International Journal of Hospitality Management 121 (2024) 103802

Contents lists available at ScienceDirect

International Journal of Hospitality Management

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

Decoding the future: Proposing an interpretable machine learning model for hotel occupancy forecasting using principal component analysis

Daniele Contessi [a], Luciano Viverit [b], Luís Nobre Pereira [c], Cindy Yoonjoung Heo [d,*]

[a] Travelbrain, Italy
[b] Hotelnet & Travelbrain, Italy
[c] Research Centre for Tourism, Sustainability and Well-being & Escola Superior de Gestão, Hotelaria e Turismo, Universidade do Algarve, Portugal
[d] EHL Hospitality Business School, HES-SO/ University of Applied Sciences and Art Western Switzerland, Switzerland

ARTICLE INFO

Keywords:
Hotel Demand Forecasting
Machine learning
Principal components analysis
Additive pickup method

ABSTRACT

Accurate hotel occupancy forecasting is vital for optimizing hotel revenue, yet interpretable machine learning tools lack extensive research. This paper presents a two-step approach utilizing historical and advanced booking data. Principal Components Analysis (PCA) groups similar patterns in booking curves, followed by a pickup forecasting model to predict occupancy. Evaluating the approach using real booking data from three European hotels (2018–2022), it outperformed two benchmarks: classical additive pickup and clustering-based pickup methods. Empirical results demonstrate the superiority of PCA-based methods across all hotels and forecasting horizons. Additionally, incorporating Average Daily Rates into PCA enhances daily hotel demand forecasts, offering potential for enhanced predictions with business operational information in a low-dimensional space.

1. Introduction

In recent years, with the continuous advancement of data processing technology, more and more industry practitioners and academic researchers have begun to analyse business data by using computer intelligence algorithms. The hospitality industry stands out as a data-rich sector, receiving substantial volumes of information from diverse systems, encompassing property management, accounting and financial, channel management, revenue management, housekeeping, and customer relationship management systems. The hospitality industry has increasingly integrated various data analysis techniques exploiting large datasets, because business intelligence can provide useful strategic insights in highly competitive environments (Mariani et al., 2018). The hospitality industry, like any other industry, is witnessing a change due to predictive analytics powered by artificial intelligence (AI). AI is an umbrella term for a range of computer science techniques used to create autonomous machines or software that can learn, think, and make decisions independently. AI-based methods can be considered a family of highly sophisticated modelling approaches that include machine learning (ML) and deep learning (DL) methods (Sun et al., 2019; Law et al., 2019; Kaya et al., 2022). ML resides within the realm of AI, emphasizing algorithms designed to glean knowledge from data. Meanwhile, DL represents a facet of ML, employing artificial neural networks to emulate the cognitive learning mechanism of the human brain. ML creates computer algorithms designed to improve their accuracy as they analyze and learn from large volumes of data. As such, ML has been widely used to design algorithms based on a data set, in order to extract insights from data in many industries, which can be used to make future predictions and suggest different actions (Sun et al., 2019). Recently a number of scholars argue the importance of interpretable ML algorithms (Gilpin et al., 2018), as they allow us to understand the mechanics of what is going on and are not black-boxes that one has to "trust".

For instance, Viverit et al. (2023's) approach diverges from the traditional method of calculating the average of past data during a trailing period. It focuses on reservation patterns, allowing for consistently high forecasting accuracy even in situations where demand deviates significantly from historical trends and remains unstable. While the clustering technique is useful in assigning data inputs with similar features to a specific group, it may not be intuitive and requires the researcher or the user to interpret the results. Further, analyzing reservation patterns relies on the availability of a certain amount of booked data before employing clustering techniques based on the observed shapes. This means that accurate forecasts may be challenging at the onset of reservations, as the booking curve's shape is unknown until a certain level of reservations has been made. But hotels need to conduct

* Corresponding author.
E-mail address: cindy.heo@ehl.ch (C.Y. Heo).

https://doi.org/10.1016/j.ijhm.2024.103802
Received 12 August 2023; Received in revised form 2 May 2024; Accepted 17 May 2024
Available online 27 May 2024

demand forecasting quickly and accurately to implement revenue management policies accordingly.

Therefore, this study proposes an interpretable two-step approach to accurately forecast daily hotel demand through the principal components analysis (PCA) to extract common information from a high number of booking curves. PCA is often considered more interpretable than some other ML methods due to its ability to reduce the dimensionality of data while preserving the majority of the variance (Jolliffe and Cadima, 2016). Clustering, on the other hand, generates results that are more difficult to interpret because it groups data points into clusters without revealing the underlying structure of the data. The principal components (PC) extracted by PCA are orthogonal to each other, meaning they are uncorrelated. This orthogonality simplifies the interpretation of each component's contribution to the overall variance in the data. Each PC is a linear combination of the original features in the dataset. This linear relationship allows for a straightforward interpretation of how each feature contributes to the PC. Applying PCA addresses this challenge by differentiating booking curves based not on their shapes but on the key features inherent in each booking transaction. This approach enables the forecasting of demand even in the early stages of reservations, providing a valuable advantage.

The first step of this study is to use PCA to reduce reduction of a high-dimensional booking data of all stay dates to a low-dimensional space of new variables. In the second step, the additive pickup method is applied to each target date by averaging booking data of stay dates close to the target date in the low-dimensional space generated by the PCA. The forecasting performance of the proposed approach was evaluated against three benchmarks through the utilization of real booking data from three European hotels over a five-year period (2018–2022).

Given the hospitality industry's sensitivity to external factors (Chow et al., 1998), it is essential for managers to have access to a reliable and accurate demand forecast for informed decision making. To this end, we propose an approach that is both easy to implement and computationally efficient, making it an ideal solution for hotels and software developers seeking to improve hotel forecasts in real time. By leveraging this approach, managers can optimise inventory levels and confidently establish proper pricing strategies.

2. Literature review

2.1. Demand forecasting in tourism and hospitality

Demand forecasting serves as a pivotal tool, furnishing essential insights to bolster decision-making processes within the travel, tourism, and hospitality sectors (Wu et al., 2017). Consequently, it remains a cornerstone subject of investigation in these domains, capturing the interest of numerous researchers across time (e.g. Athiyaman and Robertson, 1992; Song and Li, 2008; Frechtling, 2012; Peng et al., 2014; Wu et al., 2017; Song et al., 2019; Bi et al., 2022); and the papers cited therein). Since precise forecasts of demand are of the utmost importance for travel and tourism planning, as well as for destination and organisational management (Song et al., 2019; Kaya et al., 2022), researchers have proposed and tested new approaches to deal with specific problems and to improve quality of forecasts, namely, to accommodate uncertainty.

According to literature in the areas of travel, tourism and hospitality (e.g., Song and Li, 2008; Peng et al., 2014; Wu et al., 2017), demand forecasting can be split into two lines of research: high-frequency short-term demand forecasting, and low-frequency medium-term demand forecasting. The former is usually applied to forecast hotel occupancy and air travel demand (e.g., daily occupancy of rooms in a hotel, seat occupancy in a flight between origin and destination pairs), because there is interest in forecasting demand in a short time ahead of a date of the service (e.g., up to seven days, two weeks, four weeks) based on hourly or daily data. The other line of research is usually applied to forecast tourism demand (e.g., number of arrivals at a tourism destination, a region, a country), because there is interest in forecasting demand for a medium horizon period (e.g., the coming months, quarters) based on monthly, quarterly, or yearly data.

Demand forecasting has been based on time-series methods, econometric models, and AI-based models (Li et al., 2005; Song and Li, 2008; Peng et al., 2014; Wu et al., 2017). While the time-series methods and econometric models have been widely applied since the second half of the last century to forecast tourism, hotel and air travel demand, the new AI-based methods started to be applied in this century (e.g., (Song and Li, 2008; Peng et al., 2014; Sánchez et al., 2021). Several recent studies have focused specifically on reviewing literature related to tourism demand forecasting (e.g., Song et al., 2019; Li et al., 2021).

Concerning time-series approaches in hotel demand analysis, existing literature identifies three distinct methods: historical booking methods, advanced booking methods, and combined methods (Fiori and Foroni, 2020; Weatherford and Kimes, 2003). Historical booking methods focus solely on the past-day's final occupancy, whereas advanced booking methods incorporate the build-up pattern of bookings during the lead time. In addition, combined methods integrate historical and advanced booking methods through techniques such as weighted averages or regression, aiming to enhance forecast accuracy.

The time-series approaches based on advanced booking information are mostly employed in the revenue management studies and have seen extensive utilization over time (e.g., Andrew et al., 1990; Chen and Kachani, 2007; Fiori and Foroni, 2020; Heo et al., 2024; Schwartz and Hiemstra, 1997; Tse and Poon, 2015; Weatherford and Kimes, 2003; Zakhary et al., 2011). This approach, known as pickup, assumes that occupancy patterns remain consistent for corresponding calendar periods and days of the week. Pickup methods facilitate hotel occupancy by utilizing advanced booking data, which encompass cumulative reservations for future stay dates (i.e., reservations on hand). As a result, this forecasting technique amalgamates historical data encompassing daily arrival patterns with length of stay details alongside advanced booking data. A few studies have specifically concentrated on reviewing literature related to hotel demand forecasting (e.g., Wu et al., 2017; Huang and Zheng, 2023). These studies have provided valuable insights into the methodologies and approaches used in forecasting hotel demand.

2.2. Artificial intelligence for demand forecasting in tourism and hospitality

The availability of big data has stimulated the broad application of AI-based models in the field of tourism and hotel demand forecasting in an attempt to improve forecast accuracy (Peng et al., 2014; Wu et al., 2017; Sanchez-Medina and C.-Sanchez, 2020). Recent research has highlighted the remarkable predictive capabilities of ML (e.g., Hassani et al., 2015; Hassani et al., 2017; Li et al., 2018; Silva et al., 2019; Ampountolas and Legg, 2021; Zhu et al., 2021; Pereira and Cerqueira, 2022; Zhang and Wu, 2023; Zhang and Niu, 2024) when contrasted with traditional demand forecasting methods that rely on time series and econometric models. Their appeal lies in their ability to operate without stringent data assumptions and their aptitude for nonlinear prediction due to their adaptable and nonlinear nature (Li et al., 2018; Song and Li, 2008). The body of literature concerning tourism demand forecasting utilizing AI techniques has been expanding rapidly (e.g. (Wu et al., 2017; Bi et al., 2022; Li et al., 2021; Bi et al., 2022; Xing et al., 2022; Zhang et al., 2021; Zhang et al., 2022). This growth is mirrored in the realm of air travel demand forecasting, with a surge in literature centered on specific model types (e.g., (Alarfaj and AlGhowinem, 2019; Firat et al., 2021).

The domain of tourism forecasting has witnessed the emergence of various ML models, including artificial neural networks (Burger et al., 2001; Kon and Turner, 2005), rough set approaches (Au and Law, 2000; Law and Au, 2000), and support vector machines (Pai et al., 2014; Kon and Turner, 2005), all of which have gained prominence in the context

of tourist arrival forecasting (Li et al., 2018).

In particular, the neural network techniques excel in addressing nonlinear forecasting problems (Yang et al., 2015). For instance, Koutras et al. (2016) conducted a comparative analysis between an array of AI models and linear counterparts to forecast demand within the hospitality industry, ultimately demonstrating the superior effectiveness of AI models. Other researchers have employed hybrid AI-based methods in forecasting tourism demand (Chen et al., 2015; Pai et al., 2014; Sun et al., 2016). Particularly, Li et al., 2018) forecasted Beijing City's in-bound monthly tourist volume using a hybrid model combining a dimensional reduction algorithm (supported by PCA), an optimisation algorithm and a neural network. Wen et al. (2019) proposed a hybrid model that amalgamates both linear and non-linear attributes from component models to predict the monthly influx of tourists from mainland China to Hong Kong.

Several studies have started investigating the application of AI-based models in forecasting daily hotel demand (e.g., Webb et al., 2020; Phumchusri and Ungtrakul, 2020); (Wang and Duggasani, 2020; Huang and Zheng, 2021; Huang et al., 2023; Pereira and Cerqueira, 2022; Viverit et al., 2023). Webb et al. (2020) analyzed the forecasting performance of a neural network approach within the framework of dynamic booking windows. Phumchusri and Ungtrakul (2020) compared artificial neural networks with traditional time series methods, including Box-Jenkins, Holt-Winters and exponential smoothing methods for complex seasonality patterns, to forecast hotel daily demand occupancy. Wang and Duggasani (2020) forecasted constrained hotel reservations based on advanced booking data using LSTM-based recurrent neural networks. Continuing along the same research trajectory, Huang and Zheng (2021) forecasted the daily demand of multiple hotels using a new spatiotemporal deep learning LSTM model that considers the agglomeration effect among hotels. Pereira and Cerqueira (2022) evaluated the accuracy of ML models through comparative analysis, pitting them against conventional approaches like exponential smoothing and seasonal naïve methods, using time series data of daily hotel demand. Recently, Huang et al. (2023) proposed a deep learning spatial-temporal forecasting model that simultaneously predicts the daily demand of multiple hotels considering external variables (demand, price and online rating of other hotels) into the forecasting model. The summary of the literature review on hotel demand forecasting using AI-based approaches is summarized in Appendix 1.

