FORECASTING HOTEL ROOM DEMAND USING SEARCH ENGINE DATA

Bing Pan*, Ph.D.
Assistant 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

Chenguang (Doris) Wu, Ph.D.
School of Management
Sun Yat-sen University
Guangzhou, China
Telephone: 86-186-6608-0105
Fax: 86-20-8411-2686
E-Mail: wuchenguang@gmail.com

Haiyan Song, Ph.D.
Professor
School of Hotel and Tourism Management
The Hong Kong Polytechnic University, Kowloon, Hong Kong
Telephone: 852-2766-6372
Fax: 852-2362-9362
E-mail: Haiyan.Song@polyu.edu.hk

* Corresponding author (bingpan@gmail.com)

Submitted exclusively to Journal of Hospitality and Tourism Technology

March 1, 2012
AUTOBIOGRAPHICAL NOTES

Bing Pan, Ph.D., 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.

Chenguang (Doris) Wu, Ph.D., is assistant professor in the School of Management at the Sun Yat-sen University. P. R. China. Her research interests include applied econometrics and tourism forecasting.

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

Acknowledgements

The authors would like to thank Danice Ng for research assistance and the preparation of an early draft of the paper. The financial support the Hong Kong Polytechnic University (Grant No. G-U743) is also acknowledged.
FORECASTING HOTEL ROOM DEMAND USING SEARCH ENGINE DATA

ABSTRACT

Purpose
We investigated the usefulness of search query volume data in forecasting demand for hotel rooms and sought to identify the best econometric forecasting model.

Design/methodology/approach
We used search volume data on five related queries to predict demand for hotel rooms in a specific tourist city. We employed three ARMA family models and their ARMAX counterparts to evaluate the usefulness of these data. We also evaluated three widely used causal econometric models—ADL, TVP, and VAR—for comparison.

Findings
All three ARMAX models consistently outperformed their ARMA counterparts, validating the value of search volume data in facilitating the accurate prediction of demand for hotel rooms. When the three causal econometric models were included for forecasting competition, the ARX model produced the most accurate forecasts, suggesting its usefulness in forecasting demand for hotel rooms.

Research limitations/implications
To demonstrate the usefulness of this data type, we focused on one tourist city with five specific tourist-related queries. Future studies could focus on other aspects of tourist consumption and on more destinations, using a larger number of queries to increase accuracy.

Practical implications
Search volume data are an early indicator of travelers' interest and could be used to predict various types of tourist consumption and activities, such as hotel occupancy, spending, and event attendance.

Originality/Value
The study validated the value of search query volume data in predicting hotel room demand, and is the first of its kind in the field of tourism and hospitality research.

Keywords: Demand for Hotel Rooms; Search Query Volume; Time Series Analysis; Econometric Models; Forecasts; Google Trends

Paper Type: Case Study
FORECASTING HOTEL ROOM DEMAND USING SEARCH ENGINE DATA

Introduction

Because tourism products are perishable, tourism forecasting is crucial to enabling industry participants to allocate limited resources and meet tourist demand, either for a single business or for a destination as a whole (Frechtling, D 2001; Rajopadhye et al. 2001; Song, Li & Witt 2008). Traditional forecasting methods include time series analysis and econometric models (Song, Li & Witt 2008), but prior studies have shown that no single method is consistently superior to other models; depending on the evaluation criteria and data sets employed, certain models perform better than others (Song & Li 2008). Specifically, recent studies have demonstrated that combinations of forecasting methods can produce more accurate results in a tourism context (Chan et al. In Press; Chu 1998; Palm & Zellner 1992; Wong et al. 2007).

Traditional forecasting methods rely on historical data for both independent and dependent variables; the former include populations of source markets, tourist income levels, tourism prices in both the focus and competing destinations, exchange rates, and other qualitative data, as well as “one-off” events such as the Olympic Games (Song, Li & Witt 2008). In recent years, the adoption of the Internet as a travel planning and online transaction tool (TIA 2008) has made a new category of data available, that have great potential to enhance predictive power. When tourists conduct searches or book rooms or airline seats online, their behavior on the Internet can be tracked and monitored using various Internet technologies. Traces of Internet access can be captured on a variety of web servers and Internet routers. Because tourists usually plan online before actually
making a trip, aggregated traces in the form of query volumes on search engines or web access logs are early indicators of interest. Government bodies and private businesses can then use such aggregated online behavioral data to predict the future activities and consumption patterns of tourists.

In this paper, we used aggregated search volumes of five keywords related to a tourist destination to predict the demand for hotel rooms. We compared the forecasting performance of three autoregressive moving average (ARMA) family models under two scenarios: with and without search volume data as explanatory variables. The better forecasting performance under the first scenario verified the value of search volume data. We further tested three widely used causal econometric models to see whether the ARMA-type models remained superior in this forecasting exercise. These included the autoregressive distributed lag (ADL) model, the time-varying parameter (TVP) model, and the vector autoregressive (VAR) model.

