DOCTORAL DISSERTATION

Smart tourism: Artificial intelligence for adding value to tourism. A case-based study

Doctorado en Empresa, Internet y Tecnologías de las comunicaciones

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

Eleazar Caballero Sánchez

Las Palmas de Gran Canaria
2019

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

D. MIGUEL ANGEL FERRER BALLESTER COORDINADOR DEL PROGRAMA DE DOCTORADO EMPRESA, INTERNET Y TECNOLOGÍAS DE LAS COMUNICACIONES DE LA UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA,

INFORMA,

Que la Comisión Académica del Programa de Doctorado, en su sesión de fecha 31/10/ 2019 tomó el acuerdo de dar el consentimiento para su tramitación, a la tesis doctoral titulada "*Adding value to hotel management through artificial intelligence: A case-based study*" presentada por el doctorando D. Eleazar Caballero Sánchez y dirigida por el Doctor Agustín J. Sánchez Medina.

Y para que así conste, y a efectos de lo previsto en el Artº 11 del Reglamento de Estudios de Doctorado (BOULPGC 7/10/2016) de la Universidad de Las Palmas de Gran Canaria, firmo la presente en Las Palmas de Gran Canaria, a 31 de Octubre de dos mil diecinueve.

UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

ESCUELA DE DOCTORADO

Programa de doctorado:
Empresa, Internet y Tecnologías de las comunicaciones

Título de la Tesis

Smart tourism: Artificial intelligence for adding value to tourism.
A case-based study

Tesis Doctoral presentada por D/Dª Eleazar Caballero Sánchez

Dirigida por el Dr/a. D/Dª. Agustín J. Sánchez Medina

Codirigida por el Dr/a. D/Dª.

El/la Director/a,	**El/la Codirector/a**	**El/la Doctorando/a,**
SANCHEZ MEDINA AGUSTIN JESUS		CABALLERO SANCHEZ ELEAZAR -
(firma)	(firma)	(firma)

Las Palmas de Gran Canaria, a 12 de Octubre de 2019.

Education's purpose is to replace an empty mind with an open one

Malcolm Forbes

ACKNOWLEDGEMENTS

I would like to start giving my most sincere thanks to my supervisor Prof. Agustín Sánchez Medina for his guidance throughout this study. I am deeply grateful and sincerely thank you for your support and for sharing your expertise.

I would also like to express my warmest thanks to my high school colleagues: Sara, Javier, Aday, Fran … because despite the years and distance they are always there when I need them. Likewise, I would like to thank my university colleagues: Guillermo, Mario, Misael and Sonia, because I know who to phone when I need a beer and talk about life. Finally, I would like to express my deepest thanks to my work colleagues, Isabel, for teaching and guiding me from the moment I started working with her, and also to Elena, who always knows what to say to encourage me.

Last, but not least, I wish to express my warmest thanks to my maternal family, who has supported me since I can remember. Especially to my sister, Noemí, for her support when I needed it most, certainly, a person who I can blindly trust; and my mother, the person who fought for me and my sister in the most difficult years, the person I owe everything to.

Eleazar C-Sánchez

ABSTRACT

Growing competition in the tourism sector forces the companies involved to invest in new developments that generate value with the main aim of attracting new tourists and handling current reservations. Artificial intelligence-based models have proven to be very effective for this in several areas, and their application within the tourism industry is still being developed in multiple fields. This work is intended to contribute to tourism literature, as well as, explore the possibilities that artificial intelligence techniques offer when applied within the tourism industry. For this reason, this Ph. D. Thesis has been developed in a case-based format, addressing two main lines of research, the first one focuses on contributing to the development of an alternative tourism and the second on the improvement of hotel capacity management. More specifically, the first research topic addresses the forecast of night sky brightness with the aim of promoting the growing subsector of the astroturism, while the other topic attempts to forecast hotel cancellations using two different approaches, taking into consideration the cancellations made any time after the reservation was placed, and those notified very close to the entry day. The approaches presented in this work can be classified as developments within the topic of smart tourism, which is a growing area that makes use of information and communications technologies for improving aspects in the tourism industry.

The results obtained in this research are very impressive and, considering the quite different cases addressed in the document, it can be concluded that artificial intelligence-based techniques have a big future within the tourism industry.

RESUMEN

La creciente competencia dentro de la industria turística hace que las compañías del sector se vean en la necesidad de invertir en desarrollos tecnológicos capaces de atraer a nuevos turistas y gestionar las reservas ya creadas. En este sentido, los modelos basados en inteligencia artificial han probado ser muy efectivos en ámbitos muy variados y su aplicación para la resolución de problemas turísticos está aún en desarrollo en múltiples áreas. Este trabajo pretende contribuir a la literatura, así como, explorar las posibilidades de las técnicas de inteligencia artificial aplicadas a esta industria. Por esta razón, la presente tesis ha sido desarrollada en un formato basado en casos, abordando dos líneas de investigación principales, la primera con la intención de contribuir al desarrollo de un turismo alternativo, la segunda dirigida a mejorar la gestión hotelera. Más específicamente, la primera línea aborda la predicción del brillo estelar nocturno con el objetivo de promocionar el creciente sector del astroturismo, mientas el otro intenta predecir cancelaciones hoteleras a través de dos enfoques distintos; uno considerando aquellas cancelaciones a ser emplazadas en cualquier momento desde que se realiza la reserva, y otro, las cancelaciones de las que se advierte al establecimiento muy pocos días antes de la hipotética llegada del cliente. Los enfoques presentados en este trabajo pueden ser clasificados como desarrollos dentro del turismo inteligente (smart tourism), el cual se trata de un área en crecimiento basado en el uso de las tecnologías de la información y la comunicación para la mejora del turismo.

Los resultados obtenidos en este trabajo son realmente impresionantes, y teniendo en cuenta la gran diferencia existente entre las casuísticas tratadas, se puede concluir con que las técnicas basadas en inteligencia artificial tienen un gran futuro dentro del sector turístico.

Smart tourism: Artificial intelligence for adding value to tourism. A case-based study

Eleazar Caballero Sánchez

Programa de doctorado:

Empresa, internet y tecnologías de las comunicaciones

Director: Agustín. J Sánchez Medina

Las Palmas de Gran Canaria, October 2019

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Introduction – General overview

According to the World Tourism Organization, the tourism activity can be defined as "the activities of persons travelling to and staying in places outside their usual environment for not more than one consecutive year for leisure, business and other purposes”. The fact is that it is one of the most expanded industries in the world and generates income to nations through the consumptions of goods and services to tourists (e.g. importations for consumption or promotion of local products) and it allows the creation of new enterprises to provide a large number of complementary services to hotel chains, among others: transport, maintenance, information and communications systems, etc. The increasing trend of tourism is supported by a significant growth of this sector in recent decades, with the appearance of new touristic modalities and expansion to new destinations, matching or even out performing business volumes from sectors such as oil and gas or the automotive industry (World Tourism Organization, n.a). In 2017 alone the number of international tourist arrivals grew 7%, representing the highest increase since the world economic crisis in 2009 (World Tourism Organization - Tourism Highlights, 2018).

Nevertheless, the particularities of this industry force it to deal with constant uncertainty. Several external factors such as weather, natural disasters or political instability have an impact on the sector. As an example of this effect, the n*ational observatory of outbound tourism* promoted by important enterprises within the tourism sector, such as Amadeus, Iberia or Renfe among others, observes major uncertainty about the number of tourist arrivals to Spain in coming years due to the political situation in Catalonia and the impact of Brexit, as well as the effect of the recovery of tourism in traditional competing countries, such as Tunisia, Turkey or Egypt (Observatur, n.a). On the other hand, the product offered by the tourism industry is not stockpiled and it is forced to constantly adapt to demand and capacities, which complicates the planning process even more. For this reason, the involved companies have developed approaches to manage demand which have been developed over the years until the emergence of, what are known as, revenue management systems (RMS). The concept of revenue management emerged by the end of 1970 in the airline industry in order to maximise aircraft capacity, as well as, income, and thus, in turn, profit (Donaghy et al., 1995), which was then extended to other related industries, among others, the hospitality industry. Revenue management systems can be defined as "the maximisation of room revenue through the manipulation of room rates in a structured fashion, so as to take into account forecasted patterns of demand" (Jauncey et al., 1995, pp2), or "the formulation and profitable alignment of price, product, and buyer" (Donaghy et al., 1995, pp3). In the end, the main aim of RMS is to assist demand management through a detailed and structured methodology supported by a scientific approach and the required systems (Talluri & Ryzin, 2005). In this regard, management decisions relating to can be classified in three main categories (Talluri & Ryzin, 2005):

- Structural decisions: which involve the selling format (e.g. posted prices or negotiations), market segmentation, differentiation mechanisms (e.g. facilities or services) or marketing strategies and others.
- Price decisions: The price of each kind of product offered, markdown strategies (e.g. discounts related to business volume, special dates or VIPs), price presentation and others.
- Quantity decisions: which involves how to distribute the capacity across the different segments and channels of the hotel or overbooking policies, among others.

As the main aim of RMS is demand management, having relevant information in advance becomes crucial to hotels. The strong relationship between income and demand has resulted in forecasting and search for patterns becoming one of the most crucial topics within this sector and, in fact, one of the key aspects of a RMS. Forecasting in tourism has been performed using four kinds of models: time series models, econometric models, artificial intelligence (AI) models, which are all quantitative techniques, and judgmental methods that can be used for quantitative or qualitative forecasting (Song et al., 2019); however it is true that artificial intelligence based methods have gained importance in recent years. In fact, the use of IT technologies and their capacity to store a large amount of data has led to the development of advanced analysis techniques, such as data mining (Korte, 2013). Thus, the use of AI techniques has increased in recent years due to their excellent forecast capabilities (Song et al., 2019), outperforming more traditional models (Wu et al., 2017). Similarly, intelligent technologies have been also used to conduct differentiation mechanisms, as for example, the development of new tourist attractions. Enterprises are aware of the importance of creating additional value to remain competitive in an increasingly competitive marketplace, which, in addition, is one of the most important factors to encourage client loyalty (Minghetti, 2003). Hotels with smartphone applications, intelligent touch panels or augmented reality are just examples of how hotel chains have attempted to differentiate from competitors by providing new client experiences. This is known as smart tourism, which is a concept that refers to the application of information and communication technologies in order to develop innovative approaches that allow to improve tourism (Boes et al., 2015). In this regard, this work is intended to exploit one of the multiple lines of research within this area, data treatment.

For the reasons stated above, the main aim of the present Ph. D Thesis will be the creation of value within the tourism industry, in terms of providing accurate forecasts. With this goal in mind, this work will take place through the treatment of data using artificial intelligence-based techniques. For this purpose, two considerations related to tourism were studied. One activity was developed using more than 10,000 real booking records provided by a four-star hotel located in the south of Gran Canaria (Spain). Similarly, it was intended to explore new AI-based tools to help to differentiate hotels and promote growing subsectors in the tourism industry, which is the case of astrotourism.

1.2 Motivation

The fierce competition within the hotel and lodging industry, together with the emergence of probabilistic models based on artificial intelligence, which offer more accurate forecasts than the traditional ones, have forced hotel chains to use them to remain competitive in the uncertain, volatile, and changing environment in which they develop their activity. At the same time, the intense competition and the continuous diversification of tourist activities have led to the fact that traditional sun and beach tourism is no longer enough to satisfy all of the client demand, as some seek other complementary activities. In this regard, the present Ph. D Thesis is focused on using AI-based models for solving some of the problems that this sector is forced to deal with. Specifically, the main line of research of this work is based on the study of different probabilistic models and data treatment with the aim of generating an additional value to hotels by both, supporting the decision-making process with regards capacity management and contributing to the improvement of complementary demand, so that, it can be managed more efficiently.

1.3 Aims and objetives

The main challenge of the present Ph. D Thesis was to prove the usefulness that the artificial intelligence may have in the tourism sector. Accordingly, two quite distinct problems were planned from the beginning: the first one addresses the improvement of hotel booking management and the second aims to contribute to the development of an alternative tourism. Along these lines, the general objective of this Ph. D Thesis is the analysis of data through AI techniques for creating value within the tourism sector and help participating companies more competitive.

On the one hand, the idea is to support the hotel inventory management through cancellation forecasting, for which real booking records will be employed, and on the other hand, providing differentiation mechanisms able to attract clients, as is the case of nature-based tourism. Specifically, the chosen case was the astrotourism by providing probabilistic models for night sky brightness forecasts with the available conditions.

While the first line of research is intended to support the decision-making process related to hotel capacity management, the second investigates the night sky brightness in order to establish a methodology that allows forecasting the quality of the nocturnal skies, promoting this growing tourism subsector. This study would allow to create a digital information system for tourists to know the ideal places and times to enjoy the experience of contemplating celestial phenomena, thus optimising their holiday time. This last line of research makes an important contribution to literature, in that, as far as it is known, there are no precedents on the application of probabilistic models for this kind of time series.

The general objectives (G.O) and specific objectives (S.O) of the present work are detailed below:

G.O 1: Identify opportunities for implanting AI-based algorithms that allow improving hotel capacity management:

> **S.O 1.1:** Model hotel cancellations through AI techniques.
>
> **S.O 1.2:** Establish a generalist procedure for forecasting hotel cancellations using only the most common variables requested during the booking process by online travel agencies.

S.O 1.3: Establish which are the best techniques among the different AI-based used in the research.

S.O 1.4: Identify those cancellations made close to the time of service, which can be considered as critical cancellations.

G.O 2: Develop the foundations of an information system, based on AI techniques, that allow tourists to know about the best places and times to do astrotourism.

S.O 2.1: Model night sky brightness time series.

S.O 2.2: Compare traditional time series analysis techniques with AI-based techniques.

1.4 Structure of the dissertation

The present document is divided into 6 chapters, beginning with the current introduction and finishing with the conclusion of the work and future prospects. This Ph. D Thesis has been conducted by solving three main problems presented in the tourism industry; first, related with the use of AI techniques for generating additional value, addressing the case of forecasting night sky brightness for astrotourism and second, two additional cases related with hotel capacity management, one of them attending the case of forecasting general hotel cancellations and another case for forecasting hotel critical cancellations. With the aim of properly identifying each case, these have been entitled and listed below:

Case 1: Astrotourism and night sky brightness forecast: First Probabilistic Model Approach

Case 2: Predicting Hotel Booking Cancellations: An artificial intelligence approach

Case 3: Identifying Hotel Critical Cancellations using artificial intelligence

It is worth highlighting that *case 1* has been published in the indexed journal of *Sensors,* which is ranged as Q1 according to JCR requirements, while *Case 2* is currently under review in the indexed journal of International Journal of Hospitality Management, which is classified as D1 (Decile 1) according to JCR requirements.

As it can be appreciated, the cases addressed in this work are of a very different nature and have a very different way of proceeding within the hospitality industry, hence a specific introduction of each case studied in this Ph. D Thesis is detailed, and each chapter has been subdivided in order to segregate properly the content of each case. *Chapter 2* contains a specific introduction to each case, in which the particularities of each and their situation is contextualised. *Chapter 3* then presents a review of the related literature, that is, a summary of the previous models and the specific considerations for each problem. In *Chapter 4* the methodology employed for each problem is described, that is, the data treatment and how the models were trained and validated. *Chapter 5* presents the results achieved through the use of the methodology applied in each case. Tables and graphs are shown with the aim of supporting the provided explanations and results are discussed. Finally, *Chapter 6* contains the conclusions of the research, with a summary of the contributions and implications of the tourism industry, as well as, future prospects.

2. INTRODUCTION TO EACH CASE

2.1 Introduction - Case 1

Tourism is one of the most expanded industries throughout the world and one of the most changing and rapidly developing economic sectors. Recently, this industry has focused on specific branches that offer multiple tourism experiences and activities (Soleimani et al., 2018); among them is ecotourism, one of the fastest growing areas, which has drawn special interest. The amount of research into this discipline has increased in the last years with the appearance of new institutes, journal articles and books based on this topic. Its growth has even generated elective or core subjects as well as specific programmes in many universities (Weaver & Lawton, 2007). This expansion has led to the creation of specialised products for different kinds of interests in what was considered a unique concept of tourism until some years ago. “According to its definition, ecotourism can involve both cultural and environmental tourism” (Candrea & Herţanu, 2015, pp4) and so, celestial ecotourism can be considered in terms of “learning and the maximisation of positive ecological and sociocultural impacts” (Weaver, 2011, pp3). As D. Weaver notes, celestial tourism, astrotourism, astronomical tourism or, less frequently, star tourism consists in observation of celestial phenomena which can be appreciated in a natural way, excluding planetariums, which present astronomical recreations in artificial spaces, astronomy conventions or similar. Farajirad and Beiki (2015, pp2) defined astronomical tourism as “when an individual interested in the sky (an amateur or professional astronomer or anyone interested in the field) travels to a location other than their residence to study and explore the wonders of the sky”. It is also considered as nature-based tourism, where visitors find it attractive travelling to natural places, mainly for two reasons: firstly, they can experience unknown environments and places, and secondly, guests find specific facilities which cannot be found in their places of origin (Najafabadi, 2012). However, only locations with the necessary conditions can be developed for this kind of tourism, therefore, destinations with dark skies at night that are free from artificial light pollution, can be a potential area for developing related tourism business. Several organisations have appeared in favour of preserving dark skies such as the "Astronomy and World Heritage" project which was created by UNESCO (United Nation Educational, Scientific and Cultural Organization) with the aim of "raising awareness of the importance of astronomical heritage worldwide and to facilitate efforts to identify, protect and preserve such heritage for the benefit of humankind" (Portal to the heritage of the Astronomy, n.a), The Starlight foundation by the Institute of Astrophysics of the Canary Islands, The Globe at Night project funded by the National Optical Astronomy Observatory of the United States, which is an international citizen-science campaign that fights for the awareness of light pollution or the International Dark-Sky Association (IDA) founded in 1988, among others. These sources show multiple sources of information ranging from places where celestial phenomena can be observed to multimedia resources to interactive data maps. In 2001, IDA initiated a certification programme aimed at global communities, parks and protected areas mainly with the intention of preserving dark skies through six different kinds of designations. Likewise, the Starlight project started a certification programme for those places with excellent night sky vision. Both help to keep safe dark night places, but at the same time, they also encourage and promote astrotourism.

People's interest in visiting specific astrotourism destinations is motivated by different celestial phenomena, which can be classified by addressing the sun's position (Weaver, 2011). According to this, the classification of stars, skygazing, meteor showers or comets, among others, would be classified as nocturnal, while rainbows or solar/moon eclipses would be considered diurnal; finally, sunrises and sunsets would be treated as crepuscular phenomena. Astronomical activities can also be classified responding to market subsegments, specifically five types: visits to astronomical observatories, destinations for observing auroras, public astronomy parks with dark skies, amateur astronomy associations which offer public programmes and other providers that offer related products (e.g., private or public astronomical facilities) (Collison & Poe, 2013). Understanding the visitor's motivation is essential for focusing on the precise target group by providing solutions for their needs and therefore being able to develop and improve products and services in suitable destinations (Najafabadi, 2012).

Simultaneously, Information and Communication technologies (ICT) and digital technologies have revolutionised the tourism industry and have introduced opportunities for developing new applications which support tourism activities. In this manner, "smart tourism aims to apply intelligent perceptions of all kinds of tourism information, such as tourism resources, tourist economics, tourism activities, and tourism participants, among others, to develop the acquisition and adjustment of real-time tourism information through mobile Internet or Internet terminal equipment" (Guo. Y et al., 2014, pp3). Therefore, technology should be treated as an integral part of a tourism subsector and contribute to its development by adding value to this activity. It brings the opportunity to improve tourism services (e.g., personalise features such as temperature/humidity room control according to customer preferences) or even create new services in order to offer different customer experiences (e.g., augmented reality), so that the sector becomes more professional and competitive. Moreover, technology can be considered part of everyday life and its use is increasing day by day. Proof thereof is the fact that the undertaking of some daily tasks cannot be imagined nowadays without a smartphone or a terminal with Internet access. According to this reality, private and public organisations have invested in technological platforms which support a real-time system to be accessed from any end-user device. Several investigations have been carried out in this area regarding the use of such data in an effort to model and forecast future behaviour, which is a challenging task when weather-dependent variables are predicted and which is the case of the present research. With the aim of contributing to the development of astrotourism, this research proposes an accurate method for forecasting the darkness of night skies. It should be considered that visitors usually have to travel long distances, generally in the early hours of the night, to get to suitable places for celestial tourism which are typically far away, and so, if it so happens that at the time, the sky conditions are not appropriate, they leave the place with the feeling of having "wasted their time". Furthermore, some visitors carry special instruments for contemplating celestial phenomena and consequently this experience would have a negative impact in these cases. This research proposes a short-term forecast of night sky brightness (NSB) based on real field data with the aim of providing advance information about night sky quality in order to increase the guarantees for tourists being able to contemplate good celestial views and enjoy the experience, which is a crucial factor for encouraging client loyalty.

