Forecasting Destination Weekly Hotel Occupancy with Big Data

Article *in* Journal of Travel Research · September 2017

DOI: 10.1177/0047287516669050

CITATIONS
13

READS
913

2 authors:

Bing Pan
Pennsylvania State University
113 PUBLICATIONS **5,631** CITATIONS
SEE PROFILE

Yang Yang
Temple University
66 PUBLICATIONS **635** CITATIONS
SEE PROFILE

Preprint. To cite:

Pan, B. & Yang, Y. (2017). Forecasting destination weekly hotel occupancy with big data, Journal of Travel Research, In Press.

Forecasting Destination Weekly Hotel Occupancy with Big Data

Bing Pan, Ph.D.
Associate Professor
Department of Hospitality and Tourism Management
School of Business
College of Charleston, Charleston, SC 29424-001, USA
Telephone: 1-843-953-2025
Fax: 1-843-953-5697
E-mail: bingpan@gmail.com
Visiting Professor
School of Tourism and Environment Sciences
Shaanxi Normal University, Xi'an, China
Yang Yang, Ph.D.
Assistant Professor
School of Tourism and Hospitality Management
Temple University
Telephone: 1-215-204-5030
Fax: 1-215-204-8705
E-mail: yangy@temple.edu
November 29, 2015

Acknowledgements

This research is partially supported by National Natural Science Foundation of China with grant No. 41428101.

Author Bios

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.

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

ABSTRACT

Accurate forecasting of future performance of hotels is needed so hospitality constituencies in specific destinations can benchmark their properties and better optimize operations. As competition increases, hotel managers have urgent need for accurate short-term forecasts. In this study, time series models including several tourism big data sources, including search engine queries, website traffic and weekly weather information, are tested in order to construct an accurate forecasting model of weekly hotel occupancy for a destination. The results show the superiority of ARMAX models with both search engine queries and website traffic data in accurate forecasting. Also, the results suggest that weekly dummies are superior to Fourier terms in capturing the hotel seasonality. The limitations of the inclusion of multiple big data sources are noted since the reduction in forecasting error is minimal.

Keywords: Forecasting, time series, big data, search engine query volume, web traffic

Forecasting Destination Weekly Hotel Occupancy with Big Data

Introduction

The value of accurate forecasting for tourist arrivals and hotel occupancy cannot be overstated (Song and Li 2008, Kim and Schwartz 2013, Schwartz and Hiemstra 1997). Accurate forecasting is a critical component of efficient business operations and destination management. Increased competition and the adoption of revenue management practices have driven demand for accurate forecasting to maximize profits and optimize operations in the hotel industry (Schwartz and Hiemstra 1997). In recent years, hoteliers have highlighted a need for short-term and high frequency forecasting in order to stay agile in a fiercely competitive marketplace (Hayes and Miller 2011).

At the destination level, if a hotelier can anticipate an increase or decrease in average hotel occupancy in one area, she or he can benchmark the property and make appropriate marketing or hiring decisions. However, hotel occupancy in high frequency and smaller geographic areas is always harder to predict (Yang, Pan, and Song 2014). Traditional forecasting methods include time series analysis, statistical methods, neural networks, pickup methods, and simulation (Law 1998, Song and Li 2008). No single method consistently outperforms other methods, and a combination of different forecasting models could perform better than an individual model (Palm and Zellner 1992). In addition, many of these methods use historical patterns to forecast future performance. This limits the accuracy of forecasting when the economic structure changes or unexpected events occur (Yang, Pan, and Song 2014). When trying a variety of diverse models, new types of predictors might be able to help improve forecasting accuracy (Hubbard 2011).

Due to recent information technology advancements, researchers are now able to utilize the digital traces created and left behind by consumers to increase forecasting accuracy for many economic and social phenomena, including general economic indicators (Askitas and Zimmermann 2009), stock market movements (Bollen, Mao, and Zeng 2011), and election outcomes (Metaxas, Mustafaraj, and Gayo-Avello 2011). However, many scholars have also cautioned against becoming overly optimistic about the potential of forecasting using big data because quite often the accuracy is not as good as expected (Lazer et al. 2014).

In the field of tourism and hospitality, researchers have used data on internet searches and web traffic volume to forecast tourist arrivals and hotel occupancy (Bangwayo-Skeete and Skeete 2015, Pan, Wu, and Song 2012, Yang et al. 2015, Yang, Pan, and Song 2014). The results show the validity of different types of online data. However, no researchers have combined multiple data sources. One reason is that some big data sources might be similar to each other. For example, searches on Google for a destination will be highly correlated with website traffic of that destination's tourism bureau (Tierney and Pan 2012). Can multiple sources of big data be used to create a novel method that better predicts hotel occupancy?

In this study, we adopt two methods to capture the seasonality of a destination's weekly hotel occupancy. Combined with multiple sources of big data, including related search engine queries, the local tourism bureau's website traffic data and detailed weather information, we use two different models (Autoregressive Integrated Moving Average with External Variables and a Markov switching dynamic regression model) to predict weekly hotel occupancy for one destination, both with and without big data. The goal of the study is to explore the best way to

predict the weekly hotel occupancy for one destination: What methods perform the best? Does using a combination of multiple big data sources significantly increase forecasting accuracy?

Literature Review

In this section, we review forecasting literature on tourism demand related to three aspects of forecasting: the best time series models in forecasting tourism demand, methods for dealing with seasonality, and efficacy of big data sources for tourism demand forecasting. We first justify the rationale for adopting two different types of forecasting models. Then, we review the techniques that are used to model seasonality in tourism demand. Finally, we review recent literature on how big data are used to forecast economic and social phenomenon, including tourism and hospitality.

Different Models for Tourism Forecasting

Significant progress has been made in time series analysis of tourism demand over the last two decades (Peng, Song, and Crouch 2014). Song and Li (2008) described many modern time series econometric models used for forecasting in tourism and hospitality management. Peng, Song, and Crouch (2014) further classified these models into five categories: basic time series models, advanced time series models, static econometric models, dynamic econometric models, and artificial intelligence models. According to this classification, major advanced time series models include the integrated autoregressive moving average (ARIMA) model, the basic structural model (BSM) and structural time series model (STSM), the generalized autoregressive conditional heteroskedasticity (GARCH) model, and long memory models. ARIMA models incorporate the autoregressive and moving average parts of stationary data (Kulendran and Wong

2005); BSM and STSM models analyze time series by estimating different components (Cortés-Jiménez and Blake 2011, Kulendran and Wong 2011); GARCH models capture the conditional variance (volatility) for exploring the effects of external shocks (Kim and Wong 2006); long memory models apply a fractional order of integration to the data to capture the long-range dependence in time series (Assaf, Barros, and Gil-Alana 2010).

Since some exogenous variables contain valuable information on future trends of tourism and hospitality demand, they can be used as predictors in the forecasting model. Pure time series models can be modified to accommodate predictors; one of the most popular examples is Autoregressive Integrated Moving Average with External Variables models (ARIMAX), which extends the conventional ARIMA model by introducing additional independent variables (Yang, Pan, and Song 2014). Other popular econometric forecasting models with other independent variables as predictors include the autoregressive distributed lag model (ADLM), the error correction model (ECM), the vector autoregressive model (VAR) and the time varying parameter model (TVP) (Song and Li 2008).

Another dynamic econometric model that has gained popularity in the tourism forecasting literature is the Markov switching dynamic regression model (MSDR), which assumes the transition of time series over a finite set of unobserved states with a random time of transition and duration of state. Many researchers have used the MSDR model to unveil the cycles embedded in time series. By assuming that tourist markets are at different points of their lifecycles, Moore and Whitehall (2005) used a Markov switching model with a regime-dependent intercept to analyze quarterly tourist arrivals to Barbados from major source markets.

Their results confirm the different growth phases of different markets. Valadkhani and O'Mahony (2015) incorporated Markov switching components into their empirical model on inbound tourism demand to Australia. Likewise, Chen (2013) adopted a Markov switching model to investigate the tourist hotel industry cycle in Taiwan, and described two regimes: high-growth and low-growth. Using a similar method, Chen, Wu, and Su (2014) and Chen, Lin, and Chen (2015) investigated the hotel business industry cycle and tourism market cycle, respectively. Applying such a model to tourism forecasting, Claveria and Datzira (2010) employed a Markov switching threshold autoregressive model as well as a set of other time series models to forecast international tourism demand in Catalonia with the consumer confidence indicator as an additional predictor. They showed that ARIMA and Markov switching models are superior, and the latter outperforms other competitors in long-run forecasting (6 to 12 months in the future).

