Data source combination for tourism demand forecasting

Mingming Hu*
Business School
Guangxi University
Nanning 530004, China
School of Hotel and Tourism Management
The Hong Kong Polytechnic University
Hong Kong SAR, China
E-mail: mingming.hu@gxu.edu.cn

Haiyan Song
School of Hotel and Tourism Management
The Hong Kong Polytechnic University
Hong Kong SAR, China
Email: haiyan.song@polyu.edu.hk

*Corresponding Author

Acknowledgement

The authors would like to acknowledge the financial support of Hong Kong Scholars Program and the National Natural Science Foundation of China (No. 71761001)

Data source combination for tourism demand forecasting

Abstract

Search engine data are of considerable interest to researchers for their utility in predicting human behaviour. Recently, search engine data have also been used to predict tourism demand. Models developed based on such data generate more accurate forecasts of tourism demand than pure time series models. The aim of this paper is to examine whether combining causal variables with search engine data can further improve the forecasting performance of search engine data models. Based on an artificial neural network framework, 168 observations during 2005-2018 for short-haul travel from Hong Kong to Macau are involved in the test, and the empirical results suggest that search engine data models with causal variables outperform models without causal variables and other benchmark models.

Keywords. Tourism demand; forecast accuracy; artificial neural network; causal economic variables; search engine

Introduction

Due to the impossibility of stockpiling unused hotel rooms and unoccupied airline seats (Law, 2000; Chu, 2004), accurate tourism demand forecasts can provide practitioners and policy makers with useful information for formulating effective

tourism marketing and development strategies/policies (Dergiades et al., 2018). Accurately forecasting the demand for tourism services is a difficult task for practitioners and academics (Goh and Law, 2011). Reviewing the current tourism demand forecasting literature, Wu et al. (2017) divide the quantitative forecasting methods used by practitioners into three main categories: non-causal time series, econometric and artificial neural network (ANN) based methods. The data sources used for these methods are historical tourism demand series, causal variables or a combination of both.

With the increasing use of the Internet and smartphones, search engines have become efficient channels for people to obtain information. Online search histories reflecting users' interests are recorded by these search engines, and provide a rich source of data. Google, a multinational technology company that provides the most convenient and frequently used search engine services, has published search query data since 2004. The value of such data has been widely recognised; for instance, it has been successfully used to detect influenza epidemics (Ginsberg et al., 2009). Recently, search engine data have been used, sometimes in combination with historical tourism demand data, to predict tourism demand (Pan et al., 2012; Yang et al., 2015; Bangwayo-Skeete and Skeete, 2015; Önder and Gunter, 2016; Rivera, 2016; Li et al., 2017; Volchek et al., 2018). Empirical evidence shows that the use of search engine data improves tourism forecasting performance (Pan et al., 2012; Bangwayo-Skeete and Skeete, 2015). However, search query-based models do not normally include causal variables. Therefore, the question is whether introducing causal economic variables such as tourist

income, tourism prices and exchange rate data into search query-based models can further improve their forecasting performance. In other words, are causal variables still useful in search query-based tourism demand forecasting? This study addresses that question by developing a conceptual framework to interpret the roles of different data sources in tourism demand and quantitatively evaluating the usefulness of causal variables in AI-based and econometric models incorporating search engine data.

This study investigates monthly tourism demand for short-haul travel from Hong Kong to Macau during 2005-2018; in total, 168 observations are involved. Macau was chosen for three reasons. First, it is well known for its gaming industry (Lu, 2011) and fine dining (Law et al., 2019), which attract short-term visitors from surrounding regions for leisure and vacation purposes (Fong, 2017). Second, as a significant proportion of visitors are short-haul travellers, they are more likely to be influenced by economic factors and online information. Third, the completion of the Hong Kong-Zhuhai-Macao Bridge has strengthened the connections between Hong Kong and Macau. As a result, Macau has become a popular short-haul destination for Chinese tourists from neighbouring regions.

This paper makes three main contributions. First, a conceptual framework is proposed for combining data sources that clarifies the role of historical tourism demand series, causal variables and search engine data in predicting tourism demand. Second, an ANN model is used to combine causal variables, search engine data and historical tourism demand. Multiple lags for each variable are included in this model. Third, the role of causal variables is tested empirically, and the results confirm the usefulness of

causal time series in improving the accuracy of the ANN model in forecasting tourism demand.

