Published in International Journal of Hospitality Management, 2023, vol. 111, no. 103455, which should be cited to refer to this work.
DOI:10.1016/j.ijhm.2023.103455

Application of Machine Learning to Cluster Hotel Booking Curves for Hotel Demand Forecasting

1. Introduction

Demand forecasting is one of the key components of a successful revenue management (RM) tool in the hospitality industry because it is a building block of a hotel's strategic decision, as their accuracy determines the efficacy of pricing and rooms inventory optimization decisions (Haensel & Koole, 2011; Huang & Zheng, 2021; Weatherford & Kimes, 2003). Indeed, several studies showed that forecasting accuracy is critical to the success of a RM strategy (Haensel & Koole, 2011; Schwartz et al., 2016). For example, according to the Polt (2000)'s study, a decrease of forecasting error by 10~25% can lead to an increase in the airline firm's revenue by 1-2%. Similarly, Weatherford and Belobaba (2002)'s study indicated that about 1~2% of revenue of airlines can be increased by reducing forecasting error by 25%. More recently, van Leeuwen and Koole (2022) showed that it is possible to generate 3~10% more revenue for a hotel using a price optimization module based on forecasted demand curves.

However, reducing forecasting error is not an easy task for hotels. In particular, as the COVID-19 pandemic decimated the international tourism market, the tourism and hospitality industry has also been hit hard. Enforced travel restrictions and lockdowns due to the COVID-19 pandemic weighted heavily on demand and revenue of hotels worldwide. Market conditions for hotels have shifted dramatically with fundamental changes in tourists' behavior and demand patterns are continuously evolving. Tourists' anxiety has affected tourism demand and led to uneven demand patterns, which presents both short-term and long-run challenges for hotels (Zhang & Lu, 2022).

Reliable and accurate forecasts for hotel demand are critical for managing such crisis, but an unprecedented demand environment caused by the COVID-19 pandemic has made the forecasting and strategic planning process even more difficult. Historical data lost its value and forecasting has become even more complex, as demand is changing quickly and unpredictably (Kourentzes et al., 2021).

Several scholars have explored different forecasting methods for tourism and hotel demand, but research on hotel demand forecasting is not as abundant as research in tourism demand forecasting (Wu et al., 2017; Zhang & Lu, 2022). Moreover, research on demand forecasting methods in the midst of uncertainty is rare. Furthermore, a few revenue management solutions (RMS) in the market claim that machine learning has been applied to their system, but the forecasting models of most RMS are still mainly based on combined forecasting models which use historical booking records and advanced booking data (e.g., pickup methods based on trailing periods) (Sánchez, Sánchez-Medina & Pellejero, 2021). Although there has been more and more literature using machine learning to make predictions, lack of interpretability in predictive models has been raised as one of the key concerns and undermine trust in those models (Rudin, 2019). With the black box of the machine learning model, it is difficult to understand how demand is estimated or a decision is made because the model does not explain or show its methodology. Therefore, this study aims to propose a hotel demand forecasting method using an interpretable machine learning approach so that we can understand how it arrived at a specific estimation.

Schwartz (2008) argued that the shape and pattern of the time-before-the-date-of-stay determines the forecasting model predictions. Indeed, Schwartz and Hiemstra (1997) focused on the shape of past booking curves for hotels and discovered that the curves similarity model is considerably more accurate than traditional time series models (i.e., Stepwise Autoregression Model, High Order Polynomial Model and Weighted Average of AR and HOPM Predictions).

Their study applied a dissimilarity measure to identify past booking curves with a similar shape, the model is tested with 10 forecasting horizons ranging from 1 to 99 days in advance of three hotels in the United States.

To add additional knowledge to the literature on hotel demand forecasting, this study proposed a new approach by clustering stay dates generated from historical booking data using a machine learning algorithm. In this study we follow a three-step approach. First, historical booking curves are clustered based on the patterns of booking curves by a machine learning algorithm. Second, clustered booking curves are used in the additive pickup model to forecast daily occupancy in the near future (up to 8 weeks). Third, the accuracy of proposed forecasting approaches (i.e., based on the clustered booking curve) is evaluated by comparing forecasting errors of the traditional pick-up forecasting method (i.e., trailing period) with four forecasting horizons ranging from 7 to 50 days.

This study will contribute to the literature on hotel demand forecasting in several ways. First, while this approach still utilizes historical booking data, it is fundamentally different from traditional forecasting approaches that assume that the booking patterns tend to be similar to the trailing periods (e.g., Tse & Poon, 2015; Lee, 2018; Fiori & Foroni, 2020). As traditional demand forecasting models for hotels use the historical reservation data corresponding to each arrival date, several scholars argued that the selection of trailing periods is critical for hotel demand forecasting (Ma et al., 2014). But the hotel industry learned that historical data from trailing periods is no longer relevant when demand volatility and uncertainty are high in a time of crisis like the COVID 19 pandemic. Therefore, this study circumvents the basic assumption that future trends will hold similar to historical trends based on the same day last year booking data (i.e., based on trailing periods). Instead, this study calculated the average number of bookings from clustered booking curves, instead of booking curves from trailing periods. Forecasting daily hotel room demand based on clustered booking curves by machine learning

has not been adopted for short-term forecasting for hotels. Second, this study proposes a hotel demand forecasting method using an interpretable machine learning approach so that we can fundamentally understand how a specific forecast has been made. The findings of this study will help future scholars to improve our forecasting approach further by using a machine learning algorithm. Third, this study tests the ability of a new demand forecasting model during unprecedented demand uncertainty caused by the COVID-19 pandemic by analyzing daily booking data of three independent hotels between 2018 and 2020 in Italy and France.

2. Literature review

Demand forecasting is a critical tool for generation of information to support the decision-making process in the travel, tourism and hospitality industries (Wu et al., 2017). The theory of forecasting is based on the premise that past and current data or information can be used to make predictions about the future (Petropoulos et al., 2022). Researchers use past and current data as inputs to make informed forecasts that predict the direction of future trends. Although the topic of demand forecasting seems very narrow, it actually contains two important lines of research: tourism demand forecasting and hotel demand forecasting.

Tourism demand forecasting has been a popular line of research among tourism scholars (e.g., see Athiyaman & Robertson, 1992; Song & Li ,2008; Frechtling, 2012; Peng et al., 2014; Wu et al., 2017, and the papers cited therein) because tourism planning and destination management rely on accurate tourist arrival forecasting. Several studies that summarized the state of the art of tourism demand forecasting across the time (e.g., Witt & Witt, 1995; Song & Li, 2008; Frechtling, 2012; Wu et al., 2017) showed that most of the published studies aiming to forecast tourism demand have used quantitative methods dominated by time-series models and econometric approaches. They also concluded that the tourist arrivals variable is the most popular measure of tourism demand, although predictor

variables included in the tourism demand econometric models vary depending on the study's objectives (Song & Li, 2008). Time-series models focus on predicting the future path of an interest variable regarding its own historical and a random disturbance term, while the econometric models try to identify causal relationships between the tourism demand variable and its influencing factors in order to accurately forecast tourism demand (Li, Song & Witt, 2005). Since the beginning of this century, new machine learning-based methods have started to be used to forecast tourism demand (e.g. Song & Li, 2008; Peng et al., 2014).

