Research Article

Journal of Travel Research
2014, Vol. 53(4) 433–447
© 2013 SAGE Publications
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0047287513500391
jtr.sagepub.com
SAGE

Predicting Hotel Demand Using Destination Marketing Organization's Web Traffic Data

Yang Yang[1], Bing Pan[2], and Haiyan Song[3]

Abstract

This study uses the web traffic volume data of a destination marketing organization (DMO) to predict hotel demand for the destination. The results show a significant improvement in the error reduction of ARMAX models, compared with their ARMA counterparts, for short-run forecasts of room nights sold by incorporating web traffic data as an explanatory variable. These empirical results demonstrate the significant value of website traffic data in predicting demand for hotel rooms at a destination, and potentially even local businesses' future revenue and performance. The implications for future research on using big data for forecasting hotel demand is also discussed.

Keywords

tourism demand forecasting, time series, online data, hotel occupancy, big data, website traffic

Introduction

Forecasting future hotel guest arrivals and occupancy rates is a key aspect of hotel revenue management (Weatherford and Kimes 2003). Accurate forecasting is crucial to enable hoteliers to efficiently allocate hotel resources and refine pricing strategies (Weatherford and Kimes 2003). Traditional forecasting methods comprise a large spectrum of statistical, econometric, and artificial intelligence methods, including time series regression, moving average, autoregressive models, neural networks, pickup methods, exponential smoothing, expert judgment, Monte Carlo simulation, and so on (Law 1998; Weatherford and Kimes 2003; Andrew, Cranage, and Lee 1990; Song and Li 2008). However, the validity of these methods could be tempered in this context, as they are, in general, based upon historic performance or forecasts of other independent variables, both of which hinges heavily on a consistent pattern of tourist activities and a stable economic structure. Therefore, certain one-off events or dramatic changes in the economy, and the concomitant shocks to the tourism and hospitality industry, are highly likely to reduce the accuracy of these forecasting models.

Since the mid-1990s, the development of the Internet and information technologies has made available a new type of online data sets representing the behavioral traces that consumers leave behind when they engage in various online activities. The data include search engine keyword volumes, amounts of tweets, web traffic volumes, and posts from various social media platforms (Hubbard 2011; Leiner et al. 1997; Pan, Wu, and Song 2012). These so-called online pulse data have been adopted in forecasting films' box office revenue (Asur and Huberman 2010; Goel et al. 2010), flu outbreaks (Ginsberg et al. 2009), unemployment benefit claims (Choi and Varian 2009), and even stock market movement (Bollen, Mao, and Zeng 2011). In particular, for the tourism and hospitality industry, about 76% of travelers in the United States go online to plan their trips (TIA 2009) and visit airline, travel agency, and destination websites before taking their trips (Fesenmaier et al. 2011). Their online traces can be viewed as early behavioral indicators of tourist activities and, therefore, these data have great potential for increasing the accuracy of forecasting tourist activities as valuable predictors of their forthcoming travel.

However, tourist activities can actualize in many different types of indicators: the number of visitor arrivals for a destination or various attractions, length of stay, total local spending or spending at specific activities, demand for hotel rooms, hotel occupancy, etc. These indicators might be correlated but different from each other. Demand for hotel rooms in one area is specifically closely associated with the hospitality industry and the data are more readily available from commercial sources. As a result, we focus on hotel demand

[1]Temple University, Philadelphia, PA, USA
[2]College of Charleston, Charleston, SC, USA
[3]The Hong Kong Polytechnic University, Kowloon, Hong Kong

Corresponding Author:
Bing Pan, Department of Hospitality and Tourism Management, School of Business, College of Charleston, Charleston, SC 29424-001, USA.
Email: bingpan@gmail.com

in this study: we combine the use of traditional econometric methods with a new type of online data, namely, the web traffic volumes from a local Convention and Visitors Bureau (CVB), to predict demand for hotel rooms in a specific tourist area. The study confirms the value of web traffic data from local destination marketing organizations (DMOs) in predicting the demand for hotel rooms in a tourist destination.

Literature Review

Tourism demand forecasting is crucial in facilitating businesses and organizations to allocate limited resources efficiently and to stay competitive (Song, Witt, and Li 2008). It has been a widely researched area in the field of tourism and hospitality (Frechtling 1996, 2001; Palm and Zellner 1992; Song, Witt, and Li 2008), and abundant literature can be found forecasting tourist arrivals, expenditure, room nights sold, and occupancy rates. In revenue management practices in particular, accurate forecasts can facilitate decision making on pricing strategy for individual or group hotels. The following section briefly reviews the methods employed in forecasting demand for hotel rooms as well as the literature on using online data in the forecasting process.

Hotel Room Demand Forecasting

Many researchers have attempted to forecast the demand for hotel rooms and the occupancy rates of individual properties. For example, Andrew, Cranage, and Lee (1990) use the Box-Jenkins method and the exponential smoothing method to forecast the monthly occupancy of a hotel. They show that both methods produce a very low mean sum of squared residuals, and the Box-Jenkins approach outperforms the exponential smoothing model. Schwartz and Hiemstra (1997) propose a method to forecast daily occupancy rates based on previous booking curves. The results for three hotels in the United States highlight the improved forecasting accuracy compared to three benchmarking models—an autoregression (AR) model, a high-order polynomial model, and a combination of the two. The results also show that combining the three independent models improves accuracy. Weatherford and Kimes (2003) review major forecasting methods in hotel revenue management and categorize them into historical, advanced booking, and combined methods. Their empirical study of daily guest arrivals to six hotels highlights the superiority of the exponential smoothing and pickup methods. Furthermore, some researchers adopt the Mote Carlo simulation to forecast hotel occupancy and demonstrate that the model generates more accurate forecasts than traditional approaches such as the exponential smoothing and pick-up models (Zakhary et al. 2009).

Other researchers have studied the forecasting of average occupancy and demand for hotel rooms within a country or an area. Law (1998) employs a neural network approach to forecast the hotel occupancy in Hong Kong. The results suggest that this method outperforms multiple regression and naive models. By focusing on hotel occupancy trends, he further proposes a refined extrapolative hotel room occupancy rate model to forecast the annual average occupancy rates of hotels in Hong Kong. Choi (2003) identifies the leading economic indicators of the hotel industry in the United States and uses the composite of these indicators to predict hotel demand. This method generates moderately accurate forecasts, ranging from 67% to 83% in directional change.

Forecasting with Online Data

Along with the increasing availability of data related to online activities, recently, there has been a plethora of research on economic behavior prediction using online data. So-called online pulse data refers to traces of consumer behavior, unveiling their interests and purchases in real time (Hubbard 2011). These data are analogous to the pulse of a human body and can be seen as indicators of the health status of an industry. Online pulse data are available in the forms of search engine query volumes or social media postings and have been applied to various disciplines. For example, Google Trends, a public tool provided by Google, provides voluminous data for specific search queries on the Google network (Carneiro and Mylonakis 2009). An influential study by Google researchers focuses on predicting influenza outbreaks from the search volumes of certain keywords (Ginsberg et al. 2009). The researchers used 45 flu-related keywords from Google and correlated them with the number of patients who contracted influenza reported by Centers for Disease Control and Prevention (CDC). Their forecasting model is able to detect flu outbreaks two weeks ahead of the CDC, validating the substantial value of search engine volume data in helping combat health problems.

The search engine volume data has also been applied in business, finance, and economies. Choi and Varian (2009) incorporate search engine data from the keywords *jobs* and *welfare/unemployment* into an autoregressive integrated moving average (ARIMA) model to predict unemployment claims. They show that the model generates more accurate forecasts than the baseline model (without the search engine data). In addition, Askitas and Zimmermann (2009) identify a strong correlation between keyword searches and monthly unemployment rates in Germany using a simple error-correction model. In another study, Zhang, Jansen, and Spink (2009) estimate a number of ARIMA models incorporating raw search engine data from Dogpile.com, and their findings suggest that the daily log data could be used to detect users' behavioral change across different time periods.

Considered as an online form of the wisdom of crowds (Surowiecki 2004), social media content, another type of online pulse data, has also been utilized to predict various economic activities. Gruhl et al. (2005) apply automated data mining of blogs to predict the volume of book sales. Zhang, Fuehres, and Gloor (2010) investigate six months' records of Twitter feeds, especially on the daily volumes of tweets containing expressions of hope and fear. They show that the percentages of these "emotional" tweets are significantly correlated with Dow Jones, NASDAQ, and S&P 500 fluctuations, and this result has been confirmed by Bollen, Mao, and Zeng (2011). Asur and Huberman (2010) construct a linear model based on the volume of tweets associated with certain movies to predict their box-office revenues. The results show that the model outperforms the predictions generated by the Hollywood Stock Exchange. In the field of public health, Lampos and Cristianini (2010) also incorporate the volume of Twitter messages including certain keywords to track and predict flu epidemics and highlight the value of these data for early detection and geo-location of outbreaks.

