International Journal of Forecasting 37 (2021) 1049–1060

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Big data from dynamic pricing: A smart approach to tourism demand forecasting

Andrea Guizzardi [a,*], Flavio Maria Emanuele Pons [b], Giovanni Angelini [c], Ercolino Ranieri [d]

[a] Department of Statistical Science, University of Bologna, Via delle Belle Arti 41, 40126, Bologna, Italy
[b] LSCE-IPSL, CEA Saclay l'Orme des Merisiers, 91191 Gif-sur-Yvette, France
[c] Department of Economics, University of Bologna, Piazza Scaravilli 1, 40126, Bologna, Italy
[d] Phi Global Group, Via A. Gramsci 79 66016 Guardiagrele, Chieti, Italy

ARTICLE INFO

Keywords:
Regional forecasting
Daily forecasting
Leading indicator
Advance booking
Dynamic pricing
Hotelier's expectations about tourism demand

ABSTRACT

Suppliers of tourist services continuously generate big data on ask prices. We suggest using this information, in the form of a price index, to forecast the occupation rates for virtually any time-space frame, provided that there are a sufficient number of decision makers "sharing" their pricing strategies on the web. Our approach guarantees great transparency and replicability, as big data from OTAs do not depend on search interfaces and can facilitate intelligent interactions between the territory and its inhabitants, thus providing a starting point for a smart decision-making process. We show that it is possible to obtain a noticeable increase in the forecasting performance by including the proposed leading indicator (price index) into the set of explanatory variables, even with very simple model specifications. Our findings offer a new research direction in the field of tourism demand forecasting leveraging on big data from the supply side.

1. Introduction

The task of demand forecasting is often based on dynamic statistical models – mainly involving time series analysis – or model-free machine learning techniques. Such models present one main drawback that limit their application in long-term forecasting: the need to identify informative variables at very large lags or to accept increasingly wider confidence intervals in the case of dynamic forecasting. Predicting the sectoral demand is not simple even in the short term, when small area or high-frequency details are requested. In fact, official statistics are not usually published with high time-space detail or – when available – they are neither complete nor timely.

In this paper, we explore the possibility to exploit public big data to partially overcome these problems in all those industries where online dynamic pricing is spreading as a marketing-management practice. In particular, we focus on the accommodation sector, as it is important for the economies of many locations worldwide and above all, characterized by a huge number of decision makers interested in making daily forecasts about arrivals for circumscribed periods like summer, weekends, or days when fairs or other relevant events are held.

Buono et al. (2017) highlight the potential of online price data in nowcasting and forecasting inflation, though this kind of big data has not been extensively studied in macroeconomics. In the field of tourism, according to Li et al. (2018), big data are drawn from three primary sources: social media, devices, and operations. They originate mainly from the behavior of customers (even potential), while online sales prices are among the few operations generated by the supply side. Surprisingly, big data analytics have not yet considered hotels' pricing

* Corresponding author.
E-mail addresses: andrea.guizzardi@unibo.it (A. Guizzardi), flavio.pons@lsce.ipsl.fr (F.M.E. Pons), g.angelini@unibo.it (G. Angelini), ercolino.ranieri@phiglobalgroup.com (E. Ranieri).

https://doi.org/10.1016/j.ijforecast.2020.11.006

strategies – i.e., the whole set of prices that a hotel publishes on an Online Travel Agency (OTA) at different advance bookings. As we further argue, these data represent one of the most effective examples of shared knowledge regarding businesses' expectations about future demand, allowing for high-frequency prediction, even in a small geographic area.

In fact, given the strong connection between accurate forecasting of guest arrivals (occupancy rates) and successful pricing (Weatherford & Kimes, 2003), we argue that hoteliers calibrate and apply a subjective inverse demand function when they price rooms across the advance booking period. The shape of this inverse demand function is quite complex, as prices in the booking window (for a given target date) depend on both the observable (i.e., hotel's past and actual occupancy rates or seasonality/special events) and the unobservable (i.e., hotelier's expectations about future demand). Monitoring price dynamics is therefore akin to conduct a continuous survey on corporate sentiment; like thousands of virtual questionnaires revealing managers' forecasts for occupancy rates in the advance booking window, conditioned by their observations on their own past and present reservations.

We suggest to make use of a synthesis of these supply-side big data to estimate an aggregate demand function. As our focus is on forecasting, we constrain the function to employ only a set of predetermined information, e.g., deterministic variables, in addition to prices. This approach allows to make static forecasts of the daily occupancy rates, even in small areas (municipalities or neighborhoods), at – virtually – any horizon, provided that the number of decision makers who share their pricing strategies on the web is sufficient.

We believe that our approach has great potential in research on tourism demand forecasting. It can be easily generalized to all the empirical settings covered by at least one OTA; it is operable (and effective) with linear statistical models, even though we show that a more general parametrization which admits nonlinearity yields better results. In addition, it guarantees transparency and replicability, as big data from OTAs are raw data less dependent on changes in user behavior and search interfaces when compared to the big data from search engines (Lazer et al., 2014). Finally, even if we focus only on the hotel segment, a similar methodology could be employed using prices scraped from platforms promoting private accommodations.

Our findings highlight the value of online prices as publicly available and reliable data, which can link the stakeholders of tourism services to the digital environment. The remainder of the paper is organized as follows: Section 2 provides a review of the existing literature on forecasting and big data; in Section 3, we describe the dataset and the modelling/evaluation framework; and in Section 4, we present the estimation and model evaluation results. In Section 5, we discuss our findings, linking them to managerial implications that show how our approach can facilitate intelligent interactions between a geographic area and its inhabitants by providing a "smart" place to start the decision process.

2. Literature review

Predicting day-to-day tourist numbers is a very important task, as it enables public administrators to better manage both the issues of crowding/tourism sustainability (Cheung & Li, 2019) and the net economic contribution that tourists provide local businesses (Voltes-Dorta et al., 2014). These issues are increasingly important in mature destinations in light of growing anti-tourism movements (Seraphin et al., 2018) and serious space problems between visitors and residents that are also on the rise in developing countries (Huang et al., 2017a). The transient visitor population creates a surge in demand for assets that tourists share with residents, which also implies the nonsustainability of tourism on an economic level. Looking at the smallest and largest municipalities in Spain, Voltes-Dorta et al. (2014) show that revenue generated by tourism does not usually match the increased expenditures required to service the floating visitor population.

In agreement with Sánchez-Galiano et al. (2017), we believe that the lack of specific data on the transient population is one of the main factors preventing local administrators from efficiently managing public services and residents' attitudes toward tourism. An accurate and timeliness prediction of seasonal (daily) peaks would improve the scenario for sustainable development, while also limiting the negative effects of overcrowding (Martínez-Garcia et al., 2017).

In European countries, three main official investigations cover the monitoring system of tourist flows (UNWTO, 2010). These show limits in terms of completeness, timeliness, and accuracy (Guizzardi & Bernini, 2012), as they do not provide data that are actually useful to the authorities in managing local tourism (i.e., prompt submonthly data at municipal level). This information gap led many researchers to look at indirect indicators of tourism demand, and/or search for the digital tracks left by tourists and tour operators in the form of big data (Liu et al., 2018). Regarding indicators, Tang et al. (2017) demonstrate the value of using readily available OECD economic climate indicators to estimate hotel occupancy trends in Hong Kong, while Guizzardi and Stacchini (2015) show the value of business sentiment indicators to improve forecasting performance on tourism flows.

However, big data analytics is certainly the approach that offers greater opportunities to scholars (Buhalis & Law, 2008). Big data support the decision-making process, allow for a better understanding of many tourism issues (Xiang et al., 2015), are inexpensive (compared to consultant inputs), very territorially detailed and without the limitations in terms of sample size and timeliness shown by official information. The tourism literature focuses on three primary big data sources (Li et al., 2018): user-generated content (social media), devices (GPS, mobile roaming, Bluetooth, etc.), and operations (web searching and other transaction data).