Many studies on hotel demand forecasting have followed the approaches used for tourism demand forecasting, such as time-series and econometric models (Wu et al., 2017). But hotel demand forecasting cannot be parallel to tourism demand forecasting because hotel guests are not strictly only tourists. In addition, forecasts of aggregated market demand of a hotel (e.g., quarterly or monthly level), based on a time-series approach, helps hoteliers to understand seasonal patterns of demand in order to define strategic policies for the variables of the marketing mix. However, in the hospitality industry, forecasts of short-term demand (at weekly or daily levels) have a more critical impact on hotel RM operational decisions, such as pricing decisions and inventory control (Pereira, 2016; Lee, 2018; Fiori & Foroni, 2020). While annual, quarterly, or monthly data are often used in tourism demand forecasting, such data is not sufficient for short-term hotel demand forecasting to support decision-making at that operational level (Huang & Zheng, 2021). This is the key reason why the majority of the literature on demand forecasting uses daily data (e.g., see Kouprioucha et al. (2014) and Weatherford (2016) and the papers cited therein, as well as more recent papers, such as Lee (2018) and Fiori & Foroni (2020)).

In the hotel demand forecasting literature, forecasting models are mostly based on historical transaction data and on advanced booking data (i.e., reservations on hand) (Pereira, 2016). For short-term hotel demand forecasts, advanced booking data are the most important

type, because it reflects the most recent demand changes (Zakhary, Gayar & Atiya, 2008). Therefore, advanced booking models (e.g., pickup models, econometric advanced booking models) are mostly used for short-term revenue management-oriented forecasting (e.g., Weatherford & Kimes, 2003; Tse & Poon, 2015; Lee, 2018; Fiori & Foroni, 2020). While econometric models try to recognize the quantitative relationship between final bookings data and reservations on hand, pickup models identify the unique features of reservation data and estimate the reservations to receive in the future by aggregating the possible additional reservations. As pickup models use the historical reservation data corresponding to each arrival date, previous literature found that the selection of those data (i.e., trailing periods) is critical in the hotel demand forecasting (Ma et al., 2014).

While no consensus on the best forecasting model for hotel daily room demand is reached, some scholars have pointed out the limitations of traditional forecasting models (e.g. Webb et al., 2020; Huang & Zheng, 2021). As a result, a growing body of literature has focused on new approaches for hotel demand forecasting by taking different perspectives, whether including new data sources (e.g., Pan, Wu & Song, 2012) or using AI-based models (e.g., Sánchez, Sánchez-Medina & Pellejero, 2021), but it is still limited when compared with the literature dealing with new quantitative approaches of tourism demand forecasting. On one hand, Pan, Wu and Song (2012) included search query volume data from Google specifically categorized as travel queries, in econometric forecasting models for hotel occupancy forecasting. Similarly, Wu, Hu and Chen (2021) suggested mixed data sampling models for hotel occupancy rate and compared it with competitive models such as times series and econometric models. The results of their analysis showed that combining big data sources, such as daily visitor arrival and search query data, can improve forecasting accuracy especially when demand variation is high. On the other hand, Webb et al. (2020) assessed the forecasting performance of neural networks with advanced booking data. They concluded that neural

networks are suitable for forecasting hotel demand within a context of dynamic booking windows. Wang and Duggasani (2020) forecasted constrained hotel daily demand using LSTM-based recurrent neural networks. Using real time series from four hotels located in the USA, they concluded that two deep learning LSTM models (a time-based and a time-rate-based model) outperform a set of machine learning algorithms in the majority of the cases. In the same line of research, Huang and Zheng (2021) proposed a spatiotemporal deep learning LSTM model. This study revealed that the deep learning model with spatial and temporal correlations performs better than time series (ARIMA model), econometric (Vector Autoregression model) and other LSTM models in an empirical study using hotel daily demand data of 210 Chinese hotels. Pereira and Cerqueira (2022) compared machine learning models based on arbitrating with traditional methods, such as seasonal naïve and exponential smoothing methods for double seasonality using a real time series of daily demand, for a four-star hotel in Europe. Their study found that hotel demand forecasting using machine learning models outperformed traditional exponential smoothing methods.

Although recent studies have attempted to incorporate new approaches on how to improve accuracy of hotel demand forecasts, the majority of the literature about this topic is still based on traditional approaches (time series, pickup and econometric models). In addition, the COVID-19 pandemic has unpredictably changed individuals' way of life including travel behavior. As travelers' booking patterns, such as the booking window, have been dramatically changed due to high uncertainty, using only methods strongly dependent on the "same day last year" booking data is not very relevant for hotel forecasting. Indeed, Webb et al. (2020) applied a neural network approach to explore how the shift in booking windows affects forecasting accuracy. Webb et al. (2020)'s study showed that forecasting with dynamic booking windows poses significant challenges to hoteliers and that forecasting models using the booking curve tend to be less affected by shifts in booking window. Assaf and Tsionas (2019) proposed two

nonlinear Vector Autoregressive models and an extension of them by using neural networks to make the models more flexible. This new approach significantly improved the accuracy of monthly forecasts of occupancy rates of hotels belonging to the same competitive set. More recently, Zhang and Lu (2021) forecasted hotel demand within the COVID-19 pandemic context, but they used traditional regression models to forecast quarterly hotel demand.

In the context of hotel demand forecasting, there is evidence that disaggregated forecasts of hotel demand are more accurate than aggregated forecasts (Weatherford et al., 2001). Although this line of research has been unexplored, recently Bandalouski et al. (2021) disaggregated hotel demand into several categories (e.g., time of the booking, time and length of the stay, room type) in order to improve forecast accuracy. Kaya et al. (2022) argued that disaggregation of hotel demand is an attempt to use subsets of data with the same behavior for forecasting purposes. Thus, their study first grouped hotels into similar segments with additional features obtained from K-Means Clustering findings and forecasted weekly hotel demand using a LSTM model. This is an interesting approach that aims to take advantage of forecasting at disaggregated level, and using AI-based models, but it did not provide high-frequency forecasts and applied a clustering method only at the property level. Two disaggregated forecasting approaches by Bandalouski (2021) and Kaya et al. (2022) may produce a more reliable forecast compared to aggregated approaches when demand is stable, but their forecasting accuracy cannot be assured when demand uncertainty is high due to a sudden external shock. Although several studies have adopted several new forecasting approaches, no study has used a machine learning algorism for clustering booking curves based on the booking patterns, instead of trailing periods, to enhance the accuracy of forecasts of hotel daily demand.