In summary, the major advantages of the online pulse data provided by search engines and social media lie in the fact that they are real-time, high-frequency (daily and weekly instead of quarterly or annually), and sensitive to slight changes in consumer behavior. As shown in past studies, researchers in other fields have highlighted the great value of online pulse data in generating accurate socioeconomic forecasts. With these overwhelming merits, the online pulse data provide a solution to the problem of traditional forecasting models: the heavy reliance on a consistent historic pattern and the stability of the economic structure.

A considerable number of tourists will search for information on the Internet before undertaking their trips (Fesenmaier et al. 2011), and they usually visit the website of a DMO as part of information search process. Accordingly, the traffic volume of a DMO website could indicate the interests and purchase intentions of potential tourists. To our knowledge, only two studies have so far used online data to predict tourists' activities. Choi and Varian (2009) incorporate Google search volume data to predict visitor arrivals to Hong Kong from nine different countries. By combining an AR model with search volume data for the keyword *Hong Kong* from the specific country, they show that the model fits the data remarkably well with a very high R^2. In another study, Pan, Wu, and Song (2012) use Google search volume data to predict hotel room demand in Charleston, United States. By including five travel-related queries to a tourist destination in an autoregressive moving average with exogenous inputs (ARMAX) model, the results show that search engine volume data makes a significant contribution to predicting room demand by reducing the mean absolute percentage error (MAPE) (Pan, Wu, and Song 2012). However, a noticeable drawback of these two studies is rooted in the ex post nature of their forecasts: both studies use online data in current time to predict current tourist activities (Song, Witt, and Li 2008). Even though these models could help practitioners to obtain tourism statistics earlier than hotels' own reports at best, their applicability is limited. Ex ante forecasts would be more useful for practitioners, as these use past values of the independent variables to predict the future values of dependent variables. Another drawback of these two studies comes from their use of search volume of certain queries on Google instead of website traffic data. Website traffic data might be more superior to search query data in three aspects: (1) search query data are usually normalized and scaled, resulting in the limited predictive power (Pan, Wu, and Song 2012), whereas website traffic data, in general, record the raw number of visits or visitors to a website; (2) one needs to investigate many different queries and it involves trial and errors in order to pick the right ones with the strongest predictive power; in the United States, a CVB's website is usually the obvious choice for one tourist area given its importance in destination marketing (Xiang et al. 2010). Thus, its web traffic data are the apparent data choice for predicting tourist arrivals; (3) website visits are usually the subsequent step for searching through search engines and even closer to the actual conversion during the customers' purchase process (Fesenmaier et al. 2011). Therefore, website visits might possess stronger predictive power than search engine query volumes.

The data from various traveler surveys further corroborate the importance of website in trip planning. A national survey on American travelers reveals that 33% to 47% of travelers find destination websites very useful or essential for their trip planning (Fesenmaier et al. 2011). Specifically, in the context of this research, according to the most recent Charleston visitor intercept survey (Smith and Pan 2013), 32% of Charleston visitors used the CVB Website to research and plan their trips; among them, 92% stayed overnight. Thus, there is strong evidence suggesting that tourists' visitation to the CVB website is a precursor to their hotel stays in the United States. Hence, website visit data are a potential candidate variable for forecasting hotel demand and occupancy.

In this study, the web traffic data from a DMO is utilized to predict demand for hotel rooms in a tourist area. DMOs are the marketing organizations of destinations and also act as tourist service providers (Gretzel, Yuan, and Fesenmaier 2000). They are the main providers of comprehensive tourist information about the destination via the web. Accordingly, their web traffic volumes may indicate potential tourists' interests and act as leading indicators of tourist visits, which will ultimately result in hotel room bookings in the destinations. Thus, the overall research question we consider in this study is whether or not the website traffic data of a DMO could improve the forecasting accuracy of demand for hotel rooms in that area.

Data Description

In order to investigate the predictive power of DMO website traffic in forecasting the demand for hotel rooms in a destination, we chose Charleston, South Carolina, United States, as the target destination, because of our authorized access to the Charleston Area Convention and Visitors Bureau (CACVB) website traffic data and local hotel room demand data. We collected web traffic data from the Google Analytics account of the CACVB website, and Smith Travel Research, Inc. (STR), provided the hotel demand and occupancy data for the Charleston area.

Google Analytics is a free tool provided by Google Inc. that enables website owners understand how their web visitors interact with websites (Hasan, Morris, and Probets 2009; Plaza 2011). To this end, a website owner needs to create a unique Google Analytics account and embed a piece of Javascript code in every page of the site. When a web visitor browses this page, the Javascript will automatically send the visitor's information to the Google server, and the information can then be viewed and analyzed in the Google Analytics account. The aforementioned visitors' information includes the number of web visits, the number of unique visits, the time spent on the site by a visitor, the geographical location of the visitors, and the source of their visits (e.g., search engines and referral sites). Google Analytics provides two major types of web traffic volume data: *visitors* represents the number of unique users (identified by cookies) accessing a specific website and *visits* represents the number of individual sessions initiated by all visitors to the site. A session is defined as a continuous period of access to the website from a computer cookie or IP address without any time period of longer than 30 minutes between adjacent accesses. Following this definition, if a user has been inactive on the site for 30 minutes or more, any future activity is considered a new session (Google 2012). In this study, we tested the predictive power of both *visitors* and *visits*. The web traffic data covers the time span from the 21st week of 2007 (the time Google Analytics was adopted for the CACVB website) to the 17th week of 2011 (the start of the research).

STR provided weekly data on hotel occupancy rates and the hotel room nights sold for the Charleston area (Agarwal, Yochum, and Isakovski 2002). Around 190 hotels and motels are located in the Charleston area and STR collects daily occupancy and room rates data from about 110 of them. STR then generates the estimates of average hotel occupancy rates and total room nights sold based on the samples. The two hotel performance data sets are obtained to indicate the demand and occupancy rates of hotels in Charleston, South Carolina. The variable *occ* denotes the average weekly occupancy rate of hotels in the Charleston area, and *dem* represents the total number of hotel room nights sold in a week in the same area. The two hotel demand and occupancy series range from the 1st week of 2006 to the 16th week of 2011.

Research Questions

The main research objective of this study is to examine whether the web traffic data of a local DMO website helps to improve the accuracy of forecast demand for hotel rooms and hotels' occupancy rates. Since the web traffic data can take two different formats (the number of web visitors, and the volume of web visits), the first subquestion is

> *Research question 1*: Which type of data is more useful in forecasting the demand for hotel rooms and hotel occupancy rates: the volume of web visits, or the number of web visitors?

There are two candidate models which can incorporate the additional web traffic data to forecast hotel demand and occupancy. The ARMAX model extends the traditional ARMA (autoregressive moving average) model by including the explanatory variables as direct predictors; the TAR (threshold autoregressive) model incorporates the explanatory variable in an indirect way by treating it as a threshold variable. Thus, the second research question is

> *Research question 2*: Which forecasting models should be used to incorporate the web traffic data: the model incorporating the web traffic series as a direct predictor, or the one incorporating it as an indirect threshold variable?

The value of online data lies in the fact that they can provide time-sensitive leading indicators of visitor behavior, which cannot be otherwise captured. Because of the short planning horizon of most visitors, we suspect that this type of data will be more valuable in increasing the short-run forecasting accuracy than the long-run one. Thus, the third research question is

> *Research question 3*: Is the model's short-run forecasting performance superior to its long-run forecasting performance?

Since the structures of the ARMAX and TAR models vary significantly with the lag lengths of the independent and dependent variables included in the models, we are also interested in identifying the best possible model for predicting room demand. Thus, the fourth research question is

> *Research question 4*: Which configuration of the models performs best in forecasting error reduction?

As suggested in past research, existing time series models show relatively ideal forecasting accuracy (Andrew, Cranage, and Lee 1990; Weatherford and Kimes 2003). The online data will be deemed valuable only if it significantly improves the forecasting performance of such models. Hence, the fifth research question is

> *Research question 5*: How much can web traffic data contribute to the reduction of forecasting error?

Methodology

In order to answer the aforementioned research questions, four successive steps of investigation were carried out.

