Title Page (with Author Details)

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

CEO
Hotelnet, Italy
+39 02 4547 2175
luciano.viverit@hotel

*
Associate Professor
EHL Hospitality Business School
HES-SO / University of Applied Sciences and Art
+41 21 785 15
cindy.heo@ehl.ch

Assistant Professor
Research Centre for Tourism, Sustainability and Wel
Escola Superior de Gestão, Hotelaria e Turis
+ 351 28 980 0136
lmper@ualg.pt

Associate Professor
Department of Physics and Center for Complexity and B
Università degli Studi di Milano, Italy
+39 025 031 7221
guido.tiana@unimi.it

Declarations of Interest: None

Acknowledgements

Authors are thankful to Infoelba srl and Dylog SAM for providing the hotel data, and Umberto Grieco and Sammer Allamadani for the support given during the research. This paper is partially financed by Portuguese National Funds provided by FCT- Foundation for Science and Technology through project UIDB/04020/2020.

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) in the hospitality industry because it

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 RM strategy (Haensel & Koole, 2011; Schwartz et al., 2016). For example, according to Polt (2000)'s study, a d

of forecasting error by 10~25% can lead to an-2% increas

Similarly, Weatherford and Belobaba (2002)'s

airlines can be increased by reducing forecasting error by 25%.

However, reducing forecasting error is not an easy task for hotels. In particular, as the COVID-19 pandemic has heavily hit the international tourism market, the tourism and hospitality industry have been considerably affected. Enforced travel restrictions and lockdowns due to the COVID-19 pandemic had a massive negative impact on demand and revenue of hotels worldwide. Market conditions for hotels have shifted dramatically with fundamental changes in tourists' behavior a

Touri [illegible] leads to reductions in tourism demand and uneven demand pattern, which presents both short-term and long-run challenges for hotels (Zhang & Lu, 2022).

Although the tough times are not yet over, hoteliers need to remain agile and make decisions with speed and accuracy. Reliable and accurate forecasts for hotel demand are critical for managing this continuing crisis. However, an unprecedented demand environment caused by 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 demands are 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 those in tourism demand forecasting (Wu et al., 2017; Zhang & Lu, 2022) and research on demand forecasting method 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 models). Therefore, this study aims to propose a new approach for daily hotel room demand forecasting by using clusters of stay dates generated from historical booking data. This new approach is fundamentally different from traditional forecasting approaches for hotels that assume the booking curves and patterns tend to be similar during the trailing period approach. In this study we follow a two-step approach. First, historical booking curves are clustered by a machine learning algorithm using an auto-regressive manner. Second, clustered booking curves are used in the additive pickup model to forecast daily occupancy in the near future (up to 8 weeks).

The efficacy of a new forecasting approach is tested using real hotel booking data of three hotels. To the best of our knowledge, forecasting daily hotel room demand based on clustered booking curves by machine learning has not been adopted for short-term forecasting for hotels nor is used in connection with advanced booking models so far. Therefore, this study will contribute to the literature on hotel demand forecasting. In addition, the findings of this study will be useful for industry practitioners such as hoteliers and hotel RMS providers by providing a new approach to forecast demand when they face high demand uncertainty.

2. Literature review

2.1. Tourism demand forecasting

Tourism demand forecasting has been a popular topic 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 to forecast tourism demand 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 with study 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). However, since the beginning of this century new machine learning-based methods start to be used to forecast tourism demand (e.g. Song & Li, 2008; Peng et al., 2014).