Consensus has not been reached in the tourism forecasting literature on the superiority of a single model across different scenarios (Witt and Witt 1995, Peng, Song, and Crouch 2014, Kim and Schwartz 2013). Among pure time series models, when no other predictors are available, the ARIMA model family is recommended if the time series does not have any structural breaks (Peng, Song, and Crouch 2014). Kim and Schwartz (2013) conducted a meta-analysis on the forecasting accuracy of various tourism forecasting models, and found that the causal econometric model generally produces more accurate forecasts than pure time series models. Similarly, Peng, Song, and Crouch (2014) established a meta regression model to unveil factors explaining forecasting errors in the existing literature. The results show that, after controlling for other factors such as origin, destination, time period, sample size and demand measure, dynamic

econometric model forecasts are associated with a lower level of error. Hence, the results from these two meta-analytic studies highlight the importance of external predictors to improving forecast accuracy.

Dealing with Seasonality in Tourism Forecasting Models

A notable characteristic associated with tourism demand, seasonal fluctuations in quarterly, monthly and weekly data, can create complexity in tourism forecasting (Song and Li 2008). Seasonality can be incorporated in the time series model in two ways, by either the stochastic method or the deterministic method (Kulendran and Wong 2005). To treat seasonality as a stochastic component, the data can be seasonally differenced (Kulendran and Wong 2005) or modeled using a state space form with a seasonal component (Song et al. 2011). For the deterministic method, a set of independent variables are included in the model. The most popular way is to introduce a set of dummy variables (Song and Li 2008). An alternative way is to incorporate trigonometric terms such as sine and cosine terms. For example, Stynes and Pigozzi (1983) used sine and cosine terms in their regressions to model seasonality in tourism employment. Chan (1993) showed that a sine wave time series regression generates more accurate forecasts than other models, such as the ARIMA model. Wong (1997) found that a model with a linear trend and two sine functions outperforms other models in forecasting international tourist arrivals. Yang et al. (2014) incorporated the Fourier terms in their empirical model testing the stationarity of tourism demand in Taiwan. Apergis, Mervar, and Payne (2015) demonstrated that the model with Fourier transformation consistently outperforms others in forecasting tourist arrivals to different Croatian regions. Even though some statistical tests can be used to select between deterministic and stochastic methods in modeling seasonality, Kulendran and

Wong (2005) showed that these tests may yield misleading results after evaluating the forecasting performance of different models.

Forecasting with Big Data

One can improve forecasting accuracy by adopting the winning model in the competition among many diverse options. However, once the competition reaches its limit, incorporating powerful predictors is a valid method to further increase forecasting accuracy. Recently, so-called big data have emerged as powerful potential predictors. Big data refers to those data generated from human activity at volumes too large to be handled by traditional computing methods (Mayer-Schonberger and Cukier 2013). Today, travelers are in continuous interaction with information technologies during their travels as well as in their everyday lives. They search for information on the internet, make purchases on websites, bring various gadgets with them on trips, and comment about their experiences on social media. These interactions leave many types of digital traces that indicate their locations, spending habits, preferences and satisfaction levels. These data include search engine queries, website traffic, transaction records, social media posts, and geographic locations. Many studies have been performed in different fields, including tourism and hospitality, on the predictive values of these big datasets.

Search Engine Queries

Due to the large amount of information on the internet, search has become the most popular online activity in the United States (Purcell 2011). Thus, one can use search engine query content and volume to understand and predict human social and economic behavior. Researchers have used search engine query volumes to forecast disease outbreaks (Pelat et al. 2009, Helft 2008, Dugas et al. 2012, Valdivia and Monge-Corella 2010, Ginsberg et al. 2009, Althouse, Ng, and

Cummings 2011), unemployment rates (Askitas and Zimmermann 2009), housing prices and sales (Wu and Brynjolfsson 2014), film revenues (Hand and Judge 2012), and tax evasion rates (Ayers, Ribisl, and Brownstein 2011). In the tourism field, Choi and Varian (2009) used the query tool Google Trend to forecast visitor volumes to Hong Kong; similarly Gawlik, Kabaria, and Kaur (2011) forecasted tourism demand by proposing a query selection process. Pan, Wu, and Song (2012) used Google search data to forecast hotel demand for one destination; likewise, Yang, Pan, Evans, and Lv (2015) used Baidu query volumes to forecast the number of visitors to a province and achieved good results. Bangwayo-Skeete and Skeete adopted Autoregressive Mixed-Data Sampling (AR-MIDAS) models, the Seasonal Autoregressive Integrated Moving Average (SARIMA) ,and autoregressive (AR) approach to investigate the ways incorporating search trend data into the forecasting of tourist arrivals in the Caribbean and they demonstrated the superiority of the AR-MIDAS model (Bangwayo-Skeete and Skeete 2015). These studies have validated the value of using search engine query volume data as a powerful predictor for forecasting social and economic phenomena, including those related to tourism and hospitality.

Website Traffic

For many businesses and organizations, websites serve as virtual storefronts. Normally, a consumer will visit a website prior to making a purchase. Thus, a visit to a website can indicate a business's future revenue or performance. Web log software can help track visits to specific websites, through page-tagging or web server logs (Clifton 2010). Trueman, Wong and Zhang (2001) used website traffic for many internet companies from 1998–2000 to predict their revenue; likewise, Lazer, Lev, and Livnat (2001) found a significant correlation between the web traffic data of publicly-traded internet companies and their portfolio returns. In the field of

tourism, Yang, Pan, and Song (2014) used a local destination marketing organization's web traffic to forecast the average hotel occupancy of the area. They revealed that website traffic data increased the forecasting accuracy by 7 to 10%. Thus, as a precursor to purchasing activities, website traffic data can be a powerful potential predictor.

Weather Information

Weather and climate are crucial considerations in the tourism industry (de Freitas 2003, Becken 2010). It can be a key attraction and also a necessary condition for travel. Thus, forecasted weather conditions could predict future visitor volumes (Frechtling 1996). Agnew, Palutikof, and colleagues (Agnew et al. 2006, Agnew and Palutikof 2001) discovered the correlation between weather conditions and travel demand. Specifically, outbound and inbound travel for British citizens is associated with different weather conditions, such as rain or snow, temperature, and the amount of days with sunlight. Álvarez-Díaz and Rosselló-Nadal (2010) used meteorological variables to predict outbound British visitors to the Balearic islands and achieved good forecasting results. They also found an impact of the North Atlantic Oscillation on domestic demand to Galicia, Spain (Otero-Giráldez, Álvarez-Díaz, and González-Gómez 2012). Similarly, snow conditions have been shown to forecast visitor volumes to ski resorts (Falk 2013). Falk (2014) modeled the impact of weather data on monthly overnight stays in several areas in Europe over 50 years. He found that the numbers of hours with sunshine and temperatures have a significant and positive impact on domestic visitor night stays and its impact on international visitor stays has a 1-year lag. Multiple weather data were also adopted to forecast the travel population in three cities in South Korea and different levels of forecasting accuracy were

achieved (Lee et al. 2015). In general, weather data have become a significant big data source which is useful for forecasting tourism demand.

In conclusion, the ARIMA series of models and the MSDR model have exhibited superior performance compared to other models. Fourier decomposition can be used to model seasonality in tourism demand fluctuations, especially for high-frequency (e.g., monthly) data. In addition, big data sources, such as search engine queries, website traffic, and weather-related information could provide additional external variables for predicting hotel occupancy. However, no study has incorporated all the above methods and big data sources in a weekly forecasting model. In this study, we combined different big data sources and the two aforementioned time series models based on Fourier decomposition or traditional weekly dummies to forecast weekly hotel occupancy for one destination. All these can possibly contribute to the accurate forecasting of weekly hotel occupancy of on area.