Data generated by social media provide details about individual experiences and expressions with time-stamped, demographic, and evidence-based insights (Yang et al., 2015). Social media activity largely relies on geo-tagged data sources, microblogging services, and review platforms. Among others, Giglio et al. (2019) used photographs uploaded on Flickr to determine the attractiveness of various tourism sites, while Miah et al. (2017) apply the same data to decision support. Chua et al. (2016) and Brandt et al. (2017) used Twitter posts to assess users' mobility patterns on a regional scale, while Xiang et al. (2017) focus on review platforms, that show discrepancies in the representation of hotel products.

GPS, Wi-Fi Bluetooth, and beacon functionalities have been employed to track tourist movements, that provide considerable spatio-temporal big data (Hardy et al., 2019). Automatic weather station sensors have helped to collect a rich mine of meteorological data (Guo, 2016), which in turn can be exploited to gain knowledge about tourist arrivals in locations that offer weather-sensitive activities. For example, Falk (2010) investigates the relationship between the number of overnight stays and the snow depth at 28 Austrian ski resorts.

Web search data were mainly used as exogenous information to improve accuracy in the prediction of tourist flows and hotel demand (Li et al., 2018). Choi and Varian (2012)'s seminal work improved the forecasting accuracy of ARIMA models using travel-related Google search data. Pan et al. (2012) obtain the same results with multivariate ARMA models, while Bangwayo-Skeete and Skeete (2015) implement an AR-MIDAS regression. Rivera (2016) use Google Trends data in a dynamic linear model to forecast arrivals to Puerto Rico, while Gunter and Onder (2016) are among the few authors who focus on a micro area (the city of Vienna). A more recent strand of literature focuses on the most effective ways to include web search data in an econometric model to limit overfitting and multicollinearity issues. We recognize three main statistical approaches: the principal component analysis (Li et al., 2015), data shift, and combination of different types of search query data, (Yang et al., 2015), and the generalized dynamic factor models (Li et al., 2017).

Other transaction data have been introduced into tourism research in very few related articles, possibly because most transaction data are difficult to obtain due to privacy concerns (Li et al., 2018). Among them, Zervas et al. (2017) show the effect of Airbnb on the hotel market, while Saito et al. (2016) analyze visitors' choices by using online booking data from four major hotels near the Kyoto station. Sobolevsky et al. (2014) demonstrated the applicability of bank card transactions in analyzing tourist mobility patterns, while Huang et al. (2017b) measure the carbon emissions of self-driving tourism and the spatial relationship with scenic spots, using an ArcGIS-Network analysis. Even though we only cite a few examples, it is important to emphasize that few papers consider a daily time dimension (e.g., Brandt et al., 2017).

Analytics of the pricing behavior of hotels, as it appears on the OTAs, are expected to be informative about daily tourism demand. In fact, these supply-side big data directly reflect expected occupancy rates as managers' forecast of occupancy rates is one of the major inputs for most revenue management systems producing recommendations about pricing and availability (Tang et al., 2017). More complex approaches also consider customers' price sensitivity and cancellation probabilities (Antonio et al., 2019; Talluri & Van Ryzin, 2004) or variables reflecting subjective reactions to unexpected events by customers and hoteliers (Yang et al., 2014). Overall, scholars report a strong correlation between good forecasting of occupancy rates and (successful) pricing along the advance booking period (Koupriouchina et al., 2014; Tse & Poon, 2015). On the basis of these findings, it seems possible to employ the information provided by revenue managers through the OTAs to improve the forecasting accuracy of models for daily occupancy rates, even for small areas such as municipalities or neighborhoods, provided that the number of decision makers sharing their pricing strategies on the network reaches a minimum level.

To the best of our knowledge, only a very recent paper by Tsang and Benoit (2020) considers the pricing behavior of hotels to forecast daily tourism demand at a city level with a time series approach. They exploit Gaussian processes and machine learning algorithms, with orthogonalized variables as regressors. We show that our leading indicator approach is worth as it does not require multistep dynamic forecast to predict, is more robust to missing data bias, and allows to obtain good forecasting performances.

3. Data and methods

To explain the variability observed in daily demand, we construct a dataset from the best available rates (BARs) published on Expedia.com every day at 00:00 AM for a one night stay (one adult). We consider a panel of 107 hotels in Milan, over a time interval of 274 days, from January 1st to September 30th, 2016. In the following, arrival dates are denoted by an index t. For each arrival date t, 29 BARs were recorded, corresponding to the lowest offered price along a four-week advance booking period (i.e., from 28 to 0 days). The booking process was simulated from December 4th, 2015, collecting 893,664 BARs observations. The data source is Rate Tiger, a market intelligence service, which monitored the pricing activity of the selected hotels (on demand and for a fee).

As a proxy of the realized tourist demand in the hotel sector, we consider the time series of daily average occupancy rates for hotels in the city of Milan over the same period provided by the STR Share Center. Milan is a business destination in Northern Italy which, in 2017, registered 10.1 million overnight stays in the 27,519 rooms offered by 467 hotel structures (Municipality of Milan, 2019). It represents an optimal empirical setting for our purposes, given both the high hotel market share (86%) and a noticeable daily seasonality. Thus, we include dummy variables to control for the arrival day of the week, bank holidays, and a selection of fairs and events between late winter and spring, (the periods in which Milan experiences the highest occupancy rates and price changes).

As we aim to assess whether the inclusion of advanced booking prices for a set of hotels in a destination improves the prediction of the occupation rate for the same destination, the choice of the advance booking lag is a core aspect. *A priori*, we expect a trade-off between different benefits provided by large and small advance booking horizons. The former provides policy makers with a larger degree of freedom, for example, more time to manage resources to avoid a surplus or deficit of the assets that tourists share with residents. On the contrary, predictions leveraging the last-minute pricing behavior should be more accurate in light of both the progressive reduction of uncertainty for new bookings/cancellations and the increasing importance of pricing tactics over strategies (Guizzardi et al., 2017). To operationalize the choice, we should also consider that, for our purpose, any single advance booking is suboptimal respect to a price index. In fact, particularly with events inducing peaks of demand, we observed that many hotels decide to temporarily avoid offering rooms on the OTA distribution channels: in this case, the data scraping process provides many failures producing inconsistent sample sizes at certain advance bookings. Moreover, online dynamic pricing is associated with irregular price variability over the booking horizon (Guizzardi et al., 2017). For this reason, we suggest considering a price index obtained as a moving average of the BARs posted over a week. In particular, we consider four weekly non-overlapping windows along the advance booking as we aim to measure the extent that forecasting performance changes using price indices calculated on different advance booking windows.

More formally, we denote weeks along the observed advance booking with an index $i = 1, 2, 3, 4$ starting from the day before the arrival date and counting backwards. The advance booking price index $P_{i,t}$ is a moving average of all the prices, for a stay on day t, posted online by the $N_a = 107$ hotels from day $(t - 7i)$to day $[t - 7(i - 1) - 1]$. Then each advance booking price index $P_{i,t}$ is computed as follows:

$$P_{i,t} = \frac{1}{7\mathrm{N}_a} \sum_{j=7(i-1)+1}^{7i} \sum_{a=1}^{N_a} p_{j,t,a} \tag{1}$$

where $p_{j,t,a}$ is the BAR posted on Expedia by the hotel a, for the arrival date t on the day t-j and N is the number of considered BARs. We highlight that, chosen a week i, it is possible to calculate the index $P_{i,t}$ at time $[t - 7(i - 1) - 1]$. Moving the time reference forward to the present (t), we can use the price index in Eq. (1) as a leading indicator to forecast the occupancy rate at time $[t + 7(i - 1) + 1]$. Note that, when more than the 33% of the expected $p_{j,t,a}$ are missing – at least in one of the four weeks – we do not calculate the index, and the days corresponding to these values are excluded from the sample which, consequently, reduces from 274 to 262 observations.