1 In the first step, we conducted an exploratory analysis by investigating the trend of the data and using several methods to check the stationarity of each time series. We also explored the possible correlation between different series. The methods used included the modified Dickey–Fuller test for unit roots, the cross-correlogram for testing the correlation between two time series, and the impulse-response analysis for detecting the responses of the variables to external shocks within a vector autoregressive (VAR) framework.

a. The modified Dickey-Fuller test is based on an auxiliary generalized least square (GLS) regression with de-trended time series data. The test has greater power than the early versions of the augmented Dickey-Fuller test (Elliott, Rothenberg, and Stock 1996). It is known as the DF-GLS test with a GLS regression specified as

$$\Delta y_t^* = \alpha + \beta y_{t-1}^* + \sum_{j=1}^{k} \varsigma_j \Delta y_{t-j}^* + \varepsilon_t, \tag{1}$$

where y_t^* is the de-trended time series data. For this DF-GLS unit root test, the null hypothesis is H_0: $\beta = 0$, which suggests that y_t is a random walk. Several criteria can be used to select the order of lags, k, in the GLS regression (equation 1), including the Ng-Perron sequential t (Ng and Perron 1995), the Schwarz information criterion (SIC), and the Ng-Perron modified Akaike information criterion (MAIC) (Ng and Perron 2003).

b. Cross-correlogram is used to detect the cross-covariance between two different time series y_1 and y_2 at different lags, and is given by

$$\mathrm{Cov}\left\{y_{1t}, y_{2,t+k}\right\} = R_{12}(k). \tag{2}$$

With this method, the correlation between two time series at different time lags can be analyzed. More importantly, this tool is useful in determining the lagged structure of the independent variables in the ARMAX models.

c. Impulse-response analysis is another useful tool analyzing the responses of the variables to external shocks in the VAR system. Suppose there are two endogenous variables, y_1 and y_2, the corresponding VAR model can be specified as

$$\begin{cases} y_{1t} = \alpha_1 + \sum_{m=1}^{j1} \beta_{11m} y_{1,t-m} + \sum_{n=1}^{k1} \beta_{12n} y_{2,t-n} + \varepsilon_{1t} \\ y_{2t} = \alpha_2 + \sum_{m=1}^{j2} \beta_{21m} y_{1,t-m} + \sum_{n=1}^{k2} \beta_{22n} y_{2,t-n} + \varepsilon_{2t} \end{cases}. \tag{3}$$

The Wald lag-exclusion test helps to select the optimal lags of length, j_1, j_2, k_1, and k_2, in the model. Based on the estimation results of the VAR model, the impulse-response function (IRF) can be used to inspect the relationship between the two time series. It visualizes the effects of a shock to an endogenous variable on the others as well as on itself (Lütkepohl 2005).

2. In the second step, two general forecasting strategies were constructed to incorporate the web traffic data and predict demand for hotel rooms and hotel occupancy rates. The first was the ARMAX model incorporating the web traffic data as a direct predictor, and the second was the TAR model treating the web traffic data as an indirect threshold variable to indicate which regime should be used for prediction. The regime refers to the forecasting regime shown in equation 5. The forecasting performance of these two models was then evaluated and compared based on the estimation and validation samples specified at the outset.

d. The ARMAX model extends the traditional ARMA model by including explanatory variables as direct predictors and can be specified as

$$\begin{aligned} y_t &= \alpha + \sum_{i=0}^{j} \beta_i x_{t-i} + \mu_t \\ \mu_t &= \sum_{i=1}^{m} \rho_i \mu_{t-i} + \sum_{j=1}^{n} \theta_j \varepsilon_{t-j} + \varepsilon_t, \end{aligned} \tag{4}$$

where x_t is the explanatory variable. If β is set to zero, the model becomes a standard ARMA(m, n) model. Extensive efforts should be made to specify adequate m and n to guarantee the model residual ε_t to be a white noise. To determine the lag length of the explanatory variable, j, a correlogram of pre-whitened x and y is used to select lags with significant cross-correlation (Box, Jenkins, and Reinsel 2011).

e. The TAR model incorporates the explanatory variable indirectly by treating it as a threshold variable. It is specified as

$$y_t = \begin{cases} \phi_1 + \sum_{i=1}^{m} \phi_{1i} y_{t-i} + \varepsilon_t & \text{for } x_{t-1} \le \psi \\ \phi_2 + \sum_{i=1}^{m} \phi_{2i} y_{t-i} + \varepsilon_t & \text{for } x_{t-1} > \psi, \end{cases} \tag{5}$$

where the explanatory variable x_{t-1} does not enter the prediction function directly. Instead, it is used as an indicator of the forecasting regime. If the explanatory variable exceeds the threshold value ψ, a Regime 1 equation is specified, while if the explanatory variable is below the threshold value, the equation switches to Regime 2. The specification of TAR is more complicated because both lag length m and threshold value ψ should be specified. The AIC is used to choose the appropriate lag order and threshold value.

3. In the third step, we focused particularly on the improvement in accuracy obtained by including the web traffic data, especially the improvement in

the ARMAX model compared to that of the ARMA. Two measures of forecasting accuracy are used to evaluate the forecasting performance of the models: MAPE and the root mean square percentage error (RMSPE). They are specified as follows:

$$\text{MAPE} = \frac{1}{m}\sum_{t=1}^{m}\left(\frac{|\hat{y}_t - y_t|}{y_t}\right)\times 100\%, \tag{6}$$

$$\text{RMSPE} = \sqrt{\frac{1}{m}\sum_{t=1}^{m}\left(\frac{\hat{y}_t - y_t}{y_t}\right)^2}\times 100\%, \tag{7}$$

Using MAPE, the improvement can be calculated as

$$I^t_{MAPE} = \frac{MAPE^t_{ARMA} - MAPE^t_{ARMAX}}{MAPE^t_{ARMA}}\times 100\%, \tag{8}$$

where I^t_{MAPE} denotes the t-step-ahead forecasting improvement of the ARMAX model measured by MAPE. In the same way, I^t_{RMSPE} is defined to measure the t-step-ahead forecasting improvement of the ARMAX model as measured by RMSPE. Benchmarking the proposed model against a number of alternative models that are frequently used in tourism forecasting is a standard exercise in tourism demand forecasting literature (Song, Witt, and Li 2008, pp. 181-95). The purpose is to evaluate the forecasting performance of various models so as to select the best model.

4. In the last step, the estimation and validation samples were selected with a view to investigating the robustness of the forecasting results obtained in the second and third steps.

To test the forecasting performance of the different models, the data sample was split into two sub-samples: estimation and validation data. However, the way in which the sample is divided can influence the model's forecasting performance. Therefore, several different splits were performed to test the robustness of the results. At the outset, the period of week 21 of 2007 to week 46 of 2010 was chosen as the estimation sample and the period of week 47 of 2010 to week 2 of 2011 as the validation sample for the analysis in the second and third steps. It is of particular interest to understand in which situations the additional web traffic data will be more useful in reducing forecasting errors. Therefore, based on the forecasting improvement measure (equation 8) for different sample splits, we run several auxiliary regression models to unveil the factor influencing the amount of improvement.

Research Results

After a preliminary examination of the data, all four original time series, *visitors, visits, occ*, and *dem*, were found to exhibit a trend of exponential growth; thus, the logarithm of each series was necessary for modeling and forecasting purposes. Figure 1 presents the plots of these variables in the logarithm. The two hotel demand series are ln*dem* (room nights sold per week) and ln*occ* (average weekly room occupancy rates), and the two web traffic series are ln*visits* and ln*visitors.* All the four time series share very similar seasonality patterns.

Results of Exploratory Analysis

The results of the DF-GLS unit root tests showed that the null hypothesis of random walk cannot be rejected and the two web traffic time series are no-trend series. However, if they are specified with trends, the null hypothesis is rejected at the 0.10 significance level. For hotel demand time series, the statistics indicate that they do not have unit roots in either the de-trended or trended cases.

The cross-correlogram was employed to measure the cross-correlation of pairwise time series at different lags. Figure 2 presents the cross-correlogram between the two web traffic and two hotel demand series. It shows that the cross-correlation between web traffic data and hotel demand and occupancy data reaches its maximum at the lag of 4, suggesting that the web traffic volume data has its largest impact on hotel demand and occupancy at around week four.

To look further into the relationship between the web traffic data and hotel demand and occupancy, the VAR model and associated impulse-response analysis were estimated. Figure 3 gives the orthogonalized IRF of the web traffic data on the hotel demand and occupancy series. As shown in the graphs, the shocks generated by the web traffic volumes are likely to have a lasting impact on the two demand series. This impact reaches its highest level at about four weeks, which is consistent with the results from the cross-correlograms in Figure 2, suggesting that most visitors might browse the CACVB website four weeks before their visits to Charleston, South Carolina. In addition, to confirm the casual relationship between web traffic data and hotel demand, Granger causality test was conducted and the results reveal a significant reciprocal Granger causality between the two sets of variables: web traffic volume tends to Granger-cause hotel demand/occupancy, while hotel demand/occupancy would also Granger-cause web traffic volume (Granger, 1969).

Results of ARMAX and TAR Models

Before specifying the forecasting models, we examined the auto-correlation structure of each time series. Figure 4 presents the correlogram and partial correlogram of the two series to be forecasted, namely, ln*occ* and ln*dem*. The results show that except for the significant partial autocorrelation statistics at small lags, some partial autocorrelation statistics are also significant at large ones, such as at lag 52, suggesting a lag of one year. Therefore, the longer lag length should also be considered in the model. Since the unit root test

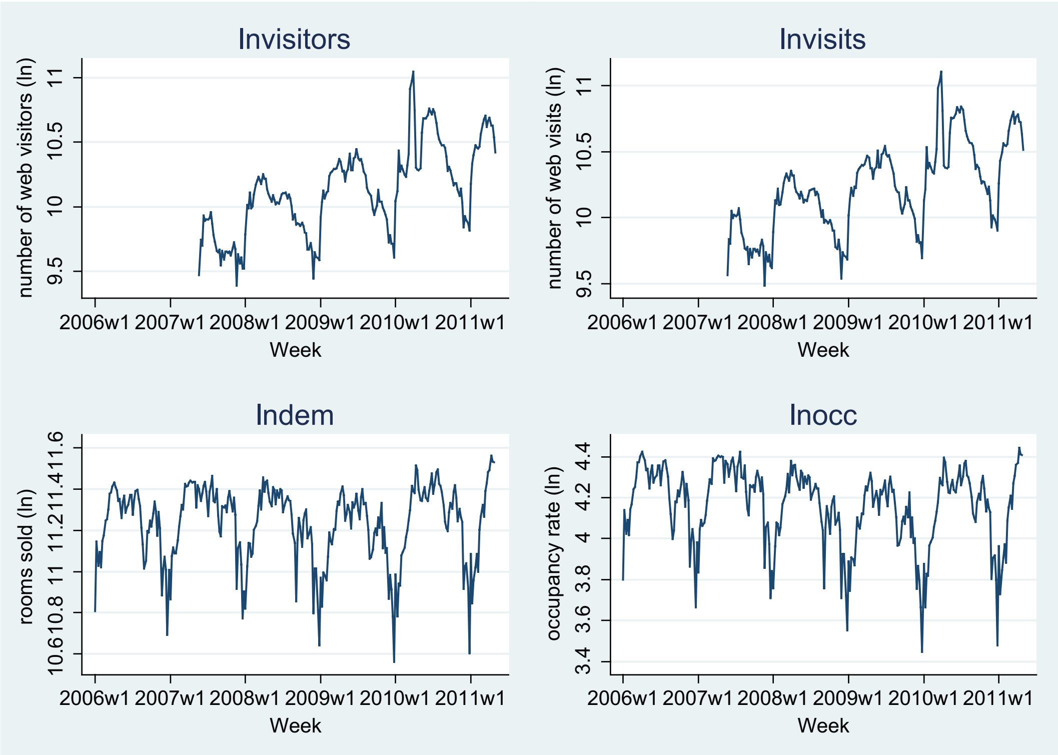

Figure 1. Plots of time series (in logarithm).

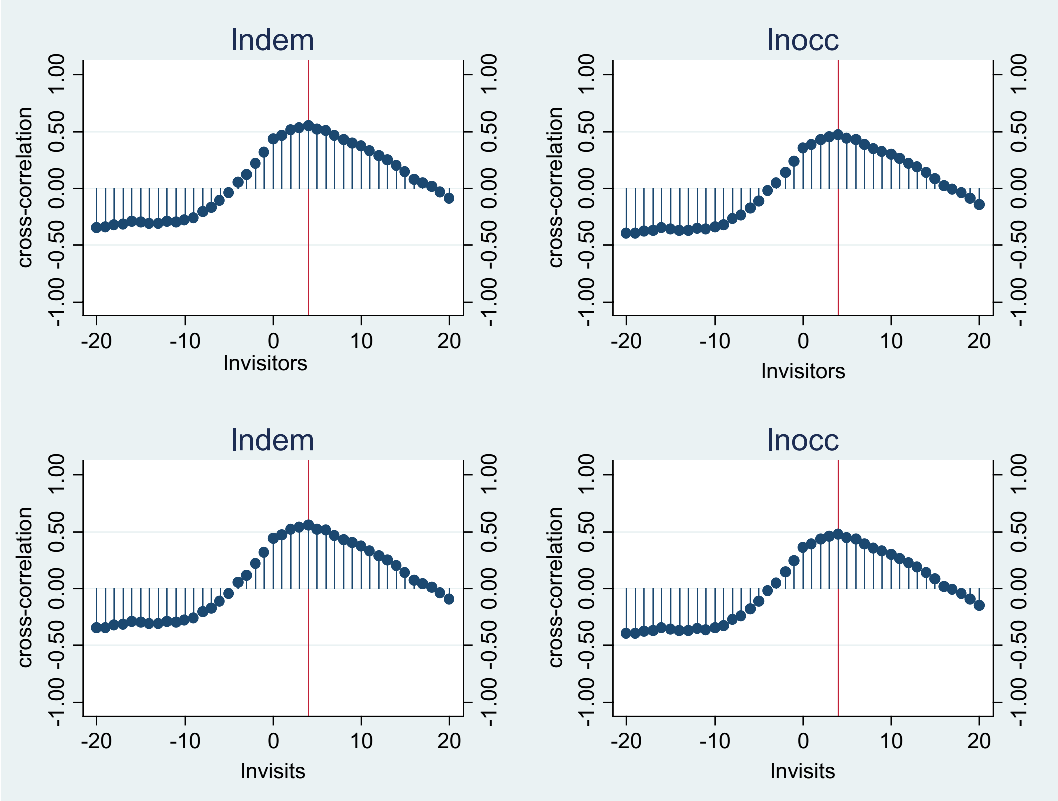

Figure 2. Cross-correlogram between web traffic and hotel demand and occupancy series.

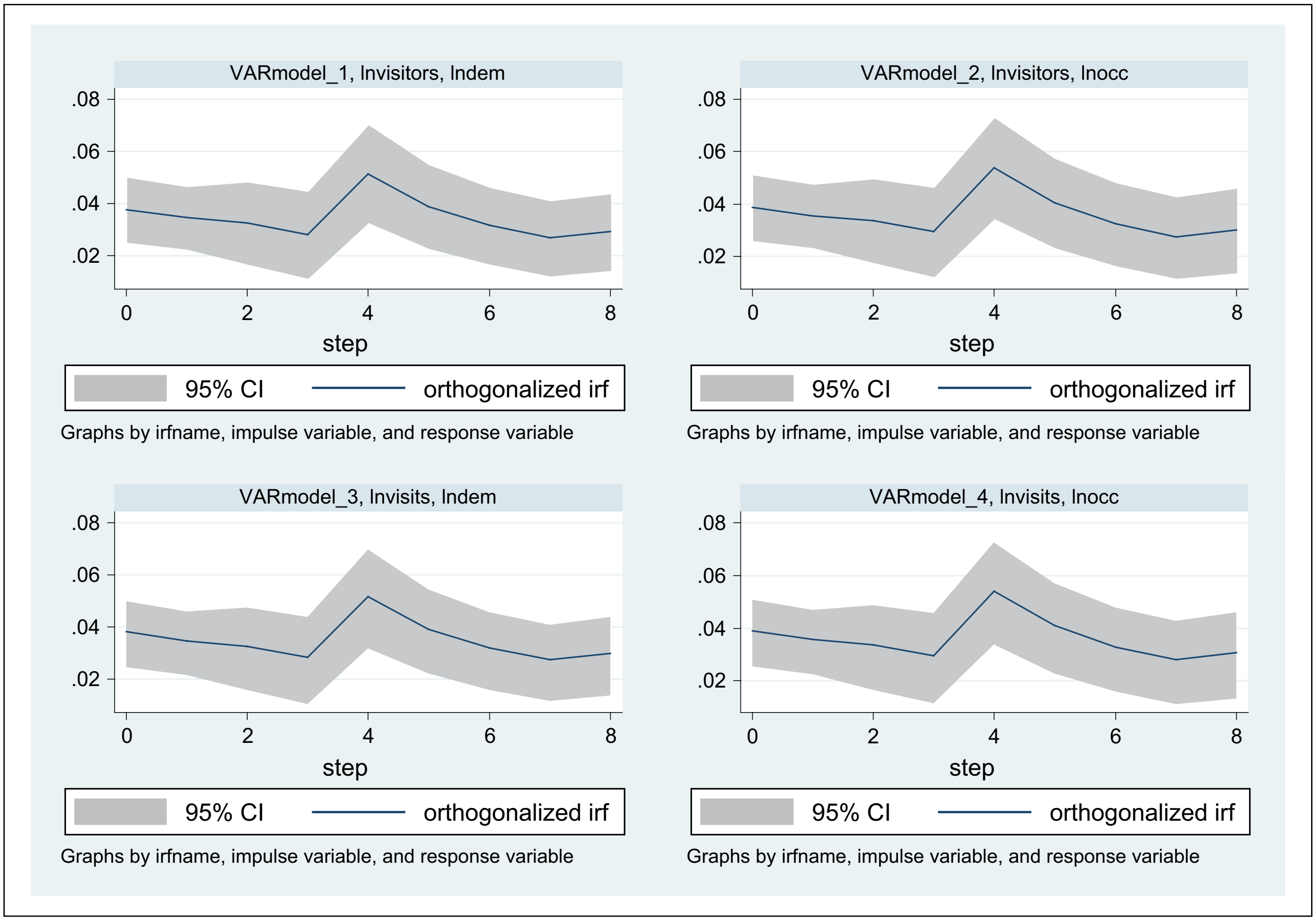

Figure 3. Orthogonalized impulse-response function of web traffic on hotel demand and occupancy.

suggests that both ln*occ* and ln*dem* are stationary, data differencing is unnecessary. The ARMA model is a natural candidate to forecast future values if additional web traffic data are not incorporated.

To specify the ARMAX model (equation 4), it is important to select the lagged structure of the additional web traffic data. To determine the lag length of the explanatory variable, a correlogram of pre-whitened x and y is used to select lags with significant cross-correlation (Box, Jenkins, and Reinsel 2011). Since the autocorrelation structure of series x contaminates the covariance of the sample estimate in the transfer function, the pre-whitening transformation filters out the autocorrelation structure and make the lag identification feasible through the sample cross-correlation function (Wei 2006, pp. 328-29). The results of these cross-correlograms are presented in Figure 5, and the significant correlations are identified at lags of 0, 1, 4, and 5.

Table 1 presents the estimation results of the ARMAX models with ln*visitors* or ln*visits* as the explanatory variable. After several preliminary estimation trials, the lag length of 1 and 52 for AR terms and 2, 3, 4, and 6 for MA terms are selected. All coefficients are estimated to be statistically significant. Bartlett's (B) statistic and the Portmanteau (Q) statistic are highly insignificant for every model, suggesting that the residuals are white noise.

In terms of the TAR model, the optimal lag length and threshold value were determined by the AIC criteria. The best model with lag lengths up to 6 was chosen. The threshold values of different quantiles were also evaluated and compared. As judged by the minimum AIC values, the TAR models with AR (6) and 10% quantile of the threshold variable as threshold value were chosen. The small threshold value suggests that a smaller web traffic volume would switch the forecasting regime of the two hotel demand series. Therefore, when the value of the web traffic volume is small, the regime used for forecasting should be changed. Based on the specification selected, Table 2 presents the estimation results of the four TAR models obtained from conditional maximum likelihood estimation.

After obtaining the estimation results, the forecasts of the two demand variables in the validation period were generated based on ex ante rather than ex post forecasts. In practice, it is more realistic to generate ex ante forecasts, which assume that independent variables (web traffic data) are not

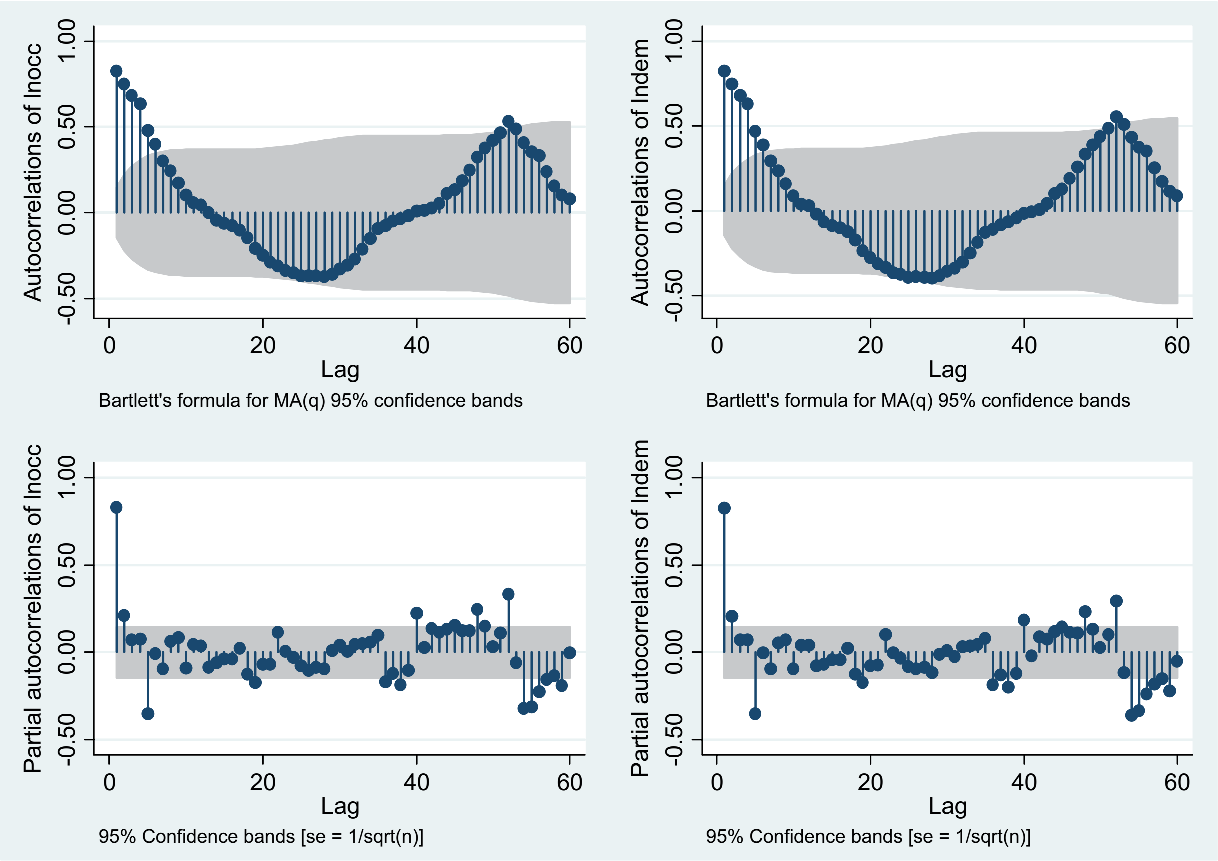

Figure 4. Autocorrelation and partial autocorrelation of hotel demand and occupancy.

available in the validation period (Song, Witt, and Li 2008). In ex ante forecasting, the forecasts from the ARMA model will be the same as in the ex post forecasts as this model does not incorporate any independent variables. For the ARMAX and TAR models, the forecasts of the explanatory variables in the validation period are generated by the univariate ARIMA model.

The ex ante forecasting performance measures of the ARMA, ARMAX, and TAR models are presented in Tables 1 and 2. Table 3 presents these results with different forecasting horizons ranging from 4 to 20 weeks ahead. The better-performing models are underlined in the table. Similar conclusions are reached if we use MAPE and RMSPE to select the best model. First, the forecasting accuracy of the TAR model is always the worst compared to the ARMA and ARMAX models. Second, it shows that the web traffic data are useful for forecasting the two hotel demand series in Charleston as indicated by the lower MAPE and RMSPE values of the ARMAX models in four-weeks-ahead forecasts (i.e., forecasting for the next four weeks). These results suggest that the ARMAX model is superior in short-run forecasting. Third, in general, ln*visits*, which denotes the total volume of web visits, is equally effective as ln*visitors* as a predictor of demand for hotel rooms and of hotel occupancy rates. However, for the 8- to 20-weeks-ahead forecasts, the ARMA model is found to be superior to the ARMAX model, suggesting that the additional web traffic data do not necessarily improve the forecasting accuracy with this sample split when the forecasting horizon is longer than 4 weeks ahead.

Rates of Error Reduction

Using equation 8, it can be shown that by including the web visit (ln*visits*) as a predictor, the accuracy of the four-weeks-ahead ARMAX forecasts of the demand for hotel rooms (ln*dem*) is improved by 14.47% according to MAPE and 0.81% according to RMSPE. This relatively large difference between MAPE and RMSPE is due to the fact that RMSPE penalizes some extreme prediction errors. In different sample splits in Table 4, this inconsistency disappears when using different sample splits.