Almost all the above-mentioned studies compared traditional forecasting methods with AI-based methods. This was possible because the data used to run empirical studies have a relatively low dimensionality. Nevertheless, in the era of big data harnessing for forecasting problems, the prospect of dimensionality reduction merits consideration. This approach was followed by Assaf and Tsionas (2019); Li et al. (2018); Fang et al. (2023) and Zhang et al. (2023) to forecast low-frequency demand (tourism demand). However, to the best of the author's knowledge, this approach has not yet been applied to forecast high-frequency demand (i.e., daily hotel demand). In addition, several unresolved methodological challenges require future research to improve the accuracy of forecasts used in modern revenue management systems (RMS). Hence, this paper endeavours to bridge this research gap by employing an AI-based approach to reduce the dimensionality of booking data, aiming to enhance the forecast accuracy of daily hotel demand while also improving interpretability. Indeed, a major contribution of this paper relies on the use of the PCA as a dimensional reduction technique of booking curves, although its use in a wide range of applications, it was not yet applied in hotel demand forecasting problems. The PCA was only used a few times in the area of tourism demand forecast (Assaf and Tsionas, 2019; Li et al., 2018).

3. Research methods

The proposed forecasting approach can be implemented in four sequential steps:

1. Data collection and storage of booking data: historical and cumulative reservation data for each stay date, ie, the accumulated number of reservations over the booking window (per each lead time day) for each past stay date; and cumulative reservations on the book for future stay dates. A graphical representation of these booking data is known as the booking curve. Since each stay date is a variable with data from each day in advance with respect to the check-in, a big dataset is generated when data come from several consecutive years.
2. Dimension reduction of booking data: application of the PCA to the booking data to reduce dimensionality of a big dataset of booking data and generation of a small number of orthogonal PCs.
3. Forecasting daily hotel occupancy: implementation of forecasting methods usually used for advanced booking data: the additive and multiplicative pickup methods, and the cluster-based method.
4. Evaluation of accuracy of the forecasts: a set of accuracy measures, along with a statistical test, were applied to evaluate the superiority of the proposed approach.

3.1. Data

The dataset utilized in this study comprises authentic reservation data spanning five consecutive years (2018–2022) from three distinct hotels. Previous literature (e.g. Lee, 2018; Webb et al., 2020; Zakhary et al., 2011) has highlighted the significance of reservation data in short-term hotel demand forecasting. This data typically includes reservation date, arrival date, length of stay, room rate, and number of rooms reserved, all of which have been shown to correlate strongly with hotel occupancy levels. To assess the effectiveness of the proposed method across diverse market targets, three independent boutique hotels situated in different locales were chosen: a summer tourism destination, a year-round tourism destination, and a major metropolitan city. The chosen hotels, situated within Europe and more precisely in Italy and France, do not adopt advanced Revenue Management Systems (RMS). Hotel 1, situated on Isola D'Elba (Italy), operates as a three-star establishment with 32 rooms and is open exclusively from April to mid-October, catering exclusively to leisure customers. Meanwhile, Hotel 2, situated in Nice (France), is a four-star property with 47 rooms, and Hotel 3, located in Paris (France), offers 26 rooms within its four-star accommodations. Both Hotel 2 and Hotel 3 operate year-round, serving both business and leisure clientele. Key information within each reservation record encompasses the booking date, arrival date, length of stay (LOS), room rate, and the number of rooms reserved. Additional particulars concerning the dataset can be found in Table 1.

The data indicates that the average daily occupancy rate and the average LOS for Hotel 1 surpass those of the other two hotels. This variance can be attributed to Hotel 1's primary emphasis on catering to leisure tourists, coupled with its operational schedule that spans exclusively during the mid and high seasons. Hotel 2 has the shortest Average LOS (i.e., 3.7 days), and it also has the lowest ADR (€132.5). On the other hand, Hotel 3 in Paris has the highest ADR (i.e., €180.7) of the three hotels.

3.2. Principal component analysis (PCA)

In order to improve the quality of hotel demand forecasts without jeopardizing the interpretability of the applied method, this study developed a new method that combines a dimensionality reduction algorithm with a well-established forecasting model for hotel demand, namely the pickup method (Weatherford and Kimes, 2003; Fiori and Foroni, 2019; Fiori and Foroni, 2020), within a scope of a data-driven automatic approach. First the PCA is applied because it is particularly useful in those frameworks where the dimensionality of the data is high, like we have in the booking data, since it is a fast and fully invertible way of reducing it for the purpose of obtaining an effective, more tractable representation while preserving as much of the data variability as possible (Jolliffe and Cadima, 2016). The PCA is often employed as an

Table 1
Profile of the hotels and dataset.

	Property Profile						Data Profile		
	Location	Number of rooms	Average occupancy rate (%)	Average LOS (days)	ADR (€)	Average booking window (days)	Number of years	Number of days in the booking horizon	Number of observations
Hotel 1	Isola D'Elba, Italy	32	89	4.9	146.6	52.58	5	180	426
Hotel 2	Nice, France	47	78	3.7	132.5	52.28	5	180	423
Hotel 3	Paris, France	26	77	4.5	180.7	76.75	5	180	450

Note: LOS-Length of Stay; ADR-Average Daily Rate.

unsupervised learning algorithm of dimension reduction in several different applications (e.g. Jolliffe and Cadima, 2016; Wetzel, 2017), including in forecasting problems (Li et al., 2018; Fang et al., 2023; Zhang et al., 2023), because it has the advantage of increasing interpretability and simultaneously removing redundant information and multicollinearity (Jolliffe and Cadima, 2016). In our specific case, PCA was applied to the booking curves of the hotels. Assume that every booking curve for a given stay date might be interpreted as a vector, or variable, $\boldsymbol{x_j}$, $j = 1, \ldots, p$, of n components (one for each day in advance with respect to the check-in). As explained by Jolliffe and Cadima (Jolliffe and Cadima, 2016), it is possible to derive an optimal set of linear combinations of the p variables with maximum variance, generating a smaller number, q ($q < p$), of new orthogonal variables, called PC. These q PCs can be written as:

$$\begin{cases} z_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p \\ z_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p \\ \qquad\qquad \ldots \\ z_q = a_{q1}x_1 + a_{q2}x_2 + \ldots + a_{qp}x_p \end{cases} \tag{1}$$

where a_{ij} ($i = 1, \ldots, q$; $j = 1, \ldots, p$) are the PC loadings. It is important to highlight that the PCs are derived by descending order of variance explained, which means that the first PC, z_1, accounts for the largest amount of variability, the second PC, z_2, contains the second largest amount of variability, and so on. It means that a small set of PCs might preserve a high proportion of the total variance.

We decided to follow the common practice to identify the number of PCs that should be retained, namely to select the first PCs that account for at least 70% of the total variance and, whenever satisfied this condition, a maximum of three PCs in order to make a graphical representation that works as a way to increase interpretability of the proposed ML model (Jolliffe and Cadima, 2016). As a result, the first three PCs were selected in the empirical study applied to each hotel, which explained 90.0%, 97.2% and 97.9% the total variance, respectively, for each hotel. Thus, an excellent graphical representation in a low-dimensional space is possible as a way to approximately visualize a large and high-dimensional dataset.

In the hotel demand forecasting problem, every stay date is a point in a highly-dimensional space which cannot be visualised. PCA maps linearly all these points to a low dimensional space and the similarities in their behaviour can be fully appreciated as vicinities among them. In other words, similar booking curves become points that are close by in the human-readable low-dimensional space. On the other hand, dates that are far apart in the principal subspace exhibit booking curves with a different pattern. This vicinity criterion provides a measure of the similarity between booking curves and is the basic building block of the proposed method for the forecasts.

3.3. Clustering vs. principal component analysis

Another approach for identifying similar booking curves is to use clustering algorithms using a distance measure between multidimensional vectors and a threshold, like it was applied by Viverit et al. (Viverit et al., 2023). We plot in Appendix 2 the results of a cluster analysis for all the booking curves of Hotel 1. Clearly, each cluster contains curves with different behaviours as the bookings are made many days in advance (e.g. cluster 6), with a constant rate (e.g. cluster 5) or mainly last minute (e.g. cluster 7). For the sake of comparison, we performed a PCA on the same full booking curves of Hotel 1 and we show the obtained low dimensional space in Appendix 3. The relative positions among the dates (points) in this space is related to the different booking patterns as can be appreciated from the colours corresponding to the same clusters of Appendix 2. Thus, the PCA is able to preserve enough information about the booking curves' behaviour so the two approaches are affine.

Drawing inspiration from Viverit et al. (2023), a favourable strategy comes to light: initially grouping historical data of stay dates displaying similar booking behaviour to the target date, thereby laying the foundation for forecasting hotel demand in forthcoming bookings. However, in that case the classification of various dates after the clustering was done following an *a posteriori* protocol, meaning that the study considered all the curves as already fully expressed until the check-in day. Although the cluster-based approach was useful for identifying interesting correlations among data, it is not suited for an actual live forecasting because of the lack of full reservation values for the target date when forecasting n days in advance. For this purpose, in this work we simulated a real live forecast for every target date where only the reservations on hand are considered until a specific current reading day. On the other hand, PCA can be easily performed only on samples (portions of the booking curves) belonging to the same space size hence having the same number of components. The comparison between the PCA approach and the clustering shown Appendix 3 is intended to give a hint to the physical meaning of PCA, and it's not black magic.

Concretely, we assumed we knew the reservations for a target date until a reading date which is t days in advance with respect to the check-in and we wanted to forecast the final occupancy at time 0 (check-in). We considered all the booking curves at our disposal corresponding to stay dates before the target date, in the booking window from t days before to the beginning of their booking horizon (180 days before). All this samples are now of the same size: $180 - t$. At this point we applied the PCA, retaining only the first three PCs for each sample and hence obtained a low-dimensional representation for every booking curve. The pickup method was applied at each target date by averaging only on those dates (samples) closest to the target date itself in the low dimensional space. Our assumption is that we can generate a good final reservation forecast by resorting to already expressed past booking curves with respect to the target date which have a similar booking behaviour up to t days before. The average was weighted according to the inverse squared distance between the target date and the neighbouring points in the low-dimensional PCA space, up to a threshold distance.

The same live forecasting can be performed also with the cluster-based approach (Viverit et al., 2023). Considering only the truncated

curves, the segmentation is done with a distance-based method (e.g. Ward or Complete Linkage methods) and the pickup average is considered among only those curves belonging to the same cluster. We compared the two approaches in the Results section. One advantage of using PCA is that, in principle, no threshold must be chosen and the average is weighted according to the metric of the distance in the low-dimensional space. The result is that really closed points to the target date, hence the ones displaying a very similar booking pattern, are taken much more into consideration with respect to further points.

Another major advantage of PCA when compared to clustering is the natural addition of *other* orthogonal components. While PCA is reasonably applied to the homogeneous data of the booking curves, other information such as ADR and the previous year final occupancy can be considered independent features of every stay date. Adding them as additional components to the low-dimensional representation of the samples (after being properly normalised) can enhance the quality of the forecasts. We expect that if those properties are considered as degrees of freedom of every stay date, the pickup average based on vicinity criterion will benefit from the extra relevant information. In Appendix 4 we show the same data as of Appendix 2 but this time we plot the first two PCs together with ADR. The dates appear to form almost clean separated clouds in such a space.

3.4. Forecasting methods

A well-established forecasting model for hotel demand, namely the additive pickup method (Weatherford and Kimes, 2003; Fiori and Foroni, 2019), was applied after dimensionality reduction of the booking data. We opted for this classical method based on the assumption of independence between the number of on-hand reservations and the subsequent bookings for rooms. In addition, it ensures a heightened sensitivity to dynamic demand fluctuations across the booking time frame (Weatherford and Kimes, 2003; Fiori and Foroni, 2019). The additive pickup method forecasts the number of rooms that will be booked for a future stay day, t, as a sum of the number of on-hand reservations until the current reading day, d, and the estimated pickup of rooms during the forecasting horizon, h=t-d. Thus, the demand forecast for a horizon of h periods ($h\geq 1$) is given by:

$$\widehat{y}_{d(h)} = b_{t,h} + \sum_{l=0}^{h-1} \overline{a}_{d(l)}, \tag{2}$$

where $b_{t,j}$ represents the number of rooms booked for the t-th check-in day (t=d+j) at least j days in advance (j=0, 1, …, J) and $\overline{a}_{d(j)}$ is a prediction made on the reading day d of the net incremental bookings between lead times j+1 and j. Assuming a moving average of k-neighbor periods, where k is an odd number, this prediction is computed as follows:

$$\overline{a}_{d(j)} = \frac{1}{k}\sum_{i=d-(k-1)+j}^{d+j} a_{i,j}, \tag{3}$$

where $a_{t,j} = b_{t,j} - b_{t,j+1}$ represents the daily net incremental rooms booked for the t-th check-in.

This study also adopts the additive pickup method as a benchmark for accuracy assessment of the proposed approach. In addition, we compare forecasting accuracy of the PCA approach with the multiplicative pickup (Fiori and Foroni, 2020) and the cluster-based approach recently proposed by Viverit et al. (2023).