The statistics in Table 1 show that both mean level and price variability increase with the approaching arrival date. This reflects a typical pricing behavior (Mauri, 2013) aimed at hindering cancel-and-rebook strategies, while allowing for increasing discounts/surcharges of the offered BAR as the target date approaches. BARs variability is in fact being increasingly used as a tool to control customers' Internal Reference Price (Choi & Mattila, 2018) and booking propensity or, ultimately, help managers to meet their goals in terms of the occupancy rate.

The sampled hotels are primarily 4-star establishments. The 54% are independent, while the remaining 46% are affiliated with a chain or franchise. They are mainly located in the city center (75% are less than 3 km from the city center) and they are mostly business hotels specialized in hosting MICE events. The hotels have an average of 110 rooms, much higher than both the Milan and the Italian national average. The *fairs* dummy variable has a value of 1 for the most visited fairs/events (detail available upon request).

3.1. Modelling and evaluating framework

In this section, we discuss how to employ a price index $P_{i,t}$ $i = 1, 2, 3, 4$ as a leading indicator to forecast the aggregate demand function for day t. To keep the approach relatively inexpensive and easy to use for policy makers, we limit exogenous variables – other than prices – to deterministic variables that can be constructed with a simple information retrieval (such as noting bank holidays and choosing a set of possibly relevant events in the chosen location, such as the fairs in our case). We also avoid using month dummies to model seasonality, as our sample is shorter than one year, making monthly statistics inconsistent.

Our main goal is to provide an effective forecasting of the daily occupancy rate over a relatively long (i.e., a few weeks) time horizon, without the need to collect every day the actual occupancy rate, as it would be required in a dynamic forecasting framework. However, for the sake of completeness, we also investigate the performance of dynamic time series models, with and without exogeneous regressors. In the following two subsections, we first describe leading indicator regression models – not relying on lags of neither the dependent nor the independent variable – and then we consider a dynamic forecasting experiment using linear time series models.

3.1.1. Leading indicator models

Let Occ_t denote the average occupancy rate in Milan at day t, *trend* a linear trend represented by a vector of indices for the days in the sample (taking value from 1 to 262), Dow_t a set of 6 dummies assigning the day of the week corresponding to date t (excluding Sundays), hol_t a dummy variable indexing, the arrival dates of which are Bank holidays and $fair_t$ a dummy variable indicating whether an important fair is taking place on the arrival date t. We consider three model specifications, *M1*, *M2*, *and M3*

$$\begin{aligned} M1\colon Occ_t &= \alpha + \beta_1 trend + \boldsymbol{\beta}_2 Dow_t \\ &\quad + \beta_3 hol_t + \beta_4 fair_t + \varepsilon_t \\ M2_i\colon Occ_t &= \alpha + \gamma \log\left(P_{i,t}\right) + \beta_1 trend + \boldsymbol{\beta}_2 Dow_t \\ &\quad + \beta_3 hol_t + \beta_4 fair_t + \varepsilon_t \quad i = 1, 2, 3, 4 \\ M3_i\colon Occ_t &= \alpha + s\left(P_{i,t}\right) + \beta_1 trend + \boldsymbol{\beta}_2 Dow_t \\ &\quad + \beta_3 hol_t + \beta_4 fair_t + \varepsilon_t \quad i = 1, 2, 3, 4 \end{aligned}$$

where $P_{i,t}$, denotes the price index in Eq. (1) calculated averaging the prices posted during the ith week of advance booking for a stay in day t. Thus $M2_i$ and $M3_i$ exist in four

Table 1
Descriptive statistics.

Continuous/discrete variables	Median	Mean	st. dev.		interq. range	
Occupation Rate (%)	65.3	65.2	16.0		(51.4-77.9)	
Average BARs 1 week adv. book. ($P_{1,t}$)	115.0	139.0	114.1		(89.0,153.2)	
Average BARs 2 week adv. book. ($P_{2,t}$)	113.4	137.4	106.5		(89.0,153.0)	
Average BARs 3 week adv. book. ($P_{3,t}$)	110.5	131.8	95.1		(85.3,149.0)	
Average BARs 4 week adv. book. ($P_{4,t}$)	110.0	129.7	88.0		(86.5,145.1)	
# rooms (*nrooms*)	89	110	65.2		(65, 143)	
# meeting rooms (*nmr)*	2	3.3	4.3		(0, 5)	
# restaurant seats (*nrs)*	0	56.2	72.2		(0, 100)	
Km from city center (*dist*)	1	2.6	4.9		(0, 3)	
Km from airport (*dista*)	6	8.9	11.1		(0, 10)	
Dummy var. (time related)	Frequency	Other descriptive statistics				
If Bank holiday (*hol*) = 1	5%	# hotel		107	3 stars	13%
If Main Fairs/Events (*fairs*) = 1	16%	Chain or franchise hotel		46%	4 stars	86%
		Independent		54%	5 stars	1%

versions each, which differ because they use as explicative variable a price index computed considering different advance bookings. This way, we point to separately assess possible differential effects of longer/shorter horizons in the prediction of demand.

M1 considers only seasonal and calendar effects (day of week, holidays, fairs), while *M2* and *M3* incorporate information about prices posted in the advance booking window *i*. In particular, *M1* and *M2* are standard linear regression models, and the parameters can then be estimated using ordinary least squares (OLS). The choice of a nonlinear transformation follows the inspection of the scatterplots of Occ_t against $P_{i,t}$ shown in Fig. 1. The logarithm is a common transformation applied to prices in economic and econometric modelling, but in this case it is not optimal as the residuals of models *M2* are not normal and heteroskedastic, which underscores the idea that a more complex specification could provide better results.

Finally, we specify *M3* as a Generalized Additive Model (GAM, see Hastie, 2017 for a comprehensive description of this class of models), including the same explanatory variables as in *M2.* Formally, the structure of a GAM is akin to an ordinary linear regression. However, the relationship between the response variable and the covariates can be more complex than linear, or linear in a simple transformation, such as the natural logarithm of the price index in *M1* and *M2.* The term $s(P_{i,T})$ in the expression of *M3* denotes a nonlinear smooth function of the regressor, represented by a penalized regression spline, whose estimation is carried out through restricted maximum likelihood (Wood & Wood, 2015). The spline transformation of the independent variable allows for a much more flexible modelling as compared to traditional and even generalized linear models. As a result, GAMs allow us to take care of the effects of (possibly multiple) seasonality, non-Gaussianity and nonlinearity between dependent and independent variables, in a setting characterized by a small number of parameters. In practice, we use the standard R function lm() to obtain estimates of the parameters of *M1* and *M2* through OLS , and the gam() function from the R package mcgv (Wood & Wood, 2015) for the smoothing parameter in *M3.*

Once we have estimated the models *M1*, $M2_i$, and $M3_i$, a (static) forecast on the horizon *h* can be easily computed as follows:

$$
\begin{aligned}
M1\colon \widehat{Occ}_{t+h} &= \hat{\alpha} + \widehat{\beta}_1 trend + \widehat{\boldsymbol{\beta}}_2 Dow_{t+h} \\
&\quad + \widehat{\beta}_3 hol_{t+h} + \widehat{\beta}_4 fair_{t+h} \\
M2_i\colon \widehat{Occ}_{t+h} &= \hat{\alpha} + \hat{\gamma}\, log\left(P_{i,T+h}\right) + \widehat{\beta}_1 trend \\
&\quad + \widehat{\boldsymbol{\beta}}_2 Dow_{t+h} + \widehat{\beta}_3 hol_{t+h} + \widehat{\beta}_4 fair_{t+h} \\
&\quad i = 1, 2, 3, 4 \\
M3_i\colon \widehat{Occ}_{t+h} &= \hat{\alpha} + \hat{s}\left(P_{i,T+h}\right) + \widehat{\beta}_1 trend \\
&\quad + \widehat{\boldsymbol{\beta}}_2 Dow_{t+h} + \widehat{\beta}_3 t + \widehat{\beta}_4 fair_{t+h} \\
&\quad i = 1, 2, 3, 4
\end{aligned}
$$

We note that specification *M1* allows to perform forecasts of occupancy rates without limits in the value of *h*, as it considers only calendar and deterministic variables (provided the dummy for fairs is known up to time *t+h*). The only limitation to the use $M2_i$ and $M3_i$ specification for forecasting purposes is that the horizon *h* has to be less or equal than $[7\,(i-1)+1]$ with $(i \geq 1)$ the maximum number of advance booking weeks available. This horizon limitation does not apply to pure time series models, as showed in Section 3.1.2.