To perform this task, open datasets, which collect real NSB data provided by the Globe at Night programme, were used for training and testing a first model approach. The Globe at Night programme is an international programme which monitors sky brightness through a network of Sky Quality Metres (SQM) worldwide. This represents a first step towards the development of a tourism application that not only allows visitors to get information about celestial touristic places, but also provides them with NSB forecasts, so that it may add value and promote this growing tourist sector. Specifically, two probabilistic models, Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN), have been trained and tested for establishing an accurate forecasting method of NSB.

2.2 Introduction - Case 2

Tourism is one of the most expanded industries in the world but it also suffers one of the most rapidly changing environments. Its importance in the global economy does not raise any doubts and it's growing trend seems to be sustained (World Tourism Organization - Tourism Highlights, 2018, pp4). In terms of economic development, the tourism industry generates income to nations through the consumption of goods and services by tourists, taxes, development of enterprises and employment opportunities, among others. However, the particularities of this sector mean it faces constant uncertainties. In this regard, the tourism industry is subject to multiple external factors that may have a significant impact on income, such as political instability, weather and natural disasters to mention but a few. Furthermore, the hotel industry offers a product, which does not become stockpiled. Therefore, organisation and planning are essential and having an accurate forecasting tool is of paramount importance. Considering three kinds of horizons, forecasting may be divided into long-term forecasting which helps to plan for global infrastructure and services (e.g. airports or highways), medium-term forecasting which may help with market analysis (e.g. establish sales strategies) and short-term forecasting which may provide more operational flexibility; (Gunter & Önder, 2015; Hassani et al., 2017). Along these lines, hotel industry revenue is strongly affected by demand so having an accurate forecasting system becomes critical. As stated by Kourentzes et al. (2017, pp1), forecasting demand plays a crucial role in modern organisations because of its impact on a variety of "business decisions, from operational, to tactical, to strategic level, such as capacity planning, resource planning, advertising and promotional planning or tactical production planning, among others". However, hotel demand is difficult to forecast (Pan & Yang, 2017), because of its complex and dynamic behaviour (Ostaijen et al., 2017), which makes it a very challenging task (Li et al., 2018). As long as hotel accommodation demand varies and reservations may be cancelled, an accurate method to determine effective hotel occupancy is crucial for decision making in terms of successful revenue management (Haensel & Koole, 2011). Management decisions are clearly influenced by customer demand, but not just regarding arrivals as both arrivals and cancellations are the main components of effective demand or net-demand (Rajopadhye et al., 2001). In this regard, it is estimated that about 20% of income is lost because of not considering cancellations as part of room reservation management systems (Sierag et al., 2015). However, literature available on cancellation forecasting within the industry is underdeveloped (Zakhary et al., 2011; Antonio et al., 2017a; 2019b) and little is known about the reasons that lead guests to cancel or how to avoid it (Hajibaba et al., 2016). Moreover, Antonio et al. (2019b) noted that only two publications focused on the importance of identifying which individual is likely to cancel. Following these two publications, only one more by Antonio et al. (2019a) has been released.

In recent decades the cancellation problem has become even more dramatic with the increasing use of Internet for placing bookings. Among the multiple reasons for cancelling a reservation, such as illness, bad weather or natural disasters, the ease of looking for new opportunities and cancelling previous reservations, as well as the possibility of booking several places at the same time to keep options open until making a final decision, have a big impact on cancellations (Antonio et al., 2017b; Chen et al., 2011).

This research contributes to the hotel and lodging industry by proposing and testing an empirical model based on artificial intelligence for forecasting hotel cancellations using personal name records (PNR). One of the main aims of this case is to shed light on the usefulness of PNR data for forecasting individual hotel cancellations after academic discussion arose on the research carried out by Romero Morales & Wang (2010), who concluded that it was not possible and Antonio et al. (2017b) & Antonio et al. (2019a), who concluded the opposite. In addition, another novelty of this approach is that individual hotel booking cancellations are forecast using a reduced number of variables in comparison with other researches (Antonio et al., 2017a; 2019a), achieving a similar high success rate, thus avoiding the need for large historical booking datasets by either building complementary variables through additional calculation based on database queries or getting access to external databases. Furthermore, it should be noted that this study has been performed with the most commonly-used variables that web booking platforms such as Booking, Tripadvisor or Trivago among others, may collect for the hotel industry. In fact, one of the main assets of this approach is the simplification of the procedure for building the dataset, as well as, the dataset itself, which makes for a faster and simpler training phase of the probabilistic models, meaning models can be trained more frequently, thus allowing a better following of market trends. For this reason, additional variables that could be extracted from the individuals after historical analysis of the database were not considered, as this would have increased the complexity for the hotels to use this procedure. Finally, another novelty in the proposed model is that genetic algorithms are used for configuring the structural parameters of the artificial neural network.

2.3 Introduction - Case 3

One of the most significant issues that hotel managers face is trying to match the hotel's capacity with demand. In the hospitality industry the product cannot be stored, which forces hoteliers to deal with demand and the limited number of rooms in a specific time window in such a way that each unoccupied room does not result in a loss of revenue (Chu, 2009). This leads to hoteliers attempting to increase their revenue by maximising occupancy, which involves dealing with future demand. However, as already cited, demand is subject to several external factors, such as weather, political stability, high competitiveness and others, which make it difficult to forecast. Reservations already placed may be cancelled, and this adds even more complexity to the planning and organising process of the hotel's capacity. Indeed, the importance that cancellations have within the hotel and lodging industry is not just reflected in terms of inventory management, but also in pricing strategies (Chen & Xie, 2013). It is important to remember that, in reference to incomes, cancellations suppose a revenue loss that at times may account for about 20% (Sierag et al., 2015). The huge importance that cancellations have for the lodging industry has led some authors to talk about the analysis of "net-demand", instead of just demand (Rajopadhye et al., 2001). Bearing in mind that hotels work with reservations for future services which may be cancelled, requested bookings do not reflect the real number of services to be provided by the hotel, therefore, "net-demand" represents the number of reservations requested minus the number of cancellations (Romero Morales & Wang, 2010; Antonio et al., 2017a; 2019b). This approach allows segregating the problem of demand, so that, instead of treating it as a whole, apparent demand and cancellations can be studied separately. Therefore, researches can focus on one part of the problem, in this case, the analysis of cancellations, which is a critical aspect for the hotel and lodging industry.

The impact that cancellations have on hotel chains has resulted in the development of strategies designed specifically with the goal of reducing cancellations. One such strategy is overbooking, which consists in accepting more bookings than the hotel has capacity for relying on the fact that some cancellations will take place. The main aim of this strategy is to match the number or clients who do not appear at the time of service (no-show), last minute cancellations as well as other cancellations notified in advance (Ivanov & Zhechev, 2011), so that hotels can avoid having idle capacity as far as possible. If the number of arrivals is below the inventory, hotels lose income because of unsold rooms, while in the opposite case, a hotel's capacity is not large enough to assume the demand and the establishments incur a loss of revenue because of having to relocate guests, but this strategy can also affect their reputation and corporate image. Other strategies, such as cancellation policies attempt to encourage customers to cancel in advance (Chen & Xie, 2013) through the imposition of restrictions within a specific period before the time of service (Law & Wong, 2010). This is common practice in the hospitality industry (Chen et al., 2011) and helps to reduce the number of no-shows and last minute cancellations (Zakhary et al., 2011). On the other hand, cancellation policies have a negative impact on the hotel's corporate social reputation, and at the same time may create a discouraging effect on clients (Antonio et al., 2017a). For these reasons, having information in advance about cancellations is crucial for hotel management and, different approaches have been developed with this in mind (Antonio et al., 2017a; 2019a; Falk & Vieru, 2018; Romero Morales & Wang, 2010). Accurate cancellation forecasts may support management in the decision-making process, as well as in the design of optimal strategies to reduce the impact that cancellations have on income. Nevertheless, the literature that address hotel cancellations is underdeveloped (Antonio et al., 2017a; 2019b; Zakhary et al., 2011) and this situation becomes even more critical for literature about short-term hotel demand forecasting, which has less presence than other revenue management related researches (Pereira, 2016).

The period of notice for a cancellation must be considered a critical aspect, as cancellations placed close to the time of service produce a particularly high loss in revenue (Chen et al., 2011; Koide & Ishii, 2005). These cases leave hotel management with little margin to react and may result in unsold inventory or substantial discounts having to be made on price (Antonio et al., 2017a). This has become even more critical in recent years because of the increasing use of online travel agencies for hotel room bookings, which has led to customers making several reservations before finally choosing one and cancelling the rest (Antonio et al., 2017a; Chen et al., 2011). Moreover, e-commerce allows customers to easily compare different offers and even read about the experience of previous clients, thus increasing the risk of cancellation (Koide & Ishii, 2005). Likewise, the growth of last-minute bookings encourages customers to take advantage of more economical offers to the detriment of previous reservations that are cancelled, with this having a direct impact on cancellations.

Previous research has managed to forecast which individuals are likely to cancel with a very high level of accuracy, above 90% which is the case of Antonio et al. (2017a) or the *Case 2* of present work, however there is no research addressing the forecasting of individual hotel cancellations made close to the time of service. In order to fill this gap and, considering that this kind of cancellation generates high revenue loss (Chen et al., 2011; Koide & Ishii, 2005), the aim of this research is to forecast individual hotel cancellations made close to the time of service through artificial intelligence techniques based on real personal name records (PNR) provided from a four-hotel partner located in one of the most touristic places in Spain, the island of Gran Canaria. As in the previous case, for this propose the most commonly requested data by booking platforms (e.g. Booking, Tripadvisor or Trivago) will be used, avoiding complementary variables to be extracted from database queries; client's identification will not be used, as this would require more time and computational resources. Neither will external data sources be employed for the same reasons. Therefore, the methodology presented in this work allows us to simplify the forecasting process in terms of building the dataset and the data itself, which represents one of the main advantages of this approach. It also leads to more frequent training so that hoteliers are better able to follow the market trend, which of course is especially important for short-horizon forecasts.

In this regard, this approach attempts to forecast individual cancellations likely to be made very close to the entry day, from 4 day to 7 days in advance, and which can be considered “critical cancellations” in that they leave management with no time to react.

3. LITERATURE REVIEW

3.1 Literature review – Case 1

Weather-dependent variables are challenging to forecast because of so many uncertainties that are often not expected. In this specific case, nocturnal sky brightness can be affected by clouds (Kocifaj, 2007), dust particles in the atmosphere (Garstang, 1991; Ściężor & Kubala, 2014) or even wild animals in uncontrolled remote areas which can block the sensor view. Moonlight also affects the SQM measurements by reducing the NSB levels, which is not appropriate for night sky observations. Puschnig et al. (2013) studied the effect of the moonlight and cloud conditions on the NSB measurements, concluding that moonlight affects the NSB values considerably, but that cloudy conditions are even brighter, especially in those places near bright cities. This means that NSB values are also affected by the lunar cycle and therefore periodicity in the NSB related to the lunar phase was observed in the measurements carried out in this study. Additionally, the effect of the clouds on the NSB values has reversed the previous concept of darkness because of the rapid change in the use of artificial light at night in the Anthropocene epoch (Kyba et al., 2015).

In this section, previous physically based models for sky brightness radiation are reviewed, as well as some probabilistic models for seasonal time series.

3.1.1 Physically-based models

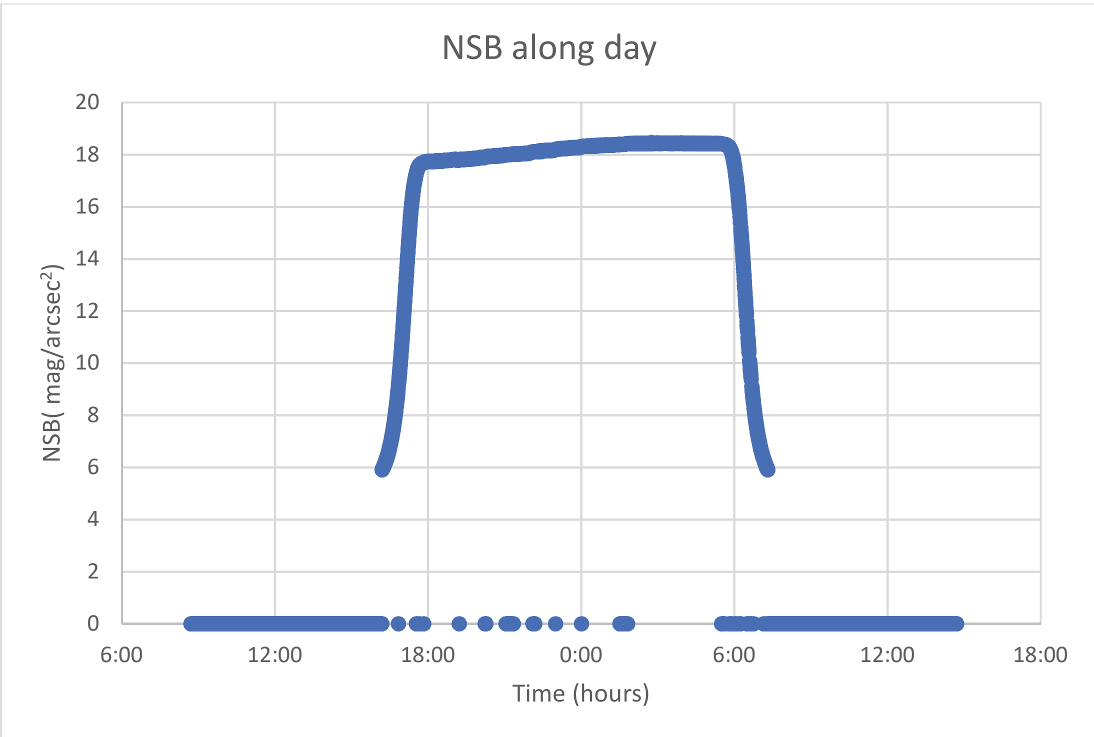
In 1986, Garstang published a general physical model that calculates the sky brightness at a specific point on earth based on multiple light sources. This research concludes that NSB predictions are satisfactory for cities outside observers, as well as for small city centres, but the model does not perform well for big cities. Two years later, the same author published a model with the purpose of predicting NSB variation due to seasonal changes, concluding that the "results give an idea of the size of likely seasonal effects on NSB for good astronomical sites", but more data should be used for understanding how its behaviour can be attributed to seasonal effects or short-term variations because of local causes (Garstang, R. H., 1988, pp2). Garstang's light propagation model has been successfully tested (Garstang, 1986; Garstang, 1988; Garstang, 1991) and used in other research, such as the creation of the first World Atlas of artificial NSB (Cinzano et al., 2001) in which light propagation models are used in conjunction with satellite radiance data and other similar ones (Duriscoe et al., 2018). Garstang's model is a good approximation, but results can be improved (Solano-Lamphar, 2018) and more complex models have been developed in this area. In 2007, Miroslav Kocifaj proposed a light-pollution model which considers cloudy and cloudless night skies. As this author noted, common models are focused on astronomical observations, which are usually carried out in cloudless conditions so that parameters such as altitude and spectral reflectance of a cloud layer are contemplated. The same author proposed further theoretical models for light pollution under clear sky conditions, which were successfully verified with experimental data (Kocifaj, 2014). It was concluded that except for big cities nearby or multiple bright cities around the measuring point, urban light could be deducted considering a single scattering. Later, a sky glow model was published for determining the city's emission function, which was verified with synthetical data, achieving consistent results (Kocifaj, 2017). Some restrictions of this model are that only clear sky conditions can be considered and radiance measurements should be in a range of 1–2 city diameters and several tens of city diameters.

These generalist physically based models can be used in multiple scenarios, such as the preliminary evaluation of new observatories, however, the continuous change of local conditions does not permit them to establish an accurate short-term forecast. In fact, as Slingo & Palmer (2011) stated, minimal changes in weather and climate predictions can dramatically affect the solution so that predictions must be based on probabilistic models instead of a single deterministic solution.

3.1.2 Probabilistic model

To the best of our knowledge, there is no literature available on the use of probabilistic models applied to NSB time series forecasting. For this reason, solar radiation time series forecasting has been reviewed as a homologous case; as long as both temporal series are weather-dependent, the seasonal component plays a crucial role and the curves that represent its behaviour present large similarities (Figure 1). It can be appreciated that during diurnal hours, NSB values cannot be measured, while during nocturnal hours, NSB values rise, becoming stabilised and decreasing again.

Figure 1. NSB curve of a random day in 2017 taken from the station at the National Astronomical Observatory of Japan.



Traditionally, research in solar irradiance time series models has been developed using Autoregressive (AR), Autoregressive Moving Average (ARMA), Autoregressive

Integrated Moving Average (ARIMA) and Markov chain methods, which are based on probability estimation. However, these methods do not estimate well when stochastic components become higher, so other artificial intelligence (AI)-based methods have been applied in those cases, achieving better results (Chen et al., 2013).

Some authors proposed and tested several solar radiation models, such as ARIMA, Unobserved Components Models (UCM), regression models, transfer functions, neural networks and hybrid models, and have found that the ARIMA model shows better results than the rest because of its improved capacity to capture the diurnal cycle, though other methods show better results when using higher resolutions (around 5 minutes) (Reikard, 2009). Other investigations used historical data of solar radiation, temperature and sky conditions by applying fuzzy logic first and then neural networks in different sky conditions (Chen et al., 2013). They achieved good results but concluded that the sky condition does not seem to have a significant impact on the forecast. Smart persistence, Artificial Neural Network (ANN) and random forest have been compared in order to predict solar radiation hourly within the subsequent six hours with the random forest showing the best performance (Benali et al., 2019). As might be expected, the more hours one tries to forecast, the more inaccurate the prediction.

Some authors proposed a forecast method using solar data collected in three different locations with diverse meteorological variabilities (Fouilloy et al., 2018). They established a comparison of several methods, ARMA, ANN or regression tree-based methods, such as random forest, among others. It was concluded that Autoregressive Moving Average and multilayer perceptron were the most efficient for weak variabilities, for medium variabilities the Autoregressive Moving Average and multilayer perceptron were the most efficient, while for high variabilities, tree-based methods showed the best results.

Other research presented a short-term cloud forecast based on irradiance historical data collected in a radiometric network in Almeria (Spain) (Caballero et al., 2018). When applying a recurrent neural network known as Long Short-Term Memory (LSTM), they concluded that this method shows improvement in comparison with other common methods used for time series analysis.

3.1.3 Sky Quality Metre: Sensor Technology

For this research, NSB behaviour over the period of a year was modelled using field data. The characteristics of the technology sensor are presented in this section.

Research related to the use of artificial light during night periods and its consequences has become far more common in recent years. This trend "can be attributed to several factors, including: recognition of the impacts of artificial light on ecology and health, increasing amounts of artificial light in the environment, improved quality of imagery from space, the current global change in lighting technology, and increasing quality, ease, instruments, and methods for measuring light at starlight intensities" (Hänel et al., 2018, pp1). The most common indicator of the night darkness condition is given by the NSB, which refers to the narrow-band spectral radiation of the night sky generally measured in the astronomical magnitude system, instead of SI standard units, mag/arcsec^2 (magnitude per square arc-second) (Duriscoe et al., 2018). This is a logarithmic scale in which values increase with darker skies.

The SQM (Sky Quality Metre) is a commercially available NSB photometer with a pocket-size design that allows the quantification of the quality of the night sky (Cinzano, 2005). This is a portable low-cost device originally developed by the Canadian manufacturer Unihedron with the aim of bringing to the general public, mostly amateur astronomers, the possibility of quantifying the sky conditions (Cinzano, 2005; Hänel et al., 2018), however, it has been widely used for astronomy-related research (Bará, 2016; Birriel & Adkins, 2010; Hänel et al., 2018; Posch et al., 2018). The SQM "determines the amount of light scattered in the direction of the observer by the air column located above" (Bará, 2016, pp3), thus atmospheric conditions affect the measurement as well as the impacts from natural sources such as moonlight and other celestial bodies. It is known for having a rough linear response between 6 and 19 mag/arcsec^2, its deviation is under 2.6%, the field of vision is about 20 degrees and the maximum measurable value is 22 mag/arcsec^2, but it is not especially designed to measure properly in extremely low intensity environments (Kocifaj et al., 2018). As a reference, the darkest place on earth shows values of about 22 mag/arcsec^2, values beyond 19 mag/arcsec^2 may be found in rural areas, while bright cities show values between 16 and 17 mag/arcsec^2. More examples are collected in the following references: Bará (2016), Hänel et al. (2018) & Kocifaj et al. (2018).