The remainder of this paper begins with a review of prior literature on forecasting methods in the field of tourism and hospitality and studies on forecasting using search volume data in relevant fields. We then detail the data and specific methodology used in this study before discussing the results and implications for future research and management.
Literature Review

Scholars have carried out numerous studies and reviews on tourism demand forecasting (Athiyaman & Robertson 1992; Chu 1998; Frechtling, DC 1996; Frechtling, D 2001; Li, Song & Witt 2005; Palm & Zellner 1992; Rajopadhye et al. 2001; Song & Li 2008). These studies have adopted various methods and different types of data. This section specifically reviews the different data sources employed, highlights the nature of search volume data, and surveys the use of search data in other research areas.

According to a few review articles summarizing the state of the art for tourism demand forecasting (Frechtling, D 2001; Li, Song & Witt 2005; Song & Li 2008; Song, Li & Witt 2008), the dependent variables traditionally used are the number of tourist arrivals, tourist expenditure, and the number of tourist nights stayed in a destination. Tourist arrivals are the most frequently used dependent variable, followed by tourist expenditure. In the hospitality area, room nights are commonly used as a surrogate for tourism demand. Data might be collected from customs, registration records at accommodation facilities, sample surveys, or bank reports. Each method has its own advantages and limitations. For example, the accommodation intercept captures overnight tourists, but misses those who stay with friends and relatives. On the other hand, the explanatory variables used in forecasting models include the population of the place of origin, income in the country or area of origin, prices in focus destinations and their competitors, exchange rates, consumer taste, marketing expenditure, and other qualitative variables such as marketing campaigns or large sporting events. Many quantitative methods also have been adopted in tourism demand forecasting, ranging from linear and nonlinear models, time series
techniques, econometric models, to artificial intelligence approaches (Song, Li & Witt 2008). Annual, quarterly, or monthly data are often used in estimating tourism demand models; annual forecasts were the most frequently produced forecasts before the 1990s, with quarterly or monthly forecasts becoming more popular since (Song, Li & Witt 2008).

Revenue management and yield management research has focused on forecasting demand for hotel rooms in a specific property (Jauncey, Mitchell & Slamet 1995; Lee-Ross & Johns 1997). Some researchers have used a special version of the exponential smoothing technique—the Holt-Winters method—to forecast daily hotel room demand in a particular property (Rajopadhye et al. 2001). Linear programming has also been widely used in hotel revenue management to maximize revenues from dynamic pricing, overbooking, and allotment of different segments of hotel assets (Baker & Collier 1999; Weatherford 1995; Weatherford, Kimes & Scott 2001). Weatherford et al. (2001) examined different ways of forecasting hotel demand and found that disaggregated forecasts, which is based on individual segments of guests with the same length of stay and room rate, outperformed all other forecasting methods, which treated all guests as a single segment.

In recent years, with the widespread adoption of the Internet for trip planning and transaction purposes (Pan & Fesenmaier 2006; TIA 2008), a large amount of online behavioral data has been made available to the tourism and hospitality industry. Internet technology provides numerous ways to capture what tourists are doing online and on
which websites they are doing it. When a tourist conducts a search or makes a booking online, traces of access can be captured. Because tourists usually plan trips before traveling, aggregated online behavioral data can be used as an indicator of the demand for travel. This data source has been employed in other research fields such as economics, the social sciences, and health research.

In the medical field, Google volume data have also been used to forecast influenza outbreaks (Ginsberg et al. 2009). In the United States, traditional methods rely on reports from the Centers for Disease Control (CDC), in which forecasts are based on physicians’ case reports. Instead, Ginsberg and colleagues (2009) used raw keyword search volumes from Google Inc. They proved that the frequencies of certain queries are highly correlated with the percentage of patients with influenza-like symptoms. An automated method identified the 45 most predictive search queries from among billions of searches. A real-time data source feed was used for the forecasting model to generate very accurate forecasts one week earlier than reports from the CDC (Ginsberg et al. 2009). In addition, Zhang et al. (2009) estimated a number of ARIMA models using raw search keyword volume data from Dogpile.com. Their study demonstrated that a time series of daily log data could be used to detect changes in user behavior across different periods (Zhang, Jansen & Spink 2009).

Google Trends (Carneiro, H & Mylonakis, E 2009; Choi & Varian 2009) is a public tool provided by Google Inc. that gives normalized and scaled search volume data for specific queries on Google. Askitas and Zimmermann (2009) demonstrated strong correlations
between keyword searches and unemployment rates in Germany using monthly data in a simple error-correction model. Choi and Varian (2009) found that Google Trends data help to improve forecasts of economic activities including retail sales, automotive sales, home sales, and international tourist arrivals. Choi and Varian (2009) also used search data on keywords *jobs* and *welfare/unemployment* in an ARIMA model to predict unemployment claims. They found that it produced more accurate forecasting results than the baseline model, in which no search data were used. Specifically, they have found that incorporating Google search volume data, exchange rates, and one-off events into a univariate seasonal autoregressive (AR) model has greatly improved forecasting performance for international tourist arrivals, with a high $R^2$ of 0.98 (Choi & Varian 2009).