We assess and compare the predictive performance of each model using a Monte Carlo (or repeated random subsampling) cross-validation. We perform K = 3000 repetitions; each time, we randomly draw a test sample with size T_1 estimating model parameters on the remaining T-T_1 observations. We perform the cross-validation twice, for $T_1 = 30$ and $T_1 = 50$. Then, we predict $\widehat{Occ}_{t,Mj_i}$ on the test sample and we compute the forecasting error of the *j*-th model $e_{t,Mj_i} = Occ_t - \widehat{Occ}_{t,Mj_i}$. Then, we calculate the Mean Absolute Error: $MAE_{Mj_i} = \frac{\sum_{t=1}^{T_1}\left|e_{t,Mj_i}\right|}{T_1}$ and the Mean Squared Error: $MSE_{Mj_i} = \frac{\sum_{t=1}^{T_1}\left(e_{t,Mj_i}\right)^2}{T_1}$ which weights the largest forecasting errors, in absolute value, more than proportionally (Clark & McCracken, 2013). We finally take into account the fact that occupation rate could display negatively skewed distributions through the Mean Absolute Percentage loss function: $MAPE_{Mj_i} = \frac{\sum_{t=1}^{T_1}\left|e_{t,Mj_i}\right|/Occ_t}{T_1}$ 100. MAPE weights the errors more in correspondence

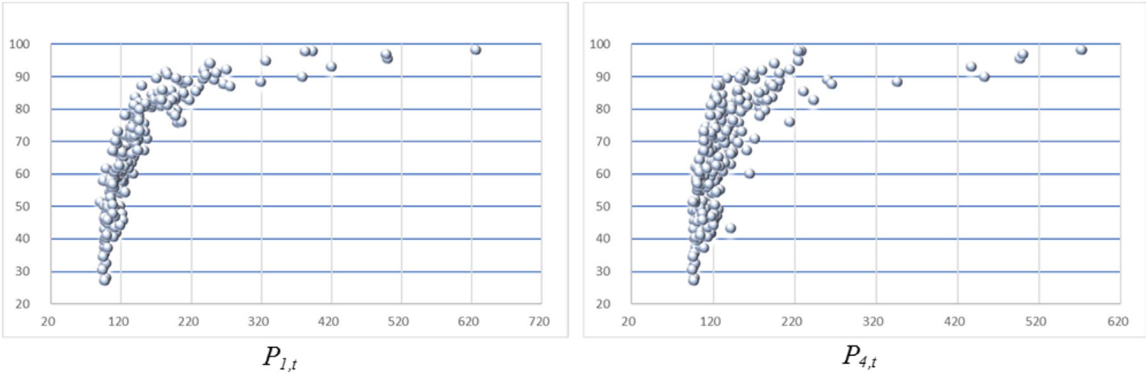

Fig. 1. Nonlinear relation between average occupancy rate in Milan at day t (y axis) and the price index, calculated considering short ($i = 1$) and long ($i = 4$) advance booking.

with low values for the observed occupancy rate, implying that correctly predicting the lower occupancy rate is equally important as predicting higher values.

As we simulate K=3000 random test samples, we are able to draw the distributions of the considered loss functions. We rank rival models using the expected value, identifying the "best predictor" with the one associated with the lowest. Looking at the whole distribution (shape and the size of the overlapping areas – if any), we are also able to assess the relative strength or weakness in forecasting extreme values. Finally, we more formally assess, the significance of differences in the expected performance between two specifications with a Kolmogorov–Smirnov test comparing the cumulative distributions of the loss functions associated to two rival models by means of the supremum function *sup* of their distance $D_{1,2} = sup_x|F_1(x) - F_2(x)|$.

3.1.2. Dynamic models

Time series modelling has been popular in tourism demand forecasting for a long time. In a recent review, Song et al. (2019) summarize 211 key studies in the field of tourism economics. Seasonal Auto-Regressive Moving-Average (SARIMA) models have been employed since the 1970s (Geurts & Ibrahim, 1975) until the present to forecast tourism expenditure, revenue, and particularly in the most recent decades, demand. It is worth mentioning that the majority of these studies rely on monthly or quarterly data, as opposed to the daily frequency in our case.

Building on this literature, we test a set of different SARIMA specifications. Giving a comprehensive review of these models is beyond the scope of this article, particularly because they have been a classic forecasting tool for a long time; for technical details we refer to Chapter 6 in Brockwell and Davis (2016).

In brief, we say that Occ_t follows a SARIMA(p,d,q)(P,D,Q)s, if:

$$(1-B)^d\left(1-B^s\right)^D Occ_t = \sum_{i=1}^{p}\phi_i Occ_{t-i} + \sum_{j=1}^{P}\Phi_i Occ_{t-js} + \sum_{h=1}^{q}\theta_h\varepsilon_{t-h} + \sum_{k=1}^{Q}\Theta_k\varepsilon_{t-ks}$$

where B is the backshift operator, such that $B^k Occ_t = Occ_{t-k}$, d is the level of integration, D the level of seasonal integration (respectively, the number of nonseasonal and seasonal differences to take to obtain stationarity), s the period of the seasonal cycle and ε_t a sequence of independent and identically distributed Gaussian random variables. In our case, $d = 0$, while we admit that $D = 1$ and $s = 7$ days. The two sums including ϕ and θ coefficients represent the autoregressive and moving average part of the process, respectively; the terms involving Φ and Θ introduce analogous terms for the seasonal component.

4. Results

4.1. Leading indicator models

In Table 2, we display the results of the parameter estimation of all models, using the entire sample of 262 days. The price index is always significant ($\alpha = 0.05$) and clearly increases the goodness of fit of models $M2$ and $M3$, with values of the adjusted R2 all higher than $\sim$0.7 as opposed to the value of 0.56 associated with M1. As expected, for both $M2_i$ and $M3_i$ the explanatory power of the index decreases for increasing i. However, the model fit is better than for M1 even with a price index computed at the fourth week of advance booking.

The coefficients of the leading indicator $P_{i,t}$ in models $M2_i$ are always positive, and the spline term *s()* resulting from the estimation of $M3_i$ also assumes positive values. In both cases, price "elasticities" increase at decreasing i, pointing to a stronger association between pricing and the realized occupancy rate close to the last minute. This result is expected and reflects an increasing capability of decision makers to correctly predict the occupancy rate at very short advance bookings, when the probability of unexpected cancellations or new reservations decreases. Symmetrically, the coefficients of the other explanatory variables tend to increase in absolute value at larger advance bookings, while the variable Fair is never significant when prices are considered.

Overall, the above results imply that, over short time horizons (up to one month), prices posted by decision

Table 2
Model estimates and goodness of fit.