The equipment is composed of a crystalline silicon (c-SI) photodiode TSL237 manufactured by Texas Advance Optoelectronic Solutions (TAOS), which combines a silicon photodiode and a current-to-frequency converter. The output signal is a square signal whose frequency is proportional to the intensity of the incident light. Additionally, it is equipped with an infrared blocking filter (HOYA CM-500) and only visible light is measurable (Pravettoni et al., 2016; Posch et al., 2018). Among multiple versions, the SQM-LE model is provided with an Ethernet connection, which allows measuring and saving data to a terminal. Regarding the measuring stability of the SQM, the device has been used for long periods of time, resulting in a small annual drift (den Outer et al., 2015). Moreover, periodical maintenance tasks are usually focused on the glass window cover, more than on the optical sensor (Bará et al., 2019).

One limitation of the SQM is that it gives an average value of the receiving light in the measurable frequency range, no matter what the colour of the light is. Sánchez de Miguel et al. (2017) identified that SQM readings vary depending on the colour of the receiving light, and although these authors conclude that SQM is a good instrument, they suggest that colour of light should be considered in the light pollution studies. In addition, the increasing use of solid-state lighting technology (LED) for street illumination has led to an increase of “blue” light pollution against the “orange” light generated from classical high-pressure sodium lamps, which may cause a variation in the perception of NSB measured by the SQM (Kyba et al., 2017; Kyba, 2018).

Other authors have proposed some improvements to the original design, which is the case of TESS-W (Telescope Encoder Sky Sensor – Wireless). This device was created within the EU-funded STARS4ALL project and proposes the addition of a dichroic filter, which is a very accurate colour filter, with the aim of extending the band pass to the red range (Zamorano et al., 2017). Additionally, a WIFI module was included so that readings may be taken from a moving vehicle. Another case is the PePSS (Photomultiplier-equipped Portable Sky Scanner) developed by the Slovak Academy of Sciences and Comenius University, which uses a sophisticated photomultiplier technology that is able to read in extreme dark conditions (Kocifaj et al., 2018).

Other one-dimension alternatives, such as dark sky metre phone applications which use the camera device or solar-cell-based light metres based on the measurements of photoelectric currents, may also be used for general NSB measurements (Hänel et al., 2018). Similarly, there exists two-dimensional equipment based on image capturing. Depending on the application, different kinds of camera are available; as an example, commercial digital cameras have been used for tracing the dynamics of skyglow by differential photometry, with the conclusion that it is a remarkable tool for this application (Jechow et al., 2018). In this regard, further analysis and comparison were carried out in a later paper with several light pollution measurement techniques (satellite image, one-dimension devices, image-based measurement), concluding that vertical plane images provide valuable additional information in comparison with other all-sky imagery (Jechow et al., 2019).

In the context of this research, SQM equipment is more suitable for tourism destinations in terms of cost and ease of installation and its accuracy is precise enough for forecasting the night sky quality, more so in suitable destinations for astrotourism, which are usually in remote areas. The NSB dataset measured with an SQM-LE network (Figure 2) provided by the Globe at Night programme was used for training and testing a probabilistic model.

Figure 2. Sky Quality Meter - LE, source: Unihedron company.

3.2 Literature review – Case 2

This section is composed of two subsections. In the first the difficulties and needs within the lodging industry are explained, while in the second, the different forecasting techniques used within the hospitality industry are presented.

3.2.1 Difficulties and needs for the hotel and lodging industry

There is a strong association between demand forecast and revenue management (Tse & Poon, 2015). As long as hotels are required to manage room occupancy within an uncertain environment, they are exposed to unclear incomes and forced to assume business risks. Moreover, unexpected reduced demand often generates a crisis in the hospitality industry because of the high sensitivity to fluctuations in demand (Yüksel, 2007). In addition, demand uncertainty does not only affect the organisation and scheduling of occupancy, but also internal issues such as budget planning, which is highly dependent on future demand forecasting (Tang et al., 2017). However, demand may not be properly forecasted if the cancellation rate is not taken into account, hence, the variable net-demand should be considered (Rajopadhye et al., 2001), which is defined as “the number of demand requests minus the number of cancellations” (Romero Morales & Wang, 2010, pp1). Consequently, hotels affront this situation by implementing their own approaches in order to handle the associated risk management, such as overbooking strategies, cancellation policies or pricing strategies. On the one hand, overbooking strategies consist in accepting reservations over and above the capacity of the establishment, assuming that some bookings will fail. Nevertheless, extra costs may be incurred when the actual hotel occupancy exceeds the capacity due to guest compensation or relocation, which may also lead to a negative impact on the reputation. On the other hand, cancellation policies try to mitigate revenue loss because of cancellations, which are specially high when referring to last minute cancellations and no-shows (Chen et al., 2011). According to Zakhary et al. (2011) cancellations are found to decrease dramatically if penalties are imposed for cancelling beyond a certain day. However, imposing rigid cancellation policies can affect not only, corporate social reputation but also income because it has a discouraging effect on clients or, due to the application of significant discounts on price (Antonio et al., 2017a). Other strategies, such as price wars are discouraged due to the fact that they may affect the business strategy in the long-term (Gehrels & Blanar, 2013). In any case, the effectiveness of the different prevention approaches varies across the tourist segments (Hajibaba et al., 2016).

A reliable and accurate cancellation forecast may help in the managerial decision taking process by reducing the risk of cancellation as well as helping establish a proper cancellation policy or pricing strategy. Moreover, Pan & Yang (2017) explain that the tourism sector needs accurate forecast performance in specific destinations to benchmark their properties and optimise operations. Additionally, they state that as competition increases, accurate short-term forecasts become essential for hotel managers. Accordingly, Koupriouchina et al. (2014, pp2) state than when hotels face elevated levels of risk and distress (e.g. intensified competition or highly volatile markets), “more pressure is placed on the revenue manager to ensure that the forecasts are accurate” and reliable. So, by reducing uncertainty in future net-demand, occupancy rates can be increased while costs depending on idle capacity can be handled more efficiently. Likewise, an accurate cancellation model may prevent hotels from implementing rigid cancellation policies and overbooking strategies which have a negative influence on revenue and reputation (Antonio et al., 2017a).

Finally, it should be highlighted that some issues have arisen in recent decades that have given even more importance to accurate cancellation forecasting. Information technologies have changed customer behaviour and have made it even more difficult to predict future demand and cancellation rates. Now customers have more information about the establishments and the services they offer, for example, they can read previous customer experiences which makes it easier to compare different offers. Web portals also make it easier to book and cancel hotel services, which has encouraged people to place several bookings on similar dates in different hotels, looking for the most convenient options, only to finally choose one of them and cancel the rest (Chen et al., 2011). Consequently, demand performed by websites seems to increase, but cancellation rates do also. Secondly, another phenomenon that has influenced net-demand forecast is the growth of last-minute bookings whereby customers attempt to take advantage of these kinds of opportunities by postponing their booking. This causes a reduction in the length of the booking window and has an impact on forecasting accuracy (MacCarthy et al., 2016). In the same manner, such policies lead to other reactions; the probability of cancellations by guests who have already booked may also increase as they may be inclined to change to a more economic option, which generates more cancellations. In this way, if last-minute chances are offered by the hotel itself or by competing companies a few days before the time of service, the probability of cancelling increases as the time gets closer.

3.2.2 Methods and techniques for demand and cancellation forecasting

Looking for patterns and forecasts is always supported by previous experience in the service, the challenge is how to perform them in the best way considering “data availability, time horizons and objectives” (Lee et al., 2008, pp2). In this section, previous approaches regarding demand and cancellation forecasting methods applied within the hospitality industry are reviewed. These techniques are classified in qualitative and quantitative techniques; the main characteristics of both are presented in the following subsections. In this section, the most relevant literature is presented, while for a more extensive review, specific research in this field such as Song & Li (2008) or more recently Song et al. (2019) could be consulted.

3.2.2.1 Qualitative techniques.

Qualitative techniques employ a team of experts who determine tendencies and probabilities based on available data, own experience and knowledge in the field. These techniques are recommended when unprecedented changes are to come and therefore, historical data does not contain information about future events, are unsuitable or not sufficient enough to perform an appropriate forecast. For example, long-term forecasts with large and/or extraordinary changes (Lee et al., 2008) such as the growing interest in nature-based tourism or short-term forecasts in which unprecedented events are expected to have an impact on the business (Uysal & Crompton, 1985), such as the emergence of new competitors or natural disasters. Among the most relevant techniques, Delphi and scenario writing are the most popular (Lin & Song, 2015). However, qualitative techniques have less presence in literature (Song & Li, 2008) and do not get accurate results if they are not based on quantity (Yüksel, 2005). As an example of qualitative techniques applied within the hotel and lodging industry, Moutinho & Witt (1995) used a consensus approach to forecast long-term tourism environments. More recently, Kaynak & Cavlek (2007) applied the Delphi technique in order to forecast the potential tourism market in Croatia, while Lee et al. (2008) forecast the demand for the International Expo tourism held in Korea in 2012.

Other authors have proposed mixed approaches, which attempt the combination of quantitative and qualitative techniques in a “quasi-Delphi process” by the integration of statistical methods and judgement of experts with the aim of forecasting tourism arrivals (Tideswell et al., 2001). Other related research papers propose the use of quantitative methods in order to forecast hotel demand and average nights of stay, which is subsequently adjusted by experts periodically (Yüksel, 2005).

3.2.2.2 Quantitative techniques.

Quantitative approaches require the existence of sufficient and appropriate historical data (Uysal & Crompton, 1985). These techniques are an optimal option if past information can be quantified and past patterns can be reasonably extrapolated to the future (Lee et al., 2008). This section reviews the most relevant quantitative forecasting models, which are categorised, according to Peng et al. (2014) and Song & Li (2008), in non-causal time-series models, econometric models and models based on artificial intelligence. It is worth mentioning that most of the cases reviewed are focused on tourism destination forecasting, while research focused on forecasting hotel booking arrivals have less presence in literature (Lee, 2018).

Non-causal time-series models attempt to reveal future patterns based on historical data. The Integrated Autoregressive Moving Average model (ARIMA) has been the most widely time-series model used for demand forecasting in the past decades, although seasonal ARIMA (e.g. SARIMA) models have increased in popularity over the years because of the strong relationship between tourism and seasons (Song & Li, 2008). Chu (2009) used the ARMA-based methods in the context of predicting tourist demand, as represented by the number of world-wide visitors to Hong Kong, Japan, Korea, Taiwan, Singapore, Thailand, the Philippines, Australia and New Zealand. In this research, ARIMA-based models were applied, concluding that all methods provided a good performance using monthly and quarterly data sets. Cho (2003) evaluates the application of three time-series forecasting techniques: exponential smoothing, univariate ARIMA, and Elman’s Model of Artificial Neural Networks (recurrent neural networks); to predict travel demand from different countries to Hong Kong. This study concludes that neural networks are the best method for forecasting visitor arrivals, especially those series without obvious patterns. Claveria & Datzira (2010) introduce consumer expectations in time-series models with the aim of analysing their usefulness in the forecast of tourism demand applied to the four main visitor markets (France, the UK, Germany and Italy) in Catalonia. The paper uses combining qualitative information with quantitative models: Auto Regressive (AR), Auto Regressive Integrated Moving Average (ARIMA), Self-Exciting Threshold Auto Regressions (SETAR) and Markov Switching Regime (MKTAR) models. In addition, models are evaluated for different time horizons (one, two, three, six and 12 months). Conclusions support that ARIMA and Markov Switching Regime models outperform the rest of the models and models that consider consumer expectations do not give best results for all time horizons analysed. With regards hotel demand forecasting, Pfeifer & Bodily (1990) applied a space-time ARMA (STARIMA) approach with the aim of forecasting hotel arrivals for 8 different hotels belonging to the same hotel chain in a single metropolitan US city. STARIMA assumes that a special dependency among multiple points exists and attributes more weight to the closer ones. They finally concluded that STARIMA outperformed one single ARIMA time series model.

On the other hand, econometric models can be classified into static models, such as the traditional regression method, gravity models or the static Almost-Ideal Demand System (AIDS); and dynamic, such as Vector Auto Regressive (VAR), Time Varying Parameters (TVP) or the Error Correction Models (ECM) (Peng et al., 2014). Song et al. (2011) examine the factors that influence the demand for hotel rooms in Hong Kong to generate quarterly forecasts of demand and to assess the impact of the financial/economic crisis. The paper uses econometric approaches to calculate the demand elasticity and its corresponding confidence intervals. Both indicators are then used to generate interval demand predictions.

As examples of cancellation forecasting, Falk & Vieru (2018, pp2) studied the factors that influence cancellation behaviour with respect to hotel bookings. In this study, variables such as length of stay, hotel, category, booking time or arrival month, among others, were used. The probability of cancellation is estimated by a probit model with cluster adjusted standard errors at the hotel level. Results show that cancellation rates are higher for online bookings than offline bookings and travel agency bookings. Additionally, they found that "booking lead time and country of residence play the largest role, particularly for online bookings".

As noted by Wu et al. (2017), artificial intelligence techniques have been used in tourism and hotel demand forecasting with satisfactory performance. These authors note that Artificial Neural Networks (ANN) are the most frequently used method, although Support Vector Machines (SVM) or fuzzy methods have also been applied in this field. Along these lines, Claveria et al. (2015) propose three different architectures of artificial neural networks with the aim of forecasting tourist arrivals to Catalonia attending different time horizons (one, three and six months) and main visiting nationalities, concluding that multi-layer perceptron and radial basis function networks have a good performance. Huang (2014) also applied Artificial Neural Networks with a back-propagation architecture for forecasting tourism demand at a resort in Taiwan. This research used several local and international variables for the purpose, such as unemployment rate, international oil prices or the number or foreign visitors to Taiwan among others, concluding that it was an excellent method. Later, Huang & Hou (2017) applied similar methodology adding genetic algorithms for optimising ANN settings with the aim of forecasting the sales revenue of a travel agency, achieving good results. Other related studies, for instance the research conducted by Moutinho et al. (2008) uses a neural network based fuzzy time series with the aim of forecasting the arrival of Chinese tourists to Taiwan. They conclude that this method outperformed previous research in which only fuzzy time-series were applied (Jeng-Ren et al., 1998) and others in which the same methodology was applied but only with the maximum degree of memberships (Huarng et al., 2007). Along these lines, Yu & Schwartz (2006) used two artificial intelligence forecast methods -fuzzy time series and grey theory- to predict annual

U.S. tourist arrivals. They suggest that given the complexity and cost associated with the application of these two methods, it is imperative to compare their performance with the accuracy of more traditional and easier methods of forecasting. More recently, Hu et al. (2019) modelled tourism demand by incorporating neural networks into Grey-Markov models to forecast the number of foreign tourists using historical annual data from the Taiwan Tourism Bureau and China National Tourism Administration. The paper confirms that the proposed model outperforms other Grey-Markov models.

In terms of hotel cancellation forecasting, there is less amount of literature that addresses this topic. Romero Morales & Wang (2010) forecast cancellation rates for services booking revenue management using data mining. They used 14 variables, such as price of the booking, room type, channel used to make the booking or reservation system used. They stated that tree-based methods and kernel methods are the most popular for hotel cancellation forecasting, specifically, they note that Support Vector Machine (SVM) is the most notable method used for this purpose. Huang et al. (2013) used Back Propagation Neural Networking (BPN) and General Regression Neural Networking (GRNN) for forecasting booking cancellations, concluding that both methods revealed high potential for this purpose. Later Antonio et al. (2017a) applied diverse two-class classification algorithms: Boosted Decision Tree, Decision Forest, Decision Jungle, Locally Deep Support Vector Machine and Neural Network; with the aim of forecasting cancellation rates using data sets from four hotels located in the resort region of the Algarve (Portugal). For this study they use variables such as number of previous bookings not cancelled, previous stays, distribution channel or days of week of booking dates, among others, concluding that machine learning algorithms, specifically decision forest, are good methods for modelling hotel cancellations.

3.3 Literature review – Case 3

This section is composed of two subsections. In the first one the relationship between hotel revenue management and forecasting is explained, while in the second previous forecasting approaches are presented for both, hotels and airlines, in which PNR data are used.

3.3.1 Revenue management and forecasting

One of the main aspects for maximising hotel revenue management lies in the efficiency of the organisation and planning procedures for the available rooms, not an easy task because of the uncertain environment that the sector is exposed to. Among others, economic crisis, inflation, environmental changes, wars, regulatory changes, new client demand or technological changes are factors that management must take into consideration in this industry (Yüksel, 2005).

Moreover, as already mentioned, an unexpected reduction in demand often generates a crisis in this sector because of the high sensitivity to fluctuations in demand (Yüksel, 2007). For this reason, it is essential to understand the environment in which hotels operate as well as develop a strategy for future room allocations (Mubiru, 2014). In this context, performing accurate forecasts is necessary for optimising operations, as well as, supporting the decision-making process. As a general overview, accurate forecasts help managers in medium- and long-term decisions not only for determining hotel polices, human resources required according to workloads or budget planning , but also for assisting in the development of short-term occupancy schedule (Gunter & Önder, 2015; Hassani et al., 2017). Bearing in mind that forecasting models are based on historical data (Uysal & Crompton, 1985), a number of investigations have encouraged management to consider the importance of having a reliable revenue management system through which past data can provide value to the organisation (Zhang et al., 2017). However, as mentioned earlier, modelling cancellations in the hospitality industry is a huge challenge because of the highly volatile and uncertain environment of this sector (Yüksel, 2007). In fact, little is known about cancellations drivers and how to prevent them in this industry (Hajibaba et al., 2016), and this situation becomes even more dramatic for short-horizon cancellations, because the guest's motivations are constantly changing as the service time approaches (Romero Morales & Wang, 2010).

Originally for revenue management, forecasts were based on seasonal data (e.g. the month, week or weather), mostly because it was the only information available (Romero Morales & Wang, 2010). However, more recently, personal name records (PNR) have been used for this (Tang et al., 2017). They contain a wide range of information about the customers which is collated at the time a reservation is made; such as preferences, number of customers, nationality and other personal details. In the context of hotel cancellations, this approach allows to know more about each customer and forecast not just anonymous cancellation rates, but determine which individuals are likely to cancel (Antonio et al., 2017a; 2019a). This kind of information is very valuable for hotel chains that are developing and investing in intelligent systems to provide accurate forecasts (Zhang et al., 2017). And this has prompted the use of "more mathematically sophisticated optimisation engines" (Weatherford, 2016, pp1). Indeed, while traditional methods such as explorative methods (e.g. time series analysis) have been widely used to forecast within the hotel and lodging industry, in recent years the amount of research using artificial intelligence based models has increased because of the excellent forecasting capabilities they present (Song et al., 2019), achieving better results than traditional models (Wu et al., 2017). In this regard, several investigations carried out within the industry have concluded that AI-based methods outperform traditional ones (Burger et al., 2001; Chen & Wang, 2007; Cho, 2003; Law, 2000; Li et al., 2018).

3.3.2 PRN-based forecasts in the tourist industry and ensemble methods

The use of PNR data for forecasting within the tourism industry is relatively new (Gorin et al., 2006) and research in this area concludes that when using this information a more accurate model can be built. In this section, PRN-based forecasting approaches addressing cancellation and no-shows within the airline and lodging industry are reviewed.