Traditional time-series models that have been used for tourism demand forecasting accommodate a wide range of approaches belonging to the exponential smoothing and the Autoregressive Integrated Moving Average (ARIMA) families of models (Song & Li, 2008; Frechtling, 2012; Peng et al., 2014). The first family of models have been widely used during decades, by both academics, most of the times as a benchmark for performance evaluation, and industry practitioners, due to its simplicity of implementation and relatively good performance. The double exponential smoothing, the Holt exponential smoothing and the Holt-Winters exponential smoothing are some well-known examples of approaches of the exponential smoothing family of models (Peng et al., 2014). The ARIMA family of models has been most applied to forecast tourism demand but different versions of the ARIMA models have also been adopted. Seasonal ARIMA model is one of the popular models as seasonality and the seasonal variation in tourism demand can impact on strategic decisions about tourism destination management (Song & Li, 2008). For example, Baldigara and Mamula (2015) used seasonal ARIMA to explain the tourism demand patterns of Germans in Croatia. However, previous literature found mixed results about the va

study found that the seasonal ARIMA models outperformed compared to eight other time-series models, but Smeral and Wuger (2005) found that the ARIMA or SARIMA model did not perform better than the naïve (no change) model. Although ARIMA variants can generate forecasts through time series, several limitations exist, because most of the ARIMA variants assume a linear relationship between the future and past time steps values.

The autoregressive distributed lag model, error correction model, and vector autoregression (VAR) are the widely used in the econometric family of models (Song & Li, 2008; Long, Liu & Song, 2019; Smeral, 2019). Some scholars have compared different forecasting models and identified the strengths and weaknesses of each approach. Wong, Song, and Chon (2006) evaluated the Bayesian vector autoregression (BVAR) to solve the overfitting problem and the empirical analysis showed that the BVAR models consistently outperform their unrestricted VAR models. On the other hand, Turner & Witt (2001) applied the structural time series models and showed it outperforms the naïve no change model, but the causal structural time series model does not generate more accurate forecasts than the univariate model. Song, Witt, and Jensen (2003) evaluated six alternative econometric forecasting models for tourism demand, which are special cases of a general autoregressive distributed lag specification.

While the most common approaches used in tourism demand forecasting are the timeseries models and the econometric models, new approaches have been proposed since the beginning of the millennium in a continuous attempt to improve forecasts accuracy (Song & Li, 2008; Peng et al., 2014; Wu et al., 2017). Indeed, Peng, Song, and Crouch (2014) categorized previous literature on tourism demand forecasting into three categories: time series, econometric, and artificial intelligence (AI)-based models. New quantitative methods recently proposed generated new categories of approaches as classified by Li and Law (2020) and Li et al. (2021). AI-based approaches became more popular than traditional models in forecasting tourism demand due to its high accuracy of the forecasts and simplicity of implementation (Wu el al., 2017; Law et al., 2019). Claveria, Monte, and Torra (2015) developed a multivariate neural network framework to improve forecasting accuracy of AI models. They applied that framework to real tourism demand monthly data of Catalonia from 2001 to 2012. Law et al. (2019) proposed a deep learning approach for tourism demand forecasting, based on the long short-term memory (LSTM) model, which automatically extracts relevant demand indicators and tested the model for Macau's tourism dem (2020) proposed a decomposition method that achieves high accuracy in short- and long-term AI-based tourism demand forecasting models. Scholars argued that deep learning approach has specific advantages such as powerful forecasting capability and feature engineering (LeCun, Bengio, & Hinton 2015; Law et al., 2019), but such approach in demand forecasting in the tourism and hospitality sector is still limited compared with other conventional approaches. Recently, Kourentzes et al. (2021) compared traditional models with AI-based models to forecast tourism demand in the uncertainty context of the COVID-19 pandemic. Although the growing popularity of new methods for tourism demand forecasting, this section briefly reviewed the main trends of research in this topic because the aim of this paper is to forecast hotel demand in the short-term. Thus, next section overviews the literature on hotel demand forecasting to highlight the unique features.