Estimates (%)									
	$M1$	$M2_1$	$M2_2$	$M2_3$	$M2_4$	$M3_1$	$M3_2$	$M3_3$	$M3_4$
Constant	49.16**	−112.71**	−104.51**	−100.51**	−96.65**	63.90**	61.15**	62.00**	57.72**
trend	0.02**	0.014**	0.019**	0.020**	0.031**	0.011**	0.014**	0.010**	0.024**
Monday	14.49**	6.75**	8.72**	9.4**	9.89**		3.56**	3.41**	7.1**
Tuesday	18.88**	8.4**	11.76**	13.17**	13.54**	1.99*	6.89**	7.49**	10.91**
Wednesday	18.12**	8.13**	11.54**	12.78**	13.44**	1.89*	6.37**	7.22**	10.93**
Thursday	13.68**	5.97**	7.98**	8.79**	9.29**		2.89**	3.13**	6.55**
Friday	6.87**	2.92*	3.34*	3.88*	4.39**				3.77**
Saturday	8.79**	4.14**	5.04*	5.71**	6.33**		2.96**	3.4**	6.09**
holy	−16.06**	−11.74**	−13.37**	−14.07**	−15.4**	−8.06**	−10.48**	-11.11**	−13.73**
fairs	19.19**								
Price Index $P_{i,t}$		35.09**	33.22**	32.31**	31.26**				
EDF (smooth term)						6.47**	5.88**	5.90*	4.56**
Goodness of fit									
F-test	38.0	108.4	76.8	71.7	66.2				
[p-value]	[<2.2e−16]	[<2.2e−16]	[<2.2e−16]	[<2.2e−16]	[<2.2e−16]				
adj R	0.560	0.787	0.723	0. 709	0.692	0.885	0.811	0.796	0.745
# of. Obs.	262	262	262	262	262	262	262	262	262
Res. 1Q	−0.072	−0.041	−0.052	−0.054	−0.050	−0.030	−0.035	−0.043	−0.051
Res. 3Q	0.072	−0.045	−0.052	0.055	0.053	0.031	0.040	0.040	0.047

** and * Denote statistical significance at 5% and 10% respectively. Estimates reads in %.

makers on the OTAs contain enough information to explain occupation rates even in cases of irregular fluctuation caused by fairs.

However, the differences in the goodness of fit between specifications $M2$ and $M3$ stress the limits of the simple linear approach being able to represent the dynamics of the occupancy rate in regular patterns. In other words, the better result obtained with a GAM approach points to the existence of a marked nonlinearity in the relationship between Occ_t and $P_{i,t}$ that we believe it is worth considering, notwithstanding the increased complexity of the model.

Analogous conclusions follow the assessment of the forecasting performance for all rival models over the K=3000 Monte Carlo simulation – see Table 3 and Fig. 2. As results for $T_1 = 30$ and $T_1 = 50$ are very similar, we only report the case $T_1 = 30$.

M1 shows the lowest performance in terms of all the three considered loss functions, while model $M3_1$ displays both the smallest mean value and variability (see Table 3). The overlapping area between the distributions of each loss function for the two models is only 2% for MAE and MSE and 7% for MAPE. This suggests that the forecasting performance is significantly different between the two models and allows us to conclude that the inclusion of the price index significantly increases the ability to predict unobserved demand as compared to a model that considers only seasonal and calendar effects. The fact that all the specifications including the price index perform better than M1, even with 4 weeks of advance booking, suggests that hotel managers set prices according to expected and desired demand well in advance. In other words, they have rational expectations on occupancy rates at the arrival date, meaning they could be wrong individually but they are not systematically biased over time. For the sake of readability, we have excluded the results related to the week 3 from the graphs in Fig. 2; in fact the Kolmogorv-Smirnov test described in Section 3, shows that results related to week 3 do not differ significantly from the one obtained considering week 4. It is worth noting that all the other distributions are statistically different.

Considering the advance booking effect, the smaller the *i* – i.e. the booking horizon – the better the forecasting performance of the same specification. For increasing *i*, we observe both increasing mean value of the distribution of the loss functions (corresponding to decreasing predictive power of the models) and increasing variability and interquartile range. These results suggest that the managers' forecasting process is efficient in the sense that the lower uncertainty - the closer to the arrival date they are - the more accurate the forecast will be.

The complex relationship between prices and occupancy rates is better caught by the GAM smoothing term in *M3*, which allows for a better prediction of the values far from the mean level, a core aspect of forecasting practices (Zhang et al., 2017). In particular, the GAM model including the information farthest from the arrival date *($M3_4$)* is associated with distributions of MAE and MSE with lower mean values (and lower variability for the MAPE) than $M2_2$, showing that accounting for nonlinearity makes up for the loss of information experienced extending the forecasting horizon of two weeks. Consistently the difference is even more evident for $M3_2$ and $M2_1$, particularly in the tails.

4.2. Dynamic models

The visual analysis of the time plot and of the global and partial autocorrelation functions, suggest that the time series Occ_t is stationary and characterized by a cycle with a period of *7* days. Stationarity is corroborated by the Augmented Dickey-Fuller test, which rejects the null hypothesis of a unit root at the level $\alpha = 0.05$. The test has been augmented, including lags up to *t-4*, which is the autoregressive order suggested by the model selection described below. We also check for possible seasonal unit

Table 3
Loss functions distribution statistics over K=3000 random test samples ($T_1 = 30$).

	$M1$	$M2_1$	$M2_2$	$M2_3$	$M2_4$	$M3_1$	$M3_2$	$M3_3$	$M3_4$
MAE									
Percentile 10%	7.1%	4.9%	5.5%	5.6%	5.6%	3.4%	4.1%	4.6%	5.0%
Median	8.6%	5.9%	6.7%	6.9%	6.9%	4.2%	5.2%	5.6%	6.2%
Mean	8.6%	5.9%	6.7%	6.9%	7.0%	4.2%	5.2%	5.7%	6.2%
Percentile 90%	10.1%	7.1%	8.1%	8.3%	8.3%	5.1%	6.4%	6.8%	7.5%
% overlap with M1	100%	20%	40%	44%	47%	2%	11%	16%	28%
MSE									
Percentile 10%	8.9%	6.1%	6.9%	7.1%	7.2%	4.4%	5.4%	5.7%	6.4%
Median	10.7%	7.5%	8.6%	8.8%	9.0%	5.5%	7.0%	7.2%	8.1%
Mean	10.7%	7.6%	8.6%	8.9%	9.0%	5.5%	7.0%	7.3%	8.1%
Percentile 90%	12.5%	9.1%	10.4%	10.7%	10.8%	6.7%	8.8%	9.1%	10.0%
% overlap with M1	100%	23%	45%	51%	55%	2%	18%	23%	36%
MAPE									
Percentile 10%	11.7%	7.9%	8.9%	9.2%	9.1%	5.6%	6.8%	7.5%	8.3%
Median	14.9%	10.2%	11.6%	11.9%	12.0%	7.4%	9.2%	9.7%	10.9%
Mean	15.1%	10.4%	11.8%	12.1%	12.1%	7.5%	9.4%	9.9%	11.0%
Percentile 90%	18.7%	13.1%	15.0%	15.2%	15.4%	9.5%	12.3%	12.4%	14.0%
% overlap with M1	100%	33%	5.5%	56%	57%	7%	24%	26%	42%