3. Research Method

3.1. Data used in this study

The data used in this research are real reservation data from three hotels for three consecutive years (i.e., 2018-2020). In order to verify the efficacy of the proposed approach for hotels targeting different markets, three hotels in distinct locations are selected (i.e., seasonal tourism destination, all season tourism destination, major city). The selected hotels are three independent boutique hotels in Europe (i.e., Italy and France) which do not have an advanced RMS. Hotel 1 is a three-star property with 32 rooms located in Isola D'Elba (Italy), open only from April to mid-October, with leisure customers only. Hotel 2 is a four-star property with 47 rooms in Nice (France) while Hotel 3 is a four-star property with 26 rooms in Paris (France). Both hotels 2 and 3 are open all year round with business and leisure customers. The information stored in each reservation record includes the booking date, the arrival date, the length of stay (LOS), the room rate and the number of rooms booked. Since demand behavior of these three hotels is generated by diverse determinants, and one hotel closes in the low demand season, 90 booking curves of check-in days between July 1st and September 28th were used in order to assure coherence in a comparative analysis. More details about the data used in this research are presented in Table 1. The data show that the average daily occupancy rate and the average LOS of hotel 1 are higher than in the other two hotels, because hotel 1 is targeted for leisure tourists and is open only during the mid and high season. On the other hand, the average room rate and the average booking window are lower in hotel 1 than in the others. As depicted in Table 2, the average daily occupancy rate ranges between 66% (hotel 3) and 88% (hotel 1), while the average LOS ranges between 3.49 (hotel 3) and 6.22 (hotel 1). In terms of booking window, customers on average book rooms 53.5 days before the date of arrival in the hotel with the highest average room rate (€ 285.28).

(Please insert Table 1 here)

3.2. Methods

A fundamental question among scholars and practitioners has been how to improve accuracy of hotel demand forecasts because it is well-known that demand forecasting is a critical component of any hotel RMS (Weatherford, 2016). This study aims to forecast hotel daily demand by identifying similar booking patterns in the historical daily booking curves using machine learning algorism, in which pickup methods were used at a disaggregated level. Pickup methods take into account information about reservations on hand for future stay dates, do not need a model specification (e.g., like the ARIMA methods or the econometric models), have a very low computational time and are among the most accurate forecasting methods used in RM (Weatherford & Kimds, 2003; Fiori & Foroni, 2020). All these reasons give this approach a strong appeal in terms of simplicity and robustness for RMS designed to generate high frequency short-term hotel demand forecasts.

Therefore, the research methods used in this research are the following. First, a cluster analysis is used to identify data-driven segments of stay dates with a similar booking pattern (i.e., the shape of booking curve). Second, daily occupancy forecasts are generated in each cluster using the additive pickup model. Finally, the quality of the forecasts is assessed using traditional accuracy forecasting measures and compared to those of the traditional pickup method.

3.2.1. A cluster analysis: Segmentation of booking curves

A cluster analysis was performed to identify segments of stay dates (calendar days) for each hotel. Clustering technique, which is one of the most common approaches in data mining and pattern recognition, is used to discover its underlying structure (Jain, Murty, & Flynn, 1999). Previous literature considered clustering to be an effective method of grouping data into many collections according to the similarities of data points' features and characteristics (Jain, 2010, Abualigah, Khader, & Hanandeh, 2018). Clustering techniques belong to the wide field

of non-supervised machine learning, which is the set of computational methods meant to classify data based only on the features of the data themselves, without training the algorithm with examples of pre-classified data. Unlike more complicated deep-learning techniques, clustering is computationally very fast and can work properly even with relatively small sets of data. Nonetheless, it can partition booking curves into clusters that are highly non-straightforward and that are barely perceptible manually.

The data-driven segmentation was based on data provided by the booking curves for each stay date in a time span of three years (2018, 2019 and 2020). A clustering process was followed to select the number of clusters (Dolnicar, Grun & Leisch, 2018). A hierarchical procedure using Ward's method and the squared Euclidean distance as a similarity measure was used. The number of clusters can be adjusted by tuning the threshold of Ward's distance in order to balance the homogeneity of the curves within each cluster and the portability of the results. A threshold of 200 seems appropriate according to the dendrograms of Figure 1, giving an eight-cluster solution for hotels 1 and 2, and a seven-cluster solution for hotel 3. A larger number of clusters would have more homogeneity within each segment, but it would be less parsimonious and require more resources to perform the forecasting task.

(Please insert Figure 1 here)

3.2.2. Time correlations in the data

In order to study how long in advance it is possible to make predictions about occupancy of rooms in each cluster, we calculated the correlation functions

$$g_c(t) = \frac{\sum_{i \epsilon c} \; r_i(t) \cdot r_i(0)}{\sqrt{\sum_{i \epsilon c} \; r_i^2(0)} \cdot \sqrt{\sum_{i \epsilon c} \; r_i^2(t)}}, \qquad (1)$$

where c is the index of the cluster, $r_i(t)$ is the number of rooms occupied for stay date i measured t days in advance, $r_i(0)$ is the actual number of rooms occupied, and the sums are performed on all dates belong to cluster c. This function quantifies to what extent the

information about the occupancy at time 0 is available *t* days in advance in each cluster. (Box, Jenkins, Reinsel, & Ljung, 2015). For each day in each cluster, we calculated the correlation functions in the time interval from 0 (corresponding to the same day, at which it is 1 by definition) to 180 days in advance.

3.2.3. Forecasting models: Forecasting with advanced booking data

Alternative forecasting models have been used to forecast short-term hotel demand. Forecasting models fall into one of three categories: historical booking models, advanced booking models and combined models (Lee, 1990; Weatherford & Kimes, 2003). Historical booking models concern only the final number of rooms occupied or arrivals for each stay day in the past, while advanced booking models reflect the pattern of reservations over a booking horizon for a target stay day in the future. Finally, combined models utilize both the historical and advanced booking models, applying either a weighted average or regression, to produce forecasts. The focus of this research is on advanced booking models due to the reasons indicated by Fiori and Foroni (2020). Since we are using real data from independent hotels, and given the fact that one hotel closes during the low demand season, these models are preferred to historical models because they do not rely on complete daily time series and are easy to implement in practice. In addition, Weatherford and Kimes (2003) concluded that pickup methods and regression produce the lowest error, while the booking curve and combination forecasts produced fairly inaccurate results.

Advanced booking models use a two-step approach to generate forecasts of the number of rooms occupied in future stay dates. First, these models forecast expected daily reservations until a future point in time (stay day) based on a daily known pattern of reservations that occurred over the recent past in each lead time (Zakhary et al., 2008; Fiori & Foroni, 2020). Second, a forecast of the number of rooms occupied for each future date until the stay day,

made on a specific reading day, is therefore obtained by adding the number of rooms occupied based on reservations on hand until the current reading day with those daily forecasts of reservations to come. Additive pickup methods assume that the number of on-hand reservations is independent of the number of rooms that will be booked later on, while multiplicative pickup methods assume that future bookings are positively correlated with the current level of reservations on hand.

