South, Southeastern, and Central Asia*, *20*(2), 159–182.

Weatherford, L. R., & Kimes, S. E. (2003). A comparison of forecasting methods for hotel revenue management. *International Journal of Forecasting*, *19*(3), 401–415. https://doi.org/10.1016/S0169-2070(02)00011-0

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, *26*(3), xiii–xxiii.

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures*, *11*(4), 245–265. https://doi.org/10.1080/19312458.2017.1387238

Zakhary, A., Atiya, A. F., El-Shishiny, H., & Gayar, N. (2011). Forecasting hotel arrivals and occupancy using Monte Carlo simulation. *Journal of Revenue and Pricing Management*, *10*(4). https://doi.org/10.1057/rpm.2009.42

Received: 15.07.2018

Revisions required: 20.09.2018

Accepted: 22.12.2018