Literature review

In tourism demand forecasting, causal variables, historical series and search query data are frequently used. Causal variables and lagged dependent and explanatory variables are commonly used in traditional tourism demand forecasting methods, such as pure time series methods, econometric methods, and AI-based methods. When search query data are involved, we name these models 'search query-based methods'. This literature review focuses on these two kinds of approaches.

Traditional tourism demand forecasting methods

The tourism demand forecasting literature classifies forecasting methods into the following categories: non-causal time series, econometric and artificial intelligence (AI) based methods (Wu et al., 2017; Song et al., 2019). Non-causal time series methods extrapolate historical tourism demand series to generate forecasts (Burger et al., 2001; Chu, 2009; Chang and Liao, 2010; Ramos and Rodrigues, 2014; Tsui et al., 2014; Wu et al., 2017). According to Ramos and Rodrigues (2014) and Wu et al. (2017), the most frequently used non-causal time series models are no-change models (Naïve I), constant growth rate models (Naïve II), exponential smoothing (ES) models, autoregressive moving average (ARMA) models (such as the autoregressive integrated moving

average model, ARIMA, and the seasonal ARIMA model, SARIMA) and structural time series (STS) models.

Econometric models (Song and Li, 2008; Song and Witt, 2012; Onafowora and Owoye, 2012; Ramos and Rodrigues, 2014; Wu et al., 2017; Song et al., 2019) incorporate causal variables into pure time series models. In addition to forecasting, econometric models explore the relationship between tourism demand and causal variables. Li, Song and Witt (2005) and Song and Li (2008) point out that the main causal variables of tourism demand include tourist income, tourism prices at the destination relative to those in the country of origin, tourism prices at competing destinations (substitute prices) and exchange rates. Transportation costs (Dritsakis, 2004; Lim, 1999), marketing expenses (Law, 2000; Law and Au, 1999; Law, 2001) and climate (Lise and Tol, 2002; Goh, 2012; Li et al., 2017; Li et al., 2018) are also recognised as important factors. Examples of widely used econometric methods include error correction models (ECMs; Kulendran and Wilson, 2000; Song and Witt, 2000), the autoregressive distributed lag model (ADLM, a general form of the error correction model; Song et al., 2003; Song et al., 2003), vector autoregressive (VAR) models (Gunter and Önder, 2016), time varying parameter (TVP) models (Song et al., 2003) and the mixed-data sampling (MIDAS) approach (Bangwayo-Skeete and Skeete, 2015). To improve forecast accuracy, these models are often combined; examples include the ECM-LAIDS model (Mangion et al., 2005), the TVP-LR-AIDS model (Li et al., 2006), the TVP-STS model (Song et al., 2011) and the TVP-EC-AIDS model (Wu et al., 2012).

AI-based methods (Claveria et al., 2015; Law, 2000; Law and Au, 1999; Palmer et al., 2006; Kon and Turner, 2005) aim to establish non-linear connections between tourism demand and its lagged values or explanatory variables. ANN models (Law, 2000; Law and Au, 1999), support vector regression models (Chen and Wang, 2007), Gaussian process regression (GPR) models (Wu et al., 2012) and deep learning approaches (Law et al.,2019) are also used to predict tourism demand. AI-based methods are particularly accurate in forecasting tourism demand (Song and Li, 2008; Wu et al., 2017). The data sources used for these AI-based methods are historical tourism demand series, causal variables or a combination of the two.

Search query based tourism demand forecasting

Due to the wide use of the Internet and smartphones, search engines have become an important platform for searching for information around the world. Google is one of the most powerful search engines, with 91.4% of the global market share of all search engines in 2018 (www.gs.statcounter.com). In addition, it has published search intensity data since 2004. Search query data have been used successfully to detect influenza epidemics (Ginsberg et al., 2009), predict abnormal stock returns and trading volumes (Joseph et al., 2011) and identify housing market trends (Wu and Brynjolfsson, 2015). In the tourism context, search engines help tourists obtain useful information on restaurants, hotels, transportation, attractions and retail stores at their planned destination. As a result, search engines' histories reflect tourists' preferences in terms of destinations, cuisine and accommodations.