In summary, the advantages of the new type of volume data provided by search engines are that they are real-time, high frequency (daily and weekly instead of quarterly or annual), and sensitive to small changes in user behavior. Researchers in other fields have proved that these data are very valuable in generating accurate forecasts. But with the exception of the investigation of Choi and Varian (2009), few studies of search engine data forecasting have been carried out in the field of hospitality and tourism. Unlike other explanatory variables traditionally used in tourism demand forecasting, search queries can be seen as early behavioral indicators of purchase intentions (Barry 1987; Gitelson & Crompton 1983). Search volume data can therefore be used as an “early warning” signal for aggregated tourist activities.
However, Google search volume data have its limitations. First, not all American travelers use Google to plan their trips. According to a 2007 report by Travel Industry Association, there are least 124 million American vacationers (Travel Industry Association 2007), but only 90 million American travelers plan trips online, including those travelling for business purposes (TIA 2009). Furthermore, even though a majority of online American travelers (86%) use general search engines to search for travel information, not everyone of them is doing so (Fesenmaier et al. 2011). In addition, Google only holds 66% of market share for search engine market (ComScore 2012). Thus, Google data only capture the behavioral characteristics of a portion of American travelers. Second, Google Trend data have its own limitations. As stated on its website, Google Trend may contain inaccuracies due to data sampling and approximation methods (Carneiro, HA & Mylonakis, E 2009). However, with these limitations, a significant contribution of Google search volume to forecasting room demand can still validate its usefulness in forecasting economic activities for the tourism and hospitality industry.

The main objective of this study was to forecast weekly room nights sold in a destination based on Google search volume data. But unlike Choi and Varian’s investigation, our focus was on a short time frequency (weekly data) with a different dependent variable, namely hotel room nights sold. This new variable is more directly relevant to the local hospitality industry, than international tourist arrivals.
Data Description

This study was based on Charleston, South Carolina, a tourist city located in the southeastern United States. The metropolitan statistical area of Charleston-North Charleston-Summerville includes Berkeley, Charleston, and Dorchester County, has a population of around 659,000, and was ranked the 80th largest metropolitan area in the U.S. in 2009 (U.S. Census Bureau 2010). Estimated visitor volume to this area is around 4 million per year, and the tourists come mainly from the southeast region of the U.S. (Charleston Area CVB 2009). A recent intercept survey, in which surveys were handed out to visitors in downtown Charleston, showed that around 16.4% of visitors had used Google to research this destination (Smith & Pan 2010).

Smith Travel Research, Inc. (STR) provided weekly hotel room demand data (Agarwal, Yochum & Isakovski 2002). The Charleston market data included the number of room nights sold for each of the three counties in the Charleston area. In Charleston, 110 of around total 190 hotels or motels report their average daily room and occupancy rates to STR. STR then estimates the total number of room nights sold according to the statistics received from the sampled properties. Thus, the figures produced by STR are considered to be good estimates of the total number of room nights actually sold in the area and are regarded as representative of the volume of overnight tourists.

We obtained search volume data from Google Trends (Carneiro, H & Mylonakis, E 2009; Choi & Varian 2009). Although it would be ideal to use search volume data from all major search engines, so far only Google provides these data. But if we could
demonstrate that Google search volume data are useful with only a portion of visitors using Google, it would validate the significance of the search data’s predictive power.

As a public tool provided by Google Inc., Google Trends “analyzes a portion of Google web searches to compute how many searches have been done for the terms … enter[ed], relative to the total number of searches done on Google over time” (http://www.google.com/intl/en/trends/about.html#1). The search volume data reported are normalized and scaled (http://www.google.com/intl/en/trends/about.html), and include volumes for all types of queries, including those Google specifically categorizes as travel queries. In order to protect the privacy of Google users, Google Trends displays search volumes only for keywords that have reached a certain undisclosed threshold (Askitas & Zimmermann 2009; Carneiro, H & Mylonakis, E 2009).

The search volume data were extracted from January 2008 to August 2009. We adopted this time frame since we obtained hotel occupancy data during the same period. Weekly search data and room demand data were available from January 13, 2008, to March 7, 2009, covering 81 weeks. We used the first 60 weeks of data as the training set to estimate the parameters of the models, and the last 21 weeks as the validation set.