The resulting forecasts are generally responsive to recent shifts in demand, particularly if the forecasts of reservations to come are computed using historical patterns of reservations that are very similar to the booking behavior of each future date until the stay day. Thus, we argue that demand forecasts computed with each segment of stay dates will be more accurate than forecasts computed with all available data.

3.2.4. Forecasting accuracy measures

The accuracy of alternative forecasting approaches is assessed using the following six measures: Mean Squared Error (*MSE*), Mean Absolute Error (*MAE*), Root Mean Squared Error (*RMSE*), Mean Absolute Percentage Error (*MAPE*), and Symmetric Mean Absolute Percentage Error (*sMAPE*) and Median Absolute Percentage Error (*MdAPE*). For a post-sample of h periods, $t = n+1, n+2, \ldots, n+h$, these accuracy measures are given by:

$$MSE = \frac{1}{h}\sum_{t=n+1}^{n+h} (y_t - f_t)^2, \quad (2)$$

$$MAE = \frac{1}{h}\sum_{t=n+1}^{n+h} |y_t - f_t|, \quad (3)$$

$$RMSE = \sqrt{\frac{1}{h}\sum_{t=n+1}^{n+h} (y_t - f_t)^2}, \quad (4)$$

$$MAPE = \frac{1}{h}\sum_{t=n+1}^{n+h} \frac{100|y_t - f_t|}{y_t}, \quad (5)$$

$$sMAPE = \frac{1}{h}\sum_{t=n+1}^{n+h} \frac{200|y_t - f_t|}{y_t + f_t}, \quad (6)$$

$$MdAPE = median\left\{\left|\frac{100|y_{n+1} - f_{n+1}|}{y_{n+1}}\right|, \left|\frac{100|y_{n+2} - f_{n+2}|}{y_{n+2}}\right|, \ldots, \left|\frac{100|y_{n+h} - f_{n+h}|}{y_{n+h}}\right|\right\} \quad (7)$$

where y_t represents the observed number of rooms occupied in day *t* and f_t denotes a forecast of y_t. We have decided to use scale-independent measures (*RMSE*, *MAPE*, *sMAPE* and *MdAPE*), apart from scale-dependent measures (*MSE* and *MAE*), because they are reported in percentage and, thus, easy to interpret. In addition, we have used the *sMAPE* because it is a symmetric accuracy measure with a fixed range, which avoids the problem of large errors and the *MdAPE* because it avoids forecasting errors considered outliers. Readers interested in learning more about forecasting accuracy are referred to Kouprioucchina et al. (2014) and Pereira (2016).

4. Results

4.1. Clustering booking curves

Figure 1 shows results of the cluster analysis for each hotel. The dendrograms reveal different agglomeration processes for each hotel. In general, the booking curves in each cluster of each hotel have different behaviors, revealing that there are sets of stay dates that receive the majority of bookings many days in advance (e.g., cluster E of Hotel 1), while other sets receive bookings a few days in advance (e.g., cluster F of Hotel 1), or across the entire booking horizon (e.g. cluster D of Hotel 1). Figure 1 also depicts that there are sets of stay dates that receive the majority of bookings a few days in advance, but belong to different clusters because some have high occupancy rates (e.g., cluster H of Hotel 1), while others have low occupancy rates (e.g. cluster F of Hotel 1). On the contrary, there are sets of stay dates that continuously receive bookings along the booking horizon, but belong to different clusters because some have high occupancy rates (e.g. cluster C of Hotel 1), while others have lower occupancy rates (e.g. cluster B of Hotel 1). Based on these results, we argue that hotel demand forecasting models should be applied at cluster-level, because forecasts of future bookings will be based only on historical data of stay dates with a similar behavior of the target forecasting date.

(Please insert Figure 1 here)

A profile of each cluster is presented in Table 2, using the following set of variables: daily occupancy rate, average LOS, average room rate and average booking window. While the machine learning algorithm identified 8 unique clusters (A~H) for Hotel 1 and 2, only 7 clusters (A~G) were identified for Hotel 3. Figure 1 further presents the shapes of each cluster. Hotel 1 has only one cluster of stay dates with a low occupancy rate (cluster F). This cluster also presents the lowest average LOS (4.7 days), room rate (€ 69.3) and average booking window (10.9 days). The remaining seven clusters have, in general, high daily occupancy rates, of which four have occupancy rates greater than 95% (clusters D, E, G and H). However, results presented in Table 2 reveal that these clusters are distinct. For example, clusters D and E have a distinct behavior in terms of average booking window (83.7 versus 110.2, respectively) and room rate (€ 130.0 versus € 142.3). Clusters G and H are also different in terms of average booking window (37.1 versus 31.8) and room rate (€ 163.0 versus € 129.0).

Hotel 2 also has only one cluster of stay dates with a low occupancy rate (cluster F), but it does not have the lowest average LOS (4.2 days) and average booking window (15.6 days) as it was observed for hotel 1. Table 2 shows that this hotel also has several clusters with high daily occupancy rates (five clusters have an occupancy rate greater than 90%), but it has two clusters with moderate occupancy rates (cluster G: 75.0%; cluster H: 81.4%). Although some clusters have similar occupancy rates, there are noticeable differences among them. For example, the two clusters with the highest and most similar occupancy rates (cluster D: 99.0%; cluster E: 98.5%) are also similar in terms of average LOS (4.2 versus 4.1), and reveal the highest, but significantly different, average booking windows (119.5 versus 134.5). Interestingly, the following three clusters with the highest occupancy rates (clusters A: 93.9%; B: 64.4% and C: 96.2%) are also similar in terms of average LOS (the three lowest LOS), but they are distinct in terms of average booking windows (65.6; 82.8; 46.4, respectively).

Finally, Table 2 shows that Hotel 3 has only two clusters of stay dates with occupancy rates greater than 90% (cluster A: 94.4%; cluster B: 94.7%), which are also similar in terms of average LOS (3.8 versus 3.9). These two clusters reveal the highest, but significantly different, average booking windows (114.1 versus 138.9). The majority of clusters of Hotel 3 have moderate occupancy rates (cluster C: 88.5%; cluster D: 84.0%; cluster F: 70.0%; cluster G: 85.9%) and similar average LOS (3.4-3.6 days), but they have significantly different average booking windows (89.9; 65.8; 22.3; 31.0, respectively). There is still a cluster of this hotel that joins the stay dates with the lowest occupancy rate (31.3%), average LOS (3.1) and booking window (12.2). An example of the clusters' members (stay dates) of Hotel 1, per year, is given in Figure 2. The same colors are used to represent each cluster in each hotel in all figures. Figure 2 reveals that the same stay dates in different years belong to different clusters. In addition, those booking patterns in 2020 are different when compared with the previous years, for the same stay dates.

In summary, Figure 1 and Table 2 show that there are clearly distinct clusters of booking curves in each hotel, and some of them are similar in different hotels (e.g. clusters E of hotels 1 and 2; clusters F of hotels 1 and 2 and cluster E of hotel 3). This result supports the idea that this methodology might be applied in different types of SME hotels.