2.2. Hotel demand forecasting

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). Forecasts of aggregated market demand of a hotel (e.g., at weekly or monthly levels), 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, and on a daily basis, have a more critical impact on hotel revenue management operations 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 use daily data (e.g., see Koupriouchina et al. (2014) and Weatherford (2016) and the papers cited therein, as well as more recent papers, such us Lee (2018) and Fiori & Foroni (2020)). In the hotel demand forecasting literature, we can find models based on historical transaction data and on advanced booking data (Pereira, 2016). For example, Bandalouski et al. (2021) used historical data to forecast disaggregated hotel demand, for short- and long-term horizons, using time-series models. Then, they used these forecasts of hotel demand in each category to feed dynamic pricing models. For short-term hotel demand forecasts, advanced booking data (i.e., reservations on hand) is 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), which are based on the information of reservations on hand, 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, the selection of those data 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 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 (VAR 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, pick up and econometric models). In addition, the COVID-19 pandemic has unpredictably changed individu way of life including travel behavior. As travel booking patterns such as booking window has been dramatically change due to high uncertainty, using only methods strongly dependent on the “same day last booking data is not very relevant for hotel forecasting. Indeed, Webb et al. (2020) applied a neural network approach to explore how the shift of booking windows effect on forecasting accuracy. Webb et al. (2020)'s study showed that forecasting in dynamic booking window pose significant challenges to hoteliers and the forecasting models using the booking curve tend to be less affected by shifts in booking window. 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, generated using traditional forecasting models, 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 forecasts accuracy. Indeed, disaggregation of hotel demand is an attempt to use subsets of data with the same behavior for forecasting purposes, for which the cluster analysis might have a key role. Using this argument, Kaya et al. (2022) first grouped hotels into similar segments and, second, 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 using the valuable information included in the booking curves.

To the best of authors' knowledge, there is not a study that combines cluster segmentation of booking curves to enhance the accuracy of forecasts of hotel daily demand in a booking horizon of up to 8 weeks. Therefore, this study aims to forecast hotel daily demand by identifying similar booking patterns in the historical daily booking curves using machine learning methods, in which pickup methods were used at disaggregated level.

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 any advanced RMS. Hotel 1 is a three-stars property with 32 rooms I open only from April to mid-October, with leisure customers only. Hotel 2 is a four-stars property with 47 rooms in Nice (France) while Hotel 3 is a four-stars property with 26 rooms in Paris (France). Both hotels 2 and 3 are open all year around 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.

The general characteristics of the hotels are indicated 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 1, 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

(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 revenue management (Weatherford, 2016). In this research we sought to contribute to that aim through the application of a forecasting model at a disaggregated level. 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.

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. 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 the Ward's method and the sq was used. The number of clusters can be adjusted by tuning the threshold of 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, while 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 generate more costs of performing the forecasting task.

Clustering techniques belong to the wide field of non-supervised machine learning, that 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. At variance with 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 could have hardly been spotted manually.

(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

$$(\ ) = \frac{\sum \quad (\ ) \quad (0)}{\sqrt{\sum \quad ^2(0)}\ \sqrt{\sum \quad ^2(\ )}}, \qquad (1)$$

where c is the index of the cluster, $(\ )$ is the number of rooms occupied for stay date i measured t days in advance, (0) is the actual number of rooms occupied, and the sums are performed on all dates belonging 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.

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, from which one hotel closes during the low 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 daily reservations yet to come 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 very similar with the booking behavior of each future date until the stay day. Thus, we argue that demand forecasts computed with each segment of stays 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 two well-known measures: Root Mean Squared Error (*RMSE*) and Mean Absolute Percentage Error (*MAPE*). For a post-sample of *h* periods, $= +1, +2, \ldots, +$, these accuracy measures are given by:

$$= \sqrt{\frac{1}{} \sum_{=+1}^{+} (-)^2}, \qquad (2)$$

$$= \frac{1}{} \sum_{=+1}^{+} \frac{| - |}{}, \qquad (3)$$

where represents the observed number of rooms occupied in day *t* and denotes a forecast of . Readers interested in learning more about forecasting accuracy are referred to Koupriouchina et al. (2014).

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 few days in advance (e.g. cluster F of Hotel 1), or another continuously receive bookings during all 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 few days in advance, but belonging to different clusterss 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 opposite, there are sets of stay dates that continuously receive bookings along the booking horizon, but belonging 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. 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 r
(10.9 days). The remaining seven clusters have, in general, high daily occupancy rates, from 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 r

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). Noticeable, 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 (s in Figure 2. Keeping the same colors 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, that 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 or not 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 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.