Particularly, tourism demand is denoted by \( y_t \), which we measured by the number of room nights sold in week \( t \). Five sets of Google search volume data were used as the explanatory variables: \( x_{1t} \) for the search volume for the query \textit{charleston sc} (\textit{sc} represents South Carolina), \( x_{2t} \) for the search volume for the query \textit{travel charleston}, \( x_{3t} \) for the search volume for the query \textit{charleston hotels}, \( x_{4t} \) for the search volume for the query \textit{charleston hotels},
*charleston restaurants*, and $x_{si}$ for the search volume for the query *charleston tourism* in week $t$. We adopted these five keywords to obtain Google search data because they are considered the most relevant and unique when tourists search for a destination city in the U.S. (Pan, Litvin & Goldman 2006).

### Method

We first adopted three common ARMA family models to forecast hotel demand. We then combined the Google search data with these three models to create the ARMAX models in order to produce three sets of forecasts on demand for hotel rooms. Comparing the forecasting performance between the ARMA family models and their ARMAX counterparts, we then assessed the value of the Google search data for hotel demand forecasting. Particularly, if forecasting accuracy improved when the Google search data were included, we could conclude that the search data were significant in forecasting demand for hotel rooms. Given the forecasting performance of the Google search data with ARMA/ARMAX models, we further estimated three causal econometric models, namely the ADL, TVP, and VAR models, to test the superiority of the ARMA/ARMAX models.

### ARMA models

We considered three ARMA family models (the AR, ARMA, and ARIMA models) for evaluating the forecasting performance of the Google search data. The ARMA $(p,q)$ model takes the following form:

$$
\ln y_t = \mu + \sum_{i=1}^{p} \phi_i \ln y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i},
$$

(1)
where \( y_t \) is tourism demand at time \( t \) (i.e., the number of room nights sold in week \( t \)).

The first part of the right-hand side of Equation 1 is a constant term \( \mu \) plus the autoregressive term with a lag length of \( p \). \( \epsilon_t \) is the error term at time \( t \). The last part denotes the moving average term with a lag length of \( q \); \( \phi_i \) and \( \theta_i \) are the coefficients to be estimated using nonlinear least square method.

The AR(\( p \)) model is a specific form of the ARMA model with \( q = 0 \) in Equation 1. It therefore takes the form of

\[
\ln y_t = \mu + \sum_{i=1}^{p} \phi_i \ln y_{t-i} + \epsilon_t.  \tag{2}
\]

The integrated ARMA (i.e., ARIMA) model is the generalization of the ARMA model. That is, when \( d = 0 \), ARIMA(\(p,d,q\)) becomes an ARMA model. This technique is used for time series that contain a unit root. An ARIMA(\(p,d,q\)) model can be written as

\[
\Delta^d \ln y_t = \mu + \sum_{i=1}^{p} \phi_i \Delta^d \ln y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i},  \tag{3}
\]

where \( \Delta \) denotes the difference function (i.e., \( \Delta \ln y_t = \ln y_t - \ln y_{t-1} \)), and \( d \) refers to the rank of difference, which is decided by the number of the unit roots in the demand series of hotel rooms.

When explanatory variables (i.e., Google search data) are included in the modeling process, the AR, ARMA, and ARIMA models are known as ARX, ARMAX, and ARIMAX models, respectively. They take the forms of

\[
\ln y_t = \mu + \sum_{j=1}^{p} \phi_j \ln y_{t-j} + \sum_{i=1}^{5} \alpha_i \ln x_i + u_t,  \tag{4}
\]
\[
\ln y_t = \mu + \sum_{j=1}^{\rho} \phi_j \ln y_{t-j} + \sum_{i=1}^{\omega} \alpha_i \ln x_i + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}, \quad (5)
\]

and

\[
\Delta^d \ln y_t = \mu + \sum_{j=1}^{\rho} \phi_j \Delta^d \ln Y_{t-j} + \sum_{i=1}^{\omega} \alpha_i \ln x_i + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}. \quad (6)
\]

As described previously, \( x \) represents volume data for five different search queries.

**Autoregressive distributed lag (ADL) model**

In a general ADL model, the current dependent variable is regressed on the lagged values of the dependent variable and the current and lagged values of one or more explanatory variables. In this study we adopted the general-to-specific modeling technique advocated by Hendry (1986) to derive the specific ADL model after a stepwise reduction process on the general model. The general ADL model can be written as

\[
\ln y_t = \mu + \sum_{i=1}^{\rho} \phi_i \ln y_{t-i} + \sum_{i=1}^{\omega} \alpha_{i,j} \ln x_{i,j} + \varepsilon_t. \quad (7)
\]

Once the general ADL model is specified, the stepwise reduction process is conducted to derive the final model. In particular, the most insignificant variable is repeatedly deleted from the model until all variables left are statistically significant and the model possesses desirable statistical properties (Song, Witt & Jensen 2003; Song, Witt & Li 2003).