(Please insert Table 2 here)

(Please insert Figure 2 here)

4.2. Time correlations in the data

A relevant question concerning the possibility of predicting the number of rooms occupied at a given date in advance is whether the time series contains such information, independently of the specific algorithm that will be used to extract it. One can quantify the possibility of knowing the occupancy of a hotel t days in advance with the correlation function

g(t) defined in Equation (1). When *g(t)* assumes values close to 1 it means that at that time it is possible to predict with high confidence the occupancy at time 0 (i.e., at the present time); when *g(t)* is close to 0 it means that at that time there is no information available to predict the occupancy at time 0. Of course, *g(t)* starts at 1 at small *t* and drops to 0 at long times.

In Figure 3 we plot the correlation between the occupancy at time 0 (i.e., the day of arrival) and the bookings on hand n days before arrival, for the different clusters. Each cluster displays a very different correlation behavior, some of them maintain values close to 1 several months in advance, while the fastest-decaying ones drop after approximately one month. This fact suggests that it is possible to make accurate forecasts for all dates in all clusters at least one month in advance, but for some clusters this possibility extends to much longer forecasting horizons.

(Please insert Figure 3 here)

In fact, the typical time needed by the RMSE to increase in the different clusters is strongly correlated with the decay time of the correlation function. The left panel of Figure 4 shows the number of days before arrival, after which the RMSE goes above 10% of the number of rooms in the hotel, versus the number of days after which the correlation function drops below 0.9 - the Pearson's correlation coefficient between the two quantities is 0.75. Analogously, the Pearson's correlation coefficient between decay time of the correlation function and average booking window (right panel of Figure 4) is consistently high (i.e., 0.86).

(Please insert Figure 4 here)

4.3. Forecasting accuracy

Table 3 summarizes the results of six accuracy measures per forecasting method and per hotel for a selected set of forecasting horizons (7, 14, 30 and 50 days before arrival). Results clearly show that cluster-based forecasts are generally more accurate compared to traditional pickup modesl with trailing periods, regardless of the accuracy measure used, both for all

forecasting horizons and for all hotels. For example, according to the MAPE, cluster-based demand forecasts of hotel 1 are 8.5% more accurate than classical pickup forecasts for a forecasting horizon of 14 days, while cluster-based forecasts are 32.7% more accurate for a forecasting horizon of 50 days. The accuracy gains of cluster-based demand forecasts are less marked in hotel 2 for forecasting horizons up to 14 days and in hotel 3 for forecasting horizons of 30 and 50 days. In fact, in hotel 3 the conventional forecasts are slightly more accurate than cluster-based demand forecasts for short-term forecasting horizons (7 and 14 days), which might be explained by the uncertainty generated by the COVID-19 pandemic, which changed the profile of the clusters generated with data from the pre-pandemic years. In summary, although those accuracy gains are neither uniform along the forecasting horizon nor across hotels, results of this study show that cluster-based forecasts tend to outperform the conventional forecasts that are based on the additive pickup method.

(Please insert Table 3 here)

5. Discussion and conclusion

5.1. Practical implication

Accurate demand forecasting forms an integral part of data-driven RM decisions for hotels. Econometric models based on historical booking information cannot capture the dynamic effect of unprecedented event such as the COVID-19 pandemic. Demand forecasting during unpredictable and volatile times pose significant challenges to hoteliers. Therefore, this study tries to avoid the traditional forecasting methods that assume that booking patterns tend to behave in similar ways if they refer to the same calendar period and the same day-of-week. Instead, this study proposed a new approach for forecasting daily hotel demand by clustering historical booking curves regardless of trailing periods and combining them into advance bookings information using AI. While traditional approaches usually forecast hotel demand based on the complete advanced booking data (and/or historical transaction data), the proposed

approach aims to improve forecasting accuracy by only using data from similar stay dates (i.e., belonging to the same cluster) to generate forecasts. This new forecasting approach was tested with real hotel booking data of three hotels and showed that using clustered booking curves can improve the accuracy of occupancy forecasts for hotels.

5.2. Theoretical implication

This study contributed to the body of knowledge on hotel demand forecasting by proposing a new approach for utilizing historical data (i.e., clustering booking curves). Based on the theory of forecasting, this study took a two-step process for hotel demand forecasting. First, the arrival dates are segmented based on their booking curves using a clustering algorithm. Second, the daily hotel demand forecasts are predicted using the additive pickup method. This study used both historical and advanced booking data only from objects of the same cluster to generate forecasts of hotel demand. This study assumed that objects (e.g., stay dates) with similar booking behavior are allocated in the same cluster, in which an internal homogeneity prevails, and is heterogeneous when compared with the other clusters.

Ma et al. (2014)'s study applied clustering to classify the historical reservation information for forecasting of railway passenger flow. It was an innovative study because it was the first attempt to identify similar patterns of booking curves using a machine learning algorism. Our paper contributes to knowledge in this area since it introduces a new approach for forecasting hotel daily occupancy by using the pickup additive forecasting model in each cluster of booking curves. Results show that this approach outperforms the traditional approach of using all bookings on hand, in each day before arrival, to forecast hotel daily occupancy in the majority of booking horizons up to 8 weeks.

This study also discovered interesting changes in hotel booking curves related to the COVID-19 pandemic. Booking curves during the COVID-19 pandemic (i.e., 2020) were clearly different from those before the pandemic (Figure 2). While the dates of 2018 and 2019

formed the same clusters which also had a clear seasonal behavior, the dates from 2020 lied, instead, in different and new clusters altogether, which also had different booking patterns. This finding further explains why traditional forecasting models cannot perform well when hotel demand is highly uncertain.

5.3. Future research

Although the current study tried to test a new forecasting approach with three hotels targeting different segments, future research may apply our approach to different types of hotels (e.g., chain vs. independent). Another interesting aspect is to see how our forecasting approach works when hotels face different types of demand uncertainty caused by exogenous shocks (e.g., economic crisis, natural disaster, and terrorism). We strongly encourage future research to extend our approach using other AI-based models to improve forecasting accuracy further.

To expand on this research in the future it would be of great value to consider the following concerns that arose during the research process. This study has shown that it is possible to cluster the booking curves from yearly data. A non-trivial challenge is turning this retrospective study into a predictive tool. If clusters of previous years are the same as the current year (i.e., 'stationary' case), as happened in 2018 and 2019. One can develop probabilistic methods to assign the early part of a booking curve to the predefined clusters, provided that the associated correlation function is large enough. On the other hand, if the system is non-stationary, the clusters of the current year can be very different from those of the previous years as in 2020. Here the problem is not just assigning booking curves to a cluster, but also building the reference clusters themselves. This situation can happen, for example, if some major event, like the COVID-19 pandemic, occurs, that completely disrupts the market . Future research may explore how to improve the predictions in the non-stationary situation using various machine-learning tools and exploring dynamic clustering approaches. Finally,

the prediction to which cluster belongs each stay day, in an early stage of the booking curve of each day, is an appealing line of research because it would be possible to improve the forecasting accuracy as well as to explore regression model to forecast hotel demand.

References

Abualigah, L. M., Khader, A. T., Hanandeh, E. S., 2018. Hybrid clustering analysis using improved krill herd algorithm. *Applied Intelligence 48*(11), 4047-4071.

Antonio, N., Almeida, A., & Nunes, L., 2019. An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations. *Data Science Journal 18*(1), 32. http://doi.org/10.5334/dsj-2019-032

Assaf, A.G., Tsionas, M.G. 2019. Forecasting occupancy rate with Bayesian compression methods. *Annals of Tourism Research* 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. *International Journal of Contemporary Hospitality Management* 4(4), 8-11. https://doi.org/10.1108/09596119210018864

Baldigara, T., Mamula, M., 2015. Modelling international tourism demand using seasonal ARIMA Models. *Tourism and Hospitality Management* 21(1), 19-31.

Bandalouski, A.M., Egorova, N.G., Kovalyov, M.Z., Pesch, E., Tarim, S.A., 2021. Dynamic pricing with demand disaggregation for hotel revenue management. *Journal of Heuristics* 27, 869-885. https://doi.org/10.1007/s10732-021-09480-2

Box, G. E.; Jenkins, G. M.; Reinsel, G. C., Ljung, G. M. 2015. *Time series analysis: Forecasting and control*. John Wiley & Sons.

Claveria, O., Monte, E., Torra, S., 2015. A new forecasting approach for the hospitality industry. *International Journal of Contemporary Hospitality Management* 27(7), 1520–1538. https://doi.org/10.1108/IJCHM-06-2014-0286

Dolnicar, S., Grun, B., Leisch, F., 2018. *Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful*. Springer, Singapore. https://doi.org/10.1007/978-981-10-8818-6

Fiori, A.M., Foroni, I., 2020. Prediction accuracy for reservation-based forecasting methods applied in Revenue Management. *International Journal of Hospitality Management* 84, 102332. https://doi.org/10.1016/j.ijhm.2019.102332

Frechtling, D., 2012. *Forecasting Tourism Demand*. Routledge, Abingdon, UK.

Goh, C., Law, R., 2002. Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. *Tourism Management 23*(5), 499–510. https://doi.org/10.1016/S0261-5177(02)00009-2

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

Huang, L., Zheng, W., 2021. Novel deep learning approach for forecasting daily hotel demand with agglomeration effect. *International Journal of Hospitality Management* 98, 103038. https://doi.org/10.1016/j.ijhm.2021.103038

Jain, A. K., Murty, M. N., Flynn, P. J. 1999. Data clustering: A review. *ACM Computing Survey, 31*(3), 264-323.

Jain, A. K., 2010. Data clustering: 50 years beyond K-Means. *Patter Recognition Letter, 31*(8), 651-666.

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

Kouprioucha, 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., Saayman, A., Jean-Pierre, P., Provenzano, D., Sahli, M., Seetaram, N., Volo, S., 2021. Visitor arrivals forecasts amid COVID-19: A perspective from the Africa team. *Annals of Tourism Research* 88(4), 103197. https://doi.org/10.1016/j.annals.2021.103197

Law, R., Li, G., Fong, D.K.C., Han., X., 2019. Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research* 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. *Tourism Management* 66, 62-71. https://doi.org/10.1016/j.tourman.2017.11.004

Li, X., Law, R., 2020. Forecasting tourism demand with decomposed search cycles. *Journal of Travel Research* 59(1), 52-68. https://doi.org/10.1177/0047287518824158

Li, X., Li, H., Pan, B., Law, R., 2021. Machine Learning in Internet Search Query Selection for Tourism Forecasting. *Journal of Travel Research* 60(6), 52-68. https://doi.org/10.1177/0047287520934871

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

Long, W., Liu, C., Song, H., 2019. Pooling in Tourism Demand Forecasting. *Journal of Travel Research* 58(7), 1161–1174. https://doi.org/10.1177/0047287518800390

Ma, M., Liu, J., Cao, J., 2014. Short-term forecasting of railway passenger flow based on clustering of booking curves. *Mathematical Problems in Engineering* 707636 https://doi.org/10.1155/2014/707636

Pan, B., Wu, D.C., Song, H., 2012. Forecasting Hotel Room Demand Using Search Engine Data. *Journal of Hospitality and Tourism Technology* 3(3), 196–210. https://doi.org/10.1108/17579881211264486

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

Petropoulos F., Apiletti D., Assimakopoulos V., Babai M.Z., Barrow D.K., Bergmeir C., ..., Ziel F. 2022. *International Journal of Forecasting* 38(3), 705-871.

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

Pereira, L.N., Cerqueira, V., 2022. Forecasting hotel demand for revenue management using machine learning regression method. *Current Issues in Tourism* https://doi.org/10.1080/13683500.2021.1999397

Polt, S., 2000. From bookings to demand: the process of unconstraining. In: Busutill, L. (Ed.) Proceedings of AGIFORS Reservations and Yield Management Study Group. AGIFORS, New York.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Natural Machine Intell*igence 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

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

Schwartz, Z., 2008. Time, price, and advanced booking of hotel rooms. *International Journal of Hospitality & Tourism Administration* 9(2), 128–146.

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

Schwartz, Z., Uysal, M., Webb, T., Altin, M., 2016. Hotel daily occupancy forecasting with competitive sets: a recursive algorithm. *International Journal of Contemporary Hospitality Management* 28(2), 267-285. https://doi.org/10.1108/IJCHM-10-2014-0507

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., Witt, S.F., Jensen, T.C., 2003. Tourism forecasting accuracy of alternative econometric models. *International Journal of Forecasting* 19(1), 123-141. https://doi.org/10.1016/S0169-2070(01)00134-0

Smeral, E., 2019. Seasonal forecasting performance considering varying income elasticities in tourism demand. *Tourism Economics* 25(3), 355–74. https://doi.org/10.1177/1354816618792799

Smeral, E., Wuger, M. 2005. Does Complexity Matter? Methods for Improving Forecasting Accuracy in Tourism: The Case of Austria. *Journal of Travel Research* 44(1), 100-110. https://doi.org/10.1177/0047287505276596

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

Turner, L.W., Witt, S.F., 2001. Forecasting tourism using univariate and multivariate structural time series models. *Tourism Economics* 7(2),135-147. https://doi.org/10.5367/000000001101297775

van Leeuwen, R., Koole, G., 2022. Demand forecasting in hospitality using smoothed demand curves. *Journal of Revenue and Pricing Management*. https://doi.org/10.1057/s41272-021-00364-5

Wang, J., Duggasani, A., 2020. Forecasting hotel reservations with long short-term memory-based recurrent neural networks. *International Journal of Data Science and Analytics* 9, 77-94. https://doi.org/10.1007/s41060-018-0162-6