3.5. Accuracy assessment

The evaluation of alternative forecasting methods was assessed through six widely employed accuracy metrics commonly used in previous literature to gauge the performance of forecasting models (Kouprouchina et al., 2014). These six metrics comprise Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (sMAPE), and Median Absolute Percentage Error (MdAPE). We incorporate scale-independent metrics (MAPE, sMAPE, and MdAPE) alongside scale-dependent measures (MSE and MAE) due to their percentage-based reporting, which facilitates straightforward interpretation. Moreover, inclusion of sMAPE is driven by its symmetrical nature, offering a fixed range that effectively mitigates the challenge of substantial errors. Additionally, MdAPE was used to specifically address outliers within forecasting errors. A lower value across all those metrics, computed as follows in Eqs. (4)-(9), corresponds to enhanced accuracy of the forecasting method:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - f_t)^2, \tag{4}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - f_t| \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - f_t)^2}, \tag{6}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{100|y_t - f_t|}{|y_t|} \tag{7}$$

$$sMAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{200|y_t - f_t|}{|y_t + f_t|} \tag{8}$$

$$MdAPE = median\left\{\left|\frac{100|y_1 - f_1|}{|y_1|}\right|, \left|\frac{100|y_2 - f_2|}{|y_2|}\right|, \ldots, \left|\frac{100|y_n - f_n|}{|y_n|}\right|\right\} \tag{9}$$

where n represents the number of observations (length of the forecasting period), and y_t and f_t denote the observed and the forecasted number of rooms occupied in day t, respectively.

In addition, we applied the Wilcoxon signed-rank test to evaluate if competing forecasting methods generate statistically significant differences in terms of repeated measurements of accuracy. This nonparametric test has been applied in the forecasting literature (e.g. Pereira, 2016; Zhang et al., 2022) because it does not require a Normal distribution of the accuracy measure but only symmetry of the loss differential (ie, difference between accuracy metrics of two competing forecasting methods, A and B, respectively). The null hypothesis of this test sets that the median of the loss differential is equal to zero. A forecasting method A is significantly more accurate than its competitor B if the null hypothesis is rejected and it generates the biggest number of more accurate forecasts in the repeated measurements. The Wilcoxon signed-rank test was applied to all accuracy measures, but we report results only relative to the RMSE because it is the most commonly used metric of accuracy and conclusions are coherent across all those metrics.

4. Empirical study

This section presents the results of the forecasting accuracy for each hotel comparing the performance of the PCA approach against the benchmarks used in this study. We also assess the performance of the PCA method with additional information given by ADR, generating the PCA(ADR) method. The results are presented by hotel and for different groups of years: pre-COVID years (2018–2019), all the available data relative to five years (2018–2022), and only 2022 making use of the previous years' data for the forecasting.

Figs. 1–5 display, in the first row, the RMSE curves of forecasts of the final occupancy of the hotel generated by different methods, with respect to available datasets, in function of the number of days in advance. Every colour represents the RMSE of a different forecasting method as in the legend. Each column represents the RMSE curves on the time series of the selected groups of years. The second row of those Figures shows again the RMSE curves but as a function of the final

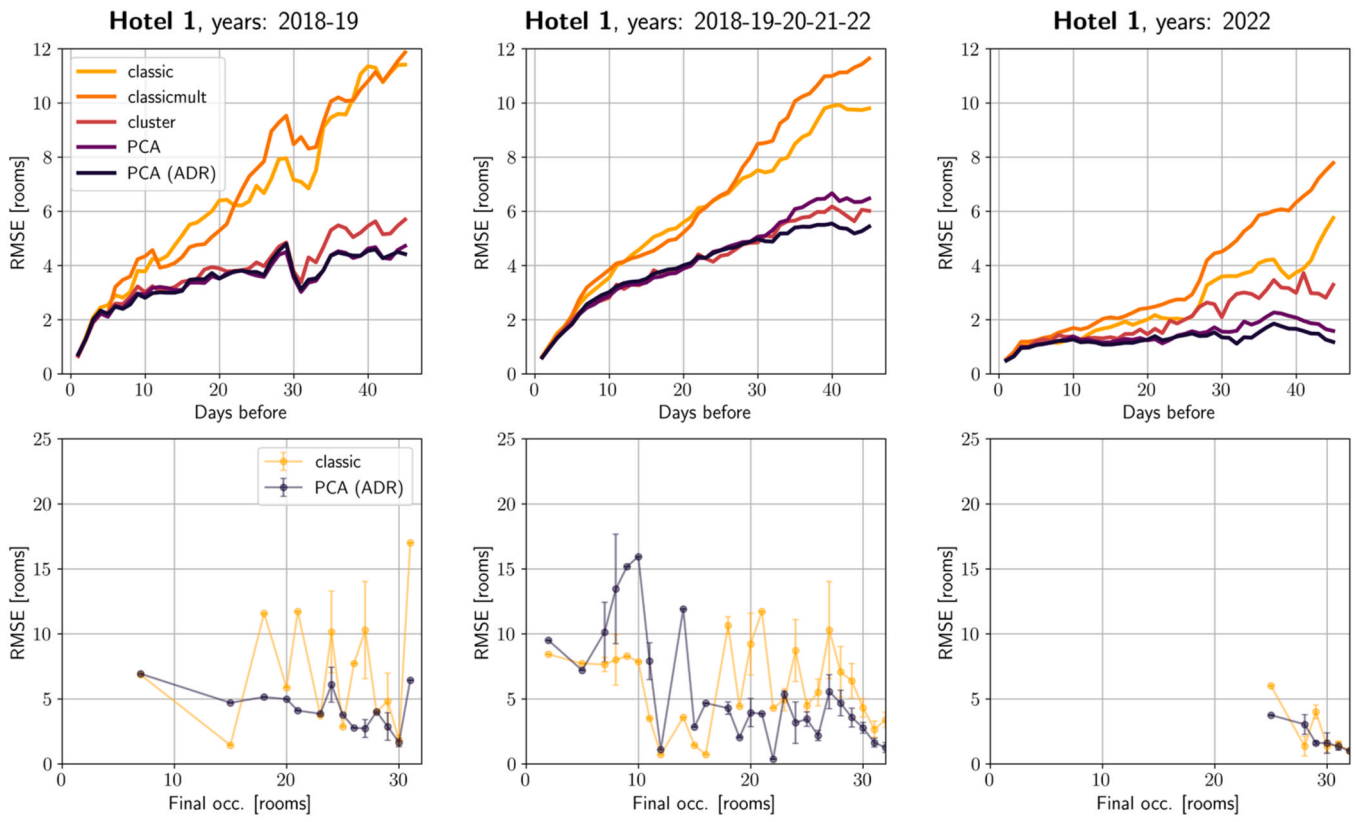


Fig. 1. RMSE of the forecasts per method and per years for Hotel 1.

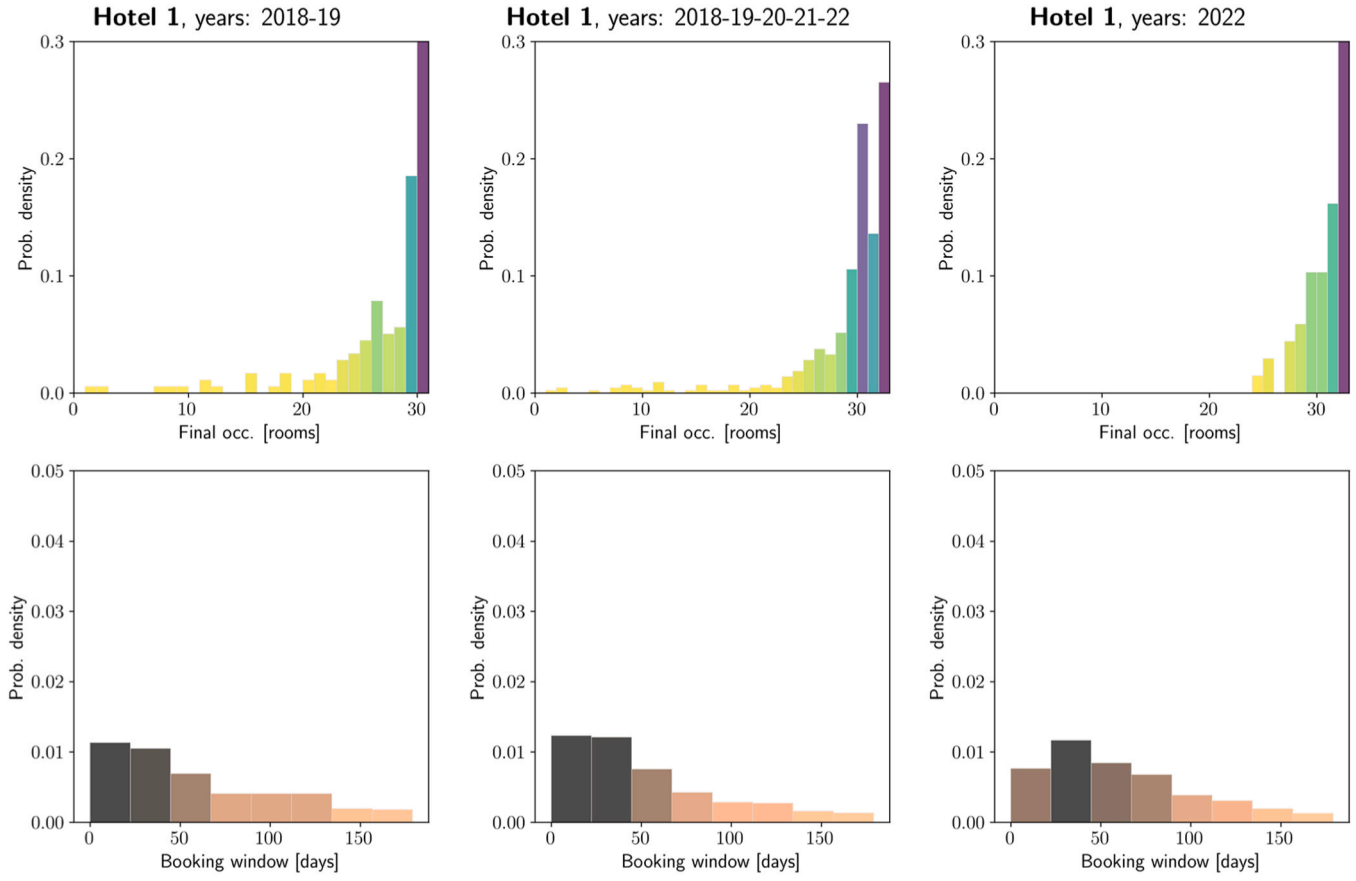


Fig. 2. Histograms of occupancy and booking window per years for Hotel 1.

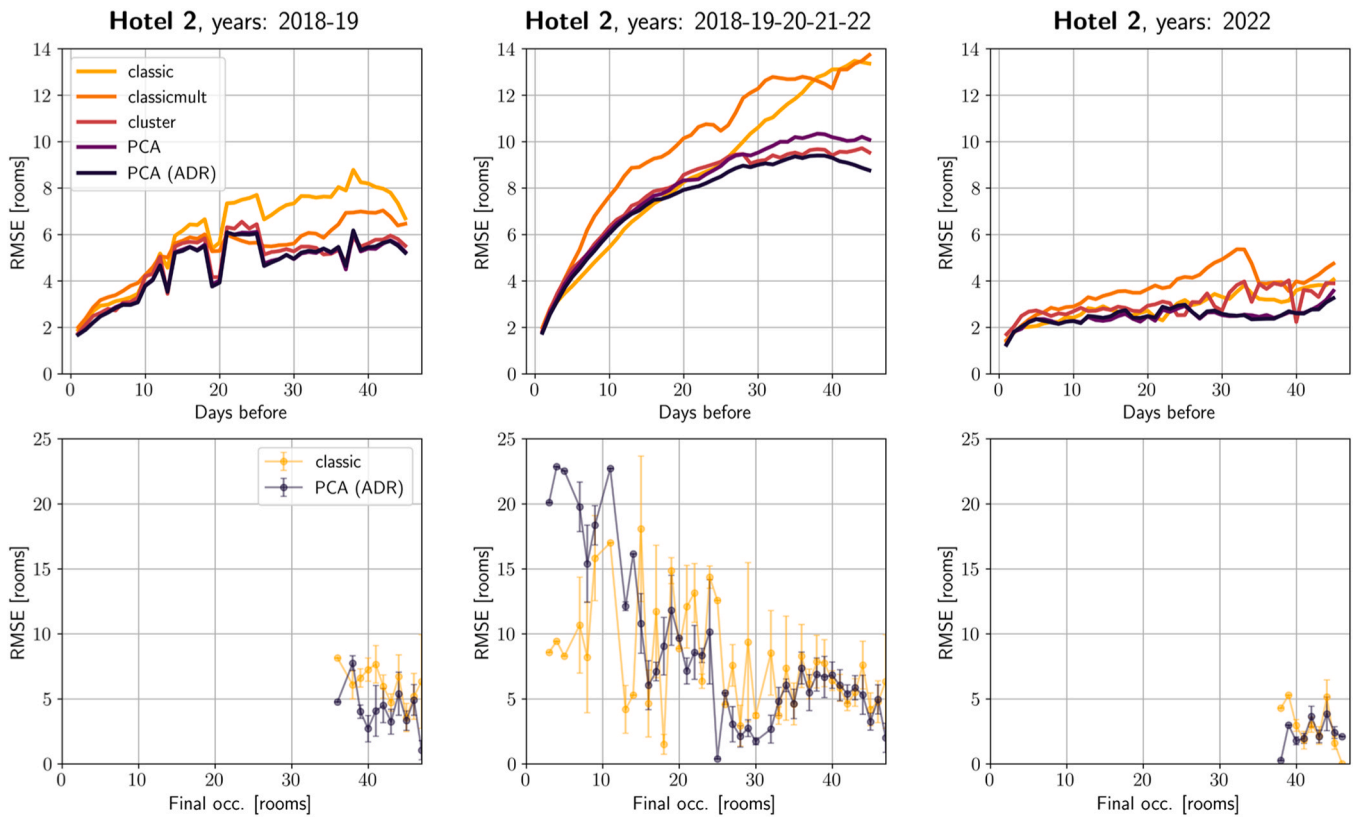


Fig. 3. RMSE of the forecasts per method and per years for Hotel 2.

occupancy of the hotel at a fixed number of days in advance (i.e., 25 days before the check-in date). We averaged over all the forecasted dates matching a given final occupancy and we reported the sample standard deviation with the error bars. Figs. 2, 4 and 6 show information about the distribution of the final occupancies (first row) and the booking windows (second row) for the respective hotel and group of years (2018–2019; 2018–2022; 2022). The histograms are normalized so that the integral is one, therefore in the form of probability density function of final occupancy and booking windows, respectively.