For the airline industry, most published research papers address the no-show problem (Antonio et al., 2019b). In this regard, Garrow & Koppelman (2004) applied a multinomial logit model using disaggregate PNR data in order to forecast no-shows, early standbys and large standbys, concluding that by using PNR data and itinerary information is possible to build more accurate models. Later, Gorin et al. (2006) propose a cost-based model to forecast no-show rates in the airline industry in which no-show rates were estimated assuming they followed a normal distribution and, in a second stage, PNR data are used to adjust these forecasts. In this research they conclude that by using this model, income could increase between 15 and 18 percent for the applied time window. Other strategies propose the mix of a traditional model, based on the analysis of non-causal time series, with PNR-based models. This is the case of the research published by Neuling et al. (2004) who compared exponential smoothing results with forecasts performed with tree-based models and a blended solution. It was found that exponential smoothing outperformed tree-based methods in the first stages, but it changed when more data were used for that matter. However, the proposed blended model improved overall. Tree-based models have been also used by Lawrence & Cherrier (2003) who applied C4.5 decision-tree, a segmented Naive Bayes algorithm and an ensemble aggregation method with the aim of forecasting no-show rates in the airline industry with the results showing the aggregation method outperformed conventional models, and improved income rates between 0.4% and 3.2%. More recently, Cao et al. (2010) compared three different machine learning techniques, using real data from a Chinese airline to forecast no-show rates. They concluded that the most accurate model was the logistic regression, followed by artificial neural network and the decision tree model, respectively.

In the hospitality industry recent research papers address the problem of hotel cancellation forecasts by employing PNR data. This is the case of Romero Morales & Wang (2010) who compared several data mining techniques with the aim of forecasting hotel cancellations by applying different time frames. In their research, they used decision tree-based methods, Naive Bayes based methods and support vector machine (SVM), concluding that the latter is a promising method for this task. Another important observation they made about their study of different time frames is that the closer the reservation is to the time of service, the less errors were found in each technique tested. Later research in the field has forecasted individual cancellations using similar AI techniques (Antonio et al., 2017a; 2019a) with a high rate of accuracy. In fact, Wu et al. (2017, pp13) note that techniques based on artificial intelligence (AI) have been applied in the hotel and lodging industry with a satisfactory performance. As stated by these authors, Artificial Neural Network (ANN) models appear the most in literature, followed by “Support Vector Regression (SVR), rough set model, fuzzy system methods, genetic algorithms and Gaussian Process Regression (GPR)”.

As it can be appreciated, advanced ensemble methods have been successfully applied in the airline industry with the aim of forecasting cancellations, however, no evidence has been found on the use of these kinds of techniques in the hospitality industry for forecasting individual cancellations, thus providing a novel factor to the present research. The goal of ensemble learning methods is to collect the output of several individual classifiers for obtaining a more accurate result. The most popular techniques for ensemble are the bootstrap aggregation (bagging) and boosting.

While the first one creates multiple versions of a predictor by sampling the training set with a replacement so that the final result is obtained by combining them; the boosting method uses the whole training set in each iteration, assigning a weight for each training instance which is adjusted during the process. As an example, Dietterich (2000) compared bagging, boosting, and randomisation as ensemble methods for decision trees, concluding that boosting got the best results in most cases while the others obtained similar results. Opitz & Maclin (1999) studied the effects of applying bagging and boosting methods for ensemble decision trees and neural networks. They noted that while the bagging method normally delivered better results than single classifiers, in some cases boosting got much better results, however bosting proved to obtain worse results than the single classifier in some tests. Despite the fact that there are multiple ensemble methods (Zhou, 2012), new ensemble models have been published. Most recently, Yu et al. (2008) proposed a multistage nonlinear radial basis function neural network ensemble for forecasting exchange rates, which uses the output of several neural networks in order to create an ensemble applying another neural network structure. They note promising results considering this proposal outperformed the rest of the methods used.

4. METHODOLOGY

4.1 Methodology – Case 1

The National Optical Astronomy Observatory (NOAO) of the United States of America and the Globe Project developed the Globe at Night programme in March 2016 with the aim of encouraging "citizen-scientists around the world to contribute simple unaided-eye observations on sky-brightness to a growing global database" (Walker et al., 2008, pp1). According to this statement, The National Science Foundation (NSF) funded the purchase of 135 sky quality metres, which were distributed along several states in the USA, as well as another five countries, so that citizen-scientists could be trained to take the readings and record them online.

More recently, the *Globe at Night-Sky Brightness Monitoring Network* (GaN-MN) project was created, which is an extension of the original Globe at Night project, with the aim of establishing a global sensor network for long-term monitoring of light pollution around the world using an SQM. Sky brightness is measured every 30s and recorded in an open dataset available at the Globe at Night website. Every month, data from the previous six months are released.

For this research, data released on the Globe at Night website was used for training and testing a probabilistic model. In this section, authors present the pretreatment of such data, as well as modelling techniques used for NSB forecasting.

4.1.1 Understanding data and pretreatment

The open dataset released by the Globe at Night contains several variables such as sensor ID, NSB, temperature, date and others from multiple stations. This research aims to find an accurate method for NSB forecasting based on the study of the NSB univariate time series, thus only the date, local time and NSB measurements of a single station were used. Due to the fact that new stations are included every year, some datasets are provided with more historical data than others, so that a preliminary review was required. As a first approach, this methodology was applied to the station nicknamed “NAOJ”, which corresponds to the National Astronomical Observatory of Japan, located in Tokyo (coordinates: 35°40'31.4"N 139°32'15.7"E), because it was one of the most complete datasets; the whole year is included except for May, and data from 2016 and 2017 are available. As mentioned before, data were collected with a very high resolution (30s) so it was necessary to average measurements into larger time slots. It must be considered that SQM measures sky brightness at a single point; hence it is exposed to rapid variation due to instantaneous external factors, which add noise to data and have a negative impact on the model. Additionally, in this kind of time series, it is recommended to limit the dataset according to some specific hours in which the variables to be predicted take a significant value. As an example, for solar radiation, it is common to restrict data to diurnal hours (Caballero et al., 2018; Chen et al., 2013). In this case, only nocturnal hours were considered, from 23 pm to 7 am the next day.

After these considerations, the result was an NSB univariate time series for the whole year (excluding May), which could be decomposed into three single time series. Decomposition aims to segregate the original time series into distinct components with different characteristics in order to analyse each one individually, therefore, it is a helpful analytical tool which allows finding some relationships that are not obvious in the original time series (Figure 3). An additive decomposition can be represented as the sum of three independent time series:

$$y = y_t + y_s + y_r \tag{1}$$

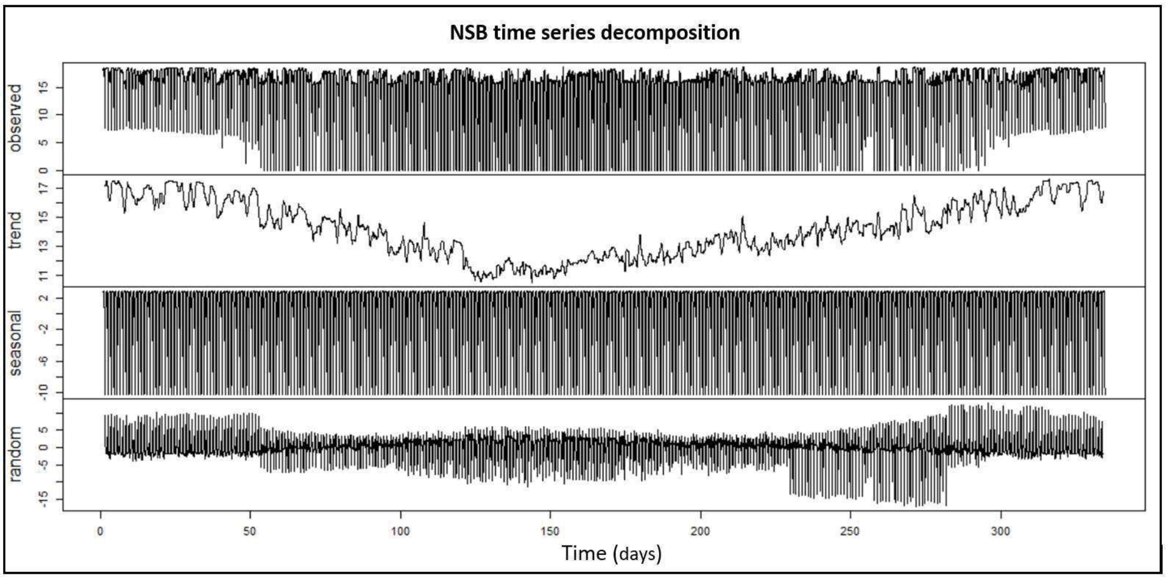
in which:

y_t: Trend – cycle component

y_s: Seasonal component

y_r: Random variation

Figure 3. Additive decomposition for NSB time series for the whole of 2017 taken in the National Astronomical Observatory of Japan.

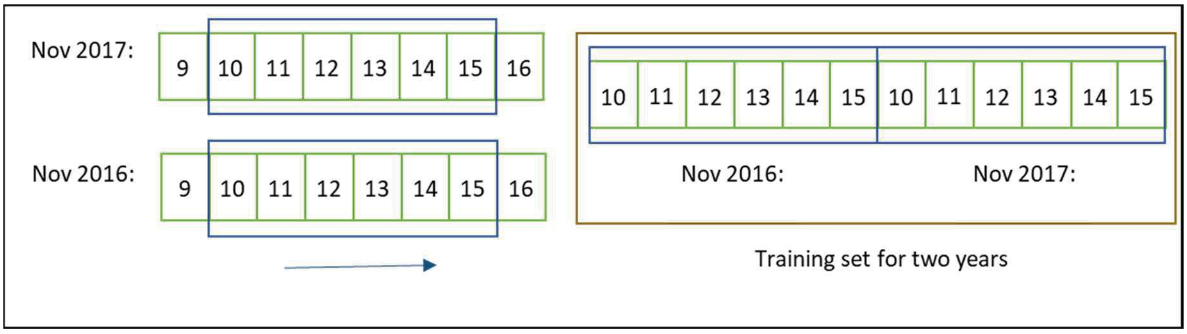


The trend component indicates that NSB values decrease during the summer while registering the highest values in winter periods. This can be explained by two main reasons: NSB peak values are higher in winter than the rest of the year and the number of sun hours also decreases in this period. The seasonal component represents the time series with a fixed and known frequency, while the random variation component is the remaining value from the original series after extracting both previous components. It is noted that the random component has mainly positive values in winter and autumn periods and negative values in summer and spring periods, which is because the solar hours are not the same throughout the year and because values are only considered from 23 pm to 7 am the next day.

4.1.2 Model development

Before applying any method, data were preprocessed using the sliding window technique; this means that the training phase was carried out with data within a specific number of days prior to those intended for the forecast. Therefore, data was segmented according to a specific number of days before a date. This window is moved along the year so that new values can be forecast. Instead of taking a whole dataset, this technique allows using data near the day intended for the forecast, so that the data are more consistent, as long as conditions are similar. As an example, if the intention were to forecast the 16th of November, data from 10th of November (five days before) would be used in the training phase (Figure 4). Full data sets were available for 2016 and 2017 (except for May) and consequently, the sliding window technique can be set with data from a single year that is to be forecast or by adding the same window period for previous years. With the aim of establishing a proper comparison, both options were tested using several sliding windows.

Figure 4. Example of sliding window with data from 2016 and 2017.



From the multiple models reviewed in the previous section, for this first approach, Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN) were selected. As stated earlier, the ARIMA model has been widely used in the fields of nature and economics, as well as other time-varying series, and its forecast relies on past values as well as previous error terms (Ohyver & Pudjihastuti, 2018). This model is a combination of autoregressive models (AR) and moving average models (MA) with an additional parameter which represents the number of times that observations are differentiated (Aasim et al., 2019; Ohyver &

Pudjihastuti, 2018). In general terms, an ARIMA (p,d,q) model can be mathematically denoted as follows:

$$\emptyset_p(B)(1-B)^d z_t = \theta_q(B)a_t \qquad (2)$$

where:

$\emptyset_p$ are unknown coefficients associated with AR model

θ_q are unknown coefficients associated with MA model

z_t is the original time series

B is the backward shift operator (Bz_t is the value of time series at $t-1$)

d is the integrating order

a_t is the white noise series

For each run, the configuration of the ARIMA model was established by multiple trial, looking for the one which presented the best Akaike Information Criterion (AIC). AIC is one of the most popular criteria for estimating the quality of statistical models given by:

$$AIC = -2Log(Ł) + 2m \qquad (3)$$

where:

$Log(Ł)$: is a measure of the model fit.

m: is the number of parameters estimated in the model

Additionally, feed-forward neural network was used for modelling NSB. This biologically inspired model has been widely used in time series forecasting (Mahajan et al., 2018) and it is composed of a specific number of nodes in each layer. Each layer receives as input the output from the previous layer in such a way that the input in each node is a weighted linear combination of the results of the previous layers (Figure 5) (Hyndman & Athanasopoulos, 2018). The input to a hidden node can be mathematically expressed by the following equation:

$$z_j = b_j + \sum_{i=1}^{n} w_{i,j} y_i \tag{4}$$

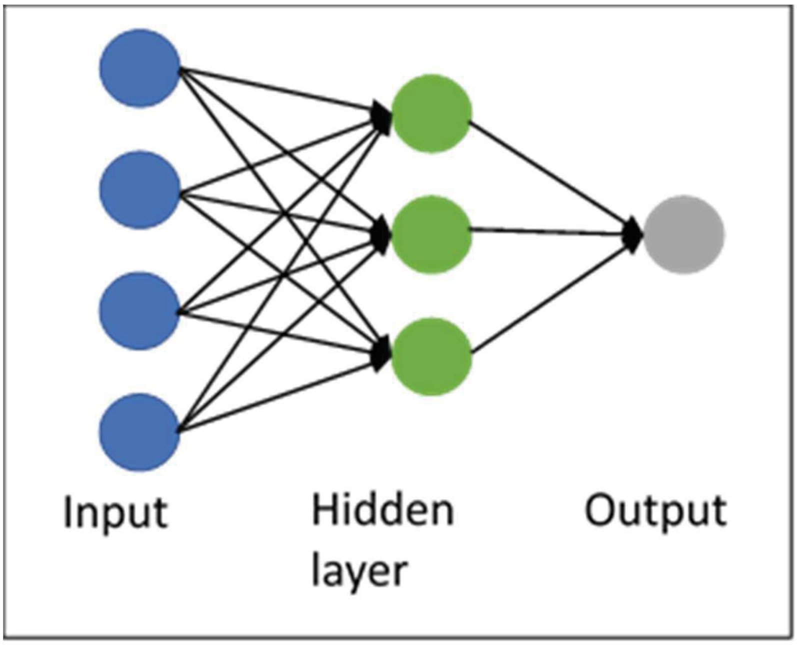
where:

b: represent the bias

$w_{i,j}$: represents the weights

y_j: represents the input values for this layer

Figure 5. ANN conceptual diagram.



A neural network also needs to be configured, however, there is not a procedure for setting its parameters, therefore, model configuration is usually carried out by trial and error (Song & Li, 2008). For the case of time series, pass values can be used as input values for modelling, which is known as Neural Network Autoregression model (NNAR) (Hyndman & Athanasopoulos, 2018) and it is denoted as $NNAR(p,P,k)_m$; where p refers to the seasonal past values inputs; P refers to nonseasonal past values inputs and k refers to the number of hidden neurons in the hidden layer. The function NNETAR implemented in the forecast package for R statistical software (R Core Team., 2013) can build an NNAR model with only one single hidden layer (Hyndman & Athanasopoulos, 2018). The training phase is run 25 times using random stating values, so that the final model is the average of previous trains. Although it is a simple neural network configuration, some investigations note that this model performs better than other more complex ones (Hassani et al., 2017).

4.2 Methodology – Case 2

This research has been developed using real booking records provided by a hotel partner located in Gran Canaria (Spain) with the aim of forecasting future cancellations. According to CRISP-DM process (CRoss Industry Standard Process for Data Mining) (Wirth & Hipp, 2000), before any data preparation takes place, it is necessary to understand the business and data itself, then, models may be built and tested. In this section, the data preparation and modelling process used for this research are detailed.

4.2.1 Understanding data and pretreatment

Although seasonal models were commonly used, mainly because it was the only information available, when revenue management systems started to include historical booking records, forecasting methods were developed using such information (Romero Morales & Wang, 2010). These data are known as Personal Name Records (PNR) that are composed of the information provided by guests at the time a reservation is placed, such as the booking channel, additional hotel services, number of customers and others. For this research only two years of booking records with more than 10,000 bookings including all of 2016 up to April 2018 for a four-star hotel partner located in Gran Canaria (Spain) were used with the aim of forecasting cancellations (Table 1). These variables are commonly requested in web booking platforms, such as Booking, Tripadvisor or Trigavo without needing to access individual historical records, which makes for a more efficient procedure in terms of timing and computational resources.

Table 1. Explanatory available variables of the database – Case2.

Name	Description	Type
Status	Booking status: in place, cancelled	Categorical

Adults	Number of adults	Numeric
Entity	Entity which trough booking was created	Categorical
Nationality	Nationality of the guest	Categorical
Advance payment	If require advance payment (1) or not (0)	Categorical
Nights	Number of nights to be spent in the hotel	Numeric
Notice period	Difference between booking date and arrival date	Numeric
Day of creation	Day in which booking was created	Numeric
Month of creation	Month in which booking was created	Numeric
Day of check in	Effective check in day	Numeric
Month of check in	Effective check in month	Numeric
Mean price	Room mean price	Numeric
Channel	Channel used for booking classified by values from 1 to 9	Categorical
Weekend	Number Saturdays and Sundays during the stay	Numeric

The number and type of available variables are crucial for this kind of research. Accordingly, new variables can be extracted from the original database, which is the case for the number of weekend days that is calculated considering entry and departure dates. Additionally, rows with omitted values were removed and only closed bookings were considered, as long as there was no evidence of the reservation being consumed for current services or future services. For this research, the booking status is the target variable, and so, the problem was treated as a binary classification in which targets may reach two possible values “cancelled” or “not cancelled”. On the other hand, the range of each variable differs from the others, so that predictors were normalised in order to make models less sensitive to scales and maintaining consistency when comparing results across them.

4.2.2 Models and validation

R statistical software (R Core Team., 2013) was used for applying different artificial intelligence techniques, belonging to the branch of supervised methods. In this regard, several packages were used: two trees decision-based algorithms, C5.0 (Kuhn et al., 2018) and random forest (Liaw & Wiener, 2002), support vector machine (SVM) (Meyer et al., 2019), artificial neural networks (ANN) (Fritsch et al., 2019) and genetic algorithm (GA) (Scrucca, 2013). In the following paragraph these methods are briefly explained.

As Rokach & Maimon (2015) noted, decision tree approaches have delivered promising results for information extraction, machine learning or pattern recognition. These authors explain that one of the most popular decision tree algorithms is the *C5.0,* an updated version of the previous *C4.5* algorithm, which uses gain ratio as splitting criteria. Another popular tree decision based algorithm is the Random Forest, which generates multiple decision trees by random selection of new data subsets in each one (Oshiro et al., 2012). On the other hand, the Support Vector Machine technique (SVM) attempts to fix a boundary surface between two classes in a multidimensional space according to the data’s features (Romero Morales & Wang, 2010), while Artificial Neural Networks (ANN) is a biologically-inspired model composed of a specific number of neurons organised in layers, so that, each layer works with the output of the previous ones (Hyndman R.J & Athanasopoulos G., 2018). These methods have been successfully applied in several areas, but there is no systematic procedure for setting architectural parameters, so they are usually adjusted by trial and error (Song & Li, 2008). For this research, tree decision models were set manually, however artificial intelligence models, which are more complex methods were set using tuning tools for SVM and genetic algorithms (GA) for ANN. The GA algorithm is an evolutionary algorithm that seeks an optimal solution for a given function. Initially, a random population is generated and during the process several reproduction, cross over and mutation operations are carried out with the aim of finding the one which best fits the function (Murali et al., 2010). The use of GA algorithm for optimising the ANN model has been used in previous publications, resulting in the combination of both methods outperforming that using conventional ANN (Arifovic & Gencay, 2001; Momeni et al., 2014; Nasseri et al., 2008). In the reviewed literature about hotel booking cancellations no previous uses of GA algorithms for setting the structural parameters have been noted, therefore we can suppose that this is a methodological innovation in this field.

With regards the method's validation, contrary to previous research in the field (Antonio et al., 2017a; Romero Morales & Wang, 2010), who use the K-fold technique, the validation of the model's performance was conducted through repeated random subsampling validation. This validation technique is specially recommended when the dependent variable is dichotomous and raw data are unbalanced (Khakifirooz et al., 2018), which is the case of this study. This technique consists in splitting whole data in a training set for building the model and the testing set for validating the model, in which the process is repeated several times and the mean value is taken as a final evaluation. Applying this method, the unreliability of one single-run training and testing is avoided (holdout method) while the number of runs is not limited, as in the K-fold method. On the other hand, some observations may not be tested, and others may be tested more than once. For this research, the whole process was iterated a hundred times using different randomly selected test and train datasets, so final performance metrics were calculated by comparing actual and forecasted values of the test set.