Weatherford, L.R., Belobaba, P.P., 2002. Revenue impacts of fare input and demand forecast accuracy in airline yield management. *Journal of the Operational Research Society* 53(8), 811-821. https://doi.org/10.1057/palgrave.jors.2601357

Weatherford, L.R., Kimes, S.E., Scott, D.A., 2001. Forecasting for hotel revenue management: testing aggregation against disaggregation. *The Cornell Hotel and Restaurant Administration Quarterly* 42(4), 53-64. https://doi.org/10.1016/S0010-8804(01)80045-8

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

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

Witt, S.F., Witt, C.A., 1995. Forecasting tourism demand: a review of empirical research. *International Journal of Forecasting* 11(3), 447-475. https://doi.org/10.1016/0169-2070(95)00591-7

Wong, K.K., Song, H., Chon, K.S., 2006. Bayesian models for tourism demand forecasting. *Tourism Management* 27(5), 773-780. https://doi.org/10.1016/j.tourman.2005.05.017

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

Wu, E.H.C., Hu, J., Chen, R., 2021. Monitoring and forecasting COVID-19 impacts on hotel occupancy rates with daily visitor arrivals and search queries. *Current Issues in Tourism* https://doi.org/10.1080/13683500.2021.1989385.

Zakhary, A., Gayar, N., Atiya, A., 2008. A comparative study of the pickup method and its variations using a simulated reservation hotel data. *International Journal of Artificial Intelligence and Machine Learning* 8, 15-21.

Zhang, H., Lu, J., 2022. Forecasting hotel room demand amid COVID-19. *Tourism Economics* https://doi.org/10.1177/13548166211035569

Zhang, Y.Z., Li, G., Muskat, B., Law, R., 2020. Tourism Demand Forecasting: A Decomposed Deep Learning Approach. *Journal of Travel Research* 60(5), 981-997. https://doi.org/10.1177/0047287520919522

Table 1. Profile of the hotels and dataset

		Hotel 1	Hotel 2	Hotel 3
	Location	Isola D'Elba-Italy	Nice-France	Paris-France
Profile	Number of rooms	32	47	26
	Daily occupancy rate	88%	72%	66%
	Average LOS (days)	6.22	4.39	3.49
	Average room rate (€)	114.80	175.85	285.28
	Average booking window (days)	37.00	40.28	53.53
Data	Number of years	3	3	3
	Number of booking curves per year	90	90	90
	Number of days in the booking horizon	180	180	180
	Number of observations	270	270	270

Table 2. Profile of the clusters in each hotel

	Clusters							
Hotel 1 (Elba)	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Daily occupancy rate	92.4%	83.3%	93.0%	95.3%	96.4%	37.0%	96.7%	96.6%
Average LOS (days)	5.4	6.3	6.4	6.5	7.3	4.7	6.9	6.3
Average room rate (€)	98.2	106.7	118.3	130.0	142.3	69.3	163.0	129.0
Average booking window (days)	22.4	46.1	53.6	83.7	110.2	10.9	37.1	31.8
Hotel 2 (Nice)	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Daily occupancy rate	93.9%	94.4%	96.2%	99.0%	98.5%	35.7%	75.0%	81.4%
Average LOS (days)	3.7	3.9	3.8	4.2	4.1	4.2	4.9	4.6
Average room rate (€)	192.3	160.1	165.1	197.0	194.3	142.3	194.0	209.3
Average booking window	65.6	82.8	46.4	119.5	134.5	15.6	14.0	29.5
Hotel 3 (Paris)	**A**	**B**	**C**	**D**	**E**	**F**	**G**	
Daily occupancy rate	94.4%	94.7%	88.5%	84.0%	31.3%	70.0%	85.9%	
Average LOS (days)	3.8	3.9	3.5	3.6	3.1	3.4	3.5	
Average room rate (€)	338.98	348.68	302.67	308.77	256.34	302.73	333.18	
Average booking window	114.1	138.9	89.9	65.8	12.2	22.3	31.0	

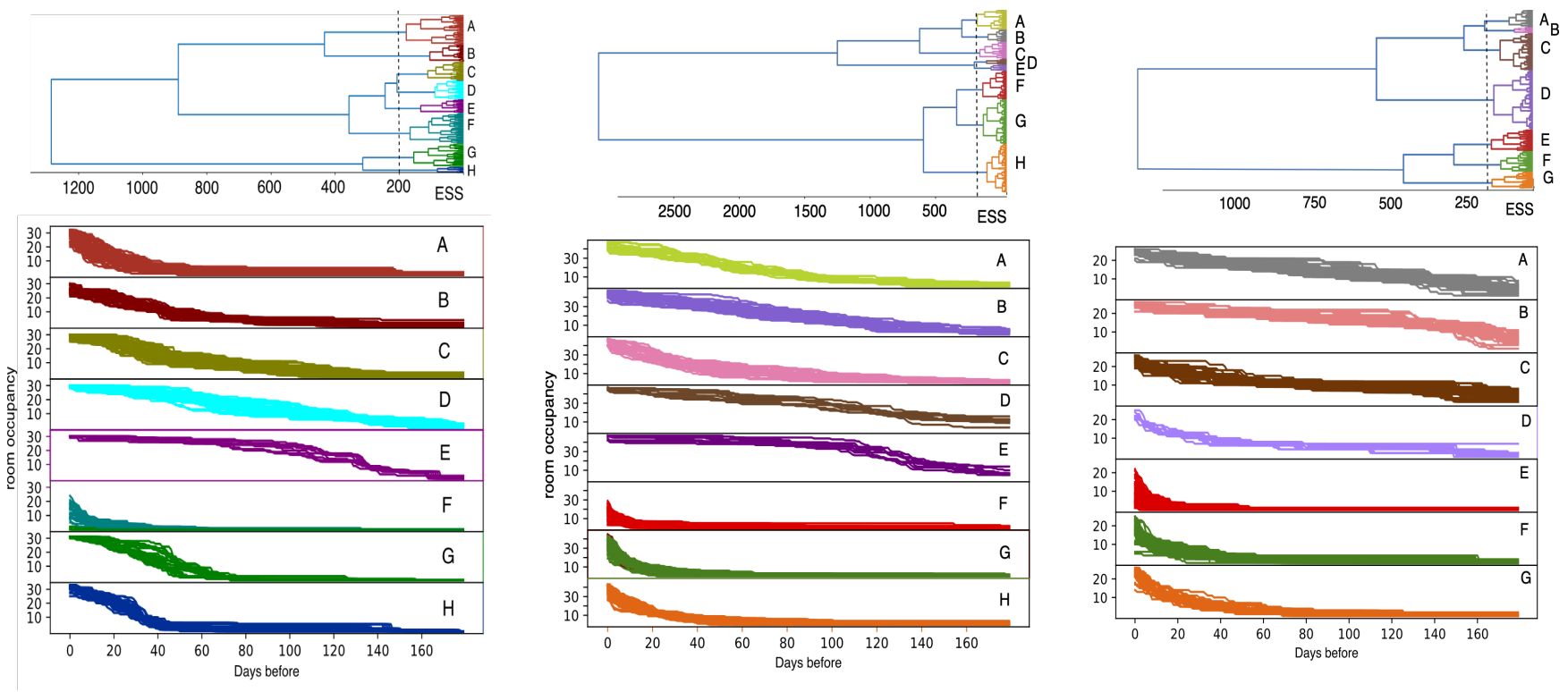
*** Note: A~H represents unique clusters of booking curves identified by machine learning algorithm based on the booking patterns.**