Determinants of an Increase in Forecasting Accuracy

In order to test the robustness of the forecasting performance of the models considered, different samples for model

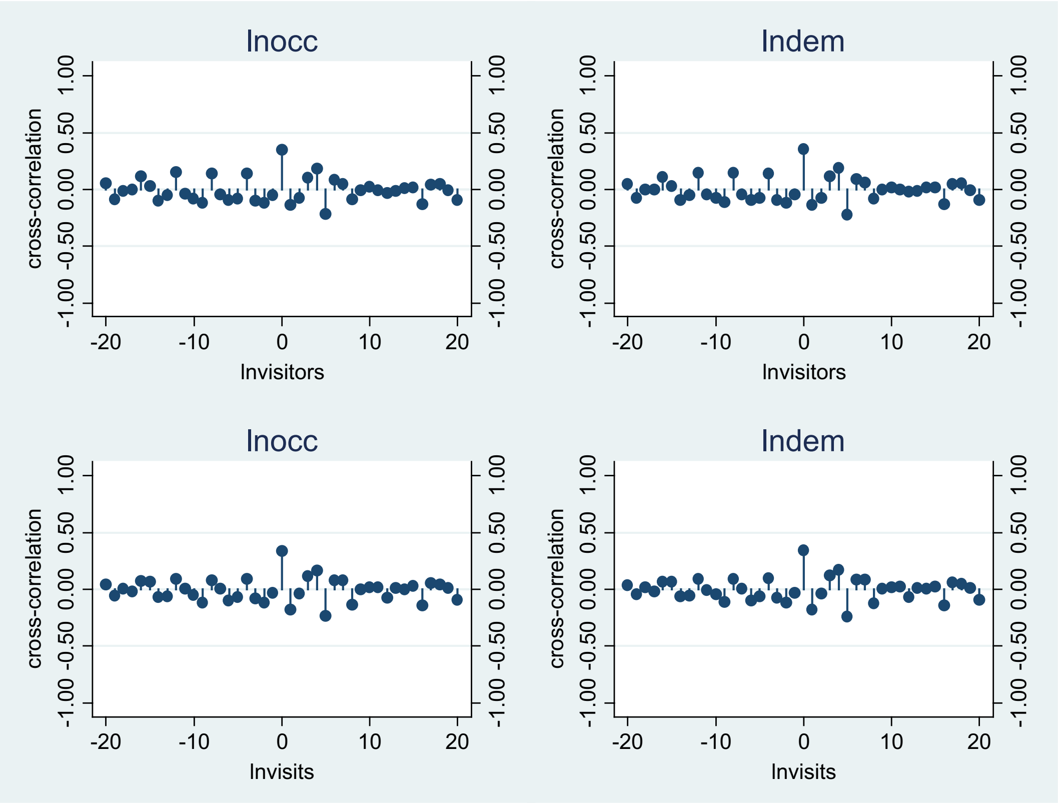

Figure 5. Cross-correlogram between pre-whitened web traffic and hotel demand and occupancy series.

estimation and forecasting were used to generate the ex ante forecasts. In this part of the data analysis, 30 different time points were used to recursively split the sample from week 32 in 2010 to week 9 in 2011. Because of the limitation of the sample size, only the four- and eight-weeks-ahead ARMAX forecasts were evaluated.

Table 4 presents the average values of the forecasting performance with 30 different sample splits. The conclusion of the exercise remains the same. In other words, the web traffic data are generally useful in improving the accuracy of forecast demand for hotel rooms (ln*dem*). By including the number of web visits (ln*visits*) in the model, the four-weeks-ahead forecasting accuracy of the demand for hotel rooms (ln*dem*) improved by an average of 7.43% according to MAPE or 5.92% according to RMSPE. For the eight-weeks-ahead forecasts, accuracy improved by an average of 10.60% and 6.32% according to MAPE and RMSPE, respectively. However, this is not true for the forecasts of hotel occupancy (ln*occ*).

Even though the ARMAX model, on average, outperformed ARMA in forecasting the demand for hotel rooms, there are still several sample splits that do not support this conclusion. Therefore, a number of auxiliary regressions were estimated to test the possible deterimants of the improvement in the ARMAX models. The dependent variable of the regression was the forecasting improvement measured by MAPE using equation 8, and the independent variable the average actual values of the demand series in the forecasting period. The results of these regressions are presented in Table 5. The models are estimated to be significant for all four- and eight-weeks-ahead forecast regression models and the β coefficients are all positive and statistically significant. This result suggests that the ARMAX models incorporating the additional web traffic data can improve forecasting accuracy in predicting large values for the hotel demand series. This means that the web traffic data are especially valuable for generating accurate forecasts during peak seasons.

In summary, the analyses have answered the five research questions presented earlier as follows: (1) *Visits* and *visitors* are two types of web traffic data that are almost equally effective in predicting demand for hotel rooms, but not in predicting occupancy rates. The difference in the dependent variables might be due to two reasons: (a) Room inventory changes over time since new hotels might have opened and old hotels might have closed. According to the STR report, from the 21st week of 2007 to the 16th week of 2011, the room inventory in the Charleston area increased from 16,292

Table 1. Estimation Results of the ARMA and ARMAX Models.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Model	ARMA	ARMAX	ARMAX	ARMA	ARMAX	ARMAX
Dependent variable	ln*occ*	ln*occ*	ln*occ*	ln*dem*	ln*dem*	ln*dem*
Independent variable		ln*visitors*	ln*visits*		ln*visitors*	ln*visits*
Lag(0)		0.327***	0.333***		0.334***	0.341***
		(0.045)	(0.044)		(0.045)	(0.044)
Lag(1)		–0.142**	–0.138**		–0.131**	–0.127**
		(0.064)	(0.063)		(0.064)	(0.063)
Lag(4)		0.249***	0.248***		0.250***	0.248***
		(0.050)	(0.050)		(0.047)	(0.047)
Lag(5)		–0.198***	–0.195***		–0.189***	–0.187***
		(0.056)	(0.056)		(0.056)	(0.056)
Constant	4.250***	1.915**	1.775*	11.30***	8.687***	8.548***
	(0.033)	(0.975)	(0.995)	(0.033)	(0.892)	(0.913)
AR(1)	0.444***	0.571***	0.571***	0.442***	0.553***	0.552***
	(0.049)	(0.050)	(0.051)	(0.049)	(0.052)	(0.053)
AR(52)	0.577***	0.511***	0.511***	0.567***	0.511***	0.511***
	(0.059)	(0.062)	(0.062)	(0.059)	(0.063)	(0.064)
MA(2)	0.194**			0.172*		
	(0.089)			(0.089)		
MA(3)	0.155**	0.192***	0.195***	0.142**	0.184**	0.189**
	(0.071)	(0.074)	(0.074)	(0.071)	(0.075)	(0.075)
MA(4)	0.478***	0.443***	0.440***	0.444***	0.417***	0.415***
	(0.064)	(0.084)	(0.085)	(0.063)	(0.086)	(0.087)
MA(6)	0.224***	0.163**	0.163**	0.212***	0.169**	0.171**
	(0.063)	(0.071)	(0.071)	(0.064)	(0.077)	(0.076)
σ^2	0.00611***	0.00506***	0.00503***	0.00610***	0.00497***	0.00493***
	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
Observed	182	177	177	182	177	177
AIC[a]	–395.4	–411.4	–412.4	–395.6	–414.7	–415.9
BIC[a]	–369.8	–376.4	–377.4	–370.0	–379.8	–380.9
Bartlett's (*B*)	0.662	0.689	0.707	0.630	0.672	0.695
Portmanteau (Q)	32.494	44.890	46.987	28.637	42.374	44.636

Note: Standard errors are in parentheses. AIC = Akaike's information criterion; BIC = Bayesian information criterion.

a. A superior alternative model with the extra variable on web traffic data.

****p* < 0.01, ***p* < 0.05, **p* < 0.10.

to 17,595 rooms, an increase of 8.0%. Thus, hotel demand might be more closely associated with web visit data. (b) Hotel occupancy ranges from 0% to 100%; though the room nights may be constrained by the capacity of the room supply, the demand for rooms has no such limit. Hotel occupancy might be less sensitive to changes in visitor numbers because of its fixed range. As a result, the two variables are highly correlated but not exactly the same: hotel demand data type in the format of room nights more closely resembles web visits, since both are aggregated indicators of the behavior of visitors during a certain time period. (2) The ARMAX model that incorporates the web traffic data as the predictive variable is more effective than the TAR model. (3) The web traffic data are most useful in predicting demand for hotel rooms four or eight weeks ahead. This is probably due to the fact that tourists prepare for their trip to Charleston, South Carolina, with a four- or eight-week lead time. (4) The ARIMA model incorporating the web traffic data is most effective in improving the forecasting accuracy of demand for hotel rooms with a specific configuration of the lagged variables. Such a configuration is more likely to be specific to the destination. And (5) the ARMAX model is most useful during peak seasons when hotel demand is high. In conclusion, the study confirms that web traffic data are useful in predicting short-run demand for hotel rooms.

Concluding Remarks

This study has validated the value of web traffic data from a DMO in predicting demand for hotel rooms. Such data is

Table 2. Estimation Results of TAR Models.

	Model 7	Model 8	Model 9	Model 10
Model	TAR	TAR	TAR	TAR
Forecasting variable	ln*occ*	ln*occ*	ln*dem*	ln*dem*
Threshold variable	ln*visitors*	ln*visits*	ln*visitors*	ln*visits*
Regime 1 model				
Constant	1.537***	1.595***	4.361***	4.552***
	(0.395)	(0.412)	(1.163)	(1.220)
AR(1)	0.281	0.302*	0.290	0.305*
	(0.204)	(0.160)	(0.206)	(0.162)
AR(2)	0.231	0.220	0.215	0.206
	(0.169)	(0.157)	(0.171)	(0.158)
AR(3)	0.189	0.179	0.180	0.172
	(0.184)	(0.175)	(0.186)	(0.177)
AR(4)	–0.230	–0.234	–0.241	–0.246
	(0.175)	(0.173)	(0.177)	(0.175)
AR(5)	0.001	–0.013	0.020	0.008
	(0.159)	(0.153)	(0.161)	(0.155)
AR(6)	0.149	0.152	0.144	0.146
	(0.126)	(0.125)	(0.127)	(0.127)
Constant	0.459**	0.461**	1.367**	1.355**
	(0.216)	(0.215)	(0.599)	(0.596)
Regime 2 model				
AR(1)	0.679***	0.711***	0.679***	0.710***
	(0.080)	(0.083)	(0.079)	(0.083)
AR(2)	0.159	0.145	0.157	0.144
	(0.097)	(0.098)	(0.097)	(0.098)
AR(3)	0.080	0.066	0.084	0.071
	(0.091)	(0.091)	(0.091)	(0.092)
AR(4)	0.425***	0.425***	0.416***	0.416***
	(0.093)	(0.093)	(0.093)	(0.093)
AR(5)	–0.449***	–0.463***	–0.450***	–0.464***
	(0.103)	(0.102)	(0.102)	(0.102)
AR(6)	–0.006	0.004	–0.008	0.003
	(0.092)	(0.092)	(0.092)	(0.091)
Observed	182	182	182	182
AIC	–866.018	–867.634	–865.335	–866.807

Note: Standard errors are in parentheses. AIC = Akaike's information criterion.
*** $p < 0.01$, **$p < 0.05$, *$p < 0.10$.

particularly useful in short-run predictions of total room nights sold using the ARMAX model. The forecasting error rate deduction reached an average of 7.43% in MAPE for four weeks ahead and 10.60% for eight weeks ahead. During peak seasons when the hotel demand is high, the rate of forecasting error reduction could reach as high as 14.5%, as indicated by the regression model. In the Charleston area, the peak seasons usually include certain large festivals or events (Pan 2010). Those festivals or events are not consistent from year to year and thus harder to predict with historic data. In this case, website traffic data as the early indicator of interest are of substantial value.

Even though the average of 7% to 10% of prediction error reduction does not seem like a large amount, because of the number of hotels worldwide, a little improvement in forecasting accuracy could lead to a large amount of saving for this industry (Chiang, Chen, and Xu 2007). This study is the first of its kind to demonstrate the value of web traffic data in predicting future business activities. Almost every business today has an online presence, be it a hotel, an attraction, a travel agency, or even a production company. The universal availability of web traffic data makes it particularly valuable for any business seeking to monitor future activities.

Our results also show that local DMOs and convention and visitors bureaus could provide their web traffic data to the local hotel markets and help them to obtain more accurate forecasting. Local hotels and motels could use the data for short-term (four- to eight-weeks-ahead) forecasts of demand for their hotel rooms. However, the forecasting

Table 3. Forecasting Performance of Various Models.

	Model	MAPE					RMSPE				
	External variable		ARMAX	ARMAX	TAR	TAR		ARMAX	ARMAX	TAR	TAR
Forecasting variable	Forecasting step	ARMA	ln*visitors*	ln*visits*	ln*visitors*	ln*visits*	ARMA	ln*visitors*	ln*visits*	ln*visitors*	ln*visits*
ln*occ*	4-week	1.552	*1.460*	*1.432*	4.338	4.218	1.715	*1.721*	*1.712*	5.052	4.863
	8-week	2.195	2.586	2.568	6.875	6.648	2.627	3.169	3.161	8.324	8.075
	12-week	2.267	2.776	2.767	6.230	5.975	2.602	3.320	3.318	7.399	7.142
	16-week	2.404	2.989	2.989	5.041	4.886	2.700	3.443	3.448	6.462	6.258
	20-week	2.388	2.947	2.952	5.151	5.085	2.636	3.327	3.335	6.314	6.203
ln*dem*	4-week	0.561	*0.462*	*0.463*	1.434	1.387	0.621	*0.616*	*0.616*	1.629	1.558
	8-week	0.756	0.795	0.795	2.187	2.100	0.916	1.025	1.024	2.625	2.532
	12-week	0.780	0.841	0.842	1.985	1.887	0.900	1.026	1.026	2.340	2.243
	16-week	0.841	0.906	0.909	1.674	1.629	0.950	1.063	1.066	2.071	2.000
	20-week	0.851	0.897	0.901	1.843	1.831	0.941	1.028	1.031	2.177	2.152

Table 4. Average Value of Forecasting Performance with Different Sample Splits.

		MAPE			RMSPE		
	Model		ARMAX	ARMAX		ARMAX	ARMAX
	External variable	ARMA	ln*visitors*	ln*visits*	ARMA	ln*visitors*	ln*visits*
Forecasting variable	Forecasting step						
lnocc	4-week	1.796	1.811	1.806	2.051	2.061	2.056
	8-week	2.073	2.082	2.080	2.333	2.396	2.395
lndem	4-week	0.646	*0.599*	*0.598*	0.738	*0.696*	*0.694*
	8-week	0.745	*0.666*	*0.666*	0.839	*0.786*	*0.786*

Table 5. Regression Analysis of Forecasting Improvement.

	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16	Model 17	Model 18
Forecasting variance	ln*occ*	ln*occ*	ln*occ*	ln*occ*	ln*dem*	ln*dem*	ln*dem*	ln*dem*
External variance	ln*visitors*	ln*visitors*	ln*visits*	ln*visits*	ln*visitors*	ln*visitors*	ln*visits*	ln*visits*
Step	4-week	8-week	4-week	8-week	4-week	8-week	4-week	8-week
Average value	0.766***	1.126***	0.783***	1.158***	0.831***	1.112***	0.853***	1.131***
	(0.195)	(0.140)	(0.195)	(0.147)	(0.188)	(0.141)	(0.187)	(0.146)
Constant	−3.111***	−4.577***	−3.179***	−4.708***	−9.214***	−12.33***	−9.453***	−12.55***
	(0.788)	(0.580)	(0.788)	(0.605)	(2.092)	(1.575)	(2.084)	(1.631)
Observed	30	30	30	30	30	30	30	30
Adjusted R^2	0.333	0.610	0.340	0.605	0.368	0.644	0.380	0.637

Note: Standards errors are in parentheses.
***p < 0.01, **p < 0.05, *p < 0.10.

horizon, be it four or eight weeks, to some extent depends on the specific type of destination. One can imagine that a destination catering mainly for business travelers will have a different planning horizon compared to a place that is more vacation-focused. Therefore, specific models need to be tested according to the characteristics of each destination. In addition, along with the ubiquitous availability of mobile technology, the planning horizons are destined to get shorter and shorter in future. Continuous monitoring and tweaking of the forecasting models will therefore be necessary.

In addition, the models could be more specific: different sets of traffic data on various sections of the DMO website, such as hotels, attractions, or events, could be used to predict more specific types of tourist activities in the destination. Other types of tourist behavior indicators, such as visitor volume, local spending, and length of stay, could also be modeled with a similar methodology. Furthermore, in the era of big data (Mayer-Schönberger and Cukier 2012), other types of online data, in the forms of tweets, blogs, or likes on social media, if combined, may achieve a further improvement in forecasting accuracy. Therefore, this study also highlights a future research direction in terms of using a variety of online data to predict tourist activities including demand for hotel rooms.

The significance of DMO web traffic data in predicting the demand for hotel rooms validates the crucial role of DMOs in promoting a destination and connecting travelers with local tourism and hospitality services. Consistent with past studies (Fesenmaier et al. 2011; TIA 2009; Xiang et al. 2010), DMO websites are one of the top information resources and trip planning tools. The attractiveness and quality of the websites will determine how likely the potential visitors will convert to actual ones. They are one of the most important, if not the most important, bridge between local tourism and hospitality industry and visitors. They can also monitor future tourist activities, analyze the online traces, and distribute the data back to the local industry for better forecasting and management. This suggests a need for more investment in DMOs, especially in terms of their crucial role in online marketing for destinations.