4.3 Methodology – Case 3

Data mining is a creative process in which data are treated to find patterns, trends or rules that explain the underlying behaviour behind them (Wirth & Hipp, 2000). As in the previous case, in order to follow an orderly process, Cross Industry Standard Process for Data Mining (CRISP-DM) was applied in this study. This is a standard created for industry data mining which describes the life cycle of a data-mining project broken down in six phases.

For this case, it is intended to build an artificial intelligence model with the aim of forecasting individual hotel critical cancellations, which means the identification of specific customers likely to cancel close to the time of service.

In this section, first, the characteristics of the data used in this research are described, as well as the treatment of the data before applying any of the methods. Second, the techniques applied in the research are detailed.

4.3.1 Understanding data and pretreatment

The same dataset employed for the *case 2* was used this time for developing *case 3*, which has been provided by a four-star hotel partner located in Gran Canaria (Spain). This dataset, composed of more than 10,000 booking records between 2016 and 2018, was treated for forecasting critical cancellations by applying different machine learning techniques.

In the same way as in the previous case, one of the most significant advantages of the proposed methodology is that it has been developed with the most commonly requested variables asked of customers when they place a reservation by online travel agencies, the hotel itself or through external platforms such as Booking or Trivago among others. Specifically, a total amount of 13 variables was used independently (Table 2), such as nationality, number of nights or channel and only the variable "*weekend*", which represents the number of weekend days within the period of stay, was calculated from the original dataset. Likewise, the state of the reservation was assigned as a dependent variable. This means if the booking was cancelled "close to the entry day" or "with sufficient time", therefore, this approach allows using two-class probability estimation methods. In addition, different time-horizons were considered, ranging from 4 days to 7 days prior to the time of service.

Considering the excellent results achieved by Antonio et al. (2017a), with an accuracy rate of above 90%, as well as, present work with an accuracy of 98% (see results - case 2) when forecasting general hotel cancellations, the present research is intended to go further and attempt to identify the cancellations that could be made close to the time of service (cancellations made between 4 and 7 days in advance) at any time after customers make a booking. For this reason, this approach will be focused only on cancelled reservations, removing any other type. It should be noted that during the construction phase of the dataset the identity of the guests was not used, therefore, avoiding the need to query the database, which significantly reduces computational timing and resources. Finally, all variables were coded into numerical numbers and normalised in order to reduce the sensitivity of the model to the different scales and maintain the consistency when comparing the outputs of the different models.

Table 2. Explanatory available variables of the database – Case 3.

Name	Description	Type
Status	Cancelled booking: critical or not	Categorial
Adults	Number of adults	Numeric

Entity	Entity through which booking was made	Categorical
Nationality	Nationality of the guest	Categorical
Advance payment	If require advance payment (1) or not (0)	Categorial
Nights	Number of nights to be spent at the hotel	Numeric
Notice period	Difference between booking date and arrival date	Numeric
Day of creation	Day in which booking was created	Numeric
Month of creation	Month in which booking was created	Numeric
Day of check in	Effective check in day	Numeric
Month of check in	Effective check in month	Numeric
Mean price	Mean room price	Numeric
Channel	Channel used for booking classified	Categorial
Weekend	Number of Saturdays and Sundays during the stay	Numeric

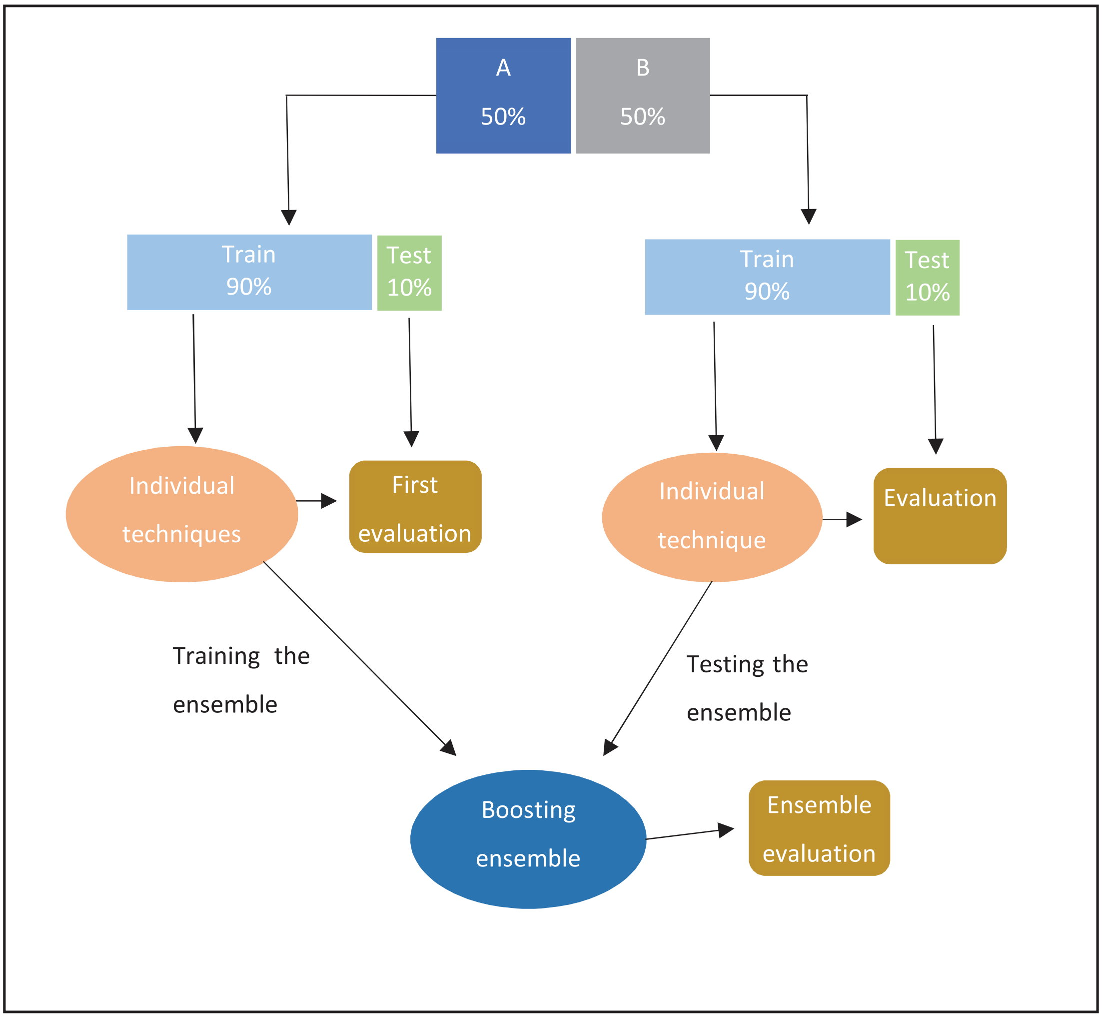
4.3.2 Methods and model

R statistical software (R Core Team., 2013) was used for the model development in which several AI techniques within the supervised area were applied. Accordingly, the following packages were used: C5.0 (Kuhn et al., 2018), Support Vector Machine

(SVM) (Meyer et al., 2019), Artificial Neural Networks (ANN) (Fritsch et al., 2019) and GBM for tree boosting ensemble (Greenwell et al., 2019).

In this approach, the boosting algorithm was used to ensemble several individual outputs, so the whole dataset was split in two sets, the first one is used to set the boosting algorithm and the second one to test the ensemble (Figure 6).

Figure 6. Flowchart of the data use



In a first stage, a balanced training and testing set is created from *dataset A*, which is used for training and forecasting critical cancellations with C5.0, R-part, SVM and ANN techniques, as well as, performing an initial evaluation of each method. This procedure is repeated a hundred times in order to avoid the unreliability of a single training and testing and in each case the training and testing datasets are randomly selected. Along this process, the result of each algorithm and the actual values are saved in order to use them later for setting the ensemble method. After that, *dataset B*, which contains the other 50% of the original dataset, is used for repeating the same procedure, but in this case, outputs of each individual technique are used for evaluating the ensemble method. Likewise, each individual method is evaluated in this second run in order to properly compare the outputs across the different techniques.

5. RESULTS

5.1 Results – Case 1

ARIMA and NNAR models were run using data from 2017 (Table 3) but also from 2017 and 2016 (Table 4). The subsequent 8 hours (night hours) were forecast every 15 min and compared with real data. Several sliding windows were tested in order to find the one that showed better performance.

While there was no standardised evaluation measurement for comparing model performance (Benali et al., 2019), the decision was taken to use the Root Mean Square Error (RMSE), as it is one of the most commonly used measures in forecasting literature (Hassani et al., 2017). However, Willmott & Matsuura (2005, pp1) indicate that Mean Absolute Error (MAE) "is a more natural measure of average error" and recommend its use. For establishing a proper comparison, both indicators were used, as well as standard deviation and the coefficient of determination. Additionally, RMSE and R^2 are represented (Figures 7 and 8) to facilitate comparative analysis.

Table 3. Forecasting results using data from 2017: Standard deviation, determination coefficient, root mean square error and mean absolute error for ARIMA and ANN - Case 1.

Data 2017					
Window	**Model**	**Sd**	**R^2**	**RMSE**	**MAE**
5 days	ARIMA	5.697	0.927	1.553	1.004
	NNAR (K = 4)	4.289	0.901	2.088	1.548
	NNAR (K = 5)	4.345	0.901	2.058	1.526
	NNAR (K = 6)	4.381	0.901	2.041	1.513
10 days	ARIMA	5.658	0.947	1.317	0.871
	NNAR (K = 4)	4.970	0.925	1.641	1.190
	NNAR (K = 5)	5.004	0.924	1.638	1.187
	NNAR (K = 6)	5.025	0.923	1.637	1.185

15 days	ARIMA	5.663	0.948	1.300	0.851
	NNAR (K = 4)	5.072	0.927	1.594	1.149
	NNAR (K = 5)	5.101	0.926	1.595	1.149
	NNAR (K = 6)	5.119	0.925	1.597	1.149

Table 4. Forecasting results using data from 2016 and 2017: Standard deviation, determination coefficient, root mean square error and mean absolute error for ARIMA and ANN – Case 1.

Data 2016 & 2017					
Window	**Model**	**Sd**	**R^2**	**RMSE**	**MAE**
5 days	ARIMA	5.629	0.943	1.363	0.889
	NNAR (K = 4)	4.723	0.874	2.095	1.574
	NNAR (K = 5)	4.757	0.873	2.090	1.570
	NNAR (K = 6)	4.774	0.872	2.092	1.571
10 days	ARIMA	5.670	0.949	1.293	0.847
	NNAR (K = 4)	5.182	0.911	1.720	1.279
	NNAR (K = 5)	5.205	0.910	1.722	1.277
	NNAR (K = 6)	5.217	0.909	1.728	1.280
15 days	ARIMA	5.693	0.951	1.278	0.837
	NNAR (K = 4)	5.354	0.921	1.620	1.194
	NNAR (K = 5)	5.371	0.920	1.625	1.196
	NNAR (K = 6)	5.380	0.919	1.632	1.199

Figure 7. Comparison of forecasting models with different windows for data 2017. Root mean square error and coefficient of determination for each model and window are shown.

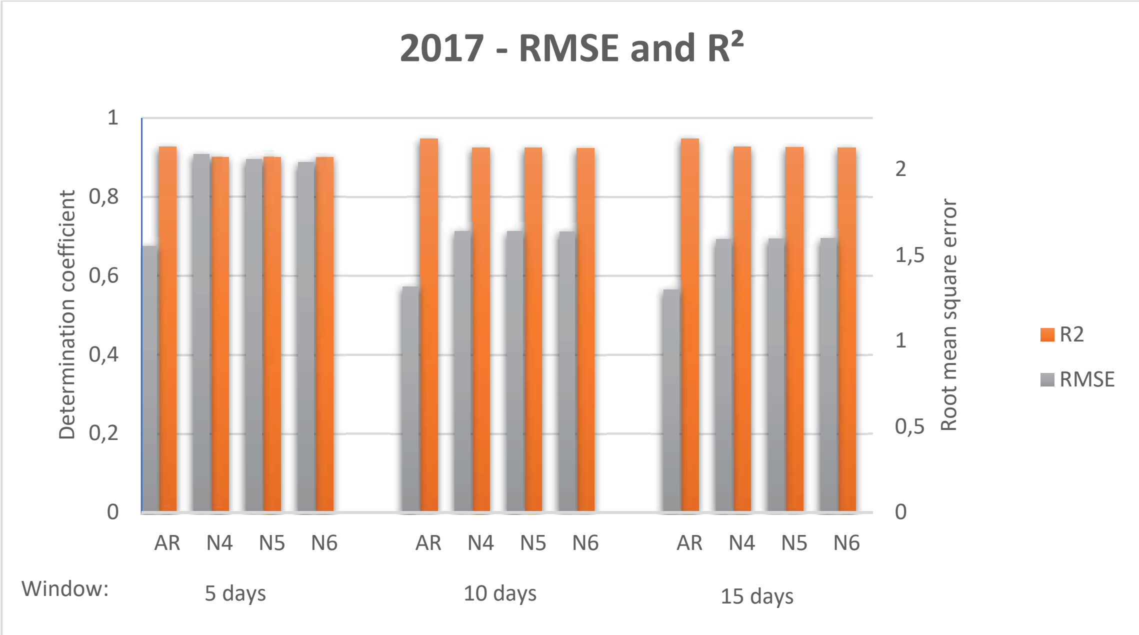
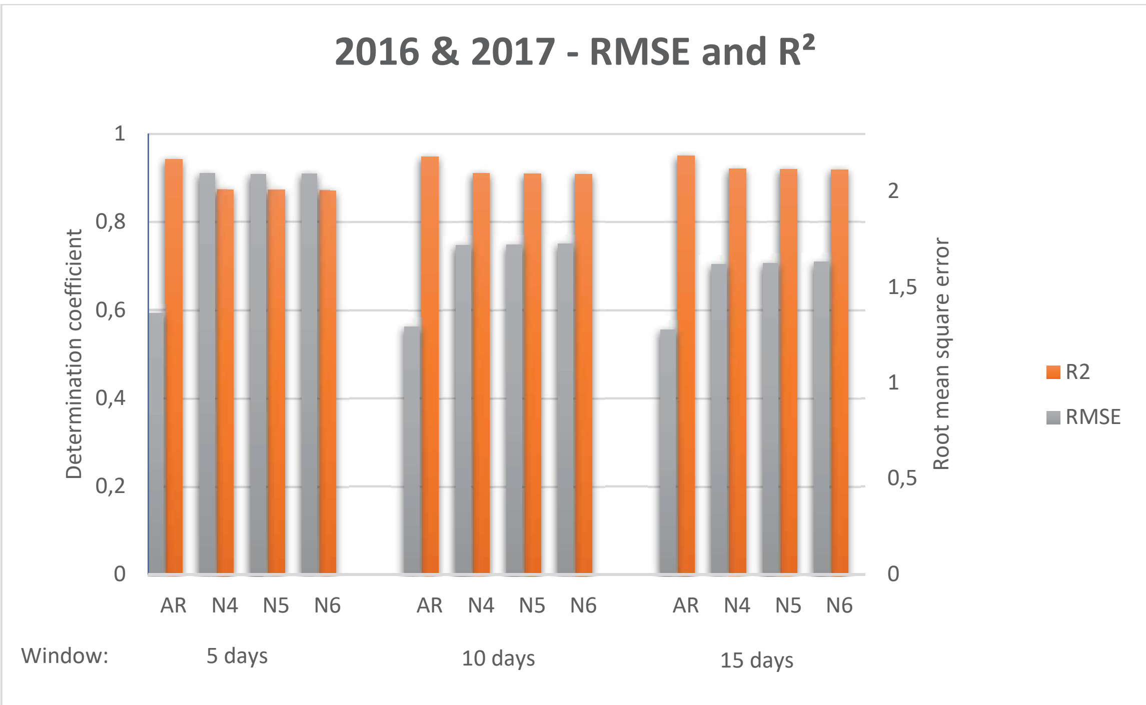


Figure 8. Comparison of forecasting models attending different windows for data 2016 & 2017. Root mean square error and coefficient of determination for each model and window are shown.



As a general impression, all models improve their results when the sliding window becomes longer; while it is true that a significant difference is found from the first window (5 days) to the second (10 days), there is no such difference from the second window to the third. Similarly, NNETAR models improve results when the number of neurons increases, but there is no significant difference among them. In all cases, ARIMA models slightly improve over the NNETAR models, however, the ARIMA models respond better when data from both 2016 and 2017 are used, whereas the best results of the NNETAR model are found when just 2017 data are used. Considering both data from 2017 and 2016 & 2017, the dispersion measure increases for NNETAR models when the sliding window is longer. For the ARIMA model, standard deviation increases with the increase of the sliding window for 2016 & 2017 data, whereas it decreases for 2017 data.

5.2 Results – Case 2

In this section the results of previous techniques are shown and discussed. First, performance measures of several techniques are presented and secondly (table 5), the models and results are discussed.

Results extracted from the classification methods are commonly analysed by the use of a confusion matrix, which is a contingency table that shows the difference between the actual class and the predicted one for the test set in a labelled table (Bradley, 1997). The relationship between these classes is expressed through two main performance measures, sensitivity which is "the proportion of true positives correctly detected by the test", and specificity which is "the proportion of true negatives correctly identified by the test" (Altman & Bland, 1994, pp1). On the other hand, the accuracy is expected to measure the reliability of the model for both categories and the precision summarise the true positives forecasted over the total positive items. An easy way to analyse forecasted results is the ROC curve (figure 9), which is a graphical support for evaluating the performance classifier giving the relationship between true positives and true negatives predicted by the model (Bradley, 1997).

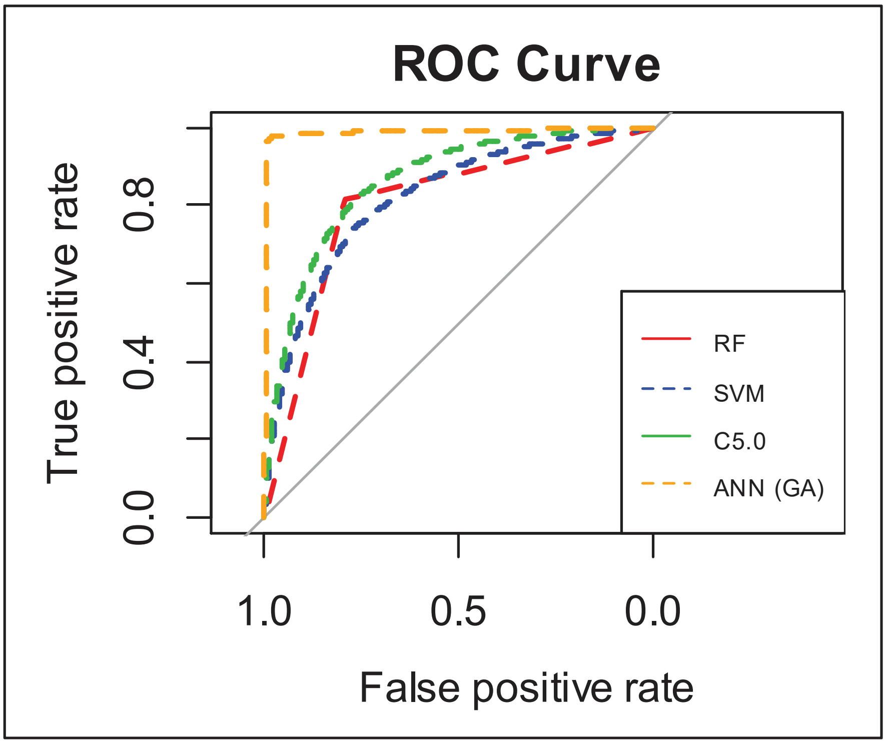
Table 5. Summary of results: performance measures – Case 2.

	Accuracy	Precision	Specificity	Sensitivity	AUC
Random forest	0.804	0.813	0.809	0.799	0.804
Support vector machine	0.753	0.733	0.743	0.764	0.820
C 5.0	0.790	0.818	0.807	0.774	0.864

ANN (GA optimized)	0.980	0.972	0.972	0.987	0.989

Results show that the SVM technique outputs good results, while tree-based methods improve the accuracy even more. While C5.0 algorithm outputs better results in terms of the area under the ROC curve (AUC), Random Forest shows better accuracy. Nevertheless, the SVM outperforms the Random Forest method in terms of AUC, but shows the lowest accuracy overall. For all three, it can be stated that they deliver good results. When ANN is optimised with GA, this method delivers all metrics with values above 0.95 and has the best performance overall. This extraordinary performance can be also seen in the ROC curve graph (Figure 9). On the other hand, in all cases specificity and sensitivity are balanced.