Tables 2 to 4 also summarise the results of the accuracy measures, where one can find a comparison of the performance of the pickup (additive and multiplicative), cluster-based and PCA (baseline PCA and PCA with ADR as an additional orthogonal component in the low dimensional space) methods, according to the six measures presented above. These accuracy measures are reported for forecasting horizons of 7, 14, 21 and 28 days before arrival (DBA).

4.1. Global results

Results presented in the first row of Figs. 1, 3 and 5 reveal that the RMSE of the selected forecasting models is affected by the groups of years and by the number of days in advance. Specifically, the RMSE of forecasts generated by all forecasting models decrease as the number of DBA decreases. This result means that occupancy forecasts are more accurate as the booking window gets shorter. However, it is empirical evidence that the rhythm of accuracy improvement may differ year after year, since it is clear that RMSE curves of the year 2022 are flatter than the curves of remaining groups of years. The flat shape of RMSE curves of 2022, whatever the forecasting method used, is explained with a highest occupancy rate of hotels assured several days in advance (Figs. 2, 4, 6). Finally, the RMSE curves presented in Figs. 1, 3 and 5 show that both PCA methods tend to perform better than both pickup methods and the cluster-based method, for all hotels. The forecasting accuracy gains obtained by the PCA methods are bigger when compared with the pickup methods than when compared with the cluster-based method.

Results presented in Tables 2 to 4 reveal that, in general, all accuracy measures coherently indicate the same type of methods as performing the worst. In this empirical study, the pickup methods perform worse than the PCA methods, for all hotels and forecasting horizons. However, relative accuracy gains of the PCA methods when compared with the pickup methods decrease with the forecasting horizons. Tables 2 to 4 also show that there are slight improvements in accuracy when ADR information is added to the basic PCA method.

The results are compatible with each other despite the different locations of the hotels. Before the COVID-19 pandemic, the final occupancy of the hotels was quite high around the maximum number of available rooms. Also, the booking windows were generally long, being the best scenario for long-term forecasting. For data belonging to those years, the PCA method performs distinctly better than the classical pickup. During the pandemic of COVID-19, especially for Hotels 2 and 3, the pattern of the distribution of final occupancy and the booking windows changed. As a consequence, the forecasting ability of the methods (even the PCA) were affected, resulting in a less pronounced difference in the relative performances. The post-COVID (2022) situation is assimilable to the pre-COVID one. A quite interesting property of the PCA method (as clear from the RMSE vs final occupancy plots) is that it performs very well for high final occupancy rates.

To verify whether the differences on the values of metrics of accuracy between each pair of forecasting methods are statistically significant per forecasting horizon, the Wilcoxon signed-rank test was employed. Due to space constraints, we report in Fig. 7 the percentage of cases (stay dates) for which the target forecasting method is significantly more accurate (p-value<0.05) than a benchmark per accuracy measure. Fig. 7 shows the performance of new forecasting methods over the additive pickup (first row), the multiplicative pickup (second row) and the cluster-based (third row) methods. Fig. 7 reveals several important findings. First, it is clear that all ML forecasting methods (cluster-based

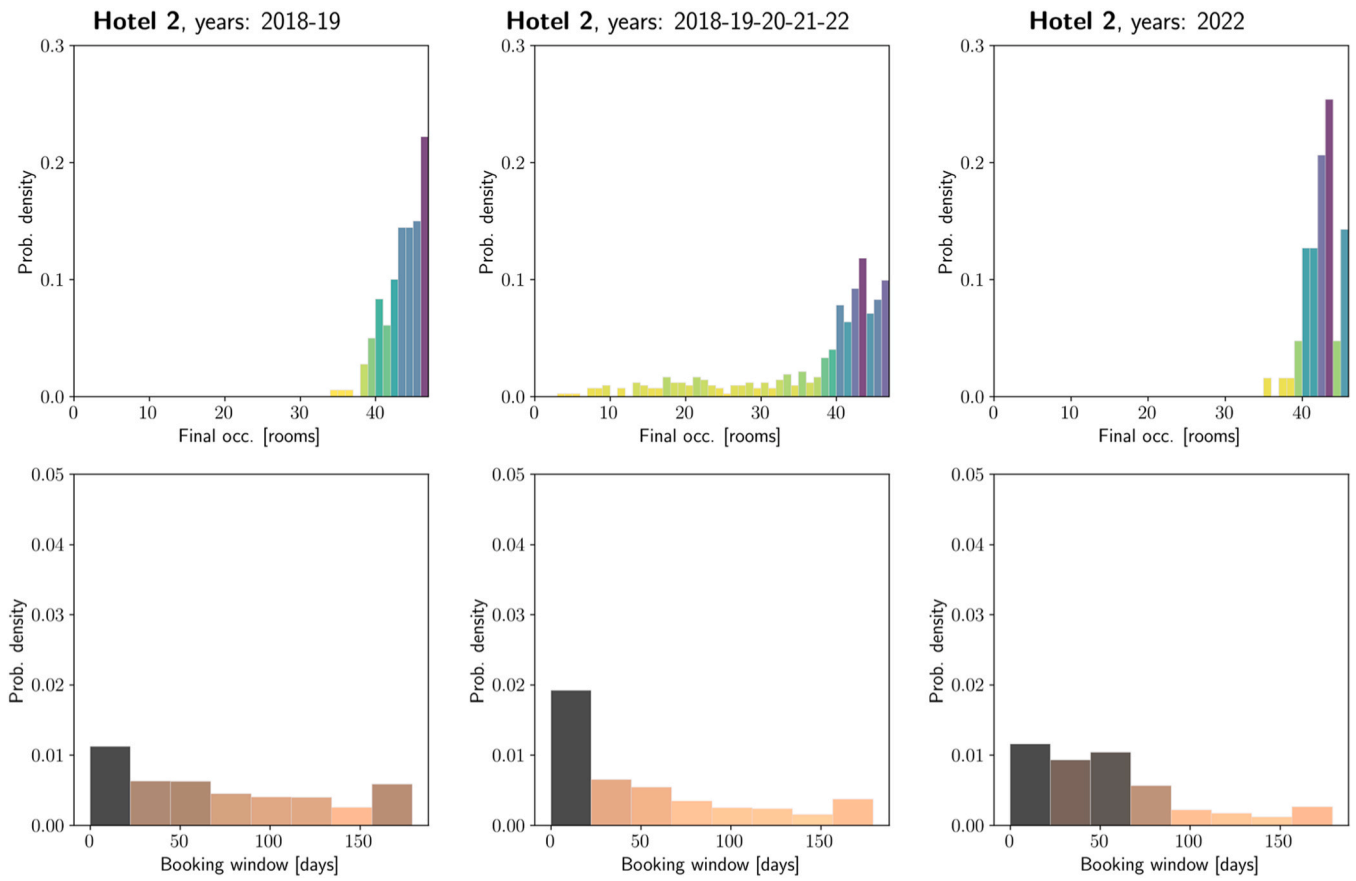


Fig. 4. Histograms of occupancy and booking window per years for Hotel 2. **Note:** Some stay dates (28–29–30–31/08/22 and 01/09/22) were removed because of some last-minute cancellations that strongly affect the forecasting performances.

and PCA methods) are significantly more accurate than both the additive and multiplicative pickup methods in almost all of the forecasts. Second, the PCA methods significantly outperform the cluster-based method in the majority of the forecasts, although the baseline PCA method when applied to data from hotel 3 provides a mild accuracy performance when compared with the cluster-based method. Finally, it is clear that the PCA(ADR) is the most accurate method for all hotels. It is an evidence that inclusion of auxiliary information in the forecasting task using the PCA approach is a key factor to significantly improve forecasting accuracy of daily hotel demand, which is a clear advantage of the PCA over the cluster approach in support of forecasting tasks.

4.2. Hotel 1

In Fig. 1 we show the results for Hotel 1 located in Isola D'Elba (Italy). The RMSE error in the forecasting of the final occupancy for the years before the pandemic (2018–19) is significantly lower when the clustering method or the PCA is applied for pickup. Also, the PCA method is more accurate than the cluster-based method, especially in the long-term forecasting. Moreover, the error with the PCA according to the final occupancy for a forecast made 25 days in advance (bottom-left panel) is in general lower and with a less noisy trend when compared to the classical pickup method. For Hotel 1, the results are unvaried even when considering all the data at disposal for the years between 2018 and 2022, therefore including also the pandemic. In the bottom central panel of Fig. 1 the RMSE is shown to be lower using our PCA method especially for the high occupied dates which are the vast majority overall (see top-centre panel of Fig. 2). The results for 2022 only can be considered as a validation of the method since the forecasting is restricted only to dates of the latter year but the dataset contains all the historical booking curves. The results clearly show how the method can become really accurate with a sufficiently large dataset: the forecasting error of the PCA method is essentially flat and for the PCA with the ADR values does not overcome an RMSE of 2.

For Hotel 1 the results of the forecasting are very promising and similar to each other for all the three time frames. In Fig. 2 we show in the top row the histograms of the final occupancy which are all peaked around the maximum capacity of the hotel (32 rooms). Likewise, the booking window distribution has very similar patterns: the first reservations are made with large advance and their number increases when approaching the arrival date. Quite interestingly, in 2022 the booking windows' distribution is shifted to a larger advance. This fact is reflected in the very accurate forecasting performance of the PCA method that remains almost constant because the booking curves' trend is already expressed.

4.3. Hotel 2

Fig. 3 shows the results for Hotel 2, located in Nice (France). As for Hotel 1, the RMSE error in the forecasting of the final occupancy for the years before the pandemic (2018–19) is lower when the cluster-based or the PCA methods are applied for pickup. In those years, the cluster-based methods' error has a similar pattern to the one of the PCA both with and without ADR. Hotel 2 was strongly affected by the pandemic: the advantage of using the aggregating methods for the pickup is partially lost when including data from 2020 and 2021, especially in the short-term forecast. Regardless, the error of the PCA method remains lower for high final occupancy dates, which are the most relevant to be considered. This is clear from the bottom plots of Fig. 3. The long-term forecasting remains better with the PCA up to an RMSE of 4 for a 45-

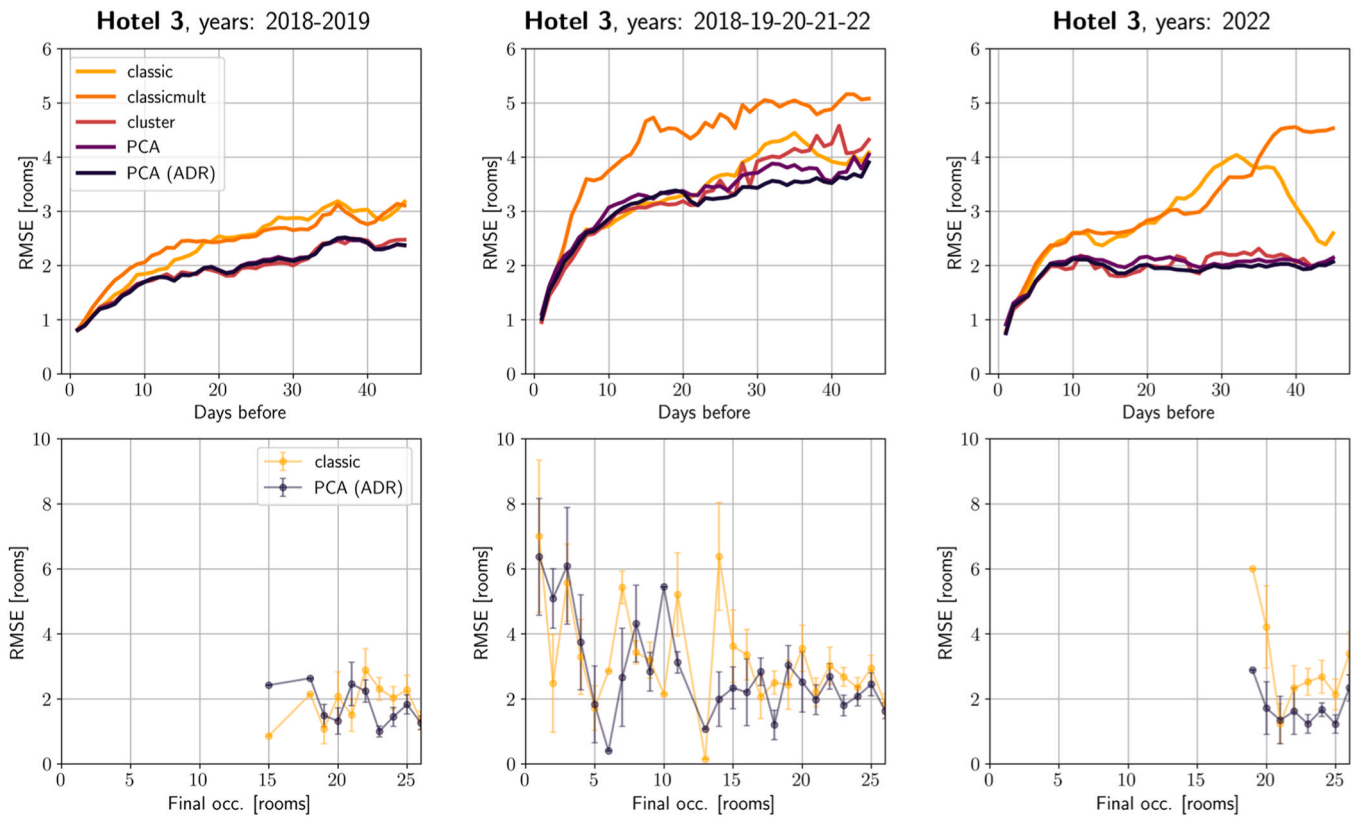


Fig. 5. RMSE of the forecasts per method and per years for Hotel 3.