Research Methodology

In order to assess the value of the two time series models and the efficacy of big data, we chose Charleston, South Carolina in the United States as our test destination. It is a historic city with around 800,000 residents living in the metropolitan area, and every year it attracts around 5 million visitors (Charleston Area CVB 2015). Located in the coastal area of South Carolina, it boasts a rich history, primarily due to its prominence during the antebellum era. We chose this destination due to convenient access to tourism demand and big data sources related to the destination.

Two Forecasting Models

In this study, we used two types of time series models and incorporated three big data sources to predict hotel occupancy. First we adopted the ARMAX models using big data-related variables as direct predictors. The ARMAX model extends the traditional ARMA model by including additional independent variables as direct predictors. It can be specified as:

$$\begin{aligned} y_t &= \alpha + \mathbf{x}_t\boldsymbol{\beta} + \mu_t \\ \mu_t &= \sum_{i=1}^{m} \rho_i \mu_{t-i} + \sum_{j=1}^{n} \theta_j \varepsilon_{t-j} + \varepsilon_t \end{aligned} \tag{1}$$

where y_t is the dependent variable, and $\mathbf{x}_t$ is a vector of exogenous independent variables (which may include lagged variables). If β is set to 0, the model becomes a standard ARMA(m, n) model. Extensive efforts should be made to specify adequate m and n to guarantee that the model residual (ε_t) is white noise. To determine the lag length of the independent variable, we selected the specification with the smallest value of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) (Liew 2004). Moreover, to capture the seasonality in the hotel demand embedded in the weekly data, we used two alternative modeling methods: 1). the trigonometric representation of seasonal components based on the Fourier series (De Livera, Hyndman, and Snyder 2011) and 2). weekly dummy variables indicating the number of weeks in a year. The Fourier terms of sine and cosine pairs with varying periods have been accepted as a method to control for seasonal variation in the outcome variable (Harvey 1990). We introduced 24 trigonometric terms as the exogenous independent variable in $\mathbf{x}_t$, $\sin(2\pi jt/52.18)$, j = 1, …, 12 and $\cos(2\pi jt/52.18)$, j = 1, …, 12, as an approximation of seasonality. We used 52.18 since a year has an average of 52.18 weeks. For the second method to add week dummies, the order of week is determined by the week's starting date. Therefore, some years may consist of 53 instead 52 weeks, and in this case, we regard 53th week as the 52nd when assigning dummy variables.

The second type of forecasting model used in this paper is the MSDR model. A general MSDR model is specified as

$$y_t = \tau_s + \mathbf{x}_t\boldsymbol{\alpha} + \mathbf{z}_t\boldsymbol{\beta}_s + \varepsilon_s \tag{2}$$

where τ_s is the state-specific intercept; $\mathbf{x}_t$ is a vector of exogenous variables (which may include lagged variables) with state-invariant coefficients; $\mathbf{z}_t$ is a vector of exogenous variables (which may include lagged variables) with state-dependent coefficients $\boldsymbol{\beta}_s$; ε_s is an *i.i.d.* normal error with a mean of 0 and a state-dependent variance of σ^2 (Hamilton 1993). This model allows variables in $\mathbf{z}_t$ with varying parameters across different states to accommodate structural breaks or other multiple-state phenomena, and these unobserved states follow a Markov process. In an aperiodic Markov chain with *J* states, p_{ij} indicates the transitional probability from stage *i* to stage *j* ($j = 1,\ldots, k$), and is specified as follows:

$$\begin{aligned} p_{ij} &= \frac{1}{1+\sum_{m=1}^{k}\exp(-q_{im})} \quad \text{if } j = k \\ p_{ij} &= \frac{\exp(-q_{ij})}{1+\sum_{m=1}^{k}\exp(-q_{im})} \quad \text{if } j \neq k \end{aligned} \tag{3}$$

where q_{ij} is the transformed parameter specified as

$$q_{ij} = -\log\left(\frac{p_{ij}}{p_{ik}}\right) \tag{4}$$

Note that seasonality terms (trigonometric terms or weekly dummies) are also included in $\mathbf{x}_t$ as independent variables. To start the model estimation, an expectation–maximization (EM) algorithm is used to generate the initial value. After that, the estimation proceeds by predicting the probabilities of each state and updating the likelihood at each time (Hamilton 1993). A key

step of this estimation is calculating the likelihood function of latent states, and this function is calculated by iterating the conditional likelihood.

Several rationales support the use of a regime-switching model in the study. First, Yang, Pan, and Song (2014) found that web traffic data, as a source of big data, is particularly useful in reducing forecasting errors during peak seasons. Second, the dynamic econometric model is capable of capturing customers' changing preferences over time (Li et al. 2006). Third, different states and regimes embedded in the MSDM model may represent different stages of the hotel industry growth cycle (Chen 2013). Therefore, it is appropriate to propose different regimes to forecast when leveraging external big data information. In addition, Li, Song, and Witt (2005) found that dynamic econometric models outperform other forecasting models in many cases. Peng, Song, and Crouch (2014) also suggested that a dynamic econometric model such as the MSDR model generates more accurate forecasts based on the results of a meta-analytic regression.

To generate the out-of-sample forecasts, we used the one-step method to predict the value for the next week. For forecasts of 2 and more weeks ahead, we used the dynamic prediction strategy. In the ARMA/ARMAX model, a recursive prediction algorithm is used to predict dependent variable values for later predictions. In the MSDR model, a non-linear filter is used to predict the probability of a state conditional on previous states (Davidson 2004). We focused particularly on improving accuracy by including big data as predictors. Also, we were interested in comparing the usefulness of two seasonality components, Fourier terms and weekly dummies, for improving weekly forecasting accuracy and comparing the forecasting performance between the ARMA(X)

and MSDR models. We used two measures of forecasting accuracy to evaluate the forecasting performance of the models: MAPE and the root mean square percentage error (RMSPE). They are specified as follows:

$$\text{MAPE} = \frac{1}{m}\sum_{t=1}^{m}\left(\frac{|\hat{y}_t - y_t|}{y_t}\right)\times 100\% \quad (5)$$

$$\text{RMSPE} = \sqrt{\frac{1}{m}\sum_{t=1}^{m}\left(\frac{\hat{y}_t - y_t}{y_t}\right)^2}\times 100\% \quad (6)$$

In the tourism forecasting literature, it is a standard practice to benchmark the proposed model against a number of competing models (Song and Li 2008). Analyzing and comparing forecasting accuracy enables forecasting practitioners to choose the proper models. To test the forecasting performance of the different types of models, we split the data sample into two sub-samples: one for model estimation and one for forecast validation. Due to the remarkable seasonality in hotel demand, we suspected that the way in which the sample was divided could have contaminated the evaluation of forecasting performance. Therefore, we performed 52 different splits to evaluate the forecasting performance throughout a year (m = 52 in Equations 5 and 6). The last 52 weeks of the data were used as the validation sub-sample for each split, and the data older than the validation sub-sample were regarded as the estimation sub-sample. It is of particular interest to understand the situations in which situations big data are more useful in reducing forecasting errors. We ran several auxiliary regression models to unveil the factors influencing the degree of improvement measured.

Data Description

In this study, we used average hotel occupancy in the Charleston area as the main time series. Smith Travel Research, Inc. (STR) collects performance data on hotel properties, aggregates them and reports the summary data back to interested properties. In Charleston County, around 70% of hotel properties report their data to STR. Thus, their hotel occupancy rate can be considered a valid surrogate for the real hotel occupancy in the area. STR provided weekly hotel occupancy data for Charleston County from week 1 of 2006 to week 30 of 2015.

There are four types of external variables in this study, three of which are big data series. First, we used Google Correlate (http://correlate.google.com) to identify the most correlated queries with hotel occupancy in Charleston County. "*Charleston hotels*" was the most highly correlated and travel-related query with the hotel occupancy series. Thus, we used the time series data for search engine queries for "*Charleston hotels*." Second, we gathered weekly session data on website traffic for the Charleston Area Convention and Visitors Bureau (CACVB) website using Google Analytics, a free tool provided by Google Inc. in order to track website visitors (Plaza 2011). A session is defined as a continuous period of access to the website from a visitor with more than 30 minutes between adjacent accesses. The web traffic data cover the time span from week 21 of 2007 (when the CACVB website installed Google Analytics) to week 30 of 2015 (the beginning of the research period). Third, we built a custom script in R language to download weather information from http://www.forecast.io. The weather data are based on public weather information gleaned from the National Weather Service for North America and Europe and thus can be considered accurate. We downloaded the daily data and aggregated them into weekly series that included the highest temperature, lowest temperature, average humidity, number of rainy days, and the number of snowy days. Fourth, events or holidays could either boost or

inhibit visitor volumes. For example, festivals can draw a large number of visitors from out of town, whereas people tend to stay home on Christmas and Thanksgiving. We manually categorized around 150 festivals, events, or holidays in Charleston as large positive events (such as the Cooper River Bridge Run and the Wine and Food Festival), small positive events (such as football game days for a local college), and negative events (the days and eves for New Year, Christmas, and Thanksgiving). During New Year, Christmas and Thanksgiving holidays, Americans usually stay with family and less likely to travel, as reflected in the data trends. We coded the number of days in a week with large positive events and small positive events, and whether or not the week had negative events as three different event variables (Table XX).

---------- Insert Table XX here ----------

Table 1 presents descriptive statistics for the dependent and independent variables. The dependent variable is the weekly occupancy rate of Charleston hotels (*occupancy*). The variable *search_engine* is based on the Google Correlate data for *"Charleston hotels*" and web traffic data from the CACVB website, and the variable *web_traffic* is the number of sessions on the CACVB website in the week. For the variable *"search_engine"*, we uploaded hotel occupancy data to Google Correlate and the tool returned the data for "*Charleston hotels*" as the most correlated search engine queries which passed face validity test. The other highly correlated queries, such as "*baseball pants*" or "*ditch witch*" were considered incidental. Thus, only search volume for "*Charleston hotels*" was used. The downloaded data was normalized by Google with a mean of 0 and a standard deviation of 1. After several preliminary runs of the forecasting model, after controlling for seasonality, only the variable *snowy_days* (number of snowy days in a week) was

found to be a valid predictor in some models among a set of weather-related variables. Also, according to the results from preliminary modeling runs and the cross-correlogram at different lags, we eliminated the number of days with large positive events and small positive events in a week because it is not correlated with hotel occupancy. Instead, we used a dummy variable, *negative_event*, to capture any negative events that happened within the week.

As shown in Table 1, the average occupancy rate of Charleston hotels was around 70.7%, with a minimum of 28.5% and a maximum of 90.9% during the research period. Figure 1 presents the time series plots of the dependent variable and major independent variables, and these three series demonstrate very similar patterns of seasonal fluctuation.

(Please insert Table 1 around here)

(Please insert Figure 1 around here)

Results

Unit root tests

Before setting up the forecasting model, we conducted two types of unit root tests, the Phillips-Perron (PP) test and the seasonal augmented Dickey-Fuller (ADF) test, to test the stationarity of major series. Table 2 presents these results. The results of both PP and seasonal ADF tests (Schwert 1989) suggested that the null hypothesis of non-stationarity can be rejected for all tested series, no matter whether a trend term is included. Therefore, first-differencing these variables is not necessary.

(Please insert Table 2 around here)

ARMA/ARMAX model estimation

Table 3 presents the estimation results of the ARMA/ARMAX models. Following the recommendation made by Athanasopoulos et al. (2011), we began by estimating pure time series models in our evaluation of the performance of forecasting models with extra independent variables. Model 1 is the ARMA model with Fourier terms (sine and cosine pairs), and Model 2 is the ARMA model with weekly dummies. The order of AR and MA terms was selected based on (partial) autocorrelation statistics and AIC/BIC values. We also predicted the residuals from these two models. An investigation of the residual correlation and normality diagnostics suggested that the residuals are white noise. As suggested by the lower AIC value of Model 2, the inclusion of weekly dummies is superior than Fourier terms in improving the model's goodness-of-fit. However, according to BIC values, Model 1 is superior with a more parsimonious set of seasonality components. To consider additional exogenous independent variables in forecasting, we use both methods to capture seasonality. Due to space limitation, we present the results of weekly dummy models only, and the estimation results of Fourier term models are available upon request to authors.

(Please insert Table 3 around here)

Before adding any big data-related variables as independent variables, we tested that *negative_event* is statistically insignificant in the ARMAX model. To keep the forecasting model parsimonious, we introduced big data-related independent variables successively. As discussed

previously, the selection of lag order of independent variables was guided by AIC and BIC values. For the ARMAX model, we found that a 2-week lag fits the data best. Model 3 is an ARMAX model with *search_engine* (in a 2-week lag) as an independent variable, and it is estimated to be insignificant. Model 4 considers *web_traffic* (in a 2-week lag), and this variable is estimated to be positive and statistically significant at the 5% level. In Model 5, both *search_engine* (in a 2-week lag) and *web_traffic* (in a 2-week lag) are included simultaneously, and only the latter is estimated to be significant. Lastly, we incorporated *snowy_days* in Model 6, and it is estimated to be insignificant, albeit negative.

MSDR model estimation

Table 4 presents the estimation results of the MSDR models. Based on the results from preliminary modelling runs, we included two autoregressive terms, AR(1) and AR(2), in all MSDR models to capture the dynamics of the dependent variables. The inclusion of more autoregressive terms in the MSDR model makes the optimization procedure hard to converge. We specified only two states/regimes to avoid potential computational difficulties. Note that since we used the MSDR model for forecasting in this study, we are not interested in discussing the transition probabilities within the Markov process based on model estimates. Models 7 and 8 estimate the MSDR models without additional big-data-related independent variables, and the former incorporates Fourier terms whereas the latter includes weekly dummies. Similar to the ARMA results, AIC/BIC values give conflicting results on selecting superior seasonality specifications. Due to space limitation, we present the results of weekly dummy models only, and the estimation results of Fourier term models are available upon request to the authors. In the MSDR model, the autoregressive terms and constant term are specified to vary across different

states. Model 9 includes *search_engine* (in a 2-week lag), and its coefficient is estimated to be significant in both states. Based on AIC and BIC values, Model 9 is superior to Model 8. Model 10 incorporates *web_traffic* (in a 2-week lag) with state-dependent coefficients, and its coefficients are found to be significant and positive. When *search_engine* (in a 2-week lag) and *web_traffic* (in a 2-week lag) are considered in Model 11, these variables are estimated to be significant in only State 1. Lastly, in Model 12, the coefficient of *snowy_days* is statistically significant and negative in State 2.

(Please insert Table 4 around here)

Comparison of forecasting performance

Using the model specifications in Tables 3 and 4, we re-estimated the model using a different estimation sub-sample from the 52 data splits, and generated forecasts for the last 52 weeks in our original data set as *ex ante* forecasts. Because many of our forecasting models cover independent variables in 2-week lags, they are able to generate forecasts up to two steps only. We also assumed that it was possible to predict the number of snowy days that would occur within a 2-week period. The *ex ante* forecasting performance measures for all models are presented in Table 5 (for MAPE measures) and Table 6 (for RMSPE measures). Also, we conducted the Diebold-Mariano (DM) test to compare the predictive accuracy of any two forecasting models. The better-performing models are underlined in the table based on certain benchmark models. Similar conclusions are reached if we use MAPE and RMSPE to select the best model.

First, models using weekly dummies (Models 2 and 8, respectively) outperforms their counterparts incorporating Fourier terms (Models 1 and 7, respectively) in terms of forecasting accuracy, and the average MAPE and RMSPE values reduce from Models 1 and 7 to Model 2 and 8, respectively, in either one-step or two-step forecasts. In particular, the DM test suggests that this improvement in accuracy is significant as measured by RMSPE in ARMA model and by both MAPE and RMSPE in MSDR models. Second, it shows that the combination of search query volume and web traffic data are particularly useful for forecasting the hotel demand series in Charleston as indicated by the lowest MAPE and RMSPE values of Model 5 across all ARMA(X) models (Models 1–6) for both one-step or two-step forecasts and that of Model 11 across all MSDR models (Models 7-12) for one-step forecasts. As indicated by the DM test, the forecasting accuracy of Model 5 is not significantly superior than that of Model 2 without big-data-related variables, but Model 11 generate significantly better forecasts than Model 7that does not incorporate any big-data-related variables. Third, the MSDR model is slightly inferior to its counterpart ARMA(X) model in terms of forecasting accuracy. Lastly, the forecasting performance data of Fourier term model with big-data-related variables are not presented but are available upon request. We find that the forecasting accuracy of Fourier term models is consistently inferior to the counterpart model with weekly dummies.

(Please insert Table 5 around here)

(Please insert Table 6 around here)

Generally, as indicated by the results in Tables 5 and 6, Models 3–5 outperform Model 2, and Models 9-11 outperform Model 8. This result highlights the usefulness of big data in improving

performance accuracy and alleviating the risk of yielding large forecasting errors. Compared to *search_query*, we found *web_traffic* is more helpful in reducing forecasting error as indicated by the lower MAPE/RMSPE values of Model 4 (Model 10) compared to Model 3 (Model 9). This result confirms the high value of web traffic data in improving forecasting accuracy of hotel demand. If both *search_query* and *web_traffic* are incorporated in the forecasting model, it generates the forecasts with lowest average errors.

Conclusion

The objective of the study is to investigate the best modeling technique for forecasting weekly hotel occupancy from big data sources. This study confirms the validity of two time series models (ARMAX and MSDR) in forecasting a destination's hotel occupancy rate (Song and Witt, 2000; Chen, Wu, and Su, 2014). Especially, ARMAX models are superior than MSDR models in forecasting accuracy in respective configuration, consistent with previous studies (Peng, Song, and Crouch 2014). Further, to capture the weekly seasonality in the hospitality time-series data, weekly dummies were found to outperform Fourier decomposition with sine and cosine terms. Combined, the ARMAX models with web traffic and search query volume information produce fairly accurate forecasts of average hotel occupancy for a destination 1 to 2 weeks in advance. The MAPE and RMSPE for out-of-sample forecasting could reach around 3.7% in both time periods for out-of-sample forecasts.

This study also demonstrates the limitations of tourism big data: individually, search engine query volume and website traffic are good predictors that help reduce forecasting errors;

combined, the rate of error reductions is not as significant as expected. Compared web traffic data, search engine query volume data only moderately reduce the forecasting error. However, in the best ARMAX model, only website traffic data is useful in reducing the error rate; search engine query volume, possibly due to high correlations between other big data sources and Fourier terms successfully capture the seasonality in tourism demand.

Regarding weather-related data, almost all series are nonsignificant in forecasting weekly occupancy. In one of the MSDR models, the number of snowy days is correlated with hotel occupancy and helpful in moderately increasing forecasting accuracy; however, over nearly 10 years, there were only 4 weeks with snowy days, since the destination is located in the southeastern United States where it rarely snows. The best models do not contain any weather-related series. Thus, weather information can hardly be described as effective in forecasting occupancy. This is different from previous studies in showing monthly weather-related information are correlated with tourism demand (Álvarez-Díaz and Rosselló-Nadal 2010, Agnew et al. 2006). The difference in this study might be due to the fact that travel decisions are usually made at least two or three weeks ahead (Yang et al. 2015). Current weather forecast is only accurate in at most five days to one week's forecast (Silver, 2012). Thus, the current weather can hardly impact the visitor volumes when they already reached their destination. This highlights the difficulty of short-term and high frequency forecasting for tourism demand. When multiple data sources are highly correlated, the increased prediction power is also limited. This is also in line with recent studies showing the limitation of big data when Google Flu Trends project lost its accuracy due to changes in user behavior and search interface (Lazer et al. 2014, Hodson 2014).

In practice, one of the authors has been producing the weekly hotel occupancy and average daily rate forecasting for the Charleston area for half a year with ARIMA models. The authors are working on incorporating big data and seasonality components to increase the forecasting accuracy. It is imaginable that other cities and areas with a strong seasonality could incorporate the methodology in this research into their forecasting practices. However, more diverse big datasets should be collected and integrated in order to further refine the forecasting models.

Discussion and Future Research

A limitation of the study lies in the single chosen destination. The destination is unique in that it has distinct seasonality patterns. Other destinations may have irregular patterns which may decrease the efficacy of our methods when modeling seasonality. Second, the cost of configuring and running the model is rather high. With additional seasonality components and big data series, it would take a rather long time to configure 12 different models. Especially for the MSDR models, a personal computer with the current configuration would run for at least a few hours. Third, the big data we chose are also rather limited: we chose only one search query to represent search engine volume data. Although the query is the most significantly correlated one, other nonobvious queries may contribute more to forecasting accuracy.

There are several opportunities for future research in this area. There has been much hype around the potential predictive power of big data in recent years. However, this study shows that big data sources do have limitations. Therefore, researchers need to seek diverse big data sources in order to further increase forecasting accuracy. For example, other big data sources that are very

different from search engine queries and website traffic data, such as social media mentions or mobile phone data, could possibly yield more valuable predictors.

In addition, considering the actual needs of local hospitality establishments, forecasting extreme values is typically more important and valuable. For example, accurately forecasting a sudden drop in hotel occupancy due to severe weather or a dramatic occupancy increase due to an unexpected event would be extremely valuable to a hotelier. Although extreme values (which might be outliers) would be more enlightening, MAPE or RMSPE represent the average accuracy across the entire forecasting period. Using MAPE or RMSPE to measure model efficacy thus might not be appropriate, in which case other types of model fit measurements need to be researched and constructed.

Finally, it is most valuable when one can improve the forecasting accuracy of demand for a single property with big data. Tens of thousands of hotels in the United States could benefit from increased accuracy because it could help them make better marketing or operational decisions. Working with individual properties and incorporating big data sources would be another fruitful research direction.

References

Agnew, Maureen D, and Jean P Palutikof. 2001. "Climate impacts on the demand for tourism." International Society of Biometeorology Proceedings of the First International Workshop on Climate, Tourism and Recreation. Retrieved from http://www. mif. uni-freiburg. de/isb/ws/report. htm.

Agnew, MD, JP Palutikof, C Hanson, J Palutikof, and AH Perry. 2006. "Impacts of short-term climate variability in the UK on demand for domestic and international tourism." *Climate Research* 31 (1):109-120.

Althouse, Benjamin M, Yih Yng Ng, and DA Cummings. 2011. "Prediction of dengue incidence using search query surveillance." *PLoS Negl Trop Dis* 5 (8):e1258.
Álvarez-Díaz, Marcos, and Jaume Rosselló-Nadal. 2010. "Forecasting British tourist arrivals in the Balearic Islands using meteorological variables." *Tourism Economics* 16 (1):153-168.
Apergis, Nicholas, Andrea Mervar, and James E. Payne. 2015. "Forecasting disaggregated tourist arrivals in Croatia: Evidence from seasonal univariate time series models." *Tourism Economics*:doi: 10.5367/te.2015.0499.
Askitas, N, and KF Zimmermann. 2009. "Google econometrics and unemployment forecasting." *Applied Economics Quarterly* 55 (2):107-120.
Assaf, A. George, Carlos Pestana Barros, and Luis A. Gil-Alana. 2010. "Persistence in the short- and long-term tourist arrivals to Australia." *Journal of Travel Research*.
Athanasopoulos, George, Rob J. Hyndman, Haiyan Song, and Doris C. Wu. 2011. "The tourism forecasting competition." *International Journal of forecasting* 27 (3):822-844.
Ayers, John W, Kurt Ribisl, and John S Brownstein. 2011. "Using search query surveillance to monitor tax avoidance and smoking cessation following the United States' 2009" SCHIP" cigarette tax increase." *PloS one* 6 (3):e16777.
Bangwayo-Skeete, Prosper F, and Ryan W Skeete. 2015. "Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach." *Tourism Management* 46:454-464.
Becken, Susanne. 2010. "The importance of climate and weather for tourism: literature review."
Bollen, J., H. Mao, and X. Zeng. 2011. "Twitter mood predicts the stock market." *Journal of Computational Science* 2 (1):1-8.
Chan, Yiu-Man. 1993. "Forecasting tourism: A sine wave time series regression approach." *Journal of Travel Research* 32 (2):58-60. doi: 10.1177/004728759303200209.
Charleston Area CVB. 2015. *2015-2016 Charleston Area Convention & Visitors Bureau Book*. Charleston: Charleston Convention & Visitors Bureau.
Chen, Ming-Hsiang. 2013. "Determinants of the Taiwanese tourist hotel industry cycle." *Tourism Management* 38:15-19. doi: http://dx.doi.org/10.1016/j.tourman.2013.01.003.
Chen, Ming-Hsiang, Chien-Pang Lin, and Brendan T. Chen. 2015. "Drivers of Taiwan's tourism market cycle." *Journal of Travel & Tourism Marketing* 32 (3):260-275. doi: 10.1080/10548408.2014.896764.
Chen, Ming-Hsiang, Kun Lun Wu, and Hung-Jen Su. 2014. "A study of the business cycle of the hotel industry in Taiwan." *Tourism Economics* 20 (3):655-664. doi: 10.5367/te.2013.0287.
Choi, Hyunyoung, and Hal Varian. 2009. Predicting the Present with Google Trends. Accessed April 2, 2009.
Claveria, Oscar, and Jordi Datzira. 2010. "Forecasting tourism demand using consumer expectations." *Tourism Review* 65 (1):18-36. doi: doi:10.1108/16605371011040889.
Clifton, B. 2010. *Advanced Web metrics with Google analytics*: Sybex.
Cortés-Jiménez, Isabel, and Adam Blake. 2011. "Tourism demand modeling by purpose of visit and nationality." *Journal of Travel Research* 50 (4):408-416.
Davidson, James. 2004. "Forecasting Markov-switching dynamic, conditionally heteroscedastic processes." *Statistics & probability letters* 68 (2):137-147.

de Freitas, Chris R. 2003. "Tourism climatology: evaluating environmental information for decision making and business planning in the recreation and tourism sector." *international Journal of Biometeorology* 48 (1):45-54.
De Livera, Alysha M., Rob J. Hyndman, and Ralph D. Snyder. 2011. "Forecasting time series with complex seasonal patterns using exponential smoothing." *Journal of the American Statistical Association* 106 (496):1513-1527.
Dugas, Andrea Freyer, Yu-Hsiang Hsieh, Scott R Levin, Jesse M Pines, Darren P Mareiniss, Amir Mohareb, Charlotte A Gaydos, Trish M Perl, and Richard E Rothman. 2012. "Google Flu Trends: correlation with emergency department influenza rates and crowding metrics." *Clinical infectious diseases* 54 (4):463-469.
Falk, Martin. 2013. "Impact of long-term weather on domestic and foreign winter tourism demand." *International journal of tourism research* 15 (1):1-17.
Falk, Martin. 2014. "Impact of weather conditions on tourism demand in the peak summer season over the last 50years." *Tourism Management Perspectives* 9:24-35.
Frechtling, DC. 1996. *Practical tourism forecasting*: Elsevier.
Gawlik, Evan, Hardik Kabaria, and Shagandeep Kaur. 2011. "Predicting tourism trends with Google Insights."
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):1012-1014.
Hamilton, James D. 1993. "Estimation, inference and forecasting of time series subject to changes in regime." In *Handbook of Statistics 11: Econometrics*, edited by G. S. Maddala, C. R. Rao and H. D. Vinod, 231-260. San Diego, CA: Elsevier.
Hand, Chris, and Guy Judge. 2012. "Searching for the picture: forecasting UK cinema admissions using Google Trends data." *Applied Economics Letters* 19 (11):1051-1055.
Harvey, Andrew C. 1990. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge, UK: Cambridge university press.
Hayes, David K, and Allisha A Miller. 2011. *Revenue management for the hospitality industry*: Wiley River Street Hoboken, NJ.
Helft, Miguel. 2008. "Google uses searches to track fluâ€™s spread." *New York Times. Retrieved January* 1:2009.
Hodson, Hal. 2014. "Google Flu Trends gets it wrong three years running." *New Scientist* 221 (2961):24.
Hubbard, D.W. 2011. *Pulse: The New Science of Harnessing Internet Buzz to Track Threats and Opportunities*: Wiley.
Kim, Namhyun, and Zvi Schwartz. 2013. "The accuracy of tourism forecasting and data characteristics: a meta-analytical approach." *Journal of Hospitality Marketing & Management* 22 (4):349-374.
Kim, Samuel Seongseop, and Kevin K. F. Wong. 2006. "Effects of news shock on inbound tourist demand volatility in Korea." *Journal of Travel Research* 44 (4):457-466.
Kulendran, Nada, and Kevin K. F. Wong. 2005. "Modeling seasonality in tourism forecasting." *Journal of Travel Research* 44 (2):163-170.
Kulendran, Nada, and Kevin K. F. Wong. 2011. "Determinants versus composite leading indicators in predicting turning points in growth cycle." *Journal of Travel Research* 50 (4):417-430.

Law, R. 1998. "Room occupancy rate forecasting: a neural network approach." *International Journal of Contemporary Hospitality Management* 10 (6):234-239.
Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The parable of Google Flu: traps in big data analysis." *Science* 343 (14 March).
Lee, Kyungmin, Bonghee Hong, Jiwan Lee, and Yangja Jang. 2015. "A Floating Population Prediction Model in Travel Spots Using Weather Big Data." Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on.
Li, Gang, Haiyan Song, and Stephen F. Witt. 2005. "Recent developments in econometric modeling and forecasting." *Journal of Travel Research* 44 (1):82-99.
Li, Gang, Kevin K. F. Wong, Haiyan Song, and Stephen F. Witt. 2006. "Tourism demand forecasting: A time varying parameter error correction model." *Journal of Travel Research* 45 (2):175-185.
Liew, Venus Khim-Sen. 2004. "Which lag length selection criteria should we employ?" *Economics Bulletin* 3 (33):1-9.
Mayer-Schonberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*: Eamon Dolan/Houghton Mifflin Harcourt.
Metaxas, Panagiotis T, Eni Mustafaraj, and Daniel Gayo-Avello. 2011. "How (not) to predict elections." Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on.
Moore, Winston, and Peter Whitehall. 2005. "The tourism area lifecycle and regime switching models." *Annals of Tourism Research* 32 (1):112-126. doi: http://dx.doi.org/10.1016/j.annals.2004.05.006.
Otero-Giráldez, María Soledad, Marcos Álvarez-Díaz, and Manuel González-Gómez. 2012. "Estimating the long-run effects of socioeconomic and meteorological factors on the domestic tourism demand for Galicia (Spain)." *Tourism Management* 33 (6):1301-1308.
Palm, FC, and A Zellner. 1992. "To combine or not to combine? Issues of combining forecasts." *Journal of Forecasting* 11:687-701.
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.
Pelat, Camille, Clement Turbelin, Avner Bar-Hen, Antoine Flahault, and Alain-Jacques Valleron. 2009. "More diseases tracked by using Google Trends." *Emerging infectious diseases* 15 (8):1327.
Peng, Bo, Haiyan Song, and Geoffrey I. Crouch. 2014. "A meta-analysis of international tourism demand forecasting and implications for practice." *Tourism Management* 45:181-193. doi: http://dx.doi.org/10.1016/j.tourman.2014.04.005.
Plaza, B. 2011. "Google Analytics for measuring website performance." *Tourism Management* 32 (3):477-481.
Purcell, K. 2011. Search and email still top the list of most popular online activities. *Pew Research Internet Project: Report*. Accessed September 1, 2014.
Schwartz, Z., and S. Hiemstra. 1997. "Improving the accuracy of hotel reservations forecasting: Curves similarity approach." *Journal of Travel Research* 36 (1):3-14.
Schwert, G. William. 1989. "Tests for unit roots: A Monte Carlo investigation." *Journal of Business & Economic Statistics* 7 (2):5-17.
Silver, Nate. 2012. *"The signal and the noise: Why so many predictions fail-but some don't"*. Penguin.