Figure 9. ROC curves. Comparison of different methods.



5.3 Results – Case 3

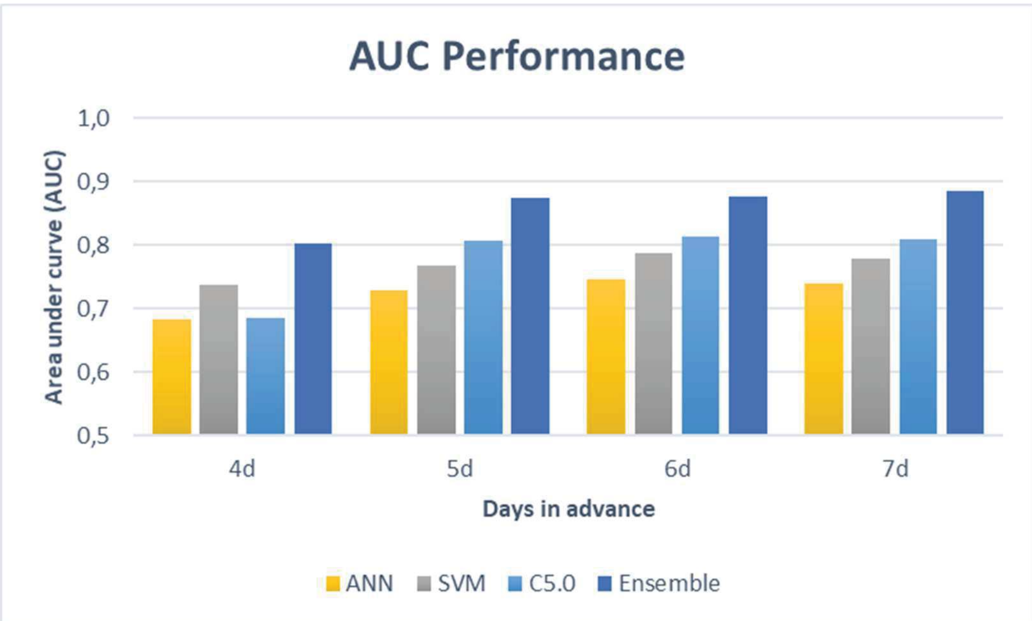
In order to analyse the results, a confusion matrix was used, as well as the same performance metrics as for the *Case 2*, which are quite common for two-class classification method: sensitivity, specificity, accuracy and precision (Table 6). Indeed, the area under the ROC curve is another parameter analysed in this research (Figure 10). It should be noted that this approach attempts to forecast cancellations just in the short horizon, which means that the data set is highly unbalanced, reaching rates of 5%. This configuration makes it more complicated to achieve accurate models because of the lack of data for training the models.

Table 6. Summary of results: performance measures – Case 3.

	KPIs	**C5.0**	**SVM**	**ANN**	**Boosting ensemble**
4 days	Accuracy	0.685	0.673	0.601	0.730
	Precision	0.568	0.551	0.242	0.658
	Specificity	0.650	0.639	0.559	0.701
	Sensitivity	0.741	0.728	0.859	0.769
	AUC	0.685	0.736	0.683	0.802
5 days	Accuracy	0.738	0.698	0.674	0.793
	Precision	0.635	0.694	0.443	0.778
	Specificity	0.697	0.696	0.619	0.785
	Sensitivity	0.799	0.699	0.825	0.803

	AUC	0.807	0.767	0.729	0.873
6 days	Accuracy	0.739	0.716	0.689	0.805
	Precision	0.634	0.718	0.488	0.797
	Specificity	0.698	0.717	0.635	0.800
	Sensitivity	0.803	0.715	0.817	0.809
	AUC	0.813	0.786	0.745	0.876
7 days	Accuracy	0.736	0.716	0.692	0.805
	Precision	0.700	0.705	0.538	0.808
	Specificity	0.720	0.711	0.647	0.807
	Sensitivity	0.755	0.720	0.776	0.804
	AUC	0.808	0.778	0.740	0.885

Figure 10. AUC performance of each method

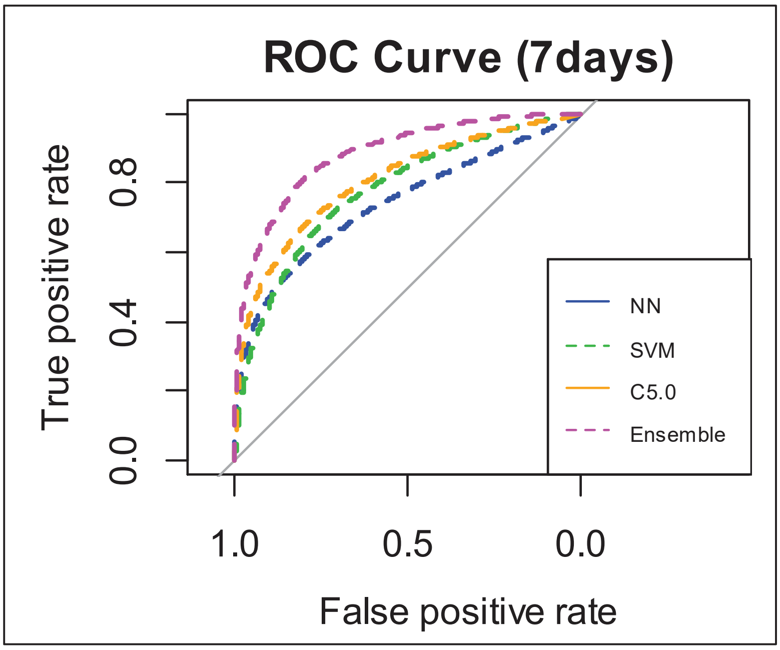


According to the presented results, the tree-based algorithm C5.0 is the individual technique that shows the best output, followed by SVM and ANN respectively. As can be appreciated, in the most unfavourable case, which is the 4-day time horizon, the least accurate rate of 60% corresponds to the ANN technique while the ensemble method achieves 73%, which is a good level of accuracy. A similar scenario is found for the case of the Area Under Curve (AUC) where the ANN technique delivers the lowest value (0.68) and the ensemble method succeeded in improving the AUC at 0.80. This can be explained because of the low number of positive cases that have a negative effect on the training phase, especially for the ANN, which is very data hungry. Regards the specificity and sensitivity, both are balanced in all cases, except for the case of ANN, in which the specificity is slightly lower than the rest however, sensitivity improves significantly.

In this research, four time-horizons were considered, ranging from those cancellations placed 4 to 7 days prior to the entry day, so that when more time is considered, the number of cancellations increases and thus, more positive cases are available for training the models. This is reflected in the results, which improve as more time prior to the entry day is considered, as well as, the specificity and sensitivity, whose trend tends to be more balanced.

On the other hand, the results confirm that the ensemble technique successfully improves the individual techniques in all cases, achieving up to 14% of AUC above the lowest value and also reaching balanced specificity and sensitivity. As an example of the significant improvements achieved with the ensemble technique in figure 11 the performance of the individual and ensemble methods is plotted for the forecasts of 7 days in advance.

Figure 11. ROC curve for each method. Forecasts within 7 days prior to the entry day



6. SUMMARY, CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

6.1 Summary and conclusions

According to the excellent results obtained in this work, it can be concluded that artificial intelligence can be extremely useful for dealing with several issues within the tourism industry, although they are of a very different nature. This work has addressed two very distinct problems related to the tourism industry, on the one hand the case of the internal process of hotel cancellation forecasts, and on the other hand, the way that AI techniques can contribute to the product/service offered, in that, the customer experience is improved, by forecasting the best time for enjoying astrotourism. In this last case, the proposed methodology would allow to create an information system in order to inform people of the best places and best time to enjoy seeing the sky phenomena so that tourists can plan their trips more efficiently and avoid wasting their holiday time.

It is worth highlighting that for *Case 1* the presented procedure has been developed supported by real data, more specifically, an open NSB dataset provided by the Globe at Night international program which monitors sky brightness along different points in the world. In the same way, for *Cases 2 & 3*, real booking records provided by a hotel partner were used to develop the proposed methodology. Therefore, all three cases were developed using real data, which provides the work with practical and realistic processes that can be easily exported to other sites.

Following the structure of the preceding chapters, the specific conclusions for each case are presented below:

6.1.1 Conclusions - Case 1

Tourism is a sector which is continuously changing, extending and adding new activities. In this regard, companies need to constantly develop new product/services in order to offer something different to clients, not just as a differentiating mechanism, but also to satisfy client requests. Activities such as canoeing, snorkeling, hiking or contemplating the skies are just examples of what is considered as nature-based tourism, which is a subsector that has undergone strong growth in recent years. Astrotourism is part of this kind of tourism and can be briefly described as when people move to different locations in order to contemplate celestial phenomena.

The present research proposes a methodology that is able to forecast the night sky brightness with a high level of accuracy in order to advise tourists about the best hours to visit the different places. In this regard, two probabilistic methods have been used for modelling night sky brightness (NSB), achieving very good results. Despite the fact that the results achieved using the ARIMA model are slightly higher, the results derived from the ANN models were very promising, all the more considering the use of one single layer with a reduced number of neurons. However, despite the ARIMA model providing better goodness of fit, ANN showed low dispersion in terms of standard deviation. In the same way, ARIMA showed better performance using data from 2016 and 2017, while ANN improved its results using data only from 2017.

This research has enabled the foundations for the creation of an information system that shows interested people this kind of forecast to be created. It should be noted that the recommended suitable places for developing celestial tourism tend to be located in remote areas and visitors must travel long distances with the hope of finding good sky conditions upon arrival. Furthermore, it is an activity that is carried out in the early hours of the day and it is commonplace to take with heavy equipment for the purpose. Therefore, if, upon arrival, the sky conditions are not good, tourists may leave with the feeling of having "wasted their time". Knowing that customer satisfaction is crucial for encouraging client loyalty, this research would assist people in planning their trip with guaranteed optimal sky conditions. It represents an additional value and helps to promote astrotourism.

Finally, it should be mentioned that this methodology could be considered as a novel contribution to the related literature, in that, no other evidence about previous application of probabilistic models for forecasting night sky brightness was found. This work was published in the JCR journal of *Sensors*, classified as Q1 (C-Sánchez et al., 2019).

6.1.2 Conclusions - Case 2 & 3

As already mentioned, adding value has become essential within the hospitality industry because of increasing competition and the uncertain environment of the sector. Taking this into account, room planning is one of the most important tasks that hotel management must face because idle capacity represents lost revenue. Due to the fact that the hotel and lodging sector must handle future demand as well as, develop future capacity schedules, and reservations that have already been made may be cancelled, cancellation management plays a crucial role. For a long time, hotel chains have adopted different strategies to fight against this situation, such as overbooking strategies or cancellation policies, however both can have a negative impact on the business, as explained in this work. For this reason, accurate cancellation forecasts are needed in order to better adjust strategies and properly plan the hotel's capacity and resources. Along these lines, the present Ph. D Thesis has contributed to providing technical solutions for adding value to hospitality companies. Two lines of research were developed: one addressing the problem of forecasting cancellations made at any time after the initial reservation (general cancellation), and another, identifying those cancellations likely to be made close to the time of service (critical cancellation). In both cases, the main goal was not just to forecast the cancellation rates, but also define better individual cancellations, that is, which specific clients are likely to cancel. The important contribution of this work in this topic should be noted, considering that cancellation forecasting is an underdeveloped topic within the hospitality industry and few researchers have addressed this issue. Moreover, there is no previous research addressing the problem of forecasting individual cancellations by applying techniques such as genetic algorithms or ensemble methods, which again represents a novelty in this area. Along these lines, *case 2* treats the data through several AI-based techniques achieving very good results, among them ANN optimised with GA outperformed the rest with a 98% accuracy rate, which generates considerable value to the organisation, because of the difficulties and revenue loss that cancellations produce. *Case 3* focusses on critical cancellations, and manages to identify up to 80% of customers that are likely to cancel up to seven days before the entry day. This case was characterised by a high unbalanced dataset with very few positive cases, which was overcome using ensemble techniques that allowed improving the provided AUC values by the individual techniques by up to 14%.

The excellent results achieved, indicate that a customer's historical records are essential for hospitality enterprises and should be treated as a key element. In fact, one of the main assets of this research was the development of a procedure using only the most common variables requested during the booking process. Reducing the number of variables to be used, makes it more complex to elaborate a proper methodology that allows achieving an accurate forecast, but it does not necessarily increase the difficulty of the methodology itself, quite the contrary, the procedure presented in this Ph. D Thesis has been developed with the aim of simplifying the data process as far as possible without sacrificing the accuracy of the model. In this regard, guest's identity, data base queries or external data sources have not been used as part of this procedure, which allows the presented methodology to simplify the forecasting process by avoiding the need of querying the existing database, assembling different databases or calculating new variables; this is one of the main advantages of this methodology. Moreover, the data employed for cancellation forecasting were provided by a hotel partner located in the south of Gran Canaria (Spain), which makes for a more practical process, as long as, in many cases, this is the only available information that hotels possess. This feature makes it easy to implant this technology in a real environment, taking into consideration that it is prepared to work with the most common variables used within the hotel and lodging industry and has been developed employing a real booking data structure. In fact, the collaboration with external enterprises is another highlight point of this Ph. D Thesis.

On the other hand, accurate cancellation forecasting leads hoteliers to take proper managerial decisions and provides organisational advantages for the industry. These techniques give management the opportunity to have information in advance, so that they can establish appropriate overbooking policies, cancellation policies and take advantage of proper pricing strategies among others. With regards overbooking policies specifically, if hoteliers have reliable information about the cancellations, they may avoid overbooking, meaning it would not be necessary to relocate guests, which may cause a profit loss and has a negative impact on reputation. In addition, the methodology presented in *case 3* is especially useful for enterprises to handle the growing trend on the market that indicates that customers place multiple reservations when planning their holidays until they finally choose only one and cancel the rest. Being aware of which individuals are likely to cancel close to the time of service is essential for hoteliers because it leaves them with very little time to react, forcing them in many cases to lose the sale or resell the room with a significantly discounted price. This situation becomes even worse when the hotels themselves or competing companies, provide “last minute” offers, because customers have the chance to cancel previous reservations and choose a more economic option.

In addition, this methodology represents a considerable competitive advantage because it can forecast the cancellation rate with a level high of accuracy, but it can also determine which customer is likely to cancel. This would allow the hoteliers to take proactive actions in order to encourage clients to maintain their reservation, such as sending reminders or contacting directly with them. Finally, it is worth highlighting that *Case 2* is currently under review in the International Journal of Hospitality Management, which is classified as a D1 (Decile 1) according to JCR requirements.

6.2 Future research and limitations

This subsection is intended to propose some future directions for research according to the experience gained in this work. In the same way, the limitations of the present work are detailed.

6.2.1 Future research and limitations – Case 1

Future research in this area may be conducted by the application of the same methodology through a new parametrisation to the used NSB time series, such as resolution, time frame considered or structural parameters of the applied techniques. Likewise, the excellent results achieved with the presented methodology, encourage including as future prospects the addition of more variables such as temperature, humidity or sky conditions.

On the other hand, it would be worthwhile to apply the presented methodology to other databases, that is, using an NSB database measured at another point. Location or weather conditions are just examples of characteristics that change the studied conditions and therefore may have an impact on the AI techniques settings, the pretreatment of data or even the methodology as it is presented. In the same way, it would be very interesting to investigate further the usefulness of using other dataset coming from data measured in other locations with the aim of taking advantage of the similarities and impacts that each element has on the others. In this regard, it is more and more common to find open NSB data in an attempt by public administrations to track the local light pollution in some cities or towns, which can be useful in this task.

Like any other research, the proposed methodologies also have limitations. On the one hand, the night hours change depending on the time of year and location, so that some of the model's parameters, such as the time frame to be considered must be adjusted in each case. On the other hand, rapid changes in the weather could be initially translated into inaccurate forecasts.

6.2.2 Future research and limitations – Case 2 & 3

Using the existing dataset, another field of research that would be interesting is the study of the drivers behind the cancellations, which is a very underdeveloped topic within the hospitality industry. In this way, a study focusing on the analysis of characteristics, for both, general and critical cancellations, could reveal significant findings in this topic, especially considering that the factors that affect customer motivations change with time (Romero Morales & Wang, 2010).

Future research could consider testing the same methodology proposed in this work in other PNR databases with different characteristics (location, weather conditions, hotel rating, prices, market segment, distribution channels, cancellation policies, etc). Even, external variables, such as weather forecasts or the economic index of the origin countries (recession index, GDP, etc.) may be interesting to use as a predictor of variables in order to improve the results. In this case, it must be considered if more complexity is added to the models and if it is worthwhile.

With regards the limitations, considering that forecasts are based on past data, rapidly unexpected changes in the market could not be initially reflected for both cancellation model presented.

7. REFERENCES

REFERENCES

A. Liaw, & M. Wiener. (2002). *Classification and Regression by randomForest. R News 2(3), 18—22.* [R News]. Available online: https://CRAN.R-project.org/doc/Rnews/ (accessed on 13 October 2019).

Aasim, Singh, S. N., & Mohapatra, A. (2019). Repeated wavelet transform based ARIMA model for very short-term wind speed forecasting. *Renewable Energy*, *136*, 758–768. https://doi.org/10.1016/j.renene.2019.01.031

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 1: Sensitivity and specificity. *BMJ*, *308*(6943), 1.

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

Antonio, N., de Almeida, A., & Nunes, L. (2017b). Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1049–1054. https://doi.org/10.1109/ICMLA.2017.00-11

Antonio, N., de Almeida, A., & Nunes, L. (2019a). Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior. *Cornell Hospitality Quarterly*. https://doi.org/10.1177/1938965519851466

Antonio, N., de Almeida, A., & Nunes, L. (2019b). Predictive models for hotel booking cancellation: A semi-automated analysis of the literature. *Tourism & Management Studies*, 16.

Arifovic, J., & Gencay, R. (2001). Using genetic algorithms to select architecture of a feedforward artiÿcial neural network. *Physica A*, 21.

Bará, S. (2016). Anthropogenic disruption of the night sky darkness in urban and rural areas. *Royal Society Open Science*, *3*(10), 1–14. https://doi.org/10.1098/rsos.160541

Bará, S., Lima, R. C., & Zamorano, J. (2019). Monitoring Long-Term Trends in the Anthropogenic Night Sky Brightness. *Sustainability*, *11*(11), 1–14. https://doi.org/10.3390/su11113070

Benali, L., Notton, G., Fouilloy, A., Voyant, C., & Dizene, R. (2019). Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. *Renewable Energy*, *132*, 871–884. https://doi.org/10.1016/j.renene.2018.08.044

Birriel, J., & Adkins, J. K. (2010). *A Simple, Portable Apparatus to Measure Night Sky Brightness at Various Zenith Angles*. *38*, 221–229.

Boes, K., Buhalis, D., & Inversini, A. (2015). Conceptualising Smart Tourism Destination Dimensions. *Information and Communication Technologies in*

Tourism 2015 (pp. 391–403). https://doi.org/10.1007/978-3-319-14343-9_29

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Burger, C. J. S. C., Dohnal, M., Kathrada, M., & Law, R. (2001). A practitioners guide to time-series methods for tourism demand forecasting—A case study of Durban, South Africa. *Tourism Management*, *22*(4), 403–409. https://doi.org/10.1016/S0261-5177(00)00068-6

Caballero, R., Zarzalejo, L. F., Otero, Á., Piñuel, L., & Wilbert, S. (2018). Short term cloud nowcasting for a solar power plant based on irradiance historical data. *Journal of Computer Science and Technology*, *18*(03), 186–192. https://doi.org/10.24215/16666038.18.e21

Candrea, A. N., & Herţanu, A. (2015). Developing ecotourism destinations in Romania. A case study approach. *Economic Sciences*, *8*, 163–174.