Acknowledgments

The corresponding author would like to thank the School of Hotel and Tourism Management of The Hong Kong Polytechnic University for financial support during his sabbatical visit. Charleston Area Convention and Visitors Bureau generously provided the data for this research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Agarwal, Vinod B., Gilbert R. Yochum, and Tatiana Isakovski. (2002). "An Analysis of Smith Travel Research Occupancy Estimates: A Case Study of Virginia Beach Hotels." *Cornell Hotel & Restaurant Administration Quarterly*, 43 (2): 9-17.

Andrew, W. P., D. A. Cranage, and C. K. Lee. (1990). "Forecasting Hotel Occupancy Rates with Time Series Models: An Empirical Analysis." *Journal of Hospitality & Tourism Research*, 14 (2): 173-82.

Askitas, N., and K. F. Zimmermann. (2009). "Google Econometrics and Unemployment Forecasting." *Applied Economics Quarterly*, 55 (2): 107-20.

Asur, S., and B. A. Huberman. (2010). "Predicting the Future with Social Media." http://arxiv.org/abs/1003.5699.

Bollen, J., H. Mao, and X. Zeng. (2011). "Twitter Mood Predicts the Stock Market." *Journal of Computational Science*, 2 (1): 1-8.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. (2011). *Time Series Analysis: Forecasting and Control*. 4th edition. Oxford: Wiley-Blackwell.

Carneiro, H. A., and E. Mylonakis. (2009). "Google Trends: A Web Based Tool for Real Time Surveillance of Disease Outbreaks." *Clinical Infectious Diseases*, 49 (10): 1557-64.

Chiang, W. C., J. C. H. Chen, and X. Xu. (2007). "An Overview of Research on Revenue Management: Current Issues and Future Research." *International Journal of Revenue Management*, 1 (1): 97-128.

Choi, H., and H. Varian. (2009). "Predicting Initial Claims for Unemployment Benefits." In *Google Technical Report*.

Choi, J. G. (2003). "Developing an Economic Indicator System (a Forecasting Technique) for the Hotel Industry." *International Journal of Hospitality Management*, 22 (2): 147-59.

Elliott, Graham, Thomas J. Rothenberg, and James H. Stock. (1996). "Efficient Tests for an Autoregressive Unit Root." *Econometrica*, 64 (4): 813-36.

Fesenmaier, D. R., Z. Xiang, B. Pan, and R. Law. (2011). "A Framework of Search Engine Use for Travel Planning." *Journal of Travel Research*, 50 (6): 587-601.

Frechtling, D. C. (1996). *Practical Tourism Forecasting*. Boston: Butterworth-Heinemann.

Frechtling, D. C. (2001). *Forecasting Tourism Demand: Methods and Strategies*. Boston: Butterworth-Heinemann.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. (2009). "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature*, 457 (7232): 1012-14.

Goel, S., J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts. (2010). "Predicting Consumer Behavior with Web Search." *Proceedings of the National Academy of Sciences*, 107 (41): 17486-90.

Google. (2012). The Difference Between Clicks, Visits, Visitors, Pageviews, and Unique Pageviews 2012 (accessed September 20, 2012).

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica*, 37 (3): 424-38.

Gretzel, U., Y. Yuan, and D. R. Fesenmaier. (2000). "Preparing for the New Economy: Advertising and Change in Destination Marketing Organizations." *Journal of Travel Research*, 39 (2): 146-56.

Gruhl, D., R. Guha, R. Kumar, J. Novak, and A. Tomkins. (2005). "The Predictive Power of Online Chatter." Paper read at KDD '05 Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, at Chicago, Illinois, USA.

Hasan, L., A. Morris, and S. Probets. (2009). "Using Google Analytics to Evaluate the Usability of E-commerce Sites." *Human Centered Design*, 5619:697-706.

Hubbard, D. W. (2011). *Pulse: The New Science of Harnessing Internet Buzz to Track Threats and Opportunities*. Hoboken, NJ: Wiley.

Lampos, V., and N. Cristianini. (2010). "Tracking the Flu Pandemic by Monitoring the Social Web." Paper read at Cognitive Information Processing (CIP), 2010 2nd International Workshop on, 14-16 June 2010.

Law, R. (1998). "Room Occupancy Rate Forecasting: A Neural Network Approach." *International Journal of Contemporary Hospitality Management*, 10 (6): 234-39.

Leiner, B. M., V. G. Cerf, D. D. Clark, R. E. Kahn, L. Kleinrock, D. C. Lynch, J. Postel, L. G. Roberts, and S. S. Wolff. (1997). "The Past and Future History of the Internet." *Communications of the ACM*, 40 (2): 102-8.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin Heidelberg: Springer-Verlag.

Mayer-Schönberger, V., and K. Cukier. (2012). *Big Data: A Revolution That Transforms How We Work, Live, and Think*. Boston:Houghton Mifflin Harcourt.

Ng, S., and P. Perron. (1995). "Unit Root Tests in ARMA Models with Data-Dependent Methods for the Selection of the Truncation Lag." *Journal of the American Statistical Association*, 90 (429): 268-81.

Ng, S., and P. Perron. (2003). "Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power." *Econometrica*, 69 (6): 1519-54.

Palm, F. C., and A. Zellner. (1992). "To Combine or Not to Combine? Issues of Combining Forecasts." *Journal of Forecasting*, 11 (8): 687-701.

Pan, B. (2010). "Charleston Lodging in 2009 by the Numbers." http://panb.people.cofc.edu/pan/Charleston_Lodging_2009_by_Numbers_BingPan.pdf.

Pan, B., D. C. Wu, and H. Song. (2012). "Forecasting Hotel Room Demand Using Search Engine Data." *Journal of Hospitality and Tourism Technology*, 3 (3): 196-210.

Plaza, B. (2011). "Google Analytics for Measuring Website Performance." *Tourism Management*, 32 (3): 477-81.

Schwartz, Z., and S. Hiemstra. (1997). "Improving the Accuracy of Hotel Reservations Forecasting: Curves Similarity Approach." *Journal of Travel Research*, 36 (1): 3-14.

Smith, K., and Pan, B. (2013). "2012 Charleston Area Visitor Intercept Survey Report." Unpublished Report, Office of Tourism Analysis, College of Charleston, Charleston, SC.

Song, H., and G. Li. (2008). "Tourism Demand Modelling and Forecasting—A Review of Recent Research." *Tourism Management*, 29 (2): 203-20.

Song, H., S. F. Witt, and G. Li. (2008). *The Advanced Econometrics of Tourism Demand*. New York: Routledge.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. New York: Random House.

TIA (Travel Industry Association of America). (2009). *Travelers' Use of the Internet, 2009*. Washington, DC: Travel Industry Association of America.

Weatherford, L. R., and S. E. Kimes. (2003). "A Comparison of Forecasting Methods for Hotel Revenue Management." *International Journal of Forecasting*, 19 (3): 401-15.

Wei, W. W. S. (2006). *Time Series Analysis: Univariate and Multivariate Methods*. Boston: Pearson Addison-Wesley.

Xiang, Z., D. R. Fesenmaier, B. Pan, and R. Law. (2010). "Assessing the Visibility of Destination Marketing Organizations in Google: A Case Study of Convention and Visitors Bureau Websites in the United States." *Journal of Travel & Tourism Marketing*, 27 (7): 519-32.

Zakhary, A., A. F. Atiya, H. El-Shishiny, and N. E. Gayar. (2009). "Forecasting Hotel Arrivals and Occupancy Using Monte Carlo Simulation." *Journal of Revenue & Pricing Management*, 10 (4): 344-66.

Zhang, X., H. Fuehres, and P. A. Gloor. (2010). "Predicting Stock Market Indicators through Twitter 'I Hope It Is Not as Bad as I Fear.'" *Procedia: Social and Behavioral Sciences*, 26:55-62.

Zhang, Y., B. J. Jansen, and A. Spink. (2009). "Time Series Analysis of a Web Search Engine Transaction Log." *Information Processing & Management*, 45 (2): 230-45.

Author Biographies

Yang Yang, PhD, is assistant professor in the School of Tourism and Hospitality Management at the Temple University. His areas of research interest include tourism demand analysis, tourist behavior, and location analysis in the hospitality and tourism industry.

Bing Pan, PhD, is associate professor in the Department of Hospitality and Tourism Management at the College of Charleston in Charleston, South Carolina. His research interests include information technologies in tourism, destination marketing, and search engine marketing.

Haiyan Song, PhD, is chair professor of tourism in the School of Hotel and Tourism Management at the Hong Kong Polytechnic University. He has a background in economics with research interests in tourism demand forecasting and tourism impact assessment.