As plotted in Figure 3, different clusters display very different correlation functions. 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)

4.3. Forecasting accuracy

Table 3 summarizes results of two 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 forecasts of hotel demand are more accurate when they are generated at cluster-level, whatever the accuracy measure used, both for all forecasting horizons and for all hotels. For example, cluster-based demand forecasts of hotel 1 are 8.5% more accurate than classic forecasts for a forecasting horizon of 14 days, while cluster-level forecasts are 32.7% more accurate for a forecasting horizon of 50 days. The accuracy gains of cluster-based demand forecasts are less pronounced in hotel 2 for forecasting horizons up to 14 days and in hotel 3 for forecasting horizons of 30 and 50 days. In summary, although those accuracy gains are neither uniform along the forecasting horizon nor across hotels, results of this case study show that cluster-based forecasts clearly outperform the classic forecasts based on the additive pickup method.

(Please insert Table 3 here)

5. Discussion and conclusion

Accurate demand forecasting is integral for data-driven revenue management decisions of hotels. Econometric models based on historical booking information cannot capture the dynamic effect of unprecedented event such as COVID-19 pandemic. Demand forecasting during unpredictable and volatile times posse significant challenges to hoteliers. Therefore, this study tries to avoid the traditional forecasting method that assumes that booking patterns tend to behave in the similar way if they refer to the same calendar period and the same day-of-week. Instead, this study proposed to cluster historical booking curves regardless of trailing periods and combine them into advance bookings information using artificial intelligence. This new forecasting method was tested with real hotel booking data of three hotels and showed that using clustered booking curves can improve the accuracy of occupancy forecast for hotels. Ma, et al (2014)'s study applied clustering to classify the historical reservation information for forecasting of railway passenger flow, but it is the first attempt to ignore the same day last year principle to utilize historical booking data in the hospitality literature, to best our knowledge.

This study discovered the interesting changes of hotel booking curves related to the COVID-19 pandemic. Booking curves of Hotel 1 during the COVID-19 pandemic were clearly different from those before the pandemic (Figure 2). While the dates of 2018 and 2019 formed the same clusters with a clear seasonal behavior, the dates from 2020 lied instead in different and new clusters altogether with different booking patterns. This finding further explains why traditional forecasting models cannot perform well when hotel demand is highly uncertain.

Although current study tried to test our forecasting model with three hotels targeting different segments, future research may apply our model to different types of hotels (e.g., chain vs. independent). Another interesting aspect is to see how our forecasting model 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 extend our model using artificial intelligence method to improve forecasting accuracy further.

To extend this research in the future it would be of a great value to consider the following concerns that arose during the research process. This study has shown above that it is possible to clustering the booking curves from yearly data. A non-trivial challenge is turning this retrospective study into a predictive tool. The easiest situation is that in which the clusters of previous years are the same as the current year, something that can be called the 'stationary' case, as happened in 2018 and 2 this case, 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 COVID-19 pandemic, occurs, changing in a consistent way the booking market. Future research may explore how to improve the predictions in the non-stational situation using various machine-learning tools. 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 models to forecast hotel demand.

References

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

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

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

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

Kourentzes, N., 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

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.

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., 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

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

	Hotel 1	**Hotel 2**	**Hotel 3**
Location	Isola-Italy D’	Nice-France	Paris-France
Number of rooms	32	47	26
Daily occupancy rate	88%	72%	66%
Average LOS (days)	6.22	4.39	3.49
Average room r	114.80	175.85	285.28
Average booking window (days)	37.00	40.28	53.53

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 ro	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 ro	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 ro	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	

Table 3: Accuracy measures per forecasting method and per hotel

		7 Days before		14 Days before		30 Days before		50 Days before	
		MAPE	**RMSE**	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**	**RMSE**
Hotel 1	**Cluster Pick-up**	11.48%	1.55	17.05%	2.32	21.48%	3.08	21.32%	2.90
	Classic Pick-up	16.92%	2.13	25.51%	3.92	32.98%	5.37	54.01%	7.86
Hotel 2	**Cluster Pick-up**	16.32%	3.31	22.06%	4.32	27.46%	5.01	28.09%	4.96
	Classic Pick-up	17.69%	3.87	28.29%	5.65	49.39%	9.25	85.67%	11.25
Hotel 3	**Cluster Pick-up**	22.80%	2.13	32.25%	2.39	62.00%	3.68	77.01%	3.91
	Classic Pick-up	27.09%	2.28	44.45%	2.85	66.25%	3.89	81.02%	4.15