Cao, R. Z., Ding, W., He, X. Y., & Zhang, H. (2010). Data mining techniques to improve no-show forecasting. *Proceedings of 2010 IEEE International Conference on Service Operations and Logistics, and Informatics*, 40–45. https://doi.org/10.1109/SOLI.2010.5551620

Chen, C. C., Schwartz, Z., & Vargas, P. (2011). The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers. *International Journal of Hospitality Management*, *30*(1), 129–135. https://doi.org/10.1016/j.ijhm.2010.03.010

Chen, C.-C., & Xie, K. (Lijia). (2013). Differentiation of cancellation policies in the U.S. hotel industry. *International Journal of Hospitality Management*, *34*, 66–72. https://doi.org/10.1016/j.ijhm.2013.02.007

Chen, K.-Y., & Wang, C.-H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management*, *28*(1), 215–226. https://doi.org/10.1016/j.tourman.2005.12.018

Chen, S. X., Gooi, H. B., & Wang, M. Q. (2013). Solar radiation forecast based on fuzzy logic and neural networks. *Renewable Energy*, *60*, 195–201. https://doi.org/10.1016/j.renene.2013.05.011

Cho, V. (2003). A comparison of three different approaches to tourist arrival forecasting. *Tourism Management*, *24*(3), 323–330. https://doi.org/10.1016/S0261-5177(02)00068-7

Chu, F. L. (2009). Forecasting tourism demand with ARMA-based methods. *Tourism Management*, *30*(5), 740–751. https://doi.org/10.1016/j.tourman.2008.10.016

Cinzano, P. (2005). Night Sky Photometry with Sky Quality Meter. *Istituto Di Scienza e Tecnologia Dell'Inquinamento Luminoso (Italy)*, *9*, 1–14.

Cinzano, P., Falchi, F., & Elvidge, C. D. (2001). The first World Atlas of the artificial night sky brightness. *Monthly Notices of the Royal Astronomical Society*, *328*(3), 689–707. https://doi.org/10.1046/j.1365-8711.2001.04882.x

Claveria, O., & Datzira, J. (2010). Forecasting tourism demand using consumer expectations. *Tourism Review*, *65*(1), 18–36. https://doi.org/10.1108/16605371011040889

Claveria, O., Monte, E., & Torra, S. (2015). Tourism Demand Forecasting with Neural Network Models: Different Ways of Treating Information: Tourism Demand Forecasting with Neural Network Models. *International Journal of Tourism Research*, *17*(5), 492–500. https://doi.org/10.1002/jtr.2016

Collison, F. M., & Poe, K. (2013). “Astronomical Tourism”: The Astronomy and Dark Sky Program at Bryce Canyon National Park. *Tourism Management Perspectives*, *7*, 1–15. https://doi.org/10.1016/j.tmp.2013.01.002

C-Sánchez, E., Sánchez-Medina, A. J., Alonso-Hernández, J. B., & Voltes-Dorta, A. (2019). Astrotourism and Night Sky Brightness Forecast: First Probabilistic Model Approach. *Sensors*, *19*(13), 2840. https://doi.org/10.3390/s19132840

den Outer, P., Lolkema, D., Haaima, M., van der Hoff, R., Spoelstra, H., & Schmidt, W. (2015). Stability of the Nine Sky Quality Meters in the Dutch Night Sky Brightness Monitoring Network. *Sensors*, *15*(4), 9466–9480. https://doi.org/10.3390/s150409466

Dietterich, T. G. (2000). *An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization*. 19.

Donaghy, K., McMahon, U., & McDowell, D. (1995). Yield management: An overview. *International Journal of Hospitality Management*, *14*(2), 139–150. https://doi.org/10.1016/0278-4319(95)00013-3

Duriscoe, D. M., Anderson, S. J., Luginbuhl, C. B., & Baugh, K. E. (2018). A simplified model of all-sky artificial sky glow derived from VIIRS Day/Night band data. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *214*, 133–145. https://doi.org/10.1016/j.jqsrt.2018.04.028

Falk, M., & Vieru, M. (2018). Modelling the cancellation behaviour of hotel guests. *International Journal of Contemporary Hospitality Management*, *30*(10), 3100–3116. https://doi.org/10.1108/IJCHM-08-2017-0509

Farajirad, A., & Beiki, P. (2015). Codification of Appropriate Strategies to Astronomical Tourism Development (Seghaleh, South of Khorasan). *Applied Mathematics in Engineering, Management and Technology*, *3*, 303–312. https://doi.org/10.15584/pjsd.2017.21.2.11

Fouilloy, A., Voyant, C., Notton, G., Motte, F., Paoli, C., Nivet, M.-L., … Duchaud, J.-L. (2018). Solar irradiation prediction with machine learning: Forecasting models selection method depending on weather variability. *Energy*, *165*, 620–629. https://doi.org/10.1016/j.energy.2018.09.116

Fritsch, S., Guenther, F., Wright, M. N., Suling, M., & Mueller, S. M. (2019). Neuralnet: Training of Neural Networks (Version 1.44.2). Available online: https://CRAN.R-project.org/package=neuralnet (accessed on 13 October 2019).

Garrow, L. A., & Koppelman, F. S. (2004). Predicting air travelers' no-show and standby behavior using passenger and directional itinerary information. *Journal of Air Transport Management*, *10*(6), 401–411. https://doi.org/10.1016/j.jairtraman.2004.06.007

Garstang, R. H. (1986). Model for Artificial Night-Sky Illumination. *Publications of the Astronomical Society of the Pacific*, *98*, 364–375. https://doi.org/10.1086/131768

Garstang, R. H. (1988). Predictions of seasonal variations in night sky brightness. *The Observatory*, *108*, 159–161.

Garstang, R. H. (1991). Dust and light pollution. *Publications of the Astronomical Society of the Pacific*, *103*, 1109–1116. https://doi.org/10.1086/132933

Gehrels, S., & Blanar, O. (2013). How economic crisis affects revenue management: The case of the Prague Hilton hotels. *Research in Hospitality Management*, *2*(1–2), 9–15. https://doi.org/10.1080/22243534.2013.11828284

Globe at Night project. (n.a). Available online: https://www.globeatnight.org/ (accessed on 13 October 2019).

Gorin, T., Brunger, W. G., & White, M. M. (2006). No-show forecasting: A blended cost-based, PNR-adjusted approach. *Journal of Revenue and Pricing Management*, *5*(3), 188–206. https://doi.org/10.1057/palgrave.rpm.5160039

Greenwell, B., Boehmke, B., & Cunningham, J. (2019). gbm: Generalized Boosted Regression Models (Version 2.1.5). Available online: https://CRAN.R-project.org/package=gbm (accessed on 13 October 2019).

Gunter, U., & Önder, I. (2015). Forecasting international city tourism demand for Paris: Accuracy of uni- and multivariate models employing monthly data. *Tourism Management*, *46*, 123–135. https://doi.org/10.1016/j.tourman.2014.06.017

Guo. Y, Liu. H, & Chai. Y. (2014). The embedding convergence of smart cities and tourism internet of things in China—An advance perspective. *Advances in Hospitality and Tourism Reserach*, *2*, 54–69.

Haensel, A., & Koole, G. (2011). Booking horizon forecasting with dynamic updating: A case study of hotel reservation data. *International Journal of Forecasting*, *27*(3), 942–960. https://doi.org/10.1016/j.ijforecast.2010.10.004

Hajibaba, H., Boztuğ, Y., & Dolnicar, S. (2016). Preventing tourists from canceling in times of crises. *Annals of Tourism Research*, *60*, 48–62. https://doi.org/10.1016/j.annals.2016.06.003

Hänel, A., Posch, T., Ribas, S. J., Aubé, M., Duriscoe, D., Jechow, A., … Kyba, C. C. M. (2018). Measuring night sky brightness: Methods and challenges. *Journal of Quantitative Spectroscopy & Radiative Transfer*, 278–290. https://doi.org/10.1016/j.jqsrt.2017.09.008

Hassani, H., Silva, E. S., Antonakakis, N., Filis, G., & Gupta, R. (2017). Forecasting accuracy evaluation of tourist arrivals. *Annals of Tourism Research*, *63*, 112–127. https://doi.org/10.1016/j.annals.2017.01.008

Hu, Y.-C., Jiang, P., & Lee, P.-C. (2019). Forecasting tourism demand by incorporating neural networks into Grey–Markov models. *Journal of the Operational Research Society*, *70*(1), 12–20. https://doi.org/10.1080/01605682.2017.1418150

Huang, H., Chang, A. Y., & Ho, C. (2013). *Using Artificial Neural Networks to Establish a Customer-cancellation Prediction Model*. (1), 178–180.

Huang, H.-C. (2014). A study on artificial intelligence forecasting of resort demand. *Journal of Theoretical and Applied Information Technology*, *Vol. 70*, 265–272.

Huang, H.-C., & I Hou, C.-. (2017). Tourism Demand Forecasting Model Using Neural Network. *International Journal of Computer Science and Information Technology*, *9*(2), 19–29. https://doi.org/10.5121/ijcsit.2017.9202

Huarng, K.-H., Moutinho, L., & Yu, T. H.-K. (2007). An Advanced Approach to Forecasting Tourism Demand in Taiwan. *Journal of Travel & Tourism Marketing*, *21*(4), 15–24. https://doi.org/10.1300/J073v21n04_03

Hyndman, R.J, & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). Otexts. Available online: https://otexts.com/fpp2/index.html (accessed on 10 October 2019).

International Dark-Sky Association. (n.a). Available online: https://www.darksky.org/ (accessed on 10 October 2019).

Ivanov, S., & Zhechev, V. S. (2011). Hotel Revenue Management – A Critical Literature Review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1977467

Jauncey, S., Mitchell, I., & Slamet, P. (1995). The meaning and management of yield in hotels. *International Journal of Contemporary Hospitality Management*, *7*(4), 23–26. https://doi.org/10.1108/09596119510083229

Jechow, A., Kyba, C., & Hölker, F. (2019). Beyond All-Sky: Assessing Ecological Light Pollution Using Multi-Spectral Full-Sphere Fisheye Lens Imaging. *Journal of Imaging*, *5*(4), 1–17. https://doi.org/10.3390/jimaging5040046

Jechow, A., Ribas, S. J., Domingo, R. C., Hölker, F., Kolláth, Z., & Kyba, C. C. M. (2018). Tracking the dynamics of skyglow with differential photometry using a digital camera with fisheye lens. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *209*, 212–223. https://doi.org/10.1016/j.jqsrt.2018.01.032

Jeng-Ren Hwang, Shyi-Ming Chen, & Chia-Hoang Lee. (1998). Handling forecasting problems using fuzzy time series. *Fuzzy Sets and Systems*, *100*(1–3), 217–228. https://doi.org/10.1016/S0165-0114(97)00121-8

Kaynak, E., & Cavlek, N. (2007). Measurement of Tourism Market Potential of Croatia by Use of Delphi Qualitative Research Technique. *Journal of East-West Business*, *12*(4), 105–123. https://doi.org/10.1300/J097v12n04_05

Khakifirooz, M., Chien, C. F., & Chen, Y.-J. (2018). Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0. *Applied Soft Computing*, *68*, 990–999. https://doi.org/10.1016/j.asoc.2017.11.034

Kocifaj, M. (2007). Light-pollution model for cloudy and cloudless night skies with ground-based light sources. *Applied Optics*, *46*(15), 3013–3022. https://doi.org/10.1364/AO.46.003013

Kocifaj, M. (2014). Night sky luminance under clear sky conditions: Theory vs. experiment. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *139*, 43–51. https://doi.org/10.1016/j.jqsrt.2013.12.001

Kocifaj, M. (2017). Retrieval of angular emission function from whole-city light sources using night-sky brightness measurements. *Optica*, *4*(2), 255–262. https://doi.org/10.1364/OPTICA.4.000255

Kocifaj, M., Kómar, L., & Kundracik, F. (2018). PePSS - A portable sky scanner for measuring extremely low night-sky brightness. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *210*, 74–81. https://doi.org/10.1016/j.jqsrt.2018.02.017

Koide, T., & Ishii, H. (2005). The hotel yield management with two types of room prices, overbooking and cancellations. *International Journal of Production Economics*, *93–94*, 417–428. https://doi.org/10.1016/j.ijpe.2004.06.038

Korte, D. (2013). Business Intelligence in the Hospitality Industry. *International Journal of Innovation, Management and Technology*. https://doi.org/10.7763/IJIMT.2013.V4.435

Koupriouchina, L., van der Rest, J.-P., & Schwartz, Z. (2014). On revenue management and the use of occupancy forecasting error measures. *International Journal of Hospitality Management*, *41*, 104–114. https://doi.org/10.1016/j.ijhm.2014.05.002

Kourentzes, N., Rostami-Tabar, B., & Barrow, D. K. (2017). Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? *Journal of Business Research*, *78*, 1–9. https://doi.org/10.1016/j.jbusres.2017.04.016

Kuhn, M., Weston, S., Culp, M., Coulter, N., & Quinlan, R. (2018). C50: C5.0 Decision Trees and Rule-Based Models (Version 0.1.2). Available online: https://CRAN.R-project.org/package=C50 (accessed on 10 October 2019).

Kyba, C. C. M. (2018). Is light pollution getting better or worse? *Nature Astronomy*, *2*(4), 267–269. https://doi.org/10.1038/s41550-018-0402-7

Kyba, C. C. M., Kuester, T., Sánchez de Miguel, A., Baugh, K., Jechow, A., Hölker, F., … Guanter, L. (2017). Artificially lit surface of Earth at night increasing in radiance and extent. *Science Advances*, *3*(1–8), e1701528. https://doi.org/10.1126/sciadv.1701528

Kyba, C. C. M., Tong, K. P., Bennie, J., Birriel, I., Birriel, J. J., Cool, A., … Gaston, K. J. (2015). Worldwide variations in artificial skyglow. *Scientific Reports*, *5*(1), 1–5. https://doi.org/10.1038/srep08409

Law, R. (2000). Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. *Tourism Management*, *21*(4), 331–340. https://doi.org/10.1016/S0261-5177(99)00067-9

Law, R., & Wong, R. (2010). Analysing Room Rates and Terms and Conditions for the Online Booking of Hotel Rooms. *Asia Pacific Journal of Tourism Research*, *15*(1), 43–56. https://doi.org/10.1080/10941660903310102

Lawrence, R. D., & Cherrier, J. (2003). *Passenger-Based Predictive Modeling of Airline No-show Rates*. 10. https://doi.org/10.1145/956750.956796

Lee, C. K., Song, H. J., & Mjelde, J. W. (2008). The forecasting of International Expo tourism using quantitative and qualitative techniques. *Tourism Management*, *29*(6), 1084–1098. https://doi.org/10.1016/j.tourman.2008.02.007

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

Li, S., Chen, T., Wang, L., & Ming, C. (2018). Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management*, *68*, 116–126. https://doi.org/10.1016/j.tourman.2018.03.006

Lin, V. S., & Song, H. (2015). A review of Delphi forecasting research in tourism. *Current Issues in Tourism*, *18*(12), 1099–1131. https://doi.org/10.1080/13683500.2014.967187

MacCarthy, B. L., Blome, C., Olhager, J., Srai, J. S., & Zhao, X. (2016). Article information : *International Journal of Operations & Production Management*, *36*(12), 1696–1718. https://doi.org/10.1108/02656710210415703

Mahajan, S., Chen, L.-J., & Tsai, T.-C. (2018). Short-Term PM2.5 Forecasting Using Exponential Smoothing Method: A Comparative Analysis. *Sensors*, *18*(10), 1–15. https://doi.org/10.3390/s18103223

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chih-Chung, C., & Chih-Chen, L. (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien (Version 1.7-2). Available online: https://CRAN.R-project.org/package=e1071 (accessed on 13 October 2019).

Minghetti, V. (2003). Building customer value in the hospitality industry: towards the definition of a customer-centric information system. *Information Technology & Tourism*, *6*(2), 141–152. https://doi.org/10.3727/109830503773048246

Momeni, E., Nazir, R., Jahed Armaghani, D., & Maizir, H. (2014). Prediction of pile bearing capacity using a hybrid genetic algorithm-based ANN. *Measurement*, *57*, 122–131. https://doi.org/10.1016/j.measurement.2014.08.007

Moutinho, L., Huarng, K.-H., Yu, T. H.-K., & Chen, C.-Y. (2008). Modeling and forecasting tourism demand: The case of flows from Mainland China to Taiwan. *Service Business*, *2*(3), 219–232. https://doi.org/10.1007/s11628-008-0037-3

Moutinho, L., & Witt, S. F. (1995). Forecasting the Tourism Environment Using a Consensus Approach. *Journal of Travel Research*, *33*(4), 46–50. https://doi.org/10.1177/004728759503300407

Mubiru, K. P. (2014). *A Markov Decision Model to Optimize Hotel Room Occupancy under Stochastic Demand*. *3*(4), 6.

Murali, R. V., Puri, A. B., & Prabhakaran, G. (2010). *GA-Driven ANN Model for Worker Assignment into Virtual Manufacturing Cells*. 5.

Najafabadi, S. S. (2012). Astronomical Tourism (Astro-Tourism) in Cebu, Philippines: Essential Features In Selected Destinations and Its Complementing Visitor Attractions. *International Conference on Trade, Tourism and Management*, 5.

Nasseri, M., Asghari, K., & Abedini, M. J. (2008). Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network. *Expert Systems with Applications*, *35*(3), 1415–1421. https://doi.org/10.1016/j.eswa.2007.08.033

Neuling, R., Riedel, S., & Kalka, K.-U. (2004). New approaches to origin and destination and no-show forecasting: Excavating the passenger name records treasure. *Journal of Revenue and Pricing Management*, *3*(1), 62–72. https://doi.org/10.1057/palgrave.rpm.5170094

Observatur. (n.a). Available online: http://www.observatur.es/observatur/ (accessed on 13 October 2019).

Ohyver, M., & Pudjihastuti, H. (2018). Arima Model for Forecasting the Price of Medium Quality Rice to Anticipate Price Fluctuations. *Procedia Computer Science*, *135*, 707–711. https://doi.org/10.1016/j.procs.2018.08.215

Opitz, & Maclin. (1999). Popular ensemble methods - an empirical study. *Journal or Artificial Intelligence Research, 11, 169*–198.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? *Machine Learning and Data Mining in Pattern Recognition*, *7376*, 154–168. https://doi.org/10.1007/978-3-642-31537-4_13

Ostaijen, V., Bruno, F., Ostaijen, T. Van, & Santos, B. F. (2017). Delft University of Technology Dynamic Airline Booking Forecasting. *Proceedings of the 21st Air Transport Research Society World Conference*.

Pan, B., & Yang, Y. (2017). Forecasting Destination Weekly Hotel Occupancy with Big Data. *Journal of Travel Research*, *56*(7), 957–970. https://doi.org/10.1177/0047287516669050

Peng, B., Song, H., & Crouch, G. I. (2014). A meta-analysis of international tourism demand forecasting and implications for practice. *Tourism Management*, *45*, 181–193. https://doi.org/10.1016/j.tourman.2014.04.005

Pereira, L. N. (2016). An introduction to helpful forecasting methods for hotel revenue management. *International Journal of Hospitality Management*, *58*, 13–23. https://doi.org/10.1016/j.ijhm.2016.07.003

Pfeifer, P. E., & Bodily, S. E. (1990). A test of space-time arma modelling and forecasting of hotel data. *Journal of Forecasting*, *9*(3), 255–272. https://doi.org/10.1002/for.3980090305

Portal to the heritage of the Astronomy. (n.a). Available online: https://www3.astronomicalheritage.net/index.php (accessed on 13 October 2019).

Posch, T., Binder, F., & Puschnig, J. (2018). Systematic measurements of the night sky brightness at 26 locations in Eastern Austria. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *211*, 144–165. https://doi.org/10.1016/j.jqsrt.2018.03.010

Pravettoni, M., Strepparava, D., Cereghetti, N., Klett, S., Andretta, M., & Steiger, M. (2016). Indoor calibration of Sky Quality Meters: Linearity, spectral responsivity and uncertainty analysis. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *181*, 74–86. https://doi.org/10.1016/j.jqsrt.2016.03.015

Puschnig, J., Schwope, A., Posch, T., & Schwarz, R. (2013). The night sky brightness at Potsdam-Babelsberg including overcast and moonlit conditions (Note). *Journal of Quantitative Spectroscopy and Radiative Transfer*, *Volume 139, pp. 76–81*.

R Core Team. (2013). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, *Available online: http://www.R-project.org/* (accessed on 13 October 2019).

Rajopadhye, M., Ghalia, M. B., Wang, P. P., Baker, T., & Eister, C. V. (2001). Forecasting uncertain hotel room demand. *Information Sciences*, 11.