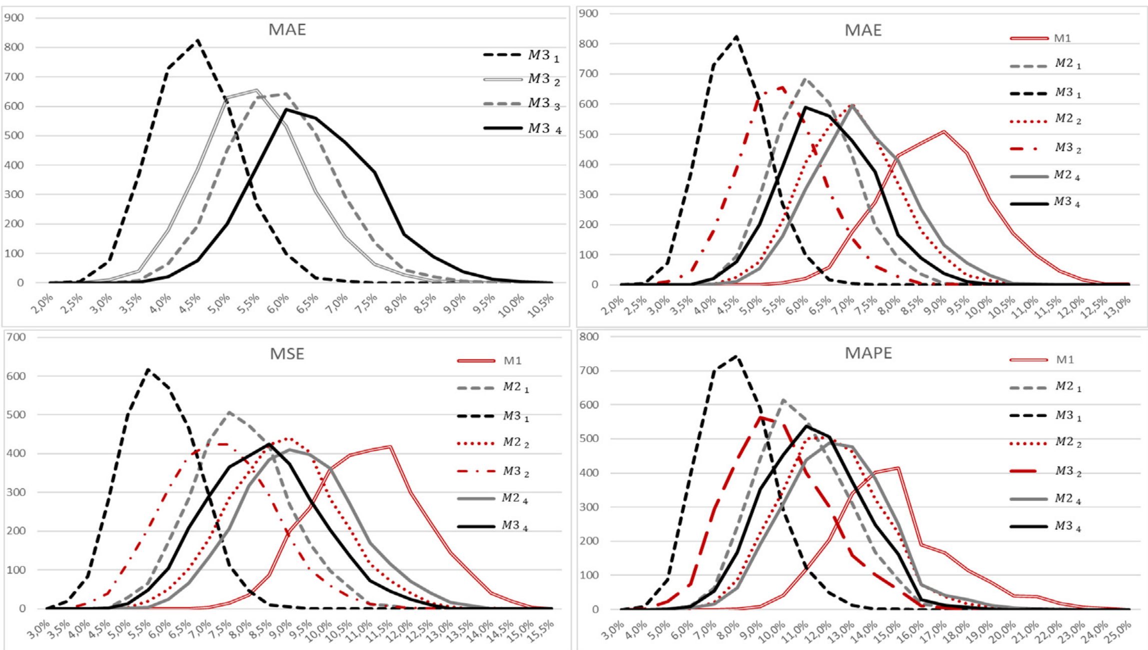

Fig. 2. Estimated probability density functions of the error measures. Upper panels: increasing mean and variability of the MAE for the GAM model (left) and the comparison of the MAE for considered models (right). Lower panels: the comparison of the MSE (left) and MAPE (right) for considered models.

roots using the Hylleberg, Engle, Granger, and Yoo test (Hylleberg et al., 1990); however, this test did not provide a clear result due to numerical problems. We then admit the possibility of seasonal unit root among the considered models, but not of a unit root in the nonseasonal part of the model.

We use the *auto.arima* function from the R package "forecast" (Hyndman & Khandakar, 2007) to determine the optimal orders *p*, *P*, *q*, and *Q*. This function estimates several combinations of these parameters and suggests the best fit based on the Akaike Information Criterion (AIC). The suggested model is a SARIMA(4,0,7)(0,0,0), that is stationary and nonseasonal. We consider a further ensemble of models for Occ_t, including seasonality and exogenous variables:

$$M4\colon Occ_t = ARIMA\,(4,0,7)\,(0,0,0)$$
$$M5\colon Occ_t = ARIMA\,(4,0,7)\,(0,1,0)$$
$$M6\colon Occ_t = ARIMA\,(4,0,7)\,(1,0,1)$$
$$M7\colon Occ_t = ARIMA\,(4,0,7)\,(0,1,0) + \beta_1 trend + \beta_3 hol_t + \beta_4 fair_t$$

Table 4
Loss functions comparison on the last 21 observations (regression and time series models).

	$M1$	$M2_4$	$M3_4$	$M4$	$M5$	$M6$	$M7$	$M8_4$
AIC	1926.1	1763.9	1725.3	1798.0	1664.9	1717.2	1651.2	1542.4
MAE	8.78%	4.85%	2.88%	19.4%	17.5%	21.5%	15.0%	12.0%
MSE	121.0	30.5	14.6	451.5	353.2	536.7	257.9	167.0
MAPE	10.7%	5.9%	3.8%	22.7%	20.8%	25.4%	17.9%	14.3%

$$M8_4{:}\, Occ_t = ARIMA\,(4,0,7)\,(0,0,0) \\ \quad + \beta_1 trend + \beta_3 hol_t + \beta_4 fair_t + \gamma \ log\left(P_{4,t}\right)$$

Time series models allow to calculate dynamic forecasts for any horizon h. However, for reasons of comparison with our linear regression and GAMs, we only consider the maximum possible horizon ($h=21$) fixing $i=4$. As in the case of time series modelling, the chronological order of the data must be respected, we split the available information in an estimation set made by the first 253 daily occupancy rates in our time series leaving the last 21 days for forecasting assessment (test set).

We do not report and comment the estimated coefficients of these additional models, as these are of limited interest in the evaluation of the forecasting performance. Results are reported in Table 4. The best goodness of fit (AIC) is associated with $M8_4$, the seasonal time series model with exogenous covariates. However, when the forecasting is assessed, the augmented linear regression and the GAM model, including the leading indicator turn to be the best choice. This result was expected as with models $M2_4$ and $M3_4$, we are able to perform a static forecast that uses the actual values of the explanatory variable $P_{t,4}$ while time series models use the forecasted values of the dependent variable to perform the 21 step ahead dynamic forecast. It is worth to note that augmenting the specification $M7$ with $P_{t,4}$ (i.e., $M8_4$) produces a decrease in the value of all the three loss functions.

In Fig. 3, we show the observed occupancy rates (black line) as compared to the predictions obtained with $M1$, $M2_4$, $M3_4$, $M4$, and $M8_4$ specifications. For the sake of clarity, we only show the simplest and the most complex time series models, as the performances of the others fall in between. We can notice that models $M1$, $M2_4$, and $M3_4$, provide the visually best forecasting, despite not being dynamical time series models.

While the accuracy of M1 is in line with Monte Carlo simulated results in Table 3, the performances of the models $M2_4$ and $M3_4$ are definitely better. Therefore, the hoteliers' ability to forecast future occupancy rate – and to reflect their prediction in an online price – appears to be much higher than expected.

Specifically, the forecasting starts from the second week of September, a very active period in Milan after a minimum in occupancy rate characterizing the month of August. This is a temporary shift from the long run equilibrium that affects the accuracy of time series models, which relies on dynamic forecasting to span a horizon of 21 days. As it is clear from Fig. 3, after a few days, models M4 and M7 are no longer able to follow the trend. Moreover, they are unable to intercept the impact of occasional public events (Kamola & Arabas, 2020). This aspect highlights another potential problem in evaluating the forecasting results of approaches where the chronological order of the data has to be respected. In fact, the number of demand fluctuations in the chosen test set could condition the assessment, as reported in Tsang and Benoit (2020).

5. Conclusion

The accuracy of traditional time series forecasting methods for the occupancy rate is largely affected by changes in the economic environment or unexpected events (Pan et al., 2012; Yang et al., 2014). To improve forecasting accuracy, Zhang et al. (2017) suggest focusing on irregular fluctuation, converting it into regular patterns – a task that is not particularly easy, especially in a high-frequency setting. Moreover, such an approach requires the continued observation of demand up to the present, to produce a forecast for an established number of steps ahead.

In this paper, we have proposed a completely different approach, presenting a method to forecast tourist demand by improving the set of available information, rather than complicating the modelling strategy. In particular, we show how it is possible to exploit the expectations of hotel managers on occupancy rates made public on OTAs through the prices asked at different advance bookings. In this sense, the diffusion of OTAs has made dynamic pricing strategies, if not transparent, at least understandable through the analysis of data that are public and can be easily collected. This way they become a source of shared knowledge that allows for significantly improved forecasting performance, in particular over small geographic areas and at high frequency (daily). This spatio-temporal scale is, at the same time, the least (realistically not at all) covered by official statistics, and the most important for the majority of tactical decisions about public resource allocation. Thus, it is imaginable that studies concerning other cities and areas – even with a strongly seasonal demand – can take advantage of our methodology by incorporating big data on dynamically pricing it into their forecasting practices.

The core of the paper resides in the use of a leading indicator – constructed using best available rates published on OTAs at different advance bookings by a sample of hotel managers – to predict daily occupancy rates at a fine grain time-space setting. We consider these supply-side big data as originated from virtual daily surveys regarding the sentiment of the manager about occupation rate at a certain arrival date. In other words, prices are treated as a measure of corporate expectations regarding future demand, given the information set and the pricing model.

As a main conclusion, we find that it is possible to obtain a sensible increase in the forecasting performance by including such a price index into the set of explanatory variables, even with very simple model specifications. This implies that hoteliers have rational expectations on occupancy rates at the arrival date or in other words, they could be wrong individually but they are not systematically biased over time.

Our results also confirm the existence of a trade-off between forecasting accuracy and time horizon. A large

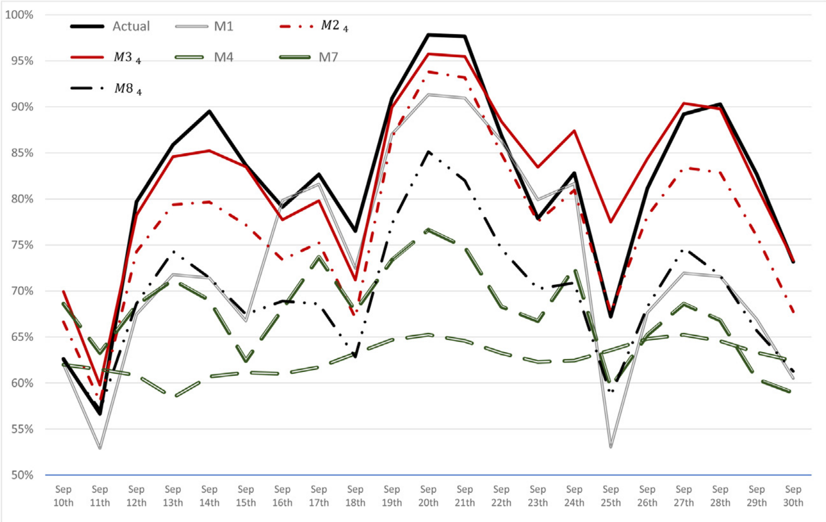

Fig. 3. Out-of-sample actual and forecasted daily occupancy rates.

horizon (i.e., the forecast relies on prices posted with large advance bookings) increases the time span over which stakeholders can take informed decisions. Unfortunately, it also results in larger forecasting errors than using data from the week closest to the arrival date, when uncertainty about new reservations and cancellations is lower. We then conclude that managers use information efficiently, in the sense that the more reliable information they have the closer the prices offered on the OTA reflect their occupation rate. However, we show that the decrease in forecasting accuracy is limited considering forecasting horizons higher than a week, making it possible to inform stakeholders even on broad forecasting horizons. From a revenue management perspective, the above results provide (indirect) evidence that – in Milan – occupancy rate guessing is an important input to the dynamic pricing algorithms at the firm level.

Moreover, we show that the information on managers' expectations contained in advanced prices is more effective as a predictor of occupancy when we introduce some flexibility to the model specification. The best performance is obtained with a GAM model, indicating that the relationship between occupancy and the price index is complex and nonlinear.

Finally, we compare our method to exploit hotel managers' expectations on occupancy rates with a more standard SARIMA approach – assuming we have access to up-to-date sampling of occupancy rates. The forecasting performance of time series models over a horizon of 21 days is overall worse than the one obtained by static regressions augmented with the price index. We also test if the inclusion of the price index as an explicative variable in a SARIMAX frame may be convenient. While the forecasting performance generally improves as compared to time series models not including the leading indicator, these models are still outperformed by the augmented linear and GAMs. This may be due to both a trend shift and local irregularities in the time series of occupancy rates between August and September, a problem that does not affect models that do not rely on lags of the dependent variable.

The possibility of exploiting big data from the supply side to forecast daily tourism demand for limited areas, hints at some managerial implications. In fact, while almost all travel destinations seek to increase tourists, the potential hazards of a massive and uncoordinated influx of tourists to popular destinations worldwide has become a very general problem (Butler, 2017; D'Alisa et al., 2014). An accurate forecasting of daily tourist demand peaks would facilitate short-term operational tactics able to avoid asset shortfalls that tourists share with residents (i.e., transportation, security, water, or urban space), minimizing the impact on the quality of life in the host area and reducing the likelihood of a permanent visitor-resident conflict (Andereck et al., 2005). Symmetrically, the temporary reduction in the volume of services offered can be planned in response to forecasted low daily demand, allowing for a more efficient management of budget and public resources. This makes it possible to think about a sustainable territorial management, while improving residents' attitudes toward tourists.

However, we believe that the impact of our findings is not only a matter of forecasting. This paper suggests that public data from OTAs offer policy makers an alternative to acquire expensive data and/or consultant inputs. As an example, local governments must frequently choose funding allocation for special events, conventions, or initiatives. In these circumstances, a framework of economic and social assessment is rarely performed (Wood, 2005), so there is no systematic and objective manner to determine the extent of support, if any, to be given to alternative events (Dwyer et al., 2000). We demonstrate the merit of online prices as an input for assessing (ex post) the impact of events and conventions. A system able to record and store BARs at small advance booking, for example, could be used to link a measure of tourist demand to any past event even if its impact is assumed to be constrained in time and space. This indirect measure can also be used to integrate official information (when available), as it suffers less from underreport bias, sometimes generated by the opportunity to obtain a tax advantage (Guizzardi & Bernini, 2012).

In summary, we demonstrate that the predictive and descriptive power of big data from the supply side may become a powerful vehicle to link the stakeholders of tourism services with the digital environment. Our methodology could be employed to facilitate intelligent interactions between the city and its inhabitants, which is a critical component of the smart city (Harrison et al., 2010). We transform data from the digital environment into business value-propositions with a clear focus on efficiency, sustainability, and experience enrichment. This would enable the delivery of intelligent tourist services, characterized by intensive information sharing, and value cocreation: in other words, we provide an asset for a smart tourism ecosystem, Gretzel et al. (2015).

A limitation of these findings lies in the choice of a single destination (a business destination). Other destinations may have a smaller number of high category hotels, and public data from OTAs might not be able to realistically predict expected occupancy. Smaller hotels with no revenue management department may rely on fixed prices, or have little experience in dynamic pricing, which would limit the effectiveness of our approach or recommend a nonrandom sampling of the structures active on the OTAs. Furthermore, even though the proposed models are simple and can be estimated in just a few seconds with any desktop or laptop computer, acquiring the price dataset requires setting up an algorithm of scraping, which is not immediate and may involve privacy issues or content licensing that is not yet resolved. Moreover, although we believe the choice of log-linear and GAM regression is sufficient to show the efficacy of our price index in forecasting occupancy rates, a future line of research consists of exploring the potential of more complex model specifications, or even machine learning techniques, to better represent the relation between posted prices and tourist demand. Finally, we did not dig deeper into the possible use of different big data sources in increase forecasting accuracy: for example, search engine queries and website traffic data, social media mentions or mobile phone data, could show synergies with our index that could lead to a further increase in forecasting accuracy.

We agree with Brandt et al. (2017) that a holistic and integrative perspective on how digitization affects all stakeholders in tourism activities is still in its infancy. However, what we suggest in this paper, adds a new direction of research to the use of big data in the field of tourism demand forecasting.

References

Andereck, K. L., Valentine, K. M., Knopf, R. C., & Vogt, C. A. (2005). Residents' perceptions of community tourism impacts. *Annals of Tourism Research*, *32*(4), 1056–1076.

Antonio, N., De Almeida, A., & Nunes, L. (2019). *Big data in hotel revenue management: Exploring cancellation drivers to gain insights into booking cancellation behavior*. Cornell Hospitality Quarterly.

Bangwayo-Skeete, P. F., & Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management*, *46*, 454–464.

Brandt, T., Bendler, J., & Neumann, D. (2017). Social media analytics and value creation in urban smart tourism ecosystems. *Information & Management*, *54*(6), 703–713.

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting*. Springer.

Buhalis, D., & Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the Internet-The state of eTourism research. *Tourism Management*, *29*(4), 609–623.

Buono, D., Mazzi, G. L., Kapetanios, G., Marcellino, M., & Papailias, F. (2017). Big data types for macroeconomic nowcasting. *Eurostat Review on National Accounts and Macroeconomic Indicators*, *1*, 93–145.