Song, H, and G Li. 2008. "Tourism demand modelling and forecasting—A review of recent research." *Tourism Management* 29 (2):203-220.
Song, Haiyan, Gang Li, Stephen F. Witt, and George Athanasopoulos. 2011. "Forecasting tourist arrivals using time-varying parameter structural time series models." *International Journal of forecasting* 27 (3):855-869.
Stynes, Barbara White, and Bruce William Pigozzi. 1983. "A tool for investigating tourism-related seasonal employment." *Journal of Travel Research* 21 (3):19-24.
Tierney, Heather L.R., and Bing Pan. 2012. "A poisson regression examination of the relationship between website traffic and search engine queries." *NETNOMICS: Economic Research and Electronic Networking* 13 (3):155-189.
Trueman, Brett, MH Franco Wong, and Xiao-Jun Zhang. 2001. "Back to basics: Forecasting the revenues of Internet firms." *Review of Accounting Studies* 6 (2-3):305-329.
Valadkhani, Abbas, and Barry O'Mahony. 2015. "Identifying structural changes and regime switching in growing and declining inbound tourism markets in Australia." *Current Issues in Tourism*:1-24. doi: 10.1080/13683500.2015.1072504.
Valdivia, Antonio, and Susana Monge-Corella. 2010. "Diseases tracked by using Google trends, Spain." *Emerg Infect Dis* 16 (1):168.
Witt, Stephen F., and Christine A. Witt. 1995. "Forecasting tourism demand: A review of empirical research." *International Journal of forecasting* 11 (3):447-475.
Wong, Kevin K. F. 1997. "The relevance of business cycles in forecasting international tourist arrivals." *Tourism Management* 18 (8):581-586. doi: http://dx.doi.org/10.1016/S0261-5177(97)00073-3.
Wu, Lynn, and Erik Brynjolfsson. 2014. "The future of prediction: How Google searches foreshadow housing prices and sales." In *Economic Analysis of the Digital Economy*. University of Chicago Press.
Yang, Ginny Ju-Ann, Yung-Hsiang Ying, Koyin Chang, and Chen-Hsun Lee. 2014. "Investigating stationarity in tourist arrivals to Taiwan using panel Kpss with sharp drifts and smooth breaks." *Tourism Analysis* 19 (5):573-580. doi: 10.3727/108354214x14116690097819.
Yang, Xin, Bing Pan, James A Evans, and Benfu Lv. 2015. "Forecasting Chinese tourist volume with search engine data." *Tourism Management* 46:386-397.
Yang, Yang, Bing Pan, and Haiyan Song. 2014. "Predicting Hotel Demand Using Destination Marketing Organization's Web Traffic Data." *Journal of Travel Research* 53 (4):433-447.

TABLES AND FIGURES

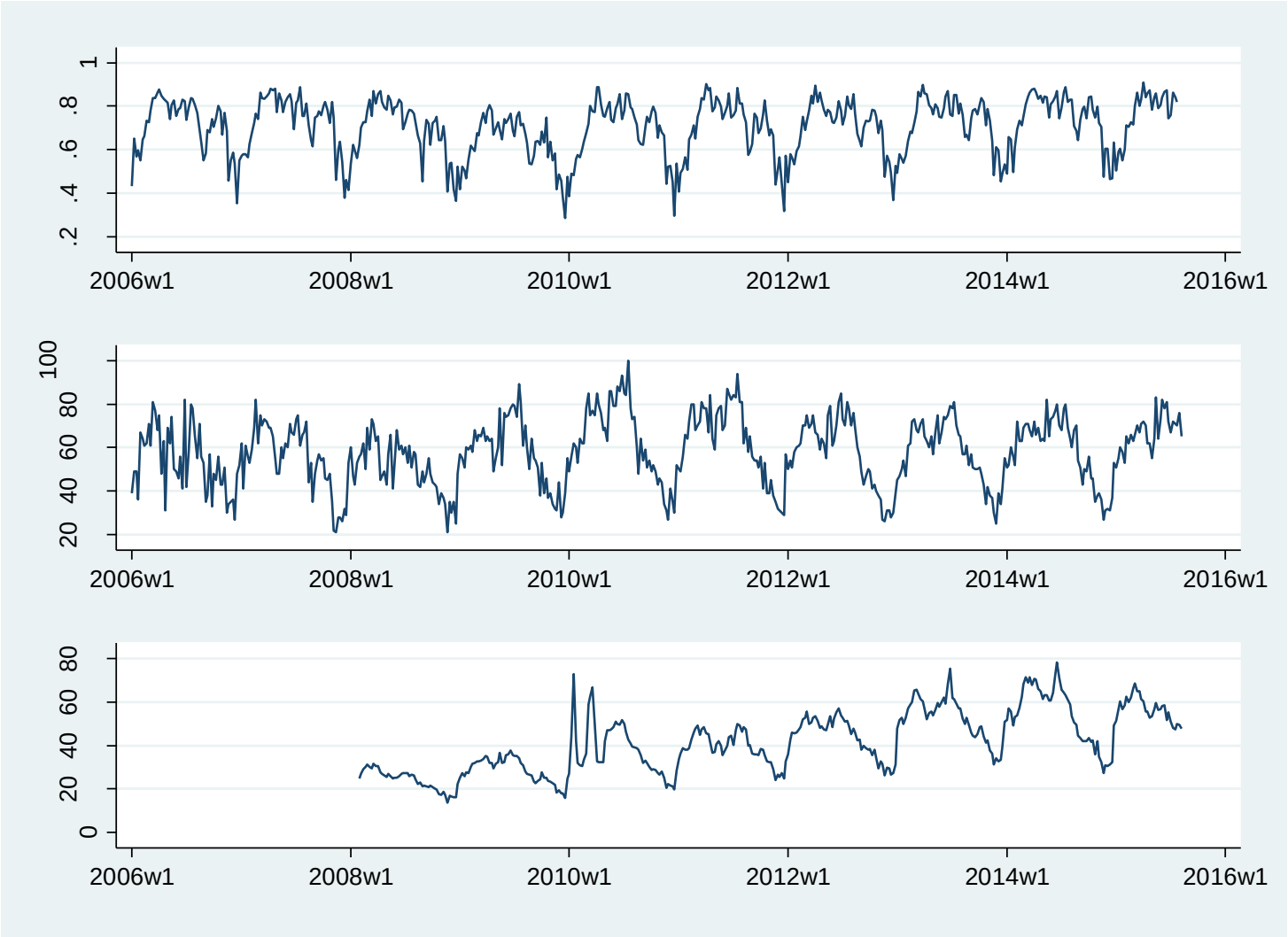

Figure 1. Time series plots of dependent and major independent variables

Table XX. Large Positive, Small Positive and Negative Events

Large Positive Events	Small Positive Events	Negative Events
Family Circle Cup, Fashion Week, HCF Festival of Houses and Garden, Independence Day, Labor Day, Memorial Day, PGA Championship, Southeast Wildlife Expo, Spoleto Festival, Wine & Food Festival.	Citadel Corps Days, Citadel Football Game, Citadel Graduation, Citadel Homecoming, Citadel Move In, Citadel Parents Weekend, College of Charleston Accepted Students Weekend, College of Charleston Family Weekend, College of Charleston Graduation, College of Charleston Homecoming, College of Charleston Move In, College of Charleston Orientation, Columbus Day, Easter, Fall Tour of Homes Gardens, Garden and Gun Jubilee, Historic Charleston Foundation Charleston Antiques Show, Veterans Day.	Christmas Day, Christmas Eve, New Year's Day, New Year's Eve, Thanksgiving.

Table 1. Descriptive statistics of variables

Variable	Period	Obs	Mean	Std. Dev.	Min	Max
occupancy	w1/2006-w30/2015	500	0.707	0.128	0.285	0.909
*search_engine**	w1/2006-w30/2015	500	57.970	15.501	21	100
*web_traffic***	w5/2008-w30/2015	391	41458.010	14103.170	13858	78054
snowy_days	w1/2006-w30/2015	500	0.008	0.109	0	2
negative_event	w1/2006-w30/2015	500	0.058	0.234	0	1

**Search_Engine* is scaled search volume where the largest number in a time series is set at 100;
** *Web_traffic* is the number of user sessions on Charleston Area Convention and Visitors Bureau website.

Table 2. Results of unit root tests

Variable	PP test statistic		Seasonal ADF test statistic	
	non-trend	trend	non-trend	trend
occupancy	-6.955***	-6.969***	-6.207***	-6.233***
search_engine	-6.607***	-7.942***	-4.962***	-5.543***
web_traffic	-3.331***	-4.563***	-3.540***	-4.965***
snowy_days	-18.955***	-18.997***	-14.890***	-14.940***
negative_event	-16.091***	-16.081***	-8.765***	-8.770***

Notes: *** indicates significance at the 1% level. Lag number in unit root tests is determined by BIC values.