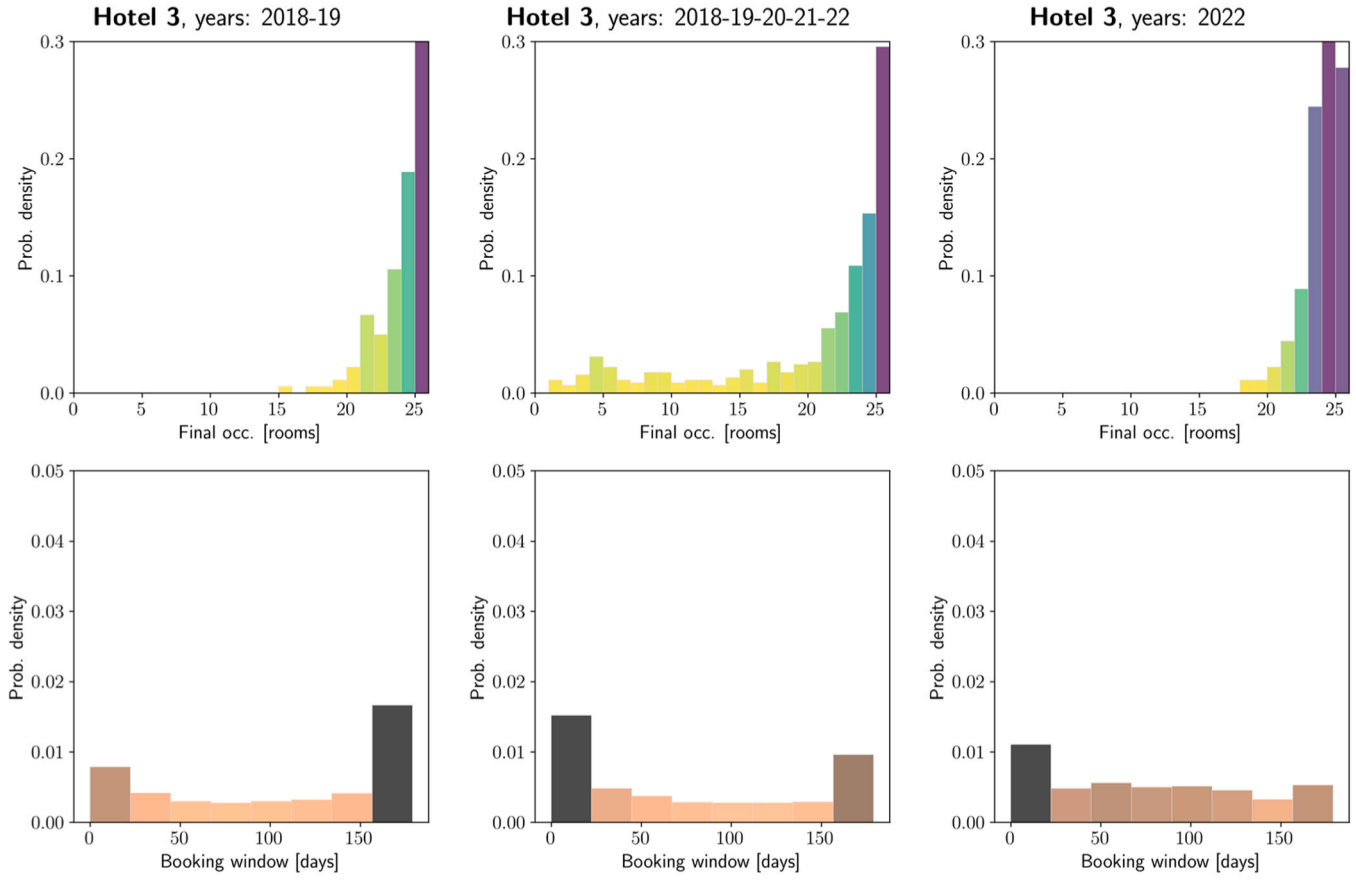


Fig. 6. Histograms of occupancy and booking window per years for Hotel 3.

Table 2
Accuracy measures per forecasting method and per years for Hotel 1.

Methods		Additive Pickup				Multiplicative Pickup				Cluster-based				PCA				PCA(ADR)			
DBA		7	14	21	28	7	14	21	28	7	14	21	28	7	14	21	28	7	14	21	28
2018/19	MSE	7.9	22.5	41.3	62.7	12.0	16.7	30.6	85.9	6.6	10.3	14.3	22.0	5.9	9.6	13.8	19.3	5.7	9.0	13.3	20.6
(Before COVID-19)	MAE	2.0	3.3	4.3	5.8	2.4	2.8	4.0	7.0	1.9	2.4	2.7	3.3	1.8	2.3	2.9	3.3	1.7	2.3	2.9	3.4
	MAPE	14.8	15.2	18.3	23.8	11.5	12.4	18.3	36.7	13.1	10.9	12.1	15.1	10.4	10.0	12.9	15.8	10.1	10.1	12.8	16.6
	sMAPE	9.5	15.6	21.6	30	10.4	12.4	17.8	31.2	9.1	10.5	11.9	15.1	8.5	10.1	12.3	15.1	8.3	9.9	12.4	15.5
	MdAPE	5.0	5.8	10.1	13.9	6.4	6.7	11.9	19.0	5.3	6.6	7.2	5.9	4.6	6.7	8.3	7	4.4	6.7	8.1	7.1
	RMSE	2.8	4.7	6.4	7.9	3.5	4.1	5.5	9.3	2.6	3.2	3.8	4.7	2.4	3.1	3.7	4.4	2.4	3.0	3.6	4.5
2018–2022	MSE	8.3	20.9	33.3	52.1	10.1	18.8	30.0	58.3	6.2	10.7	15.8	22.8	5.9	10.9	15.8	23.6	6.6	11.7	16.7	23.0
(Entire Period)	MAE	1.8	3.0	3.8	5.0	1.9	2.8	3.6	5.0	1.6	2.1	2.6	3.2	1.5	2.1	2.5	3.0	1.5	2.1	2.5	3.0
	MAPE	13.0	14.9	18	23.2	9.6	12.8	16.2	24.1	12.2	12.0	14.9	19.3	9.4	10	12.2	16.4	9.9	11	13.8	18.0
	sMAPE	8.6	13.4	17.6	23.8	9.1	13.1	17.0	23.4	7.8	9.7	11.7	14.5	7.1	9.4	11.1	13.3	7.4	9.5	11.3	13.5
	MdAPE	3.2	5.4	6.7	8.9	3.5	4.9	7.6	9.8	3.2	4.5	4.5	6.6	3.2	3.6	5.0	5.5	3.2	3.5	4.5	5.5
	RMSE	2.9	4.6	5.8	7.2	3.2	4.3	5.5	7.6	2.5	3.3	4.0	4.8	2.4	3.3	4.0	4.9	2.6	3.4	4.1	4.8
2022	MSE	1.3	2.8	4.7	10.7	1.9	4.1	6.1	17.5	1.5	1.7	2.8	6.9	1.3	1.3	1.6	2.4	1.3	1.2	1.9	2.0
(Post COVID-19)	MAE	0.9	1.2	1.5	2.2	1.0	1.6	1.9	2.9	0.8	1.0	1.2	1.9	0.9	1.0	0.9	1.3	0.9	0.8	1.0	1.0
	MAPE	2.8	4.1	5.1	7.5	3.2	5.2	6.3	9.8	2.7	3.2	4.1	6.1	2.9	3.1	3.0	4.1	3.0	2.8	3.3	3.4
	sMAPE	2.8	4.2	5.3	8.1	3.2	5.3	6.5	11.0	2.7	3.2	4.1	6.4	2.8	3.1	2.9	4.1	3.0	2.7	3.2	3.3
	MdAPE	2.2	3.1	3.1	3.2	2.7	4.2	5.5	5.1	1.7	2.0	3.1	3.4	2.2	3.1	2.7	3.1	2.9	3.1	3.1	3.1
	RMSE	1.2	1.7	2.2	3.3	1.4	2.0	2.5	4.2	1.2	1.3	1.7	2.6	1.1	1.2	1.3	1.5	1.2	1.1	1.4	1.4

Note: DBA-Days Before Arrival

Table 3
Accuracy measures per forecasting method and per years for Hotel 2.

Methods		Additive Pickup				Multiplicative Pickup				Cluster-based				PCA				PCA(ADR)			
DBA		7	14	21	28	7	14	21	28	7	14	21	28	7	14	21	28	7	14	21	28
2018/19	MSE	10.2	35.3	53.8	50.1	12.7	31.5	35.9	30.6	9.3	30	39.8	27.9		27.4	36.9	24.4		27.1	37.1	24.3
(Before COVID-19)														9.0				8.8			
	MAE	2.5	4.4	5.3	5.4	2.7	4.4	4.6	4.3	2.3	3.6	4.3	3.6	2.2	3.3	3.9	3.5	2.2	3.3	3.9	3.5
	MAPE	5.7	10.2	12.1	12.5	6.4	10.3	10.6	9.9	5.3	8.2	9.8	8.5	5.1	7.7	9.0	8.1	5.1	7.6	8.9	8.1
	sMAPE	5.8	10.8	13.3	13.2	6.5	10.9	11.2	10.3	5.4	8.9	10.8	9.1	5.2	8.3	10.0	8.6	5.2	8.2	9.9	8.6
	MdAPE	4.4	7.1	8.9	9.3	5.3	8.7	8.0	8.3	3.8	4.4	6.6	5.1	3.9	4.3	5.2	5.9	3.7	3.9	5.2	6.1
	RMSE	3.2	5.9	7.3	7.1	3.6	5.6	6.0	5.5	3.1	5.5	6.3	5.3	3.0	5.2	6.1	4.9	3.0	5.2	6.1	4.9
2018–2022	MSE	19.9	46.1	71.1	99.9	38.2	79.1	105.6	141.0	26.4	54.2	75.8	89.6		51.8	69.6	89.3		49.4	64.0	80.3
(Entire Period)														26.5				24.4			
	MAE	3.4	5.0	5.9	7.4	4.4	6.6	7.4	8.4	3.8	5.4	6.4	7.1	3.7	5.1	5.8	6.9	3.6	5.1	5.8	6.8
	MAPE	13.0	21.3	25.0	33.2	16.6	24.7	28.4	40.9	16.1	24.3	31.7	38.5	16.3	23.7	30.1	38.0	15.5	23.1	28.9	36.2
	sMAPE	11.9	17.3	19.9	24.6	15.3	23.2	25.6	31.1	13.4	18.9	22.4	25.0	13.6	17.9	20.5	24.0	13.0	18.0	20.7	23.9
	MdAPE	7.6	10.2	10.3	13.6	8.7	11.9	13.2	14.6	7.5	9.3	11.5	14.4	7.1	9.1	9.3	11.9	6.8	8.4	9.3	12.2
	RMSE	4.5	6.8	8.4	10.0	6.2	8.9	10.3	11.9	5.1	7.4	8.7	9.5	5.1	7.2	8.3	9.4	4.9	7.0	8.0	9.0
2022	MSE	5.0	8.4	5.9	10.4	8.1	11.2	14.5	20.8	6.2	7.9	8.9	12		5.2	5.2	7.0		5.8	5.9	7.2
(Post COVID-19)														5.2				4.9			
	MAE	1.7	2.3	1.9	2.6	2.2	2.7	3.1	3.3	2	2.3	2.5	2.9	1.7	1.9	1.8	2.2	1.7	2.0	2.0	2.2
	MAPE	4.0	5.5	4.5	6.3	5.3	6.5	7.3	7.9	4.7	5.6	5.9	7	4.1	4.5	4.4	5.2	4.2	4.8	4.7	5.2
	sMAPE	4.1	5.6	4.5	6.2	5.4	6.6	7.5	8.2	4.6	5.4	5.7	6.9	4.1	4.5	4.4	5.3	4.2	4.8	4.7	5.2
	MdAPE	3.2	5.1	3.7	4.8	3.7	6.1	6.7	6.0	3.5	4.5	5.2	6.4	3.5	4.3	3.6	5.3	3.5	4.5	3.9	4.9
	RMSE	2.2	2.9	2.4	3.2	2.8	3.4	3.8	4.6	2.5	2.8	3	3.5	2.3	2.3	2.3	2.6	2.2	2.4	2.4	2.7

Note: DBA-Days Before Arrival

Table 4
Accuracy measures per forecasting method and per years for Hotel 3.