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

Hotel 1	Hotel 2	Hotel 3

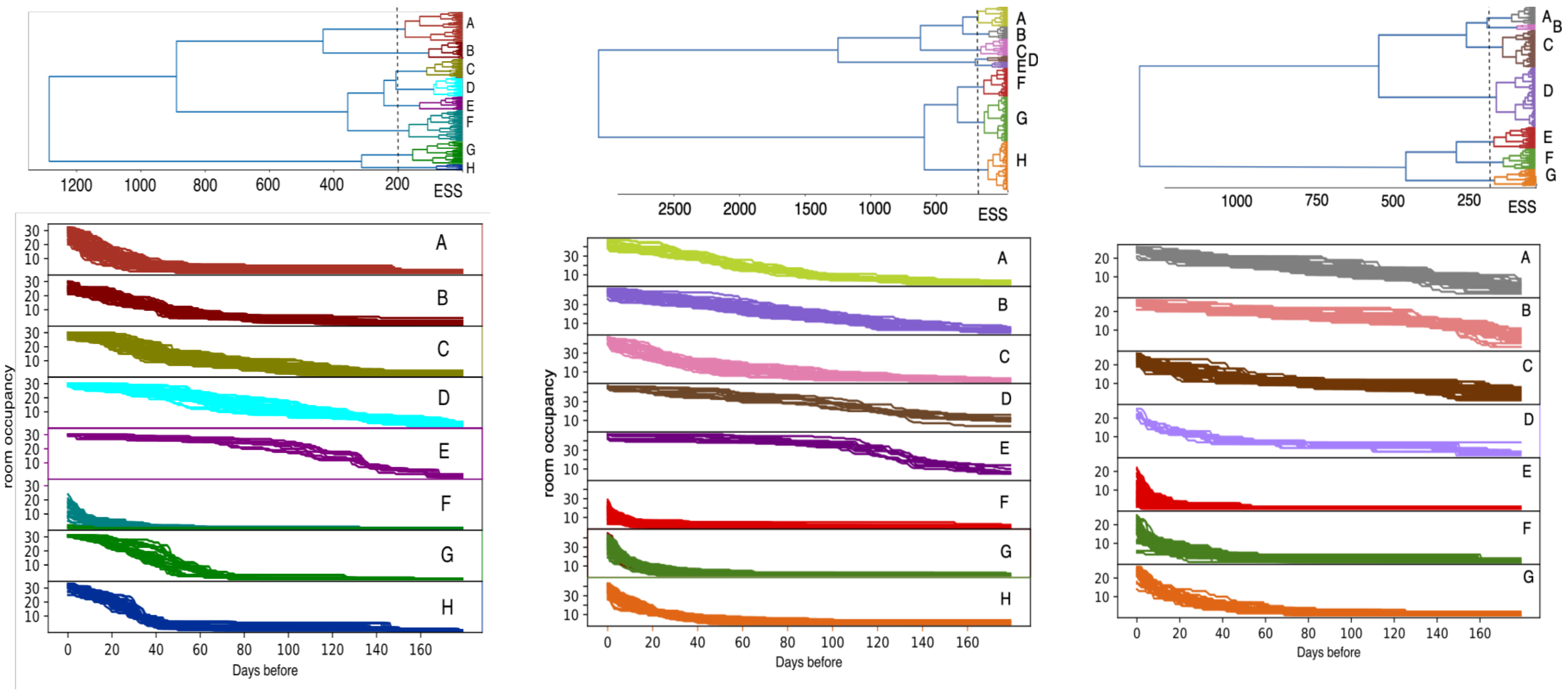

Figure 2: Calendars based on clusters (Hotel 1)

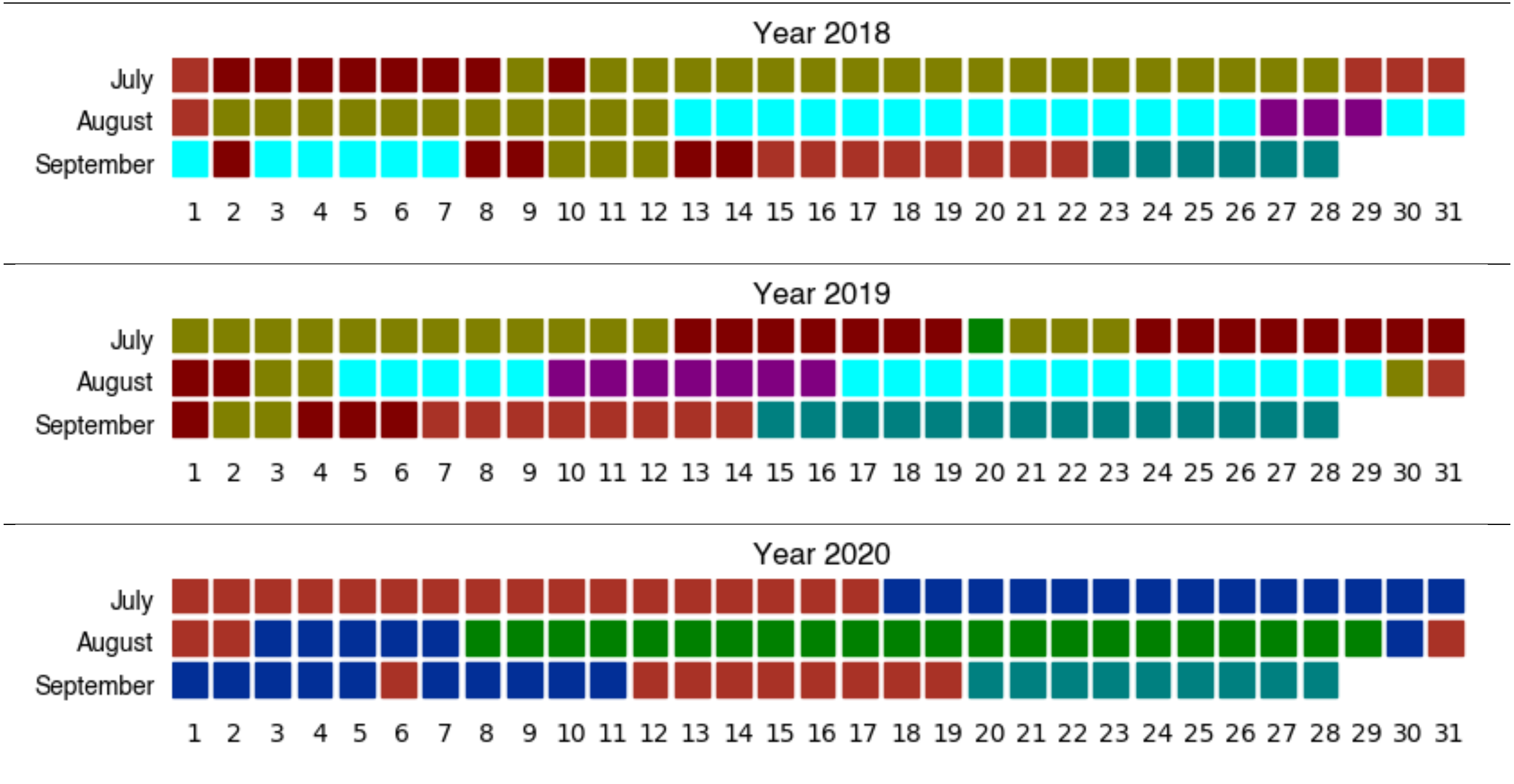

Figure 3: Time correlations

Hotel 1	Hotel 2	Hotel 3