Pan et al. (2012) demonstrate that search engine data can be used to accurately predict demand for hotel rooms. Yang et al. (2015) use search query volume to predict the number of visitors to Hainan Province and compare the predictive power of forecasting models based on two search engines, Google and Baidu. The results show that both types of search engine data significantly improve forecast accuracy, and that Baidu query data perform better due to its larger market share in China. Similarly, Bangwayo-Skeete and Skeete (2015) conduct a composite search for 'hotels and flights' from source countries to popular destinations in the Caribbean to test the performance of autoregressive MIDAS models using search query data. The results show that search engine data have significant benefits for forecasting tourism demand. Önder and Gunter (2016) evaluate the predictive power of Google Trends by focusing on Vienna as a destination and using seasonal and seasonally adjusted data. The results confirm that forecast error is reduced when Google Trends data are used. Rivera (2016) treats search query volume data as a representation of an unobservable process and uses a dynamic linear model to forecast tourism demand in Puerto Rico, taking non-resident hotel registrations as a proxy variable. The results suggest that the search query-based model only outperforms its competitors when the forecast horizon is greater than six months. Li et al. (2017) propose a composite search index using the generalised dynamic factor model (GDFM) to forecast tourism demand and compare its forecast performance with two benchmark models. The results show that the GDFM outperforms competing benchmark models. Volchek et al. (2018) use time series, econometric and ANN models with the Google Trends index to forecast the number of visits to five London

museums. They find that the inclusion of this index in pure time series models generates the most accurate forecasts, and that no other model outperforms its competitors in all situations. All of these studies confirm that search engine data can improve the forecast accuracy of forecasting models if they are properly integrated.

Traditional tourism demand forecasting methods link tourism demand with its causal variables, whereas search engine data methods examine the association of search engine data with tourism demand. To the best of our knowledge, the question of whether causal variables can be useful if they are introduced into modern search engine data methods in tourism forecasting remains unanswered.

As an AI-based method, ANN models highlight the non-linear relationship between tourism demand and input variables. They are widely used to predict tourism demand with causal variables as inputs, and are known to produce accurate forecasts of tourism demand (Uysal and El Roubi, 1999; Law and Au, 1999; Pai and Hong, 2005). To identify the linear or non-linear relationship for high forecast accuracy, the ANN model is used in this research.

Methodology

Conceptual framework

Researchers generally consider three data sources to forecast tourism demand: historical tourism demand series, historical series of causal variables and search query series. Researchers recognise that tourism demand series can be short memory series or

long memory series, reflecting cyclical or seasonal changes in tourist behaviour (Odaki,1993; Gil-Alana, 2005). Forecasting using lagged demand series depends on the intensity of short or long memories. Morley (2009) argues that a simple lagged demand term is not sufficient to account for the dynamics of tourism demand models, and finds that causal variables help to specify the demand model. From a socio-psychological and economic perspective, causal variables determine tourists' demand and search motivation when they search for travel information online (Heung et al., 2001). Search volume data generally record tourists' behaviour by indicating their search frequency. Bangwayo-Skeete and Skeete (2015) provide evidence that search query information offers significant benefits in forecasting. As a result, causal variables are dynamic factors that form unobservable tourism demand, and both historical series and search queries help to evaluate potential tourism demand. Thus, we propose a conceptual framework (Figure 1) to describe the process of tourism demand realisation, which can help us specify the behavioural models in the forecasting exercise. The primary influencing factors (causal variables) of tourism demand include tourist income, tourism prices at the destination relative to those in the source markets, tourism prices at competing destinations, exchange rates, transportation costs and marketing promotion expenditures. These determine the type of holiday tourists are interested in, and tourists then search online for information that matches their demand. The search frequency reveals their preference. Search query data, together with historical series (continuing historical patterns), are two dimensions that account for the dynamics of tourism demand. Thus, both contribute to accurately forecasting the demand for tourism.

[Insert Figure 1 Here]

Variables

To specify the tourism demand (TD) model, we include the following key causal variables based on other studies: tourist income (TI), tourism prices (TP) in Macau relative to those of Hong Kong, tourism prices in competing destinations (substitute prices, SP) and exchange rates (ER; Li et al., 2005; Song and Li, 2008). As the exchange rate between the Hong Kong dollar and the Macau pataca remained roughly the same between January 2001 and December 2018, ER is not considered in this study. When the ADLM is specified, lagged (L) tourism demand is included (Song et al., 2003). Li et al. (2017) develop a forecasting framework using search engine data, with search query keywords related to Dining (QD), Lodging (QL), Shopping (QS), Transportation (QTR), Tours (QT) and Recreation (QR). They find that these search queries provide useful information for forecasting tourism demand. However, Macau is a well-known international gaming (Lu, 2011) and fine-dining destination (Law et al., 2019). Therefore, shopping and sightseeing are not the main motivations for visitors from Hong Kong. For this reason, QS and QT are excluded from the search query list.