Butler, R. W. (2017). *Tourism and resilience*. Wallingford, UK: CABI.

Cheung, K. S., & Li, L. H. (2019). Understanding visitor–resident relations in overtourism: developing resilience for sustainable tourism. *Journal of Sustainable Tourism*, 1–20.

Choi, C., & Mattila, A. S. (2018). The effects of internal and external reference prices on travelers' price evaluations. *Journal of Travel Research*, *57*(8), 1068–1077.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*, 2–9.

Chua, A., Servillo, L., Marcheggiani, E., & Moere, A. V. (2016). Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tourism Management*, *57*, 295–310.

Clark, T., & McCracken, M. (2013). Advances in forecast evaluation. In *Handbook of economic forecasting (Vol. 2)* (pp. 1107–1201). Elsevier.

D'Alisa, G., Demaria, F., & Kallis, G. (2014). *Degrowth: A vocabulary for a new era*. London: Routledge.

Dwyer, L., Mellor, R., Mistilis, N., & Mules, T. (2000). Forecasting the economic impacts of events and conventions. *Event Management*, *6*(3), 191–204.

Falk, M. (2010). A dynamic panel data analysis of snow depth and winter tourism. *Tourism Management*, *31*(6), 912–924.

Geurts, M. D., & Ibrahim, I. (1975). Comparing the Box-Jenkins approach with the exponentially smoothed forecasting model application to Hawaii tourists. *Journal of Marketing Research*, *12*(2), 182–188.

Giglio, S., Bertacchini, F., Bilotta, E., & Pantano, P. (2019). Using social media to identify tourism attractiveness in six Italian cities. *Tourism Management*, *72*, 306–312.

Gretzel, U., Werthner, H., Koo, C., & Lamsfus, C. (2015). Conceptual foundations for understanding smart tourism ecosystems. *Computers in Human Behavior*, *50*, 558–563.

Guizzardi, A., & Bernini, C. (2012). Measuring underreporting in accommodation statistics: Evidence from Italy. *Current Issues in Tourism*, *15*(6), 597–602.

Guizzardi, A., Pons, F. M. E., & Ranieri, E. (2017). Advance booking and hotel price variability online: Any opportunity for business customers? *International Journal of Hospitality Management*, *64*, 85–93.

Guizzardi, A., & Stacchini, A. (2015). Real-time forecasting regional tourism with business sentiment surveys. *Tourism Management*, *47*, 213–223.

Gunter, U., & Onder, I. (2016). Forecasting city arrivals with Google Analytics. *Annals of Tourism Research*, *61*, 199–212.

Guo, X. (2016). Application of meteorological big data. In *2016 16th international symposium on communications and information technologies* (pp. 273–279). IEEE.

Hardy, A., Aryal, J., & Wells, M. P. (2019). Comparing techniques for tracking: the case of Tourism Tracer in Tasmania, Australia. *E-review of Tourism Research*, *16*(2/3), 84–94.

Harrison, C., Eckman, B., Hamilton, R., Hartswick, P., Kalagnanam, J., Paraszczak, & Williams, P. (2010). Foundations for smarter cities. *IBM Journal of Research and Development*, *54*(4), 1–16.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S* (pp. 249–307). Routledge.

Huang, Z., Cao, F., Jin, C., Yu, Z., & Huang, R. (2017a). Carbon emission flow from self driving tours and its spatial relationship with scenic spots. a traffic-related big data method. *Journal of Cleaner Production*, *142*, 946–955.

Huang, X., Zhang, L., & Ding, Y. (2017b). The Baidu Index: Uses in predicting tourism flows–A case study of the Forbidden City. *Tourism Management*, *58*, 301–306.

Hylleberg, S., Engle, R. F., Granger, C. W. J., & Yoo, B. S. (1990). Seasonal integration and cointegration. *J. Econom.*, *44*, 215–238.

Hyndman, R. J., & Khandakar, Y. (2007). *Automatic time series for forecasting: The forecast package for R (No. 6/07)*. Clayton VIC, Australia: Monash University, Department of Econometrics and Business Statistics.

Kamola, M., & Arabas, P. (2020). Improving time-series demand modeling in hospitality business by analytics of public event datasets. *IEEE Access*, *8*, 53666–53677.

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.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), 1203–1205.

Li, G., Law, R., Vu, H. Q., Rong, J., & Zhao, X. R. (2015). Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism Management*, *46*, 311–321.

Li, X., Pan, B., Law, R., & Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism Management*, *59*, 57–66.

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, *68*, 301–323.

Liu, Y. Y., Tseng, F. M., & Tseng, Y. H. (2018). Big data analytics for forecasting tourism destination arrivals with the applied vector autoregression model. *Technological Forecasting and Social Change*, *130*, 123–134.

Martínez-Garcia, E., Raya, J. M., & Majó, J. (2017). Differences in residents' attitudes towards tourism among mass tourism destinations. *International Journal of Tourism Research*, *19*(5), 535–545.

Mauri, A. G. (2013). *Hotel revenue management: Principles and practices*. Pearson Italia Spa.

Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A big data analytics method for tourist behaviour analysis. *Information & Management*, *54*(6), 771–785.

Municipality of Milan (2019). http://sisi.comune.milano.it/. (Accessed 21 March 2019).

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.

Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. *Tourism Management*, *57*, 12–20.

Saito, T., Takahashi, A., & Tsuda, H. (2016). Optimal room charge and expected sales under discrete choice models with limited capacity. *International Journal of Hospitality Management*, *57*, 116–131.

Sánchez-Galiano, J. C., Martí-Ciriquián, P., & Fernández-Aracil, P. (2017). Temporary population estimates of mass tourism destinations: The case of benidorm. *Tourism Management*, *62*, 234–240.

Seraphin, H., Sheeran, P., & Pilato, M. (2018). Over-tourism and the fall of venice as a destination. *Journal of Destination Marketing & Management*, *9*, 374–376.

Sobolevsky, S., Sitko, I., Des Combes, R. T., Hawelka, B., & Arias, J. M. (2014). Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. The case of residents and foreign visitors in Spain. In *2014 IEEE international congress on big data*. Anchorage, USA.

Song, H., Qiu, R. T., & Park, J. (2019). A review of research on tourism demand forecasting. *Annals of Tourism Research*, *75*, 338–362.

Talluri, K. T., & Van Ryzin, G. J. (2004). *The theory and practice of revenue management*. Boston, MA: Springer.

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

Tsang, W. K., & Benoit, D. F. (2020). Gaussian processes for daily demand prediction in tourism planning. *Journal of Forecasting*, *39*(3), 551–568.

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.

UNWTO (2010). *International recommendations for tourism statistics 2008, Studies and methods, Series M No.83/Rev.1*. New York: World Tourism Organization.

Voltes-Dorta, A., Jiménez, J. L., & Suárez-Alemán, A. (2014). An initial investigation into the impact of tourism on local budgets: A comparative analysis of spanish municipalities. *Tourism Management*, *45*, 124–133.

Weatherford, L. R., & Kimes, S. E. (2003). A comparison of forecasting methods for hotel revenue management. *International Journal of Forecasting*, *19*(3), 401–415.

Wood, E. H. (2005). Measuring the economic and social impacts of local authority events. *International Journal of Public Sector Management*, *18*(1), 37–53.

Wood, S., & Wood, M. S. (2015). Package 'mgcv'. *R Package Version*, *1*(29).

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, *58*, 51–65.

Xiang, Z., Schwartz, Z., Gerdes, J. H., Jr., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management*, *44*, 120–130.

Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management*, *46*, 386–397.

Yang, Y., Pan, B., & Song, H. (2014). Predicting hotel demand using destination marketing organization's web traffic data. *Journal of Travel Research*, *53*, 433–447.

Zervas, G., Proserpio, D., & Byers, J. W. (2017). The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. *Journal of Marketing Research*, *54*(5), 687–705.

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