**Time-varying parameter (TVP) model**

The TVP model differs from the constant-parameter models in that it is specified as a state space form, and the coefficients of the explanatory variables are normally specified as a random walk process (Song & Witt 2000). Here we used the TVP model to establish
the relationship between the demand for hotel rooms and the Google search data in order to produce forecasts of hotel room demand based on the established relationship. The TVP model is specified as a state space (SS) form that includes two equations, the measurement equation and the transition equation:

\[
\ln y_t = \alpha_{0t} + \sum_{i=1}^{5} \alpha'_i \ln x_{it} + \varepsilon_t, \quad (8)
\]

\[
\alpha_{jt} = \alpha_{j,t-1} + \mu_{jt} \quad (j = 0,\ldots,5). \quad (9)
\]

Equation 8 is the measurement equation, reflecting the relationship between the demand for hotel rooms and the explanatory variables of the search data. Equation 9 is the transition equation, in which the unobserved variables \( \alpha_{jt} \) \((j = 0,\ldots,5)\) are specified in a random walk process. This allows the coefficients of explanatory variables to vary over time. \( \varepsilon_t \) and \( \mu_{jt} \) are disturbance terms. The model is recursively estimated using the Kalman filter algorithm, where the current state is derived from the estimated state of the previous time step and the current independent variables. Song, Witt, and Li (2009) have shown that the TVP model is able to produce reliable short-run forecasts.

**Vector autoregressive (VAR) model**

The econometric models previously discussed are limited to the case where tourism demand \( y_t \) is determined by a set of independent variables, namely \( x_1, x_2,\ldots, x_j \). These independent variables are assumed to be exogenous, but this assumption might be too restrictive and unnecessary (Sims 1980). The VAR model addresses this problem by treating all variables, including tourism demand and its determinants, as endogenous,
except for deterministic variables such as trend, intercept, and dummy. Lagged variables are included in the VAR model to capture the dynamic nature of demand. The VAR is expressed as

$$Y_t = C + \sum_{i=1}^{p} A_i Y_{t-i} + e_t , \quad (10)$$

where $Y_t$ is a $(6 \times 1)$ vector of endogenous variables $\ln y_t$ and $\ln x_a (i = 1, \ldots, 5)$, $C_t$ is a $(6 \times 1)$ vector of constants, $A_i$ is a $(6 \times 6)$ matrix, and $e_t$ is a $(6 \times 1)$ vector of error terms. $p$ is the lag length.

The lag length $p$ must be chosen carefully, since too many lags can lead to over-parameterization, whereas too few can cause loss of forecasting information. We adopted the Bayesian information criterion (BIC) to determine lag length (De Mello & Nell 2005; Song & Witt 2006).

**Measuring forecasting performance**

We used two measures of forecasting accuracy to evaluate the forecasting performance of the models: the mean absolute percentage error (MAPE) and the root mean square percentage error (RMSPE). We calculated these two error measures based on the following formulas:

$$\text{MAPE} = \frac{1}{m} \sum_{t=1}^{m} \left( \frac{|\hat{y}_t - y_t|}{y_t} \right) , \quad (11)$$

$$\text{RMSPE} = \sqrt{\frac{1}{m} \sum_{t=1}^{m} \left( \frac{\hat{y}_t - y_t}{y_t} \right)^2} , \quad (12)$$
where $\hat{y}_t$ for $t = 61, \ldots, 81$ represents the room demand in week $t$ predicted by a specific model, and $y_i (i = 61, \ldots, 81)$ is the observed room demand in week $t$. These two measures have been widely applied to evaluate the forecasting performance of tourism demand models (Song & Witt 2006; Vu & Turner 2006). All analyses were carried out using the econometric modeling software EViews.

**Empirical Results**

**Unit root tests and model estimation**

It is essential to explore the properties of the time series data under consideration before model estimation. Table 1 reports the unit root test results based on the augmented Dickey-Fuller (ADF) test (for a detailed explanation of the ADF test, see Song and Witt, 2000, pp. 59-63). The results show that the dependent variable (demand for hotel rooms) is non-stationary with one unit root, whereas some independent variables are stationary with no unit root, and others are non-stationary with one unit root. The error term refers to the residual series from the estimated model containing the five independent variables in addition to the dependent variable. The ADF statistics indicate that the error terms associated with the three level models are stationary, suggesting that the cointegration relationship exists between demand for hotel rooms and the Google search variables (for a detailed explanation of cointegration, see Song and Witt, 2000, pp. 53-68).

------------ Insert Table 1 here --------------
The lag lengths for the ARMA-type models are determined by BIC, i.e., \( p = 1 \) for the AR(\( p \)) process; \( p = 1 \) and \( q = 1 \) for the ARMA(\( p, q \)) process; and \( p = 2, \ d = 1 \) and \( q = 2 \) for the ARIMA(\( p, d, q \)) process (see Equations 1-6 for reference). Table 2 reports the estimation results of the AR(1), ARMA(1,1), and ARIMA(2,1,2) models and their ARMAX counterparts when Google search variables are included. The diagnostic statistics suggest that the ARIMA(2,1,2) model may have a series correlation problem, while AR(1) and ARMA(1,1) may have a normality problem at the 1% significance level but not at the 5% level. Table 3 shows the estimation results of the three causal econometric approaches, namely the ADL, TVP, and VAR models. According to the various diagnostic statistics in Table 3, no series correlation, heteroskedasticity, or normality problems are identified for these three models at the 1% significance level.