In addition, a public holiday variable (HOLIDAY) is included as a causal variable. Based on a survey of 406 Japanese leisure travellers in Hong Kong, Heung et al. (2001) find that 'enjoying their holidays' is one of the most important motives for holidaying.

In this study, the tourism demand function is written as

$$TD = f\begin{bmatrix}(\mathrm{L})TD, (\mathrm{L})TI, (\mathrm{L})TP, (\mathrm{L})SP, HOLIDAY, \\ (\mathrm{L})QD, (\mathrm{L})QL, (\mathrm{L})QTR, (\mathrm{L})QR\end{bmatrix}$$

(i) where TD is visitor arrivals (Arr) from Hong Kong to Macau between January 2005 and December 2018. Data are collected from the Department of Statistics and Census Service of Macau (See Figure 2).

[Insert Figure 2 Here]

(ii) TI is Hong Kong's tourist income, as represented by Hong Kong's gross domestic product (GDP). Hong Kong's quarterly GDP data between 2005Q1 and 2018Q4 are collected from the Hong Kong Census and Statistics Department. Monthly tourism demand is affected by lagged quarterly GDP, which can be directly included in the artificial neural network.

(iii) TP corresponds to tourism prices in Macau relative to those in Hong Kong, as measured by the ratio of Macau's consumer price index (CPI) to that of Hong Kong. Macau's monthly CPI is collected from the Department of Statistics and Census Service of Macau (January 2005 to December 2018). Hong Kong's monthly CPI (January 2005 to December 2018) is obtained from the Hong Kong Census and Statistics Department. Therefore, tourism prices in Macau relative to those of Hong Kong are calculated by

$$CPI_{new} = \frac{CPI_M}{CPI_{HK}}. \quad (1)$$

(iv) SP is the substitute price at competing destinations. We choose the Chinese mainland (CM), Chinese Taipei (T), Korea (K) and Japan (J) as competing destinations

for Macau. Indeed, the departures of Hong Kong residents to these destinations accounted for 87% of Hong Kong's total departures in 2017. Therefore, the substitute price index is calculated as

$$CPI_{sub} = \frac{w_{CM}CPI_{CM}/ER_{CM}+w_TCPI_T/ER_T+w_KCPI_K/ER_K+w_JCPI_J/ER_J}{CPI_{HK}/ER_{HK}} , \quad (2)$$

where w_{CM}, w_T, w_k, w_J are the relevant marketing shares of the competing destinations. The monthly CPI of these destinations is collected from the statistical bureaus of the respective source markets between January 2005 and December 2018, January 2001 being the base year. All prices are converted to US dollars (US$). ER_{CM}, ER_T, ER_K, ER_J and ER_{HK} are the average exchange rates over the last five years (2014-2018) between the Chinese yuan (CNY), Taiwanese dollar (TWD), Korean won (KRW), Japanese yen (JPY), Hong Kong dollar (HK$) and US$, respectively, collected from the Fusion Media Limited website (www.investing.com).

(v) QD, QL, QTR, QR are search query data related to Dining, Lodging, Transportation and Recreation, respectively. As Hong Kong is the tourists' place of origin, monthly search queries are obtained from Google Trends between January 2005 and December 2018 (https://trends.google.com/trends/). The languages used are English and traditional Chinese. The search query keywords are shown in Table 1 and the monthly search query data in Figure 3.

[Insert Table 1 Here]

[Insert Figure 3 Here]

(vi) $HOLIDAY$ is a dummy variable representing public holidays in Hong Kong. Monthly holiday data (January 2005 to December 2018) are collected from the Hong Kong online calendar (http://m.calendar411.com).

(vii) (L) is a lag operator. The number of lags for each variable is determined by the Akaike Information Criterion (AIC) index (Sakamoto, Ishiguro & Kitagawa, 1986).

AI Models

As previously mentioned, AI-based models are frequently used to predict tourism demand (Law, 2000; Law and Au, 1999; Palmer et al., 2006; Kon and Turner, 2005; Claveria et al., 2015). They are known for their greater accuracy in tourism forecasting than regression and time series models. However, their main disadvantage is that the relationship between input and output variables is unknown (they are 'black boxes'), so they cannot be used to inform decisions (Li and Song, 2008). In contrast, econometric models require a careful selection of explanatory variables to avoid the problem of collinearity (Dormann et al., 2013).

One AI-based method involves back-propagation neural networks (BPNN; Law, 2000; Wang et al., 2015) that aim to connect input variables and output variables. A given network (Figure 4) consists of an input layer, an output layer and one or more hidden layers. Each layer contains artificial neurons (nodes) connected to the artificial neurons (nodes) of the adjacent layer(s). Each connection between a pair of artificial

neurons, like the synapses in a biological brain, can transmit signals to one another. The strength of the connection is expressed by the weight, which automatically adjusts according to the error between outputs and actual values based on the training set. After the model is trained, non-linear relationships are identified between the input and output variables. When values of variables are input into the model, the trained neural network can output forecast values.

[Insert Figure 4 Here]

The training process is carried out according to the following steps. We take a three-layer neural network as an example.

(1) Initialise the network by assigning random numbers to the weights of the connections between the input layer and hidden layer w_{ij}, and the hidden layer and output layer w'_{jk}.

(2) Transfer values forward

(2.1) Transfer input values from input variables (I_i) to the neurons (nodes) (y_j) on the hidden layer by

$$y_j = f(\sum_{i=1}^{n} I_i w_{ij} + \theta_j), j = 1,2, \dots, m. \quad (3)$$

where $f(\cdot)$ is the activation function, which represents the rate of action potential firing neuron. A sigmoidal activation function is commonly used,

$$f(x) = \frac{1}{1+e^{-x}} \quad (4)$$

(2.2) Transfer neurons' values (y_j) on hidden layer to output neuron O by

$$O = f(\sum_{j=1}^{m} y_j w_j' + \mu)$$

(5)

(3) Estimate the error, transfer error back and adjust the weight of connections.

(3.1) Estimate the error (E) between the tourism demand forecast (O) and the actual demand (Arr) on the training data set; $E = \frac{1}{2}\sum(Arr - O)^2$.

(3.2) Update the connection weights w_j' between the node j on the hidden layer and the output node of the output layer by

$$w_j' = w_j' + \eta \frac{\partial E}{\partial w_j'}, \quad (6)$$

and update the connection weights w_{ij} between the node i on input layer and the node j on the hidden layer by

$$w_{ij} = w_{ij} + \eta \frac{\partial E}{\partial w_{ij}}, \quad (7)$$

where η is the learning rate. Its value is between 0 and 1. $\frac{\partial E}{\partial w_{ij}}$ is the marginal utility of connecting weight w_{ij} on the error.

(4) Repeat (2) and (3) until the error E is within an acceptable range.

With a trained neural network, a prediction can be made by inputting the input variable values.

A neural network toolbox in MATLAB (Version R2018b) is available for conducting these processes. newff() helps to create a neural network with specified layers and numbers of nodes in each layer, train() is used to train the network on the training data set and net() can be used to apply the trained network by inputting the input variable values and outputting the prediction value.

Benchmark forecasting models

A univariate ARIMA model and multivariate ADL model are estimated as benchmark models. The ARIMA model generates tourism demand forecasts based solely on tourism demand series, while the ADL model includes lagged output, causal variables and the search query variable to predict tourism demand.

ARIMA model

The ARIMA model, a univariate model proposed by Box and Jenkins (1970), is the latest generation of models in the ARMA family. It integrates the autoregressive (AR) model and the moving-average (MA) model. The specificity of this model is that it depends only on historical data. It has become extremely popular in recent years (Song & Li, 2008). The ARIMA (*p, d, q*) model can be written as

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

where y_t is tourism demand at time *t* (i.e. the number of tourist arrivals in month *t*). Δ is the difference function (i.e. $\Delta y_t = y_t - y_{t-1}$) and *d* refers to the difference rank. μ is a constant term. ε_t is the error term at time t. ϕ_i and θ_i are coefficients. p and q are the lag length.

ADL model

The ADL model (Pesaran et al., 1996; Pesaran and Shin, 1998) is one of the main econometric forecasting methods (Song and Li, 2008). In the model, current tourism demand is regressed on lagged values of tourism demand and on current and lagged values of one or more explanatory variables. These explanatory variables are normally

economic variables, such as tourist income, tourism prices at the destination relative to those in the countries/regions of origin, tourism prices at competing destinations (substitute prices) and exchange rates (Song and Li, 2008; Gunter and Onder, 2015). In addition, Pan, Wu and Song (2012) use search engine data to predict hotel room demand and incorporate online big data as explanatory variables in econometric models. Online big data are widely used as explanatory variables (Yang et al., 2015; Bangwayo-Skeete and Skeete, 2015; Rivera, 2016; Huang et al., 2017; Li et al., 2017). The general ADL model can be written as

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{u} \sum_{j=0}^{p} \alpha_{ij} x_{i,t-j} + \varepsilon_t,$$

where $x_{i,t}$ indicates the value of the *i*th explanatory variable at time *t* (i.e. $x_{1,t}$ is tourist income in month *t*, $x_{2,t}$ is tourism prices in month *t*) and *u* is the number of explanatory variables. α_{ij} is the coefficient of the *i*th explanatory variable at lag *j*.