Methods		Additive Pickup				Multiplicative Pickup				Cluster-based				PCA				PCA(ADR)			
DBA		7	14	21	28	7	14	21	28	7	14	21	28	7	14	21	28	7	14	21	28
2018/19 (Before COVID-19)	MSE	2.3	4.4	6.3	8.3	3.3	5.4	6.1	7.2	2.1	3.1	3.3	4.2	2.1	2.9	3.5	4.6		2.9	3.4	4.5
	MAE	1.2	1.6	1.9	2.2	1.4	1.7	1.8	1.9	1.1	1.3	1.5	1.6	1.1	1.3	1.5	1.7	2.1	1.3	1.4	1.7
	MAPE	4.9	6.7	7.8	9.1	5.9	7.0	7.7	7.9	4.6	5.7	6.2	6.9	4.6	5.4	6.2	7.2	1.1	5.4	6.1	7.1
	sMAPE	5.0	6.8	8.2	9.7	6.0	7.3	8.2	8.5	4.6	5.7	6.3	7.1	4.6	5.5	6.4	7.3	4.6	5.4	6.2	7.2
	MdAPE	4.0	5.5	6.8	7.4	4.6	4.4	6.0	5.3	4.0	4.5	5.6	6.2	3.9	4.3	4.7	6.3	3.9	4.1	4.3	6.3
	RMSE	1.5	2.1	2.5	2.9	1.8	2.3	2.5	2.7	1.5	1.8	1.8	2.0	1.4	1.7	1.9	2.2	1.4	1.7	1.9	2.1
2018–2022 (Entire Period)	MSE	7.1	9.5	10.6	15.2	12.9	18.3	18.8	24.6	6.5	9.4	10.0	16.1	6.9	10.9	11.1	12.8	6.8	10.0	10.5	13.5
	MAE	2.0	2.3	2.5	3.0	2.5	2.9	3.1	3.4	1.9	2.3	2.4	3.0	1.9	2.4	2.5	2.7	1.9	2.3	2.4	2.6
	MAPE	15.7	20.7	24.3	33.8	22.6	29.0	29.2	34.8	19.3	27.7	32.4	40.3	17.2	25.1	32.4	38.8	17.5	25.9	30.6	36.7
	sMAPE	14.8	16.3	18.4	22.3	19.9	24.8	29.1	31.3	14.3	17.7	18.7	22.9	14.4	17.8	19.3	21.5	14.1	17.6	18.1	19.8
	MdAPE	7.8	8.1	10.2	12.4	8.3	9.2	10.5	12.3	6.6	7.9	8.3	9.4	7.2	8.1	9.1	9.6	6.5	8.0	8.6	8.7
	RMSE	2.7	3.1	3.3	3.9	3.6	4.3	4.3	5.0	2.6	3.1	3.2	4.0	2.6	3.3	3.3	3.6	2.6	3.2	3.2	3.7
2022 (Post COVID-19)	MSE	5.4	5.6	7.8	13.3	5.6	6.7	8.2	9.7	3.9	4.0	3.9	4.6	4.1	4.4	4.4	3.9	4.1	4.1	3.8	3.8
	MAE	1.8	1.8	2.3	3.1	1.7	2.0	2.2	2.4	1.5	1.5	1.6	1.7	1.6	1.6	1.7	1.6	1.6	1.6	1.6	1.6
	MAPE	7.6	7.8	9.7	12.9	7.4	8.5	9.2	10.4	6.5	6.6	6.7	7.1	6.9	6.9	7.2	6.7	6.7	6.7	6.8	6.8
	sMAPE	7.9	8.1	10.1	13.6	7.6	8.8	9.6	10.8	6.7	6.6	6.7	7.2	7.0	7.0	7.3	6.7	6.8	6.7	6.9	6.8
	MdAPE	6.8	6.1	8.3	11.8	6.3	7.4	7.4	10.0	5.5	5.2	5.2	5.8	6.4	5.3	7.3	5.5	6.0	5.6	6.5	6.0
	RMSE	2.3	2.4	2.8	3.6	2.4	2.6	2.9	3.1	2.0	2.0	2.0	2.2	2.0	2.1	2.1	2.0	2.0	2.0	1.9	1.9

Note: DBA-Days Before Arrival

days-before-arrival forecast. After the pandemic, the forecasting accuracy for the 2022 dates using all historical data is again increased with PCA and remains very good even for long-term forecasts, with an RMSE always lower than 4.

Fig. 4, in analogy to Hotel 1, shows in the top row the histograms of the final occupancy. The effects of COVID-19 are really evident on this hotel where the final occupancy is shifted from the maximum possible in 2018–19 to a lower value considering also 2020–21, meaning that during the pandemic the maximum occupancy was rarely reached. The booking window for Hotel 2 is generally quite short even though, as for Hotel 1, there is a shift to longer booking windows in 2022.

4.4. Hotel 3

The results for Hotel 3 (Paris, France) are displayed in Fig. 5. For this hotel the behaviour of the error curves is very similar to the Hotel 2 case with some peculiar minor differences. The forecasting accuracy does not seem to benefit from the addition of ADR in the low dimensional space of the PCA and, in general, the aggregating methods' performance are very similar. Also for Hotel 3, the forecasting ability was compromised during the pandemic and recovered afterwards in 2022. The error of the PCA method always remains lower for high final occupancy dates (bottom panels of Fig. 5).

The booking window distribution for Hotel 3 is quite unique: in 2018–19 it peaked for a very long advance. Indeed, the errors in the forecasting of the dates belonging to those years is markedly better when adopting the PCA and cluster-based methods. During the pandemic the situation changes: for many dates the hotel is not completely occupied, and the booking window distribution is peaked to very short advance. In 2022, it is almost flat and clearly this also affects the performance in the forecasting with the aggregation methods (Fig. 6).

5. Conclusion

In this paper we propose a two-step approach to forecast daily hotel demand based on advanced booking data. This approach applies a pickup forecasting model to similar stay days identified in a low-dimensional space generated by a PCA. Our empirical study revealed that the PCA methods perform better than the benchmarks for all hotels and all forecasting horizons. Furthermore, the results indicate that the PCA(ADR) method generates forecasts that are slightly more accurate than the baseline PCA method.

This paper contributes to the hospitality management field in both methodological and practical dimensions. From the methodological point of view, this paper proposes a forecasting approach that first applies a PCA to group booking curves, by reducing a high-dimensional space of booking data, including incomplete booking curves with different number of observations, to a low-dimensional space of points. This approach innovates by opening a window for adding auxiliary information (e.g. ADR), as an orthogonal component in that low-dimensional space, to enhance accuracy of high-frequency forecasts.

From a practical point of view, this paper tries to facilitate the interpretation of how ML algorithms work for practitioners. Interpretable ML algorithms are essential for building trust between users (e.g. hotel managers, revenue managers) and data scientists that develop ML models. By making the inner workings of the models more transparent, users can gain an understanding of why a particular result was reached, and how the model works. This interpretability, or transparency, can help to boost users' confidence in recommendations made by this type of AI model. By using an interpretable ML approach, this study aims to address black-box issues associated with data analysis algorithms such as ML by applying PCA to hotel occupancy forecasting. This approach enables businesses to better understand the outcomes of their analysis and look for any potential bias in the results. Furthermore, this novel approach has been adopted by firms across the world to gain valuable insights from customer data.

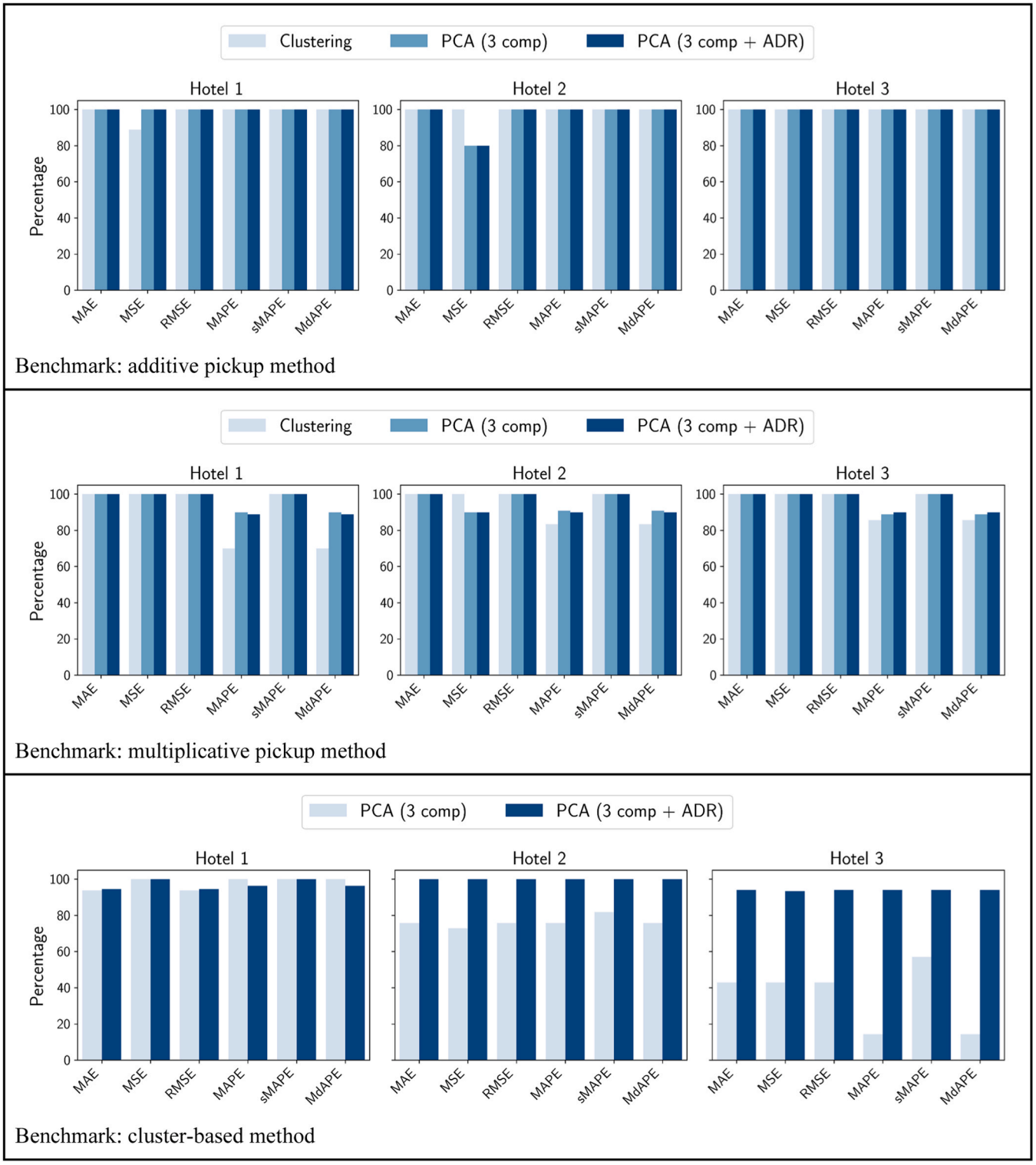


Fig. 7. Results of the Wilcoxon-signed rank tests on the differences in accuracy measures between pairs of forecasting methods.

The empirical findings from our study are constrained by the datasets obtained solely from three hotels. To strengthen the generalizability of our approach, further research should explore its applicability across various types of hotels (e.g., independent vs. chain hotels). Furthermore, an examination of the approach's effectiveness in times when hotels encounter varying demand uncertainties arising from exogenous shocks holds the potential to yield valuable insights. Exogenous shocks within the tourism industry encompass unforeseen and external occurrences or influences that considerably affect the sector's functions, demand dynamics, and overarching performance (e.g., economic crises, geopolitical conflicts, natural disasters, regulatory changes, and health epidemics). These shocks can precipitate abrupt alterations in travel trends, consumer conduct, and market circumstances. In addition, we enthusiastically encourage future research endeavours to assess the approach's efficacy through the integration of additional supplementary data, including insights into competitive dynamics, diverse booking

channels, and comprehensive customer profiles. This concerted effort has the potential to significantly elevate the accuracy and overall robustness of the forecasting model.

This study showed the enhancement of demand forecasting accuracy through the incorporation of ADR information as an orthogonal component within a low-dimensional space. For future scholars, avenues of exploration could encompass investigating how the frequency of special promotions or adjustments in prices among competitor hotels within the set, alongside other factors like economic dimensions, might provide additional enhancements to the precision of demand forecasting.

CRediT authorship contribution statement

Cindy Yoonjoung Heo: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization. **Luís Nobre Pereira:** Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation, Conceptualization. **Luciano Viverit:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Conceptualization. **Daniele Contessi:** Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

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

Acknowledgements

This paper is partially financed by the Portuguese Foundation for Science and Technology (FCT) through project UIDB/04020/2020 with DOI: 10.54499/UIDB/04020/2020.

Appendix A

Appendix Table 1
Literature review on hotel demand forecasting using artificial intelligence

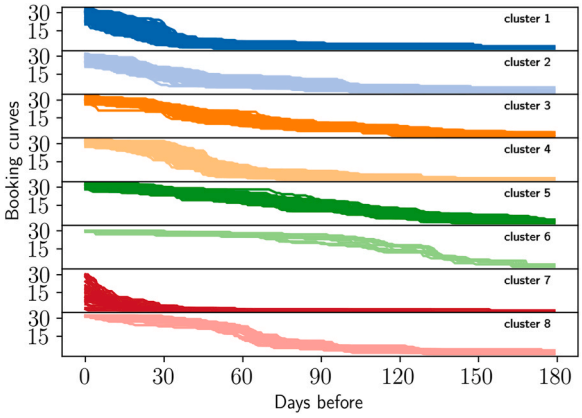
Author(s) (Year)	AI methods used	Key findings
Zhang and Niu (2024)	A novel long short-term memory interaction-based convolutional neural network (LICNN). Other AI-based models (XGBoost, ANN, LSTM, GRU, A-LSTM).	- The proposed LICNN model achieved the lowest RMSE, MAE, and MAPE, when compared with other AI-based models. - Incorporating online review features reduced the RMSE by at least 2.2% and at most 46.6%, and the MAE by at least 3.2%.
Huang et al. (2023)	A new integrated Deep Learning model based on spatial-temporal graph convolutional networks and gated recurrent units (STGCN-GRU). Other AI-based models (ANN, LSTM, LSTM-AN, DLM-ST).	- The proposed approach performed better than a general model that includes only spatiotemporal dependence, as well as than other AI-based models (ANN, LSTM, LSTM-AN, DLM-ST). - A Box-Jenkins's forecasting method (SARIMAX) performed worse than all AI-based models, except the ANN.
Viverit et al. (2023)	A new approach based on clustering booking curves by a ML algorithm, supported by the additive pickup method.	- Historical booking curves were clustered based on the patterns of booking curves by a ML algorithm. - The proposed approach generates more accurate forecasts than the classical pickup method. - Forecasts of hotel demand was more accurate when they are generated at cluster-level for all forecasting horizons.
Zhang and Wu (2023))	A novel dual attention-based long short-term memory convolutional neural network (DA-LSTM-CNN) model to optimize the model effectiveness. Other AI-based models (XGBoost, SVR, ANN, LSTM, and attention-LSTM).	- The proposed model is consistently more accurate than all AI-based benchmarks in terms of RMSE, MAE, and MAPE. - Results showed the value of integrating features derived from online reviews for enhancing hotel demand forecast performance.
Kaya et al. (2022)	A new approach using Attention-Long Short-Term Memory (Attention-LSTM) using two different features: a NN architecture and a clustering method. Other AI-based models (e.g. XGBoost, GRU, LSTM).	- The proposed approach outperforms the benchmark methods used (ML). - NN Embeddings and K-Means findings are used to improve prediction model performance.
Pereira and Cerqueira, 2022)	Several ML methods (e.g. SVR, XGBoost, RF), including a new dynamic ensemble method based on arbitrating (ADE).	- ML methods systematically outperform classical forecasting methods, for different forecasting horizons, in the framework of a high-frequency time series with a double seasonal pattern. - The use of ML models can reduce the RMSE up to 54% for a 1-day forecast horizon, and up to 45% for a 14-days forecast horizon, when compared with traditional exponential smoothing methods. - The ADE presented the overall best forecasting performance.
Ampountolas and Legg (2021)	A segmented ML approach of leveraging hierarchical clustering tied to machine learning techniques. The extreme gradient boosting (XGBoost) was used as benchmark.	- All ML algorithms improve the forecasts over the ARIMA and the Naïve forecasts for all time horizons. - Inclusion of sentiment analysis provides a moderate improvement of demand forecasts. - Using big data samples and incorporating many parameters created challenges in running ML forecasting models, which can be minimized by using a dimensionality reduction method.
Huang and Zheng (2021)	A novel Deep Learning Model with Spatial and Temporal correlations (DLM-ST). Other AI-based models (VAR, LSTM).	- This study introduced the agglomeration effect among hotels and integrating the attention mechanism and Bayesian optimization algorithm. - The proposed model is significantly better than the benchmarks (VAR, LSTM and the traditional ARIMAX) for forecasting daily hotel demand.
Zhu et al. (2021)	A set of ML methods (BPNN, SVM, LSTM).	- The ML techniques are more suitable for multi-horizon hotel demand forecasting than classical time series models. - The LSTM has particular advantages in long-horizon forecasting and handling data with complex structure.
Phumchusri and Ungtrakul (2020).	Support Vector Regression (SVR) and Artificial Neural Network (ANN).	- ML techniques studied outperform the advanced time series methods designed for complex seasonality data like BATS and TBATS.

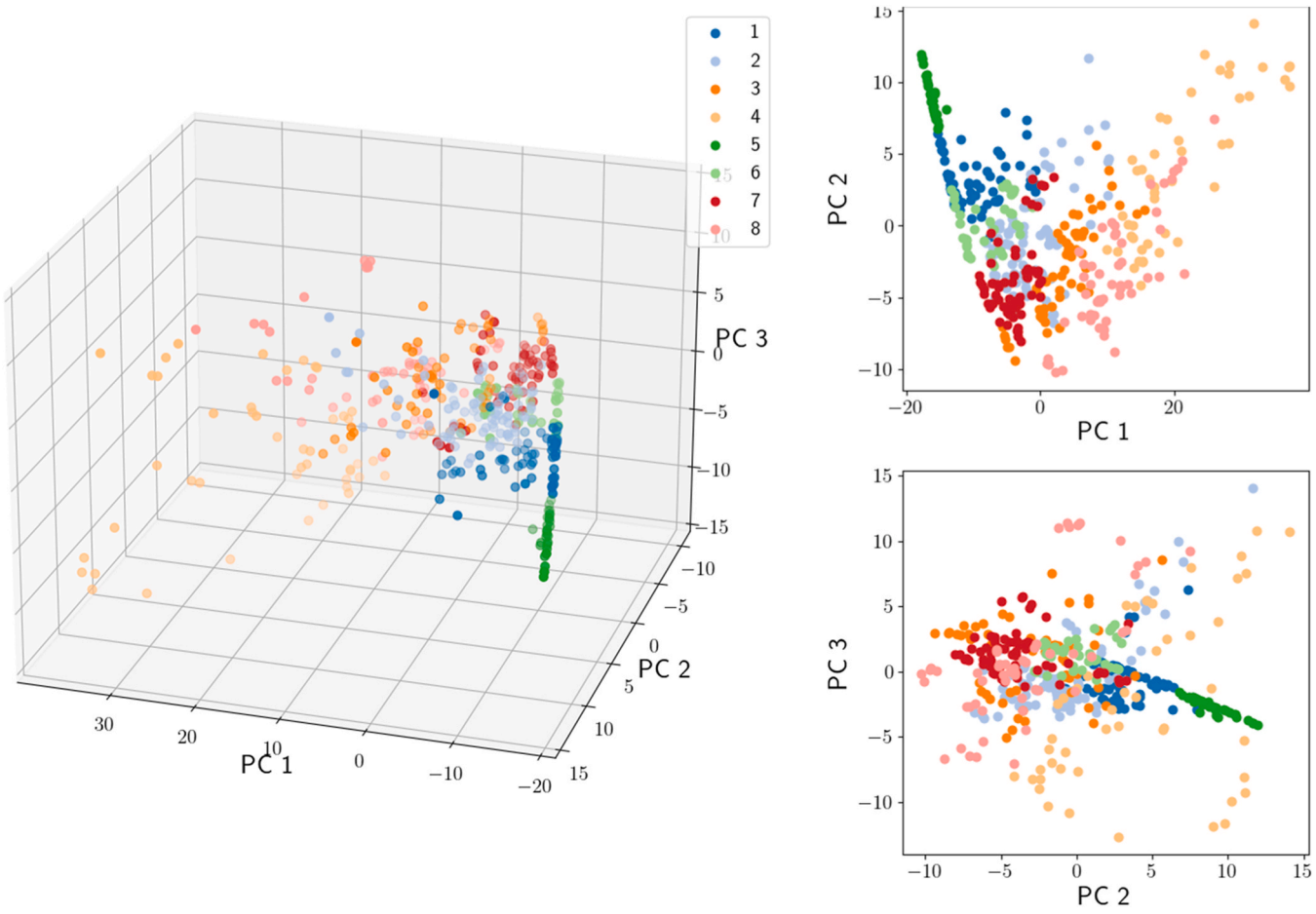
(*continued on next page*)

Appendix Table 1 (*continued*)

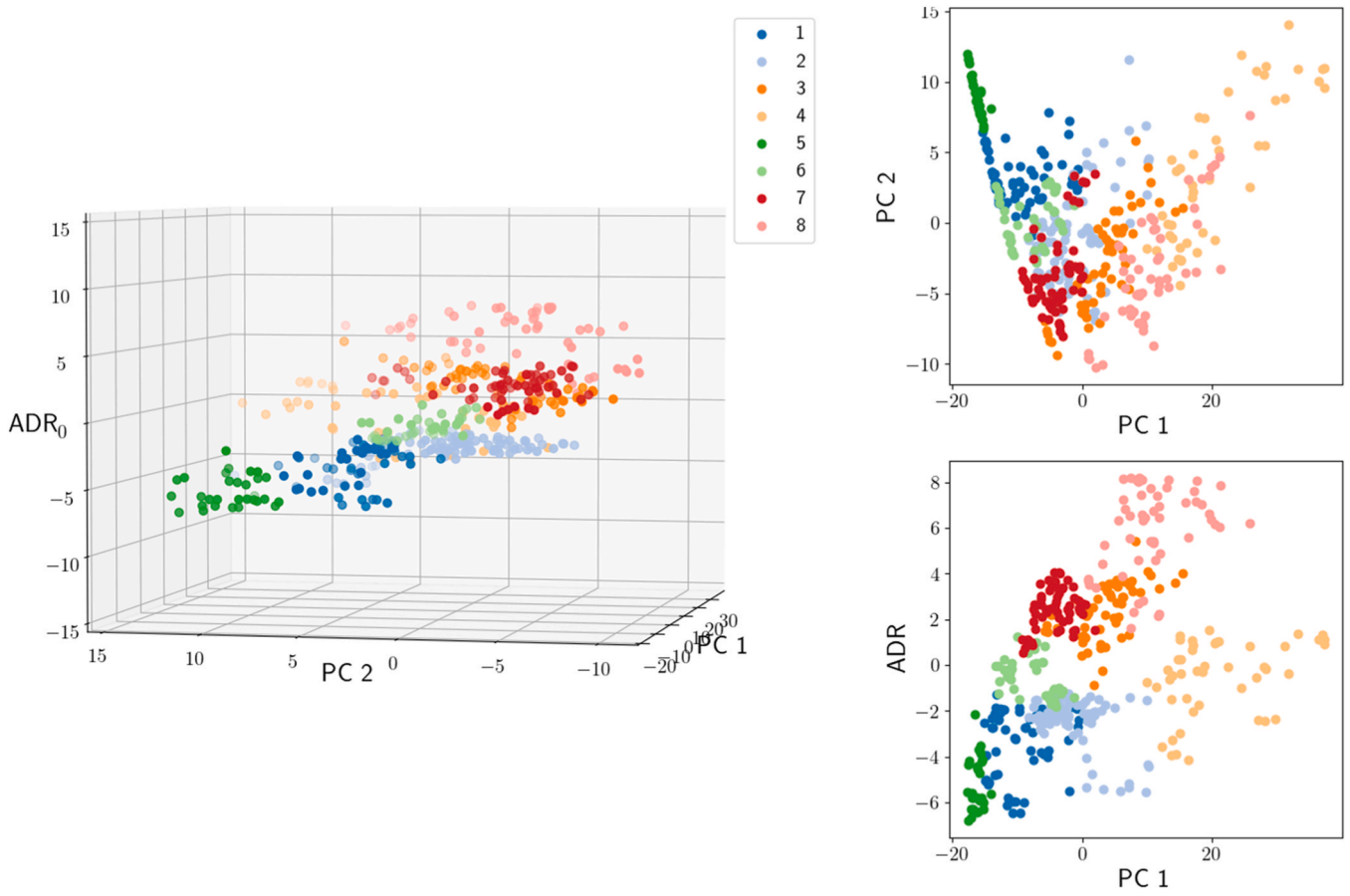
Author(s) (Year)	AI methods used	Key findings
		- Findings suggested that ANN outperformed other models with the lowest MAPE. - SVR did not significantly differ compared to ANN.

Appendix Table 2
Booking curves of Hotel 1 divided by cluster and average of key variables

	Number of dates	ADR (€)	LOS (days)	Booking window	Number of guests	Occupancy rate (%)
	83	110	5.9	30	2.3	86
	81	119	6.3	49	2.3	74
	58	168	6.3	61	2.2	84
	73	164	6.5	45	2.3	86
	56	172	6.5	82	2.2	91
	10	133	7.5	111	2.3	86
	38	76	4.7	13	2.1	41
	27	194	6.9	68	2.2	96



Appendix Fig. 1. Visualisation of the three-dimensional space after the PCA was performed on full booking curves data of Hotel 1. Note: Every point is a date. On the left the 3D representation, on the right some projections of the same plot to emphasise the clusterization. Different colours correspond to different clusters of Appendix 2.



Appendix Fig. 2. Visualization of a reduced space after the PCA was performed on booking curves and ADR for data of Hotel 1. Note: ADR is added to the first two components the ADR as an orthogonal feature.

References

Alarfaj, E., AlGhowinem, S., 2019. Forecasting Air Traveling Demand for Saudi Arabia's Low Cost Carriers. In: Arai, K., Kapoor, S., Bhatia, R. (Eds.), Intelligent Systems and Applications: Proceedings of SAI Intelligent Systems Conference. Springer, pp. 1208–1220. https://doi.org/10.1007/978-3-030-01054-6_84.