Table 3. Estimation results of ARMA/ARMAX models

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
search_engine(L2)			0.000362		0.000434	
			(0.000288)		(0.000399)	
web_traffic(L2)				0.000000937**	0.000000858*	
				(0.000000477)	(0.000000475)	
snowy_days						-0.0327
						(0.0201)
constant	0.709***	0.381***	0.550***	0.542***	0.721***	0.381***
	(0.0165)	(0.0177)	(0.0302)	(0.0327)	(0.0537)	(0.0177)
AR(1)	-0.160***	0.0537	0.0546	0.161***	0.159***	0.0619
	(0.0420)	(0.0441)	(0.0442)	(0.0595)	(0.0596)	(0.0442)
AR(2)	0.306***	0.446***	0.442***	0.478***	0.485***	0.455***
	(0.0633)	(0.0891)	(0.0893)	(0.123)	(0.117)	(0.0903)
AR(3)	0.309***	0.131**	0.135**	0.0534	0.0479	0.123**
	(0.0426)	(0.0544)	(0.0546)	(0.0712)	(0.0701)	(0.0543)
AR(4)	0.473***	0.299***	0.303***	0.247***	0.257***	0.292***
	(0.0404)	(0.0556)	(0.0545)	(0.0707)	(0.0689)	(0.0568)
MA(2)	-0.355***	-0.380***	-0.397***	-0.403***	-0.432***	-0.383***
	(0.0772)	(0.0984)	(0.0996)	(0.126)	(0.121)	(0.0991)
Fourier terms	Yes	No	No	No	No	No
Weekly dummies	No	Yes	Yes	Yes	Yes	Yes
σ	0.0393***	0.0357***	0.0355***	0.0311***	0.0310***	0.0355***
	(0.00132)	(0.00120)	(0.00119)	(0.00129)	(0.00130)	(0.00118)
N	473	473	471	362	362	473
AIC	-1656.1	-1693.5	-1689.5	-1366.7	-1366.4	-1696.7
BIC	-1527.2	-1452.3	-1444.4	-1137.1	-1132.9	-1451.3

Notes: *** indicates significance at the 1% level; ** indicates significance at the 5% level; * indicates significance at the 10% level. Standard errors are presented in parentheses. *L2* indicates the variable in a 2-week lag.

Table 4. Estimation results of MSDR models

	Model 7	**Model 8**	**Model 9**	**Model 10**	**Model 11**	**Model 12**
State 1						
AR(1)	0.0105	0.589***	0.587***	0.361***	0.303***	0.594***
	(0.0542)	(0.0702)	(0.0696)	(0.0482)	(0.0560)	(0.0698)
AR(2)	0.240***	0.151***	0.124**	0.247***	0.321***	0.155***
	(0.0553)	(0.0573)	(0.0602)	(0.0472)	(0.0555)	(0.0565)
search_engine(L2)			0.000780**		-0.000633**	
			(0.000320)		(0.000285)	
web_traffic(L2)				0.000000851***	0.00000107***	
				(0.000000173)	(0.000000226)	
snowy_days						-0.0347
						(0.0223)
constant	0.552***	0.118***	0.105***	0.282***	0.279***	0.114***
	(0.0416)	(0.0290)	(0.0299)	(0.0210)	(0.0228)	(0.0290)
State 2						
AR(1)	0.500***	0.0960*	0.177***	0.358***	0.703***	0.0954*
	(0.0541)	(0.0578)	(0.0549)	(0.114)	(0.0836)	(0.0576)
AR(2)	0.196***	0.449***	0.484***	0.360***	0.0511	0.460***
	(0.0481)	(0.0555)	(0.0537)	(0.104)	(0.0729)	(0.0551)
search_engine(L2)			-0.000933***		0.000253	
			(0.000352)		(0.000403)	
web_traffic(L2)				0.00000118**	0.000000265	
				(0.000000571)	(0.000000388)	
snowy_days						-0.0620**
						(0.0244)
constant	0.193***	0.290***	0.272***	0.133***	0.126***	0.285***
	(0.0397)	(0.0259)	(0.0254)	(0.0413)	(0.0324)	(0.0257)
Fourier terms	Yes	No	No	No	No	No
Weekly dummies	No	Yes	Yes	Yes	Yes	Yes
σ	0.0382***	0.0338***	0.0330***	0.0271***	0.0275***	0.0335***
	(0.00169)	(0.00181)	(0.00145)	(0.00159)	(0.00140)	(0.00178)
N	471	471	471	362	362	471
AIC	-1493.9	-1593.3	-1604.3	-1315.4	-1315.1	-1596.5
BIC	-1356.8	-1344.0	-1346.7	-1074.1	-1066.1	-1338.9

Notes: *** indicates significance at the 1% level; ** indicates significance at the 5% level; * indicates significance at the 10% level. Standard errors are presented in parentheses. *L2* indicates the variable in a 2-week lag.

Table 5. Forecasting performance (MAPE) of models

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
Big data predictor			*a*	*b*	*a,b*	*c*
Fourier terms	Yes	No	No	No	No	No
Weekly dummies	No	Yes	Yes	Yes	Yes	Yes
MAPE (one-step)	4.276%	3.851%	3.804%	3.687%	*3.669%*	3.879%
DM test †	0.818		-1.0742	-0.745	-0.856	0.307
MAPE (two-step)	4.399%	3.860%	3.806%	3.716%	*3.698%*	3.878%
DM test †	0.969		-1.419*	-0.748	-0.851	0.712
	Model 7	**Model 8**	**Model 9**	**Model 10**	**Model 11**	**Model 12**
Big data predictor			*a*	*b*	*a,b*	*c*
Fourier terms	Yes	No	No	No	No	No
Weekly dummies	No	Yes	Yes	Yes	Yes	Yes
MAPE (one-step)	7.159%	4.589%	4.714%	4.090%	*4.025%*	4.642%
DM test ††	2.121**		0.446	-1.407*	-1.729**	0.764
MAPE (two-step)	6.962%	4.605%	4.480%	*3.930%*	4.285%	4.586%
DM test††	1.574*		-0.350	-2.437***	-1.426*	-0.531

Notes: Smallest forecasting error in each row is underlined; *a* indicates *search_engine(L2)*; *b* indicates *web_traffic(L2); c* indicates *snowy_days*. † indicates the result compared to Model 2, and †† indicates the result compared to Model 8. *** indicates significance at the 1% level; ** indicates significance at the 5% level; * indicates significance at the 10% level based on one-side test.

Table 6. Forecasting performance (RMSPE) of models

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
Big data predictor			*a*	*b*	*a,b*	*c*
Fourier terms	Yes	No	No	No	No	No
Weekly dummies	No	Yes	Yes	Yes	Yes	Yes
RMSPE (one-step)	6.102%	5.018%	4.996%	4.787%	*4.781%*	5.066%
DM test†	1.441*		-0.837	-0.417	-0.533	0.864
RMSPE (two-step)	6.248%	5.015%	4.992%	4.821%	*4.809%*	5.063%
DM test†	1.497*		-0.829	-0.288	-0.434	0.906
	Model 7	**Model 8**	**Model 9**	**Model 10**	**Model 11**	**Model 12**
Big data predictor			*a*	*b*	*a,b*	*c*
Fourier terms	Yes	No	No	No	No	No
Weekly dummies	No	Yes	Yes	Yes	Yes	Yes
RMSPE (one-step)	13.251%	5.988%	6.053%	5.239%	*4.925%*	6.001%
DM test††	1.708**		0.820	-1.025	-1.526*	-0.0935
RMSPE (two-step)	13.459%	6.440%	5.966%	*5.105%*	5.697%	6.406%
DM test††	1.299*		-0.717	-1.826**	-1.627*	-1.1819

Notes: Smallest forecasting error in each row is underlined; *a* indicates *search_engine(L2)*; *b* indicates *web_traffic(L2); c* indicates *snowy_days*. † indicates the result compared to Model 2, and †† indicates the result compared to Model 8. *** indicates significance at the 1% level; ** indicates significance at the 5% level; * indicates significance at the 10% level based on one-side test.