Forecasting performance of the Google search data

After estimating the models, we also generated ex post forecasts based on these estimated models. We first examined the forecasting performance of the ARMA-type and ARMAX models based on MAPE and RMSPE. We also included forecasts of the naïve model in the evaluation. Table 4 indicates that the ARMA-type models were outperformed by their ARMAX counterparts with search volume data included; this suggests that including the Google search data did indeed improve forecasting accuracy. Particularly, among the three ARMAX models, the ARX(1) model performed best, followed by ARMAX(1,1). The benchmark or naïve model outperformed only the ARIMAX(2,1,2) model.
Given the value of the Google search volume data in improving forecasting performance, one may ask whether using modern econometric techniques could further improve forecasting accuracy. Thus, we further compared the forecasting performance of the econometric models with that of the ARMAX models. Similar to the previous exercises, we used the first 60 observations of the variables for model estimation and the remaining 21 for evaluating forecasting accuracy. Table 5 shows the MAPEs and RMSPEs of these econometric models. The results clearly indicate that the ARX model produced the most accurate forecasts, with the lowest MAPE and RMSPE values of 5.529 and 6.896, respectively, whereas the ADL and TVP models produced the poorest forecast accuracy. We thus conclude, based on the forecasting evaluation among all models considered, that the ARMAX-type models are superior to other models on forecasting accuracy.

Conclusion

This study used different time series and econometric models to model and forecast hotel room demand in a tourist city based on search volume data. The models tested were three ARMA, three ARMAX, and three causal econometric models, while room demand was represented by weekly hotel occupancy data provided by STR. The empirical results indicate that when five Google search data variables were included in the ARMA model, forecasting accuracy improved significantly. This provides strong support for the use of search engine data in predicting demand for hotel rooms. Especially considering that the
forecasting period was at the beginning of the economic recession in the U.S., many traditional forecasting methods that assume consistent explanatory variables or a stable economic structure may not provide accurate estimates. For future studies, destinations at different levels (country, state, or city) or different industry segment (accommodations, restaurants, or attractions) could also use search volume data to increase forecasting accuracy in planning. The forecasting competition with three commonly used econometric methods indicates that model complexity does not necessarily improve forecasting accuracy.

One limitation of this study is that only five tourism-related queries were included in the models, far fewer than the 45 queries used to predict flu epidemics in a prior investigation (Ginsberg et al. 2009). Since travel involves a complex decision-making process that is affected by many social, economic, cultural, and environmental factors, including the number of search queries would likely increase the forecasting accuracy of the models. The use of Google Trends data is particularly important, as they are freely available online and could be used to help improve forecasting accuracy at very low cost. However, as discussed earlier, Google search volume data have its own limitations: it only capture a certain segment of American travelers; the data may contain inaccuracy due to its sampling and approximation methods (Carneiro, HA & Mylonakis, E 2009). One should exercise precaution when forecasting tourism and hospitality economic activities with Google search volume data.
References


Li, G, Song, H & Witt, S 2005, 'Recent developments in econometric modeling and forecasting', *Journal of Travel Research*, vol. 44, no. 1, p. 82.

Palm, F & Zellner, A 1992, 'To combine or not to combine? Issues of combining forecasts', *Journal of Forecasting*, vol. 11, pp. 687-.


Weatherford, LR, Kimes, SE & Scott, DA 2001, 'Forecasting for hotel revenue management', *Cornell Hotel and Restaurant Administration Quarterly*, vol. 42, no. 4, pp. 53-64.

Figure 1: Preliminary Plot of Dependent and Independent Variables
### TABLES

Table 1 Unit root test results of the level and differenced variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Level</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>One trend &amp; one intercept</td>
<td>One intercept</td>
</tr>
<tr>
<td>ln $y_t$</td>
<td>-3.36</td>
<td>-2.65</td>
</tr>
<tr>
<td>ln $x_{1t}$</td>
<td>-3.64*</td>
<td>-3.45*</td>
</tr>
<tr>
<td>ln $x_{2t}$</td>
<td>-5.13**</td>
<td>-4.78**</td>
</tr>
<tr>
<td>ln $x_{3t}$</td>
<td>-1.25</td>
<td>-1.40</td>
</tr>
<tr>
<td>ln $x_{4t}$</td>
<td>-4.99**</td>
<td>-4.92**</td>
</tr>
<tr>
<td>ln $x_{5t}$</td>
<td>-4.91**</td>
<td>-3.99**</td>
</tr>
<tr>
<td>Error term</td>
<td>-3.96*</td>
<td>-3.88**</td>
</tr>
</tbody>
</table>

Note: The figures in the table are ADF statistics. *
* and ** indicate significance at the 5% and 1% levels, respectively.
Table 2 Estimation results for the ARMA-type models and their ARMAX counterparts

<table>
<thead>
<tr>
<th></th>
<th>AR(1)</th>
<th>ARMA (1,1)</th>
<th>ARIMA (2,1,2)</th>
<th>ARX(1)</th>
<th>ARMAX (1,1)</th>
<th>ARIMAX (2,1,2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>11.26</td>
<td>11.27</td>
<td>0.00</td>
<td>9.59</td>
<td>9.67</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.88)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.85)</td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.79</td>
<td>0.89</td>
<td>0.17</td>
<td>0.81</td>
<td>0.83</td>
<td>-0.53</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.01)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.30)</td>
</tr>
<tr>
<td>AR(2)</td>
<td></td>
<td></td>
<td></td>
<td>-0.89</td>
<td></td>
<td>-0.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00)</td>
<td></td>
<td>(0.13)</td>
</tr>
<tr>
<td>MA(1)</td>
<td>-0.29</td>
<td>-0.45</td>
<td>-0.08</td>
<td>0.99</td>
<td></td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.00)</td>
<td>(0.64)</td>
<td></td>
<td></td>
<td>(0.51)</td>
</tr>
<tr>
<td>MA(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.43)</td>
</tr>
<tr>
<td>ln x_{1t}</td>
<td></td>
<td></td>
<td></td>
<td>-0.16</td>
<td>-0.18</td>
<td>-0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.52)</td>
<td>(0.49)</td>
<td>(0.38)</td>
</tr>
<tr>
<td>ln x_{2t}</td>
<td></td>
<td></td>
<td></td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>ln x_{3t}</td>
<td></td>
<td></td>
<td></td>
<td>0.29</td>
<td>0.29</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>ln x_{4t}</td>
<td></td>
<td></td>
<td></td>
<td>0.20</td>
<td>0.19</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.18)</td>
</tr>
<tr>
<td>ln x_{5t}</td>
<td></td>
<td></td>
<td></td>
<td>-0.02</td>
<td>-0.03</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.58)</td>
<td>(0.54)</td>
<td>(0.44)</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.63</td>
<td>0.65</td>
<td>0.30</td>
<td>0.76</td>
<td>0.76</td>
<td>0.42</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>48.31</td>
<td>49.86</td>
<td>52.75</td>
<td>61.52</td>
<td>61.62</td>
<td>58.01</td>
</tr>
<tr>
<td>BIC</td>
<td>-1.50</td>
<td>-1.48</td>
<td>-1.50</td>
<td>-1.60</td>
<td>-1.54</td>
<td>-1.33</td>
</tr>
<tr>
<td>Jarque-Bera</td>
<td>20.12</td>
<td>35.53</td>
<td>0.14</td>
<td>5.53</td>
<td>6.98</td>
<td>1.00</td>
</tr>
<tr>
<td>test</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.93)</td>
<td>(0.06)</td>
<td>(0.03)</td>
<td>(0.61)</td>
</tr>
<tr>
<td>White</td>
<td>2.84</td>
<td>1.38</td>
<td>0.37</td>
<td>5.66</td>
<td>13.06</td>
<td>18.38</td>
</tr>
<tr>
<td>test</td>
<td>(0.09)</td>
<td>(0.71)</td>
<td>(0.83)</td>
<td>(0.46)</td>
<td>(0.11)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Breusch-Godfrey</td>
<td>2.19</td>
<td>0.00</td>
<td>9.47</td>
<td>0.16</td>
<td>0.25</td>
<td>0.00</td>
</tr>
<tr>
<td>test (rank=1)</td>
<td>(0.14)</td>
<td>(0.95)</td>
<td>(0.00)</td>
<td>(0.69)</td>
<td>(0.62)</td>
<td>(0.99)</td>
</tr>
<tr>
<td>Breusch-Godfrey</td>
<td>4.16</td>
<td>3.93</td>
<td>9.66</td>
<td>0.40</td>
<td>0.44</td>
<td>0.55</td>
</tr>
<tr>
<td>test (rank=2)</td>
<td>(0.12)</td>
<td>(0.14)</td>
<td>(0.01)</td>
<td>(0.82)</td>
<td>(0.80)</td>
<td>(0.76)</td>
</tr>
</tbody>
</table>

Notes: (1) Numbers in parentheses denote probability. (2) The Jarque-Bera test is for normality, the White test is for heteroscedasticity, and the Breusch-Godfrey tests are for serial correlations (for a detailed explanation, see Song, Witt, and Li, 2008, pp. 52-55).
Table 3 Estimation results for three causal econometric models

**ADL model**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>11.50</td>
<td>10.40</td>
<td>0.00</td>
</tr>
<tr>
<td>ln $x_{1,t-1}$</td>
<td>-1.05</td>
<td>-3.80</td>
<td>0.00</td>
</tr>
<tr>
<td>ln $x_{1,t-5}$</td>
<td>0.50</td>
<td>2.10</td>
<td>0.04</td>
</tr>
<tr>
<td>ln $x_{2,t-2}$</td>
<td>-0.20</td>
<td>-2.32</td>
<td>0.03</td>
</tr>
<tr>
<td>ln $x_{3,t-1}$</td>
<td>0.51</td>
<td>3.93</td>
<td>0.00</td>
</tr>
<tr>
<td>ln $x_{3,t-6}$</td>
<td>0.30</td>
<td>3.41</td>
<td>0.00</td>
</tr>
<tr>
<td>ln $x_{4,t-1}$</td>
<td>-0.20</td>
<td>-1.49</td>
<td>0.14</td>
</tr>
<tr>
<td>ln $x_{4,t-4}$</td>
<td>0.33</td>
<td>2.84</td>
<td>0.01</td>
</tr>
<tr>
<td>ln $x_{4,t-6}$</td>
<td>-0.30</td>
<td>-2.21</td>
<td>0.03</td>
</tr>
<tr>
<td>ln $x_{5,t-1}$</td>
<td>0.09</td>
<td>1.68</td>
<td>0.10</td>
</tr>
</tbody>
</table>

*R-squared* 0.75  
*Log likelihood* 52.75  
*BIC* 2.97

**Jarque-Bera test**  (0.23)  
**White test**  (0.02)  
**Brensch-Godfrey test** (rank=1) 3.82  
**Brensch-Godfrey test** (rank=2) 5.68

**TVP model**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>z-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>9.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln $x_{1,t}$</td>
<td>-0.12</td>
<td>11.66</td>
<td>0.00</td>
</tr>
<tr>
<td>ln $x_{2,t}$</td>
<td>0.11</td>
<td>-0.50</td>
<td>0.62</td>
</tr>
<tr>
<td>ln $x_{3,t}$</td>
<td>0.26</td>
<td>1.90</td>
<td>0.06</td>
</tr>
<tr>
<td>ln $x_{4,t}$</td>
<td>0.19</td>
<td>2.06</td>
<td>0.04</td>
</tr>
<tr>
<td>ln $x_{5,t}$</td>
<td>-0.03</td>
<td>2.18</td>
<td>0.03</td>
</tr>
</tbody>
</table>

*R-squared* 0.56  
*Log likelihood* 4.01  
*BIC* 7.94

**Jarque-Bera test**  (0.02)

**VAR model**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>5.24</td>
<td>3.80</td>
<td>0.00</td>
</tr>
</tbody>
</table>

29
\[
\begin{array}{c|c|c|c}
\text{ln } y_{t-1} & 0.62 & 6.91 & 0.00 \\
\text{ln } x_{1t-1} & -0.28 & -1.32 & 0.10 \\
\text{ln } x_{2t-1} & -0.15 & -1.91 & 0.03 \\
\text{ln } x_{3t-1} & 0.36 & 2.99 & 0.00 \\
\text{ln } x_{4t-1} & -0.18 & -1.66 & 0.05 \\
\text{ln } x_{5t-1} & 0.02 & 0.44 & 0.33 \\
\hline
\text{R-squared} & 0.72 \\
\text{Log likelihood} & 56.66 \\
\text{BIC} & -1.44 \\
\text{Jarque-Bera test} & (0.32) \\
\text{White test} & 286.40 (0.07) \\
\text{Brensch-Godfrey test} & 48.10 \\
\text{(rank=1)} & (0.09) \\
\text{Brensch-Godfrey test} & 29.70 \\
\text{(rank=2)} & (0.76) \\
\end{array}
\]

Notes: Same as for Table 3.
Table 4 Forecasting competition between the ARMA and ARMAX models

<table>
<thead>
<tr>
<th></th>
<th>Naïve</th>
<th>AR(1)</th>
<th>ARMA(1,1)</th>
<th>ARIMA(2,1,2)</th>
<th>ARX(1)</th>
<th>ARMAX(1,1)</th>
<th>ARIMAX(2,1,2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPE</td>
<td>7.515</td>
<td>9.191</td>
<td>8.955</td>
<td>8.782</td>
<td>5.529</td>
<td>5.565</td>
<td>7.918</td>
</tr>
</tbody>
</table>
Table 5 Forecasting performance competition between causal econometric models

<table>
<thead>
<tr>
<th></th>
<th>ARX(1)</th>
<th>ARMAX (1,1)</th>
<th>ARIMAX (2,1,2)</th>
<th>ADL</th>
<th>TVP</th>
<th>VAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPE</td>
<td>5.529</td>
<td>5.565</td>
<td>7.918</td>
<td>8.238</td>
<td>8.339</td>
<td>7.874</td>
</tr>
</tbody>
</table>