Ampountolas, A., Legg, M.P., 2021. A segmented machine learning modeling approach of social media for predicting occupancy. Int. J. Contemp. Hosp. Manag. *33* (6), 2001–2021. https://doi.org/10.1108/IJCHM-06-2020-0611.

Andrew, W.P., Cranage, D.A., Lee, C.K., 1990. Forecasting hotel occupancy rates with time series models: an empirical analysis. J. Hosp. Tour. Res. *14* (2), 173–182.

Assaf, A.G., Tsionas, M.G., 2019. Forecasting occupancy rate with Bayesian compression methods. Ann. Tour. Res. *75*, 439–449. https://doi.org/10.1016/j.annals.2018.12.009.

Athiyaman, A., Robertson, R.W., 1992. Time series forecasting techniques: short-term planning in tourism. Int. J. Contemp. Hosp. Manag. *4* (4), 8–11.

Au, N., Law, R., 2000. The application of rough sets to sightseeing expenditures. J. Travel Res. 39, 70–77.

Bi, J.-W., Han, T.-Y., Li, H., 2022. International tourism demand forecasting with machine learning models: The power of the number of lagged inputs. Tour. Econ. *28* (3), 621–645. https://doi.org/10.1177/1354816620976954.

Bi, J.-W., Li, C., Xu, H., Li, H., 2022. Forecasting daily tourism demand for tourist attractions with big data: an ensemble deep learning method. J. Travel Res. *61* (8), 1719–1737. https://doi.org/10.1177/00472875211040569.

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. Tour. Manag. *22* (4), 403–409. https://doi.org/10.1016/S0261-5177(00)00068-6.

Chen, C., Kachani, S., 2007. Forecasting and optimization for hotel revenue management. J. Revenue Pricing Manag. *6*, 163–174.

Chen, R., Liang, C.-Y., Hong, W.-C., Gu, D.-X., 2015. Forecasting holiday tourism flow based on seasonal support vector regression with adaptive genetic algorithm. Appl. Soft Comput. 26, 435–443. https://doi.org/10.1016/j.asoc.2014.10.022.

Chow, W.S., Shyu, J.C., Wang, K.C., 1998. Developing a forecast system for hotel occupancy rate using integrated ARIMA models. J. Int. Hosp., Leis. Tour. Manag. *1*, 55–80.

Fang, P., Gao, Z., Tsay, R.S., 2023. Supervised kernel principal component analysis for forecasting. Financ. Res. Lett. 58, 104292 https://doi.org/10.1016/j.frl.2023.104292.

Fiori, A.M., Foroni, I., 2019. Reservation Forecasting Models for Hospitality SMEs with a View to Enhance Their Economic Sustainability. Sustainability 11 (5), 1274. https://doi.org/10.3390/su11051274.

Fiori, A.M., Foroni, I., 2020. Prediction accuracy for reservation-based forecasting methods applied in Revenue Management. Int. J. Hosp. Manag. 84, 102332 https://doi.org/10.1016/j.ijhm.2019.102332.

Firat, M., Yiltas-Kaplan, D., Samli, R., 2021. Forecasting air travel demand for selected destinations using machine learning methods. J. Univers. Comput. Sci. *27* (6), 564–581. https://doi.org/10.3897/jucs.68185.

Frechtling, D. (2012). Forecasting Tourism Demand. Routledge: Abingdon, UK.

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L., 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. 80–89. https://doi.org/10.48550/arXiv.1806.00069.

Hassani, H., Silva, E.S., Antonakakis, N., Filis, G., Gupta, R., 2017. Forecasting accuracy evaluation of tourist arrivals. Ann. Tour. Res. *63*, 112–127. https://doi.org/10.1016/j.annals.2017.01.008.

Hassani, H., Webster, A., Silva, E.S., Heravi, S., 2015. Forecasting U.S. tourist arrivals using optimal singular spectrum analysis. Tour. Manag. *46*, 322–335. https://doi.org/10.1016/j.tourman.2014.07.004.

Heo, C.Y., Viverit, L., Pereira, L.N., 2024. Does historical data still matter for demand forecasting in uncertain and turbulent times? An extension of the additive pickup time series method for SME hotels. J. Revenue Pricing Manag. *23*, 39–43. https://doi.org/10.1057/s41272-023-00421-1.

Huang, L., Li, C., Zheng, W., 2023. Daily hotel demand forecasting with spatiotemporal features. Int. J. Contemp. Hosp. Manag. *35* (1), 26–45. https://doi.org/10.1108/IJCHM-12-2021-1505.

Huang, L., Zheng, W., 2021. Novel deep learning approach for forecasting daily hotel demand with agglomeration effect. Int. J. Hosp. Manag. *98*, 103038 https://doi.org/10.1016/j.ijhm.2021.103038.

Huang, L., Zheng, W., 2023. Hotel demand forecasting: a comprehensive literature review. Tour. Rev. *78* (1), 218–244. https://doi.org/10.1108/TR-07-2022-0367.

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: A review and recent developments. Philos. Trans. R. Soc. A: Math., Phys. Eng. Sci. 374 (2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.

Kaya, K., Yilmaz, Y., Yaslan, Y., Oguducu, S.G., Cingi, F., 2022. Demand forecasting model using hotel clustering findings for hospitality industry. Inf. Process. Manag. *59* (1), 102816 https://doi.org/10.1016/j.ipm.2021.102816.

Kon, S.C., Turner, W.L., 2005. Neural network forecasting of tourism demand. Tour. Econ. *11*, 301–328.

Kouprioucha, L., van der Rest, J.-P., Schwartz, Z., 2014. On revenue management and the use of occupancy forecasting error measures. Int. J. Hosp. Manag. *41*, 104–114. https://doi.org/10.1016/j.ijhm.2014.05.002.

Koutras, A., Panagopoulos, A., Nikas, I.A., 2016. Forecasting Tourism Demand Using Linear Prediction Models. Acad. Tur. - Tour. Innov. J. *9* (1), 85–98.

Law, R., Au, N., 2000. Relationship modeling in tourism shopping: A decision rules induction approach. Tour. Manag. *21*, 241–249.

Law, R., Li, G., Fong, D.K.C., Han, X., 2019. Tourism demand forecasting: A deep learning approach. Ann. Tour. Res. *75*, 410–423. https://doi.org/10.1016/j.annals.2019.01.014.

Lee, M., 2018. Modeling and forecasting hotel room demand based on advance booking information. Tour. Manag. *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. Tour. Manag. *68*, 116–126. https://doi.org/10.1016/j.tourman.2018.03.006.

Li, X., Li, H., Pan, B., Law, R., 2021. Machine learning in Internet search query selection for tourism forecasting. J. Travel Res. *60* (6), 1213–1231. https://doi.org/10.1177/0047287520934871.

Li, G., Song, H., Witt, S., 2005. Recent development in econometric modelling and forecasting. J. Travel Res. *44* (1), 82–99. https://doi.org/10.1177/0047287505276594.

Mariani, M., Baggio, R., Fuchs, M., Höepken, W., 2018. Business intelligence and big data in hospitality and tourism: a systematic literature review. Int. J. Contemp. Hosp. Manag. *30* (12), 3514–3554. https://doi.org/10.1108/IJCHM-07-2017-0461.

Pai, P.-F., Hung, K.-C., Lin, K.-P., 2014. Tourism demand forecasting using novel hybrid system. Expert Syst. Appl. *41* (8), 3691–3702. https://doi.org/10.1016/j.eswa.2013.12.007.

Peng, B., Song, H., Crouch, G.I., 2014. A meta-analysis of international tourism demand forecasting and implications for practice. Tour. Manag. *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. Int. J. Hosp. Manag. *58*, 13–23. https://doi.org/10.1016/j.ijhm.2016.07.003.

Pereira, L.N., Cerqueira, V., 2022. Forecasting hotel demand for revenue management using machine learning regression methods. Curr. Issues Tour. *25* (7), 2733–2750. https://doi.org/10.1080/13683500.2021.1999397.

Phumchusri, N., Ungtrakul, P., 2020. Hotel daily forecasting for high-frequency and complex seasonality data: a case study of Thailand. J. Revenue Pricing Manag. *19*, 8–25. https://doi.org/10.1057/s41272-019-00221-6.

Sánchez, E.C., Sánchez-Medina, A.J., Pellejero, M., 2021. Identifying critical hotel cancellations using artificial intelligence. Tour. Manag. Perspect. *35*, 100718 https://doi.org/10.1016/j.tmp.2020.100718.

Sanchez-Medina, J.A., C.-Sanchez, E., 2020. Using machine learning and big data for efficient forecasting of hotel booking cancellations. Int. J. Hosp. Manag. *89*, 102546 https://doi.org/10.1016/j.ijhm.2020.102546.

Schwartz, Z., Hiemstra, S., 1997. Improving the accuracy of hotel reservations forecasting: Curves similarity approach. J. Travel Res. *36* (1), 3–14. https://doi.org/10.1177/004728759703600102.

Silva, E.S., Hassani, H., Heravi, S., Huang, X., 2019. Forecasting Tourism Demand with Denoised Neural Networks. Ann. Tour. Res. *74*, 134–154. https://doi.org/10.1016/j.annals.2018.11.006.

Song, H., Li, G., 2008. Tourism demand modelling and forecasting: a review of recent research. Tour. Manag. *29* (2), 203–220.

Song, H., Qiu, R.T., Park, J., 2019. A review of research on tourism demand forecasting: launching the annals of tourism research curated collection on tourism demand forecasting. Ann. Tour. Res. *75*, 338–362. https://doi.org/10.1016/j.annals.2018.12.001.

Sun, X., Sun, W., Wang, J., Zhang, Y., Gao, Y., 2016. Using a Grey–Markov model optimized by Cuckoo search algorithm to forecast the annual foreign tourist arrivals to China. Tour. Manag. *52*, 369–379. https://doi.org/10.1016/j.tourman.2015.07.005.

Sun, S., Wei, Y., Tsui, K.L., Wang, S., 2019. Forecasting tourist arrivals with machine learning and internet search index. Tour. Manag. *70*, 1–10. https://doi.org/10.1016/j.tourman.2018.07.010.

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

Viverit, L., Heo, C.Y., Pereira, L.N., Tiana, G., 2023. Application of Machine Learning to Cluster Hotel Booking Curves for Hotel Demand Forecasting. Int. J. Hosp. Manag. *111*, 103455 https://doi.org/10.1016/j.ijhm.2023.103455.

Wang, J., Duggasani, A., 2020. Forecasting hotel reservations with long short-term memory-based recurrent neural networks. Int. J. Data Sci. Anal. *9*, 77–94. https://doi.org/10.1007/s41060-018-0162-6.

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

Webb, T., Schwartz, Z., Xiang, Z., Singal, M., 2020. Revenue management forecasting: The resiliency of advanced booking methods given dynamic booking windows. Int. J. Hosp. Manag. *89*, 102590 https://doi.org/10.1016/j.ijhm.2020.102590.

Wen, L., Liu, C., Song, H., 2019. Forecasting tourism demand using search query data: A hybrid modelling approach. Tour. Econ. *25* (3), 309–329. https://doi.org/10.1177/13548166187683.

Wetzel, S.J., 2017. Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. Phys. Rev. E *96* (2), 022140. https://doi.org/10.1103/PhysRevE.96.022140.

Wu, D.C., Song, H., Shen, S., 2017. New developments in tourism and hotel demand modeling and forecasting. Int. J. Contemp. Hosp. Manag. *29* (1), 507–529. https://doi.org/10.1108/IJCHM-05-2015-0249.

Xing, G., Sun, S., Bi, D., Guo, J., Wang, S., 2022. Seasonal and trend forecasting of tourist arrivals: An adaptive multiscale ensemble learning approach. Int. J. Tour. Res. *24* (3), 425–442. https://doi.org/10.1002/jtr.2512.

Yang, X., Pan, B., Evans, J.A., Lv, B., 2015. Forecasting Chinese tourist volume with search engine data. Tour. Manag. *46*, 386–397. https://doi.org/10.1016/j.tourman.2014.07.019.

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

Zhang, C., Hu, A.-Y., Tian, Y.-X., 2023. Daily tourism forecasting through a novel method based on principal component analysis, grey wolf optimizer, and extreme learning machine. J. Forecast. *42* (8), 2121–2138. https://doi.org/10.1002/for.3007.

Zhang, Y., Li, G., Muskat, B., Law, R., 2021. Tourism demand forecasting: a decomposed deep learning approach. J. Travel Res. *60* (5), 981–997. https://doi.org/10.1177/0047287520919522.

Zhang, C., Li, M., Sun, S., Tang, L., Wang, S., 2022. Decomposition methods for tourism demand forecasting: a comparative study. J. Travel Res. *61* (7), 1682–1699. https://doi.org/10.1177/00472875211036194.

Zhang, D., Niu, B., 2024. Leveraging online reviews for hotel demand forecasting: a deep learning approach. Inf. Process. Manag. *61* (1), 103527 https://doi.org/10.1016/j.ipm.2023.103527.

Zhang, D., Wu, C., 2023. What online review features really matter? An explainable deep learning approach for hotel demand forecasting. J. Assoc. Inf. Sci. Technol. *74*, 1100–1117. https://doi.org/10.1002/asi.24807.

Zhu, M., Wu, J., Wang, Y.-G., 2021. Multi-horizon accommodation demand forecasting: a New Zealand case study. Int. J. Tour. Res. *23* (3), 442–453. https://doi.org/10.1002/jtr.2416.