Reikard, G. (2009). Predicting solar radiation at high resolutions: A comparison of time series forecasts. *Solar Energy*, *83*(3), 342–349. https://doi.org/10.1016/j.solener.2008.08.007

Rokach, L., & Maimon, O. (2015). *Data mining with decision trees: Theory and applications* (Second edition). Hackensack, New Jersey: World Scientific.

Romero Morales, D., & Wang, J. (2010). Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research*, *202*(2), 554–562. https://doi.org/10.1016/j.ejor.2009.06.006

Sánchez de Miguel, A., Aubé, M., Zamorano, J., Kocifaj, M., Roby, J., & Tapia, C. (2017). Sky Quality Meter measurements in a colour-changing world.

Monthly Notices of the Royal Astronomical Society, *467*(3), 2966–2979. https://doi.org/10.1093/mnras/stx145

Ściężor, T., & Kubala, M. (2014). Particulate matter as an amplifier for astronomical light pollution. *Monthly Notices of the Royal Astronomical Society*, *444*(3), 2487–2493. https://doi.org/10.1093/mnras/stu1577

Scrucca, L. (2013). GA: A Package for Genetic Algorithms in R (Version Journal of Statistical Software, 53(4), 1-37). Available online: http://www.jstatsoft.org/v53/i04/ (accessed on 13 October 2019).

Sierag, D. D., Koole, G. M., van der Mei, R. D., van der Rest, J. I., & Zwart, B. (2015). Revenue management under customer choice behaviour with cancellations and overbooking. *European Journal of Operational Research*, *246*(1), 170–185. https://doi.org/10.1016/j.ejor.2015.04.014

Slingo, J., & Palmer, T. (2011). Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *369*(1956), 4751–4767. https://doi.org/10.1098/rsta.2011.0161

Solano Lamphar, H. A. (2018). The emission function of ground-based light sources: State of the art and research challenges. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *211*, 35–43. https://doi.org/10.1016/j.jqsrt.2018.02.034

Soleimani, S., Bruwer, J., Gross, M. J., & Lee, R. (2018). Astro-tourism conceptualisation as special-interest tourism (SIT) field: A phenomonological approach. *Current Issues in Tourism*, *12*, 1–16. https://doi.org/10.1080/13683500.2018.1444021

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting-A review of recent research. *Tourism Management*, *29*(2), 203–220. https://doi.org/10.1016/j.tourman.2007.07.016

Song, H., Li, G., Witt, S. F., & Athanasopoulos, G. (2011). Forecasting tourist arrivals using time-varying parameter structural time series models. *International Journal of Forecasting*, *27*(3), 855–869. https://doi.org/10.1016/j.ijforecast.2010.06.001

Song, H., Qiu, R. T. R., & Park, J. (2019). A review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting. *Annals of Tourism Research*, *75*, 338–362. https://doi.org/10.1016/j.annals.2018.12.001

Talluri, K. T., & Ryzin, G. J. V. (2005). *The theory and practice of revenue management*. New York, NY: Springer. International series in operations research & management science.

Tang, C. M. F., King, B., & Pratt, S. (2017). Predicting hotel occupancies with public data: An application of OECD indices as leading indicators. *Tourism Economics*, *23*(5), 1096–1113. https://doi.org/10.1177/1354816616666670

TESS Photometer – Telescope Encoder and Sky Sensor. (n.d.). Available online: http://tess.stars4all.eu/ (accessed on 13 October 2019).

The Starlight foundation. (n.a). Available online: https://www.fundacionstarlight.org/ (accessed on 13 October 2019).

Tideswell, C., Mules, T., & Faulkner, B. (2001). An Integrative Approach to Tourism Forecasting: A Glance in the Rearview Mirror. *Journal of Travel Research*, *40*(2), 162–171. https://doi.org/10.1177/004728750104000207

Tse, T. S. M., & Poon, Y. T. (2015). Analyzing the use of an advance booking curve in forecasting hotel reservations. *Journal of Travel and Tourism Marketing*, *32*(7), 852–869. https://doi.org/10.1080/10548408.2015.1063826

Uysal, M., & Crompton, J. L. (1985). An Overview of Approaches Used to Forecast Tourism Demand. *Journal of Travel Research*, *23*(4), 7–15. https://doi.org/10.1177/004728758502300402

Walker, C. E., Pompea, S. M., & Isbell, D. (2008). GLOBE at Night 2.0: On the Road Toward IYA 2009. *ASP Conference Series*, *389*, 423–424.

Weatherford, L. (2016). The history of forecasting models in revenue management. *Journal of Revenue and Pricing Management*, *15*(3–4), 212–221. https://doi.org/10.1057/rpm.2016.18

Weaver, D. (2011). Celestial ecotourism: New horizons in nature-based tourism. *Journal of Ecotourism*, *10*(1), 38–45. https://doi.org/10.1080/14724040903576116

Weaver, D. B., & Lawton, L. J. (2007). Twenty years on: The state of contemporary ecotourism research. *Tourism Management*, *28*(5), 1168–1179. https://doi.org/10.1016/j.tourman.2007.03.004

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*, 79–82. https://doi.org/10.3354/cr030079

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*. 11.

World Tourism Organization - Tourism Highlights (2018). *World Tourism Organization*. https://doi.org/10.18111/9789284419876

World Tourism Organization (n.a). Available online: http://www2.unwto.org/ (accessed on 13 October 2019).

Wu, D. C., Song, H., & Shen, S. (2017). New developments in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management*, *29*(1), 507–529. https://doi.org/10.1108/IJCHM-05-2015-0249

Yu, G., & Schwartz, Z. (2006). Forecasting short time-series tourism demand with artificial intelligence models. *Journal of Travel Research*, *45*(2), 194–203. https://doi.org/10.1177/0047287506291594

Yu, L., Lai, K. K., & Wang, S. (2008). Multistage RBF neural network ensemble learning for exchange rates forecasting. *Neurocomputing*, *71*(16–18), 3295–3302. https://doi.org/10.1016/j.neucom.2008.04.029

Yüksel, S. (2005). An integrated forecasting approach for hotels. *The International Symposium on the Analytic Hierarchy Process (ISAHP) 2005*, 10.

Yüksel, S. (2007). An integrated forecasting approach to hotel demand. *Mathematical and Computer Modelling*, *46*(7–8), 1063–1070. https://doi.org/10.1016/j.mcm.2007.03.008

Zakhary, A., Atiya, A. F., El-Shishiny, H., & Gayar, N. E. (2011). Forecasting hotel arrivals and occupancy using Monte Carlo simulation. *Journal of Revenue and Pricing Management*, *10*(4), 344–366. https://doi.org/10.1057/rpm.2009.42

Zamorano, J., García, C., Tapia, C., Sánchez de Miguel, A., Pascual, S., & Gallego, J. (2017). STARS4ALL Night Sky Brightness Photometer. *International Journal of Sustainable Lighting*, *18*, 49–54. https://doi.org/10.26607/ijsl.v18i0.21

Zhang, G., Wu, J., Pan, B., Li, J., Ma, M., Zhang, M., & Wang, J. (2017). Improving daily occupancy forecasting accuracy for hotels based on EEMD-ARIMA model. *Tourism Economics*, *23*(7), 1496–1514. https://doi.org/10.1177/1354816617706852

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms* (1st ed.). https://doi.org/10.1201/b12207

A. Resumen en castellano

A.1 Introducción y objetivos

El turismo es una de las actividades con mayor crecimiento a lo largo de todo el mundo. Según la Organización Mundial del Turismo, ésta puede ser definida como "las actividades que realizan las personas durante sus viajes y estancias en lugares distintos a su entorno habitual durante un período de tiempo inferior a un año, con fines de ocio, negocios u otros". Lo cierto es que el turismo es una de las industrias más expandidas a lo largo de todo el mundo y genera ingresos en los países en donde se desarrolla a través de impuestos o consumo de bienes y servicios, tales como importaciones al consumo o promoción del producto local, a la vez que permite el afloramiento de empresas que proveen todo tipo de servicios complementarios a las cadenas hoteleras: transporte, mantenimiento, sistemas de la información, etc. Sin embargo, su actividad de desarrolla en un ambiente cambiante e incierto, supeditado a numerosos factores externos tales como el tiempo, desastres naturales o inestabilidad política; mientras que al mismo tiempo la gran expansión que ha experimentado ha hecho que aumente fuertemente la competitividad. Esto ha obligado a las empresas del sector a invertir en nuevos mecanismos que le permitan seguir siendo competitivos mediante la atracción de nuevos clientes, así como en la gestión de las reservas en curso. En este sentido, las tecnologías de la información y la comunicación (TIC) han aportado un nuevo campo de desarrollo para la construcción de nuevas herramientas que permitan generar valor añadido a las empresas del sector. El uso de las TIC con el objetivo de aportar enfoques innovativos dentro de la industria turística, y así contribuir a su mejora se conoce como turismo inteligente o Smart Tourism.

La presente tesis doctoral tiene como objetivo explorar las posibilidades que tienen las técnicas basadas en inteligencia artificial dentro del sector turístico. Esto ha llevado a plantear varias casuísticas de diferente naturaleza en donde se aplican múltiples técnicas basadas en inteligencia artificial, motivo por el cual esta tesis ha sido desarrollada en un formato basado en casos. La primera línea de investigación aborda el desarrollo de herramientas que permitan promocionar un turismo alternativo, más concretamente, se tomó el caso del astroturismo, enmarcado dentro del turismo natural y que se encuentra en fase de crecimiento. El desarrollo propuesto plantea la predicción de las condiciones nocturnas de los cielos con el objetivo de sentar las bases de un sistema de información para aquellas personas interesadas sobre los lugares y mejores horarios para disfrutar de esta experiencia. La segunda línea trata de contribuir a mejorar la gestión hotelera, desarrollando en esta ocasión dos casos asociados a cancelaciones hoteleras. Por un lado, se plantea la predicción de las cancelaciones que pudiera sufrir un hotel desde que el cliente emplaza su reserva. El objetivo de este caso va más allá de intentar predecir los ratios de cancelación, pretendiendo detectar qué clientes de manera individual tienen una alta probabilidad de cancelar. Por otro lado, se plantea la predicción de cancelaciones críticas, entendiéndose éstas como aquellas cancelaciones con alta probabilidad de cancelar días previos a la llegada prevista del huésped. Al igual que en el caso anterior, la intención principal es identificar qué clientes, de manera individualizada, son proclives a emplazar este tipo de cancelaciones.

Atendiendo a las casuísticas expuestas, se detallan a continuación los siguientes objetivos generales (O.G), así como objetivos específicos (O.E) propuestos para este trabajo:

O.G 1: Identificar oportunidades para la implantación de algoritmos basados en inteligencia artificial que permitan mejorar la gestión hotelera:

> **O.E 1.1:** Modelar cancelaciones hoteleras a través de técnicas basadas en inteligencia artificial.
>
> **O.E 1.2:** Establecer un procedimiento generalista para predecir las cancelaciones hoteleras usando únicamente las variables más comunes

solicitadas al cliente durante el proceso de reserva a través de plataformas de reservas online.

O.E 1.3: Establecer qué técnicas son las más adecuadas de entre las empleadas en la investigación.

E.O 1.4: Identificar aquellas cancelaciones a ser emplazadas cerca de la fecha de entrada del huésped.

O.G 2: Desarrollar las bases de un sistema de información, basado en técnicas de inteligencia artificial, que permita a los turistas saber acerca de los lugares y mejores horas para desarrollar actividades de astroturismo.

O.E 2.1: Modelar series temporales de brillo de cielo nocturno

O.E 2.2: Comparar técnicas tradicionalmente usadas para el análisis de series temporales con aquellas técnicas basadas en inteligencia artificial.

A.2 Casos estudiados

En esta sección se detalla el marco en donde se engloban los casos estudiados, así como las metodologías propuestas y los resultados derivados de los mismos. Para facilitar la correcta identificación de los mismos, cada uno de ellos llevan un título asociado.

Caso 1: Astroturismo y la predicción de brillo celeste nocturno: Primer modelo probabilístico.

El astroturismo se puede definir como el interés de un individuo, ya sea un astrónomo amateur o profesional, que viaja a otros lugares diferentes al de origen para observar los cielos. Los diferentes fenómenos estelares, tales como estrellas, cometas o lluvia de meteoritos, pasando por eclipses solares/lunares o puestas de sol, atraen a personas interesadas en su observación. En este sentido, han aparecido numerosas organizaciones cuyo objetivo principal es la conservación de la oscuridad de los cielos y evitar el incremento de la contaminación lumínica. Entre los proyectos más destacados cabe mencionar el “Astronomy and World Heritage” creado por la UNESCO o “the Globe at Night programme” fundado por el observatorio óptico astronómico nacional de Los Estados Unidos. Este último cuenta con una red de sensores de brillo de cielo nocturno a lo largo de todo el mundo, y cuyos datos han permitido el desarrollo de este estudio. Cabe destacar que, si bien existen numerosos modelos generalistas basados en la física para la predicción del brillo estelar en cualquier punto de la Tierra, hasta ahora no se conocía de la existencia de modelos probabilísticos que permitieran predecir esta magnitud. Es por tanto que, la metodología propuesta en esta tesis presenta una novedosa contribución en la materia.

La calidad de los cielos oscuros puede ser medida mediante distintos tipos de aparatos de medida, de entre los que destaca el equipo “Sky Quality Metre”, presentado por la empresa Unihedron, debido a su tamaño, precio y portabilidad (figura 1). Este equipo ha sido usado por programas como “The Globe at Night” para la medición del brillo nocturno estelar y cuyo histórico de datos ha permitido del desarrollo de este trabajo.

Figura 1. Sky Quality Meter – LE. Fuente: Unihedron.

La metodología propuesta en este trabajo de tesis presenta un tratamiento preliminar de los datos que posteriormente son usados para crear un modelo basado en AI, concretamente mediante redes neuronales artificiales, y análisis clásico de series temporales, específicamente a través de modelos ARIMA (Modelo autorregresivo integrado de media móvil). A través del procedimiento propuesto se ha conseguido predecir los valores de NSB con un alto índice de acierto, alcanzado hasta un 94% de correlación entre el valor real y el predicho.

A raíz de los resultados obtenidos se puede concluir con que este tipo de técnicas no sólo son de utilidad en la materia, sino que además presentan una alta tasa de acierto. Futuras líneas de investigación en este área podrían ir encaminadas en el uso de distintos tipos de parametrización, así como la aplicación de esta metodología a un conjunto de datos distintos a los empleados en este trabajo.

El trabajo presentado en el *caso 1* ha sido publicado en la revista *Sensors* clasificada como Q1 acorde a los requerimientos JCR.

Caso 2: Prediciendo cancelaciones de reservas hoteleras: Un enfoque desde la inteligencia artificial.

Las cancelaciones son uno de los aspectos fundamentales a tener en cuenta dentro de la gestión hotelera debido al impacto que generan en el sistema de reservas, y sin embargo, la literatura relevante en este ámbito es escasa. De hecho, muy poco se sabe sobre los motivos por los que un cliente decide cancelar o cómo poder evitar la cancelación.

El objetivo de este caso se centra en predecir aquellos clientes con una alta probabilidad de cancelar su reserva a través de datos de reservas hoteleras o PNR (personal name records), esto es, los datos recogidos del cliente a la hora de realizar la reserva. La intención fue la de establecer una metodología con las variables más comunes empleadas en la industria hotelera, pero que además permita hacerlo mediante un procedimiento que simplifique, en la medida de lo posible, el procesado de los datos. Atendiendo a esta premisa, el procedimiento propuesto no hace uso de datos de carácter personal para conocer el histórico del cliente, al igual que no emplea fuentes de datos externas, lo que hace que la metodología presentada tenga un carácter muy práctico y facilite su implantación en las empresas del sector.

Para la elaboración del caso que se presenta se empleó un set de datos con más de 10.000 reservas turísticas reales recogidas durante los años 2016 y 2018, correspondientes a un hotel colaborador ubicado en el sur de la isla de Gran Canaria. Aplicando distintas técnicas basadas en inteligencia artificial fue posible extraer resultados realmente impresionantes, logrando alcanzar hasta un 98% de exactitud en la predicción de las cancelaciones de este hotel. Cabe destacar que no solamente se obtienen ratios de cancelación, sino, además, saber qué clientes tienen una alta probabilidad de cancelar. Esto permitiría a los hoteles ajustar sus estrategias de overbooking, pricing o políticas de cancelaciones, entre otras, de tal forma que podrían reducir los efectos negativos que estas conllevan.

El trabajo presentado para el *caso 2* se encuentra en proceso de revisión en la revista *International Journal of Hospitality Management*, clasificada como D1 (decil 1) acorde a los requerimientos JCR.

Caso 3: Identificando cancelaciones hoteleras críticas mediante el uso de inteligencia artificial.

Como se comentaba en el caso anterior, las cancelaciones suponen un gran impacto a las industrias hoteleras y son un aspecto crítico del sistema de gestión. Especialmente aquellas cancelaciones que son emplazadas pocos días antes de la llegada del cliente generan importantes pérdidas en los establecimientos ya que dejan con muy poco margen de maniobra a los gestores del hotel para dar salida a este inventario. De hecho, en muchas ocasiones se ven obligados a dar por perdida la venta o conseguir revender esta habitación a un precio muy inferior del original. A esto se le suma la creciente tendencia del mercado en la que los clientes realizan varias reservas para, posteriormente, optar por una de ellas y cancelar el resto en el último momento. Esta situación se ve aún más agravada cuando salen al mercado ofertas del tipo “last minute”, bien sea por parte del hotel en cuestión o por sus competidores al fomentar el emplazamiento de cancelaciones en favor de opciones más económicas.

Atendiendo a la problemática descrita, el objetivo de este caso ha sido identificar aquellos individuos con una alta posibilidad de cancelar su reserva a pocos días de su fecha de llegada (entre 4 y 7 días antes de su ingreso) desde el mismo momento en que realiza la reserva. Al igual que en el caso anterior, se presenta una metodología capaz de lograr este objetivo sin incurrir en la necesidad de utilizar el historial particular del cliente o consultar bases de datos externas, simplificando también en esa ocasión el proceso de identificación de estos individuos. La casuística abordada en esta ocasión se caracteriza por tener un número de casos positivos muy bajo, al igual que un set de datos altamente desequilibrado.

Empleando distintas técnicas basadas en inteligencia artificial y un algoritmo de ensamblado se han conseguido identificar hasta un 80% de clientes proclives a cancelar 7 días antes de su fecha de entrada. Enfoques como el presentado permitirían a las empresas conocer de antemano aquellas cancelaciones críticas, y por lo tanto, podrían tomar acciones para intentar evitar que el cliente cancele, tales como contactar directamente con ellos por vía telefónica o email (enviando un recordatorio, por ejemplo). De forma similar al caso anterior, dan un nuevo enfoque del problema a los gestores del hotel para poder establecer un ajuste más fino de las estrategias de overbooking, pricing o cancelaciones, lo que les aporta una ventaja competitiva respecto a sus competidores.

A.3 Conclusiones

Acorde a los excelentes resultados conseguidos en este trabajo, se puede concluir con que la inteligencia artificial es extraordinariamente útil para la resolución de distintos problemas que se dan en la industria turística, independientemente de la naturaleza de la que provengan. Esto permite no solo conocer mejor sobre los clientes, sino que además permite adaptarse a sus necesidades. En este sentido, este trabajo ha abordado este problema desde dos puntos de vista completamente diferentes, por una parte, el caso de los procesos internos en las compañías hoteleras, como son la predicción de cancelaciones en establecimientos hoteleros, y por otra, la forma en la que las técnicas basadas en inteligencia artificial son capaces de contribuir a mejorar la experiencia del cliente mediante la oferta de nuevos productos/servicios. Este último caso fue abordado mediante la predicción del brillo celeste nocturno, con la intención de proporcionar a los astrónomos, aficionados o profesionales, las mejores horas para disfrutar de la experiencia del astroturismo, permitiéndoles realizar los viajes de manera más eficiente y evitando pérdidas de tiempo innecesarias. Para todos los casos, tanto para la predicción de brillo estelar nocturno como de las cancelaciones hoteleras, las metodologías propuestas parten del uso de datos reales, en el primer caso provenientes de un programa mundial de mediciones y en otro de un conjunto de datos de reservas hoteleras provistas por un hotel colaborador. Uno de los aspectos más destacados de los procesos expuestos en este trabajo es que permiten abordar las casuísticas presentadas desde un punto de vista muy práctico a la vez que generalista, lo que permite una fácil exportación de las metodologías propuestas a otros entornos.

UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA