The following publication Pan, B., Chenguang Wu, D. and Song, H. (2012), "Forecasting hotel room demand using search engine data", Journal of Hospitality and Tourism Technology, Vol. 3 No. 3, pp. 196-210 is published by Emerald and is available at https://doi.org/10.1108/17579881211264486

FORECASTING HOTEL ROOM DEMAND WITH 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)

AUTOBIOGRAPHICAL NOTES

Bing Pan, Ph.D., is assistant 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 WITH SEARCH ENGINE DATA

ABSTRACT

Purpose

The purpose of the study is to investigate the usefulness of search engine query volume data in forecasting the demand for hotel rooms and identify the best econometric forecasting model.

Design/methodology/approach

The study uses the search volume data on five related queries to predict the demand for hotel rooms in a specific tourist city. Three ARMA family models and their ARMAX counterparts considering search volume data are employed to evaluate the usefulness of the search volume data. Three widely used causal econometric models, i.e., ADL, TVP and VAR models, are also evaluated for comparison purpose.

Findings

All three ARMA models consistently outperform their ARMAX counterparts, which validates the value of search engine volume data in facilitating the accurate prediction of demand for hotel rooms. When three causal econometric models are included for forecasting competition, the ARX model has produced the most accurate forecasts than others. This suggests the usefulness of ARX model to forecast the demand for hotel rooms.

Research limitations/implications

To demonstrate the usefulness of the special type of data, the study focuses on one tourist city with five specific tourist-related queries. Future studies can focus on other aspects of tourist consumption on more destinations based on a large number of queries in order to increase accuracy.

Practical implications

Search engine query volume data are an early indicator of travellers' interests 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 engine query volume data in predicting hotel room demand, the first of its kind in the field of tourism and hospitality research.

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

Paper Type: Case Study

FORECASTING HOTEL ROOM DEMAND WITH SEARCH ENGINE DATA

Introduction

Due to the perishability of tourism products, tourism forecasting is crucial to enable 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). 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 historic data for both dependent and independent variables; the latter include populations of source markets, income levels of tourists, tourism prices in both the focus destination and competing destinations, exchange rates, and other qualitative data, and "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 available a new category of data that has 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 the 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 use such aggregated online behavioral data to predict the future activities and consumption patterns of tourists.

The study reported in this paper used aggregated search volumes for five keywords related to a tourist destination to predict the demand for hotel rooms. The forecasting performance of three ARMA family models are compared under two scenarios: with and without search volume data as explanatory variables. The better forecasting performance under the first scenario verifies the value of search volume data for forecasting the demand for hotel rooms. Three widely used causal econometric models were further tested in order to see whether the ARMA type models are still superior in this forecasting exercise. . They include 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 relevant to the tourism and hospitality field and studies in other fields on forecasting using search engine volume data. The data and specific methodology used in this study is then detailed before the results and implications for future research and management are discussed.

Literature Review

Numerous studies and reviews on tourism demand forecasting have been carried out (Athiyaman & Robertson 1992; Chu 1998; Frechtling, DC 1996; Frechtling, D 2001; Li, G, 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 recent review articles summarizing the state of the art for tourism demand forecasting (Frechtling, 2001; Li, G, 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 place of origin population, 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 a marketing campaigns or large sporting events. Many quantitative methods have been adopted in tourism demand forecasting, ranging from linear and nonlinear models, time series techniques,

econometric models, to artificial intelligence approaches such as neural networks (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 thereafter (Song, Li & Witt 2008).

Revenue management and yield management research has focused on forecasting the 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 (Weatherfold, 1995; Baker and Collier, 1999; Weatherford, Kimes, and Scott, 2001). Weatherford, Kimes, and Scott (2001) examined different way of forecasting hotel demand: they found that disaggregated forecasts 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 where

they are doing it. When a tourist conducts a search or make a booking online, traces of access can be captured. Because tourists usually plan trips before travelling, 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, social sciences, and health research. Google Trends (Carneiro & Mylonakis 2009; Choi & Varian 2009) is a public tool provided by Google Inc. that gives search volume data for specific queries on Google. Choi and Varian (2009) found that Google Trends data helps to improve forecasts of economic time series including retail sales, automotive sales, home sales, and international tourist arrivals (Choi & Varian 2009). Specifically, prior investigations have found that by incorporating Google search volume data, exchange rates, and "one-off" events into a univariate seasonal autoregressive (AR) model, the forecasting performance for international tourist arrivals has been greatly improved, with a highest R^2 of 0.98 (Choi & Varian 2009). Choi and Varian (2009) have also used search data on "jobs" and "welfare/unemployment" in an ARIMA model to predict unemployment claims and found that it produced more accurate forecasting results than the baseline model in which no search data were used. In addition, Askitas and Zimmermann (2009) demonstrated strong correlations between keyword searches and unemployment rates in Germany using monthly data in a simple error-correction model.

In medical field, Google volume data were also used to forecast influenza outbreaks (Ginsberg et al. 2009). In the United States, traditional methods rely on reports from the Center for Disease Control (CDC), in which forecasts are based on physicians' case reports. Ginsberg and colleagues instead used raw keywords search volumes from

Google—instead of normalized and scaled data from Google Trends. 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 CDC (Ginsberg et al. 2009). In addition, Zhang, Jansen, and Spink (2009) estimated a number of ARIMA models using raw search engine keyword volume data from Dogpile.com. Their study demonstrated that time series of daily log data could be used to detect changes in user behavior across different time periods (Zhang, Jansen & Spink 2009).

In summary, the advantages of the new type of volume data provided by search engines are real-time, high-frequency (daily and weekly instead of quarterly or annual), and they are sensitive to small changes in user behavior. Researchers in other fields have proved that these data are very valuable in generating accurate forecasts. With the exception of the investigation of Choi and Varian (2009), very few studies of search engine data forecasting have been carried out in the field of hospitality and tourism. Tourists usually check out their destination and plan online before making the trip (TIA, 2008). In contrast with other explanatory variables traditionally used in tourism demand forecasting, search queries can be seen as 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. The main objective of this study was to forecast weekly room nights sold in a destination based on Google search engine volume data.

Unlike Choi and Varian's investigation, the focus of this study was on a short time frequency (weekly data) and a different dependent variable: hotel room nights sold. This new variable was considered more relevant to the local hospitality industry.

Data Description

The case study reported here was based on Charleston, a tourist city located in South Carolina (SC) in the southeast of the 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). The estimated visitor volume to this area is around 4 million per year and tourists mainly come 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, for which 110 out of the total of around 190 hotels/motels report their average daily room rates and occupancy rates to STR. STR estimates the total number of room nights sold according to 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.

Search volume data were obtained from Google Trends (Carneiro & Mylonakis 2009; Choi & Varian 2009). Google is currently the largest search engine provider in the U.S., with a market share of 64.6% by the number of searches (http://searchenginewatch.com/3634991). It will be ideal to use search volume data for all major search engines; however, so far only Google provides search volume data; if the authors can demonstrate that Google search volume data are useful with only a portion of visitors using it, it will validate the power 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 you enter, 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. However, 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 & Mylonakis 2009). The search volume data employed in this study were extracted from January 2008 to August 2009. This timeframe was adopted since the authors obtained hotel occupancy data during the same time period. Weekly search data and room demand data were available from 13th January 2008 to 7th March 2009, a period covering a total of 81 weeks. The first 60 weeks of data were used as a training set to estimate the parameters of the models, with the last 21 weeks of data being used as a validation set. Particularly, tourism demand

is denoted by y_t , which is meansured by the number of room nights sold in week t . Five sets of Google engine search volumn data are used as the explanatroy variables:

x_{1t} denotes the search volume for the query of *charleston sc* (*sc* represents South Carolina);

x_{2t} denotes the search volume for the query of *travel charleston;*

x_{3t} denotes the search volume for the query of *charleston hotels;*

x_{4t} denotes the search volume for the query of *charleston restaurants*, and

x_{5t} denotes the search volume for the query of *charleston tourism* in week t , respectively.

These five keywords are adopted 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).

Methodology

This study first uses three commonly used ARMA family models to forecast the hotel demand. Then the Google search data are combined with the three ARMA family models, which is known as ARMAX models to produce three sets of forecasts on demand for hotel rooms. Comparing the forecasting performance between the ARMA family models and their ARMAX counterparts, the value of the Google search data for hotel demand forecasting is assessed. Particularly, if the forecasting accuracy is improved when the Google search data are included , we could conclude that the Google search data are useful in forecasting the demand for hotel rooms.

Given the forecasting performance of the Google search data using ARMA/ARMAX models, we further estimated three causal econometric models, i.e., ADL, TVP and VAR models in order to test superiority of the ARMA/ARMAX models.

ARMA models

In this study three ARMA family models, i.e., AR model, ARMA model and ARIMA model are considered for evaluating the forecasting performance of the Google search information. The ARMA model takes the following form:

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

y_t is tourism demand at time t, i.e., number of room nights sold in week t in the study. ln denotes the natural logarithm. It is conventional and the predominant method to transform the variables into logarithms before the modelling process (see Li et al., 2005). The variables with log transformation are more smooth with reduced order of integration, and are consistant with the real relationship between dependent and explanatory variables. The first part of the right-hand side of Equation 1 is a constant term plus the autoregressive term with a lag length of *p*. ε_t is the errer term at time t. The last part denotes the moving average term with the lag length of *q*, ϕ_i and θ_i are the coefficients to be estimated.

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

$$\ln y_t = \mu + \sum_{i=1}^{p} \phi_i \ln y_{t-i} + \varepsilon_t \qquad (2)$$

The integrated ARMA, i.e., ARIMA model is the generalisation of the ARMA model. This technqiue is used for time series which contain a unit root. An ARIMA (p,d,q) model can be written as

$$\Delta^d \ln y_t = \mu + \sum_{i=1}^{p} \phi_i \Delta^d \ln Y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \qquad (3)$$

where Δ denotes the difference function, 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 modelling process, the AR, ARMA and ARIMA models are known as o ARX, ARMAX, and ARIMAX models, respectively. They take the forms of

$$\ln y_t = \mu + \sum_{i=1}^{p} \phi_i \ln y_{t-i} + \sum_{i=1}^{5} \alpha_i \ln x_i + u_t, \qquad (4)$$

$$\ln y_t = \mu + \sum_{i=1}^{p} \phi_i \ln y_{t-i} + \sum_{i=1}^{5} \alpha_i \ln x_i + u_t + \sum_{i=1}^{q} \theta_i u_{t-i} \quad (5)$$

and

$$\Delta^d \ln y_t = \mu_t + \sum_{i=1}^{p} \phi_i \Delta^d \ln Y_{t-i} + \sum_{i=1}^{5} \alpha_i \ln x_{it} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \quad (6).$$

Autoregressive Distributed Lag (ADL) model

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

$$\ln y_t = \mu + \sum_{i=1}^{p} \phi_i \ln y_{t-i} + \sum_{i=1}^{5} \sum_{j=0}^{p} \alpha_{ij} \ln x_{i,t-j} + \varepsilon_t \quad (7)$$

Once the general ADL model is specified, the stepwise reduction process is conducted to derive the final model. Particularlly, the most insignificant variable is deleted from the model repeatedly until all variables left in the model are statistically significant and the model possess desireable statistical properties. Song, Witt and Jensen (2003) and Song Witt and Li (2003) have successfully applied this method to tourism demand forecasting.

Time-Varying Parameter (TVP) Model

The TVP model differs from the constant-parameter models as it is specified as a state space form and the coefficients of the explanatory variables are normally specified as a random walk process (Song and Witt, 2000, p128). In this study the TVP model is employed to establish the relationship between the demand for hotel rooms and the Google search data with a view to producing the forecasts of the demand for hotel rooms based on the established relationship. The TVP model is specified as a SS form that include two equations: the measurement equation and the transition equation.

$$\ln y_t = \alpha_{0t} + \sum_{i=1}^{5} \alpha_{it} \ln x_{it} + \varepsilon_t \quad (8)$$

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

Equation 8 is the measurement equation reflecting the relationship between the demand for hotel rooms and the explanatory variables of the Google search data. Equation 9 is the transition equation in which the unobserved variables α_{jt} $(j = 0,...,5)$ are specified in a random walk process. This allows the coefficients of explanatory variables to vary over

time. ε_t and μ_{jt} are disturbance terms. The model is recursively estimated using the Kalman filter algorithm (Kalman, 1960) where the current state is derived from the estimated state of the previous time step and the current independent variables. Song and Witt (2000) and Song, Witt and Li (2009) have showed that the TVP model is able to produce reliable short-run forecasts (Li, G. et al. 2006).

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, x_1, x_2,…, x_t. These independent variables are assumed to be exogenous. This assumption might be too restrictive and unneccsary (Sim, 1980). The VAR model (Sim, 1980) addresses this problem by treating all variables including tourism demand and its determinants as endogenous, except deterministic variables such as trend, intercept and dummy variables. Lagged variables are included in the VAR model to capture the dynamic nature of the demand. The VAR is expressed as:

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

where Y_t is a (6×1) vector of engogenous variables $\ln y_t$ and $\ln x_{it}$ ($i = 1,...,5$). C_t is a (6×1) vector of constants, A_i is a (6×6) matrix and e_t is a (6×1) vector of error terms. p is the lag length.

The lag length p must be chosen carefully as too many lags can lead to to over-parameterisation whereas too few lags cause loss of forecasting information. The

Bayesian information criterion (BIC) is adopted for the determination of the lag length. The VAR model has been widely used in macroeconomic modelling and forecasting since first introduced in 1980. Song and Witt (2006) and Witt et al. (2003) successfully employed this technique to forecast tourism generated employment in Denmark and tourist flows to Macau, while De Mello and Nell (2005) applied it to forecast the demand for French, Spanish and Portuguese tourism by the UK residents.

Measurement of forecasitng performance

Two measures of forecasting accuracy are used to evaluate the forecasting performance of the models: mean absolute percentage error (MAPE) and root mean square percentage error (RMSPE). These two error measures were calculated on the basis of the following formulas:

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

$$\text{RMSPE} = \sqrt{\frac{1}{m}\sum_{t=1}^{m}\left(\frac{\hat{y}_t - y_t}{y_t}\right)^2} \tag{12}$$

where $\hat{y}_t$ for $t = 61,\cdots,81$ represents the room demand in week t predicted by a specific model, and y_i ($i = 61,\cdots,81$) is the observed room demand in week t. These two measures have been widely applied to evaluate the forecasting performance of tourism demand model (Song & Li 2008; Vu & Turner 2006). All analyses were carried out using the econometric modelling 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 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, pp59-63). The results show that the dependent variable, the demand for hotel rooms, is non-stationary with one unit root, whereas some of the independent variables are stationary containing no unit root and some are non-stationary with one unit root. The error term refers to the residual series from the estimated model that contains the five independent variables in addition to the dependent variable. The ADF statistics indicate that the errer terms associated with the three level models are stationary. and this suggests that that the cointegration relationship exists between the the demand for hotel rooms and the Google search variables (for a detailed explanation about cointegration, see Song and Witt, 2000, pp53-68).

---------- Insert Table 1 here ----------

The lag lengths for ARMA type models are determined by the Byesian Information Criterion (BIC). Table 2 reports the estimation results of 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 series correlation problem whilst AR(1) and ARMA(1,1) may have normality problem at 1% siginificant level but not at 5% level. Table 3 shows the estimaiton results of the three

causal econometric approaches: 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 econometric models at 1% significant level.

--------- Insert Table 2 here ----------

--------- Insert Table 3 here ----------

Forecasting performance of the Google search data

After estimating the models, the ex post forecasts are also generated based on the estimated models. We first examine the forecasting performance of the ARMA type of models and ARMAX models based on the MAPE and RMSPE. The forecasts of the naïve model are also included in the evaulation. Table 4 indicates that the ARMA type models are outperformed by their ARMAX counterparts with search volumn data included and this suggests that the inclusion of the Google search data do improve the forecasting accuracy. Particularly, amongst the three ARMAX models the ARX(1) model performs the best, followed by the ARMAX(1,1) model. The benchmark model or naïve model, can only out-perform the ARIMAX(2,1,2) model.

---------- Insert Table 4 here ----------

Given the value of the Google search volumn data in improving the forecasting performance, one may ask whether using modern econometric techniques could further improve the forecasting accuracy. Thus, we further compared the forecasting performance of the econometric models with that of the ARMAX models. Similar to the previous exercises, the first 60 observations of the variables are used for model

estimation whilst the remaining 21 observations are used for forecasting accuracy evaluation. Table 5 shows that the MAPEs and RMSPEs of the these econometric models. The results clearly indicate that the ARX model produces the most accurate forecasts, with the lowest MAPE and RMSPE values of 5.529 and 6.896, respectively. ADL and TVP models produced poorest forecast accuracy. Based on the forecasting evaluation amongst all models considered, we could conclude that the ARMAX type models are superior to other models as far as the forecasting accuracy is concerned.

---------- Insert Tabel 5 here ----------

Conclusion

This study used different time series and econometric models to model and forecast hotel room demand in a tourist city based on search engine volume data. The models tested were three ARMA models, three ARMAX and three causual econometric models: the room demand is represented by weekly hotel occupancy data provided by Smith Travel Research, Inc.

The empirical results indicate that when five Google search data variables are included in the ARMA model, the forecasting accuarcy is improved significantly. This provides strong support for the use of search engine data to predict the demand for hotel rooms. Especially considering that the forecasting period was at the beginning of the economic recession in the United States, 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), attractions (such as Disneyland), hotels (such as Marriott), or restaurants (such as Chilli's)

could also use the search volume data to increase the forecasting accuarcy in their planning process. The forecasting competition with three commonly used econometric methods indicates that the complexity of the models will not necessarily improve the 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). Travel involves a complex decision-making process that is affected by many social, economic, cultural, and environmental factors; therefore, including more search queries is likely to increase the forecasting accuracy of the models examined here. The use of Google Trends data is particularly important, as they are freely available online and can therefore be used to help improve forecasting accuracy at very low cost.

References

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

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

Athiyaman, A. & Robertson, R. 1992. Time Series Forecasting Techniques: Short-Term Planning in Tourism, *International Journal of Contemporary Hospitality Management*, 4), 8-11.

Barry, T. 1987. The Development of the Hierarchy of Effects: An Historical Perspective, *Current issues and Research in Advertising*, 10 (2), 251-95.

Carneiro, H. & Mylonakis, E. 2009. Google Trends: A Web Based Tool for Real Time Surveillance of Disease Outbreaks, *Clinical infectious diseases*, 49), 1557-64.

Chan, C., Witt, S., Lee, Y. & Song, H. In Press. Tourism Forecast Combination Using the Cusum Technique, *Tourism Management*).

Charleston Area CVB 2009. *2009-2010 Charleston Area Convention & Visitors Bureau Book*, Charleston: Charleston Convention & Visitors Bureau.

Choi, H. & Varian, H. 2009. Predicting the Present with Google Trends, viewed April 2, 2009, <http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf>.

Chu, F. 1998. Forecasting Tourism: A Combined Approach, *Tourism Management*, 19 (6), 515-20.

Frechtling, D. 1996. *Practical Tourism Forecasting*: Elsevier.

—— 2001. *Forecasting Tourism Demand: Methods and Strategies*: Butterworth-Heinemann.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. 2009. Detecting Influenza Epidemics Using Search Engine Query Data, *Nature*, 457 (7232), 1012-4.

Gitelson, R. J. & Crompton, J. L. 1983. The Planning Horizons and Sources of Information Used by Pleasure Vacationers, *Journal of Travel Research*, 21 (Winter), 2-7.

Jauncey, S., Mitchell, I. & Slamet, P. 1995. The Meaning and Management of Yield in Hotels, *International Journal of Contemporary Hospitality Management*, 7 (4), 23-6.

Lee-Ross, D. & Johns, N. 1997. Yield Management in Hospitality Smes, *International Journal of Contemporary Hospitality Management*, 9 (2), 66-9.

Li, G., Song, H. & Witt, S. 2005. Recent Developments in Econometric Modeling and Forecasting, *Journal of Travel Research*, 44 (1), 82.

Li, G., Wong, K. K. F., Song, H. & Witt, S. F. 2006. Tourism Demand Forecasting: A Time Varying Parameter Error Correction Model, *Journal of Travel Research*, 45 (2), 175.

Palm, F. & Zellner, A. 1992. To Combine or Not to Combine? Issues of Combining Forecasts, *Journal of Forecasting*, 11), 687-.

Pan, B. & Fesenmaier, D. R. 2006. Online Information Search: Vacation Planning Process, *Annals of Tourism Research*, 33 (3), 809-32.

Pan, B., Litvin, S. W. & Goldman, H. 2006. Real Users, Real Trips, and Real Queries: An Analysis of Destination Search on a Search Engine, in *Annual Conference of Travel and Tourism Research Association (TTRA 2006)*, Dublin, Ireland.

Rajopadhye, M., Ben Ghalia, M., Wang, P., Baker, T. & Eister, C. 2001. Forecasting Uncertain Hotel Room Demand, *Information Sciences*, 132 (1-4), 1-11.

Smith, K. & Pan, B. 2010. *2009 Charleston Area Visitor Intercept Survey*, Charleston, South Carolina: College of Charleston.

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

Song, H., Li, G. & Witt, S. 2008. *The Advanced Econometrics of Tourism Demand*: Routledge.

TIA 2008. *Travelers' Use of the Internet, 2008 Edition*, Washington D.C.: Travel Industry Association of America.

U.S. Census Bureau 2010. *Annual Estimates of the Population of Metropolitan and Micropolitan Statistical Areas: April 1, 2000 to July 1, 2009 (Cbsa-Est2009-01)*, U.S. Census Bureau,, October 1, 2010, <http://www.census.gov/popest/metro/CBSA-est2009-annual.html>.

Vu, J. C. & Turner, L. W. 2006. Regional Data Forecasting Accuracy: The Case of Thailand, *Journal of Travel Research*, 45 (2), 186.

Wong, K., Song, H., Witt, S. & Wu, D. 2007. Tourism Forecasting: To Combine or Not to Combine?, *Tourism Management*, 28 (4), 1068-78.

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

FIGURE

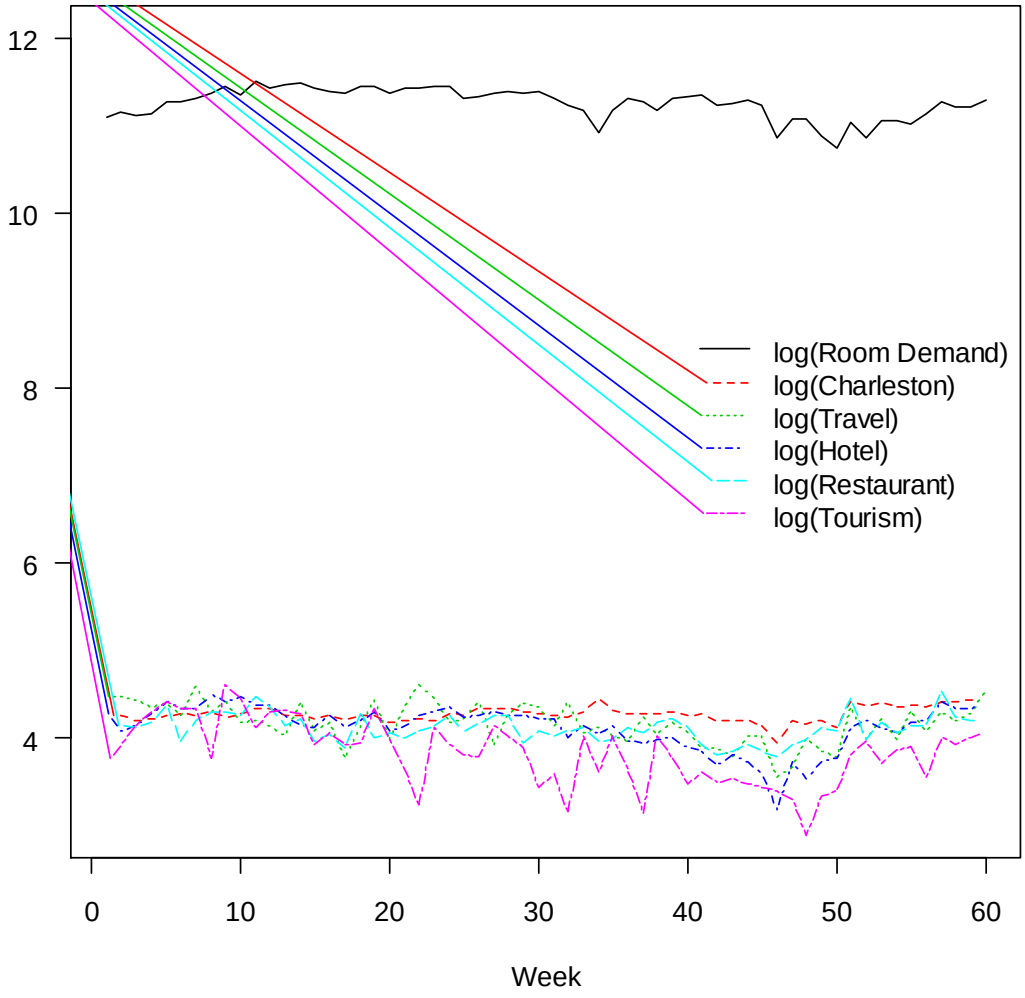

Figure 1: Preliminary Plot of Dependent and Independent Variables

TABLES

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

	Level			Difference		
	One trend & one intercept	One intercept	None	One trend & one intercept	One intercept	None
$\ln y_t$	-3.36	-2.65	0.17	-7.44**	-9.94**	-10.02**
$\ln x_{1t}$	-3.64*	-3.45*	0.35	-10.85**	-10.92**	-10.99**
$\ln x_{2t}$	-5.13**	-4.78**	-0.17	-9.11**	-9.00**	-9.08**
$\ln x_{3t}$	-1.25	-1.40	0.23	-10.85**	-10.90**	-10.98**
$\ln x_{4t}$	-4.99**	-4.92**	-0.06	-12.72**	-12.83**	-12.95**
$\ln x_{5t}$	-4.91**	-3.99**	-0.11	-11.57**	-11.67**	-11.77**
Error term	-3.96*	-3.88**	-3.91**	n.a.	n.a.	n.a.

Note: The figures in the table are ADF statistics.
* and ** indicate that the ADF are significant at 5% and 1% levels, respectively.

Table 2 Estimation results forof the ARMA type models and their ARMAX counterparts

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

Notes: (1) The numbers in parentheses denote the probablity. (2) Jarque-Bera test is a a test for normality, the White test is a test for heteroscedasticity, and Breusch-Goldfrey tests are designed to test for serial correlations (for a detailed explanation of these tests, see, for example, Song, Witt and Li, 2008, pp52-55).

Table 3 Estimation results for three causal econometric models

ADL model			
Variable	Coefficient	t-Statistic	Prob.
constant	11.50	10.40	0.00
$\ln x_{1,t-1}$	-1.05	-3.80	0.00
$\ln x_{1,t-5}$	0.50	2.10	0.04
$\ln x_{2,t-2}$	-0.20	-2.32	0.03
$\ln x_{3,t-1}$	0.51	3.93	0.00
$\ln x_{3,t-6}$	0.30	3.41	0.00
$\ln x_{4,t-1}$	-0.20	-1.49	0.14
$\ln x_{4,t-4}$	0.33	2.84	0.01
$\ln x_{4,t-6}$	-0.30	-2.21	0.03
$\ln x_{5,t-1}$	0.09	1.68	0.10
R-squared	0.75		
Log likelihood	52.75		
BIC	-1.22		
Jarque-Bera test	2.97 (0.23)		
White test	19.91 (0.02)		
Brensch-Godfrey test (rank=1)	3.82 (0.06)		
Brensch-Godfrey test (rank=2)	5.68 (0.06)		
TVP model			
Variable	Coefficient	z-Statistic	Prob.
constant	9.50		
$\ln x_{1t}$	-0.12	11.66	0.00
$\ln x_{2t}$	0.11	-0.50	0.62
$\ln x_{3t}$	0.26	1.90	0.06
$\ln x_{4t}$	0.19	2.06	0.04
$\ln x_{5t}$	-0.03	2.18	0.03
R-squared	0.56		
Log likelihood	4.01		
BIC	0.34		
Jarque-Bera test	7.94 (0.02)		
VAR model			
Variable	Coefficient	t-Statistic	Prob.
constant	5.24	3.80	0.00

$\ln y_{t-1}$	0.62	6.91	0.00
$\ln x_{1,t-1}$	-0.28	-1.32	0.10
$\ln x_{2,t-1}$	-0.15	-1.91	0.03
$\ln x_{3,t-1}$	0.36	2.99	0.00
$\ln x_{4,t-1}$	-0.18	-1.66	0.05
$\ln x_{5,t-1}$	0.02	0.44	0.33
R-squared	0.72		
Log likelihood	56.66		
BIC	-1.44		
Jarque-Bera test	2.26 (0.32)		
White test	286.40 (0.07)		
Brensch-Godfrey test (rank=1)	48.10 (0.09)		
Brensch-Godfrey test (rank=2)	29.70 (0.76)		

Notes: Same as Table 3.

Table 4 Forecasting competition between ARMA and ARMAX models

	naïve	AR(1)	ARMA(1,1)	ARIMA(2,1,2)	ARX(1)	ARMAX(1,1)	ARIMAX(2,1,2)
MAPE	7.515	9.191	8.955	8.782	5.529	5.565	7.918
RMSPE	8.591	10.205	9.962	10.209	6.896	7.021	9.783

Table 5 Forecasting performance competition amongst causal econometric models

	ARX(1)	ARMAX (1,1)	ARIMAX (2,1,2)	ADL	TVP	VAR
MAPE	5.529	5.565	7.918	8.238	8.339	7.874
RMSPE	6.896	7.021	9.783	9.766	9.773	9.019