Table 3: Accuracy measures per forecasting method and per hotel

		DBA	MSE	RMSE	MAE	MAPE	sMAPE	MdAPE
Hotel 1	**Classical Pickup**	7	11.77	3.43	2.30	18.65	13.25	4.61
		14	32.01	5.66	4.00	30.84	21.56	6.67
		30	63.76	7.99	6.15	36.94	35.77	20.69
		50	116.94	10.81	8.99	62.72	56.36	32.28
	Cluster Based Pickup	7	7.72	2.78	1.85	14.52	11.42	4.08
		14	13.79	3.71	2.75	21.53	15.69	6.67
		30	18.43	4.29	3.25	29.17	19.71	7.64
		50	24.03	4.90	3.79	45.58	26.81	9.96
Hotel 2	**Classical Pickup**	7	18.77	4.33	3.38	14.72	12.76	7.70
		14	50.76	7.12	5.34	27.79	19.93	10.71
		30	115.91	10.77	7.85	48.18	29.16	15.22
		50	235.48	15.35	11.70	91.72	43.82	19.01
	Cluster Based Pickup	7	18.38	4.29	3.21	16.33	12.68	5.92
		14	37.56	6.13	4.36	23.37	16.85	7.08
		30	54.73	7.40	5.38	36.42	22.90	9.24
		50	62.80	7.92	5.81	50.72	27.33	10.43
Hotel 3	**Classical Pickup**	7	6.42	2.53	1.80	23.47	19.00	7.29
		14	9.23	3.04	2.17	31.16	20.24	8.33
		30	17.84	4.22	3.04	59.91	29.95	11.61
		50	20.33	4.51	3.34	69.17	32.44	11.23
	Cluster Based Pickup	7	5.98	2.45	1.80	31.19	18.93	6.65
		14	9.09	3.02	2.26	43.13	22.96	8.33
		30	14.55	3.81	2.81	67.09	29.18	8.33
		50	17.08	4.13	3.14	68.66	31.90	9.50

***Note: Classical pickup refers to the traditional approach based on trailing periods**

Figure 1: Dendrograms and booking curves of each cluster per hotel



*** Note: A~H represents unique clusters of booking curves identified by machine learning algorithm based on the booking patterns.**

Figure 2: Calendars based on clusters

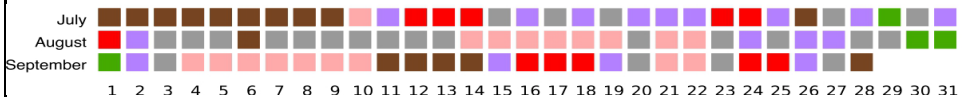
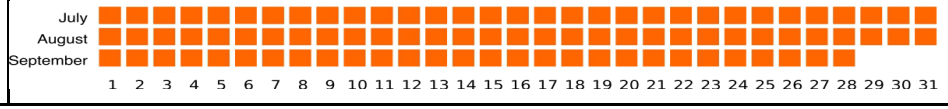
Hotel	Year	Calendars based on clusters
Hotel 1	2018	
	2019	
	2020	
Hotel 2	2018	
	2019	
	2020	
Hotel 3	2018	
	2019	
	2020	

Figure 3: Time correlations between the occupancy at a given time in the past and the final occupancy for each cluster

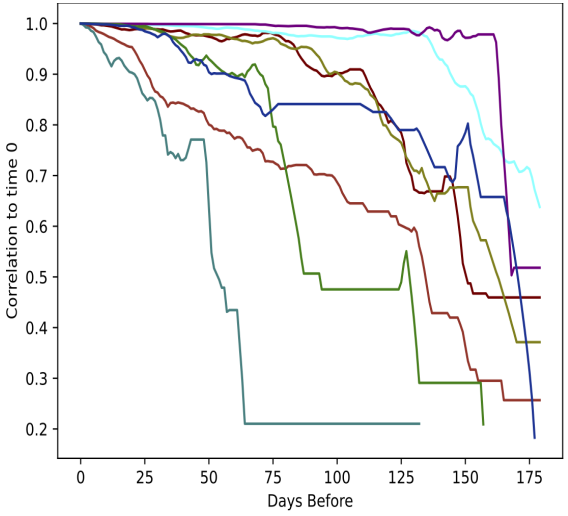
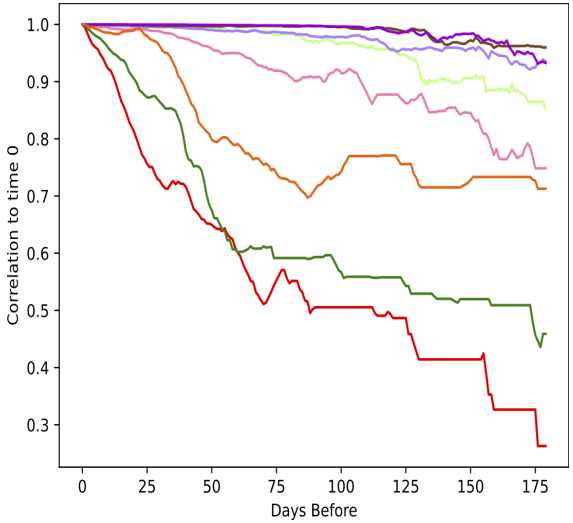
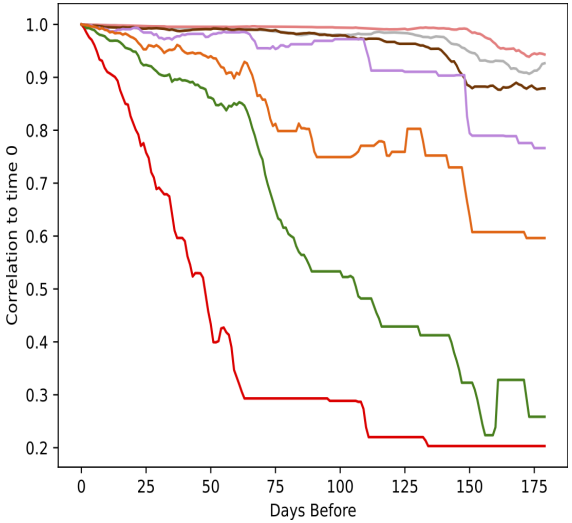
Hotel 1	Hotel 2	Hotel 3
		

Figure 4: The relation between the decay time of the correlation function and the typical increase time of the RMSE in each cluster (on the left) and the average booking window (on the right).