Accuracy measure

To verify the forecasting accuracy of the proposed models, we adopt the mean absolute deviation (MAD), mean squared error (MSE), mean absolute percentage error (MAPE), root mean square error (RMSE) and root mean square percentage error (RMSPE).

$$MAD = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}, \quad MAPE = \frac{\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i}}{n}, \quad MSE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|^2$$

$$RMSE = \left[\frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|^2 \right]^{1/2}, \quad RMSPE = \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{|y_i - \hat{y}_i|}{y_i} \right)^2 \right]^{1/2}$$

Results and implications

Preparation

The data sample covers the period from January 2005 to December 2018. The frequency of the data varies according to the variables. For example, Hong Kong's GDP is a quarterly series, while all of the other variables are monthly series. Using lags of these variables that may affect the demand variable, 156 observations from January 2005 to December 2017 are selected to train the ANN model and estimate the parameters of the ADL and ARIMA models. The period from January 2018 to December 2018 is reserved for evaluating the performance of these models for one-step-ahead forecasts. The variables included in the different models are defined as follows.

Historic tourism demand series (HTDS): Visitor arrivals to Macau from Hong Kong.

Causal variables (CV): Hong Kong's GDP; Macau's CPI relative to that of Hong Kong; a CPI index of substitute destinations; and *HOLIDAY*, a dummy variable for public holidays in Hong Kong.

Search engine data (SED): QD, QL, QTR, QR related to Dining, Lodging, Transportation and Recreation, respectively.

The AIC (Sakamoto, Ishiguro & Kitagawa, 1986) is also used to determine the lags of input variables to interpret current tourism demand.

Specifying competing models

To test the values of the causal variables, two groups of data are used. The data in the first group include HTDS, CV and SED. The data in the second group only include HTDS and SED for comparison purposes. A three-layer ANN model is trained on each set of data, and the performance is compared. According to Wanas et al. (1998), the best performance of a neural network occurs when the number of hidden nodes is equal to $\log_2 Number\ of\ Samples$. With this rule, we set $\log_2 156 \approx 8$ nodes in the hidden layer. Each group of data is used as the input to generate a new ANN model. Model 1 contains the causal variables as input for the ANN model, while Model 2 does not include any causal variables.

Model 1: $\mathrm{ANN}_1(\mathrm{HTDS\ \&CV\ \&SED})$

$$ANN_1\begin{bmatrix} (\mathrm{L}^3)Arr, (\mathrm{L}^4)Arr, (\mathrm{L}^5)Arr, (\mathrm{L}^9)Arr, (\mathrm{L}^{12})Arr, \\ (\mathrm{L}^1)GDP, (\mathrm{L}^5)GDP, (\mathrm{L}^4)CPI, (\mathrm{L}^6)CPI, (\mathrm{L}^{12})CPI, (\mathrm{L}^3)RSPCPI, HOLIDAY, \\ (\mathrm{L}^8)QD, (\mathrm{L}^{11})QD, (\mathrm{L}^1)QL, (\mathrm{L}^3)QL, (\mathrm{L}^5)QL, (\mathrm{L}^{11})QL, (\mathrm{L}^4)QTR, (\mathrm{L}^1)QR, (\mathrm{L}^{10})QR \end{bmatrix}$$

Model 2: $\mathrm{ANN}_2(\mathrm{HTDS\ \&SED})$

$$ANN_2\begin{bmatrix} (\mathrm{L}^3)Arr, (\mathrm{L}^4)Arr, (\mathrm{L}^5)Arr, (\mathrm{L}^9)Arr, (\mathrm{L}^{12})Arr, \\ (\mathrm{L}^8)QD, (\mathrm{L}^{11})QD, (\mathrm{L}^1)QL, (\mathrm{L}^3)QL, (\mathrm{L}^5)QL, (\mathrm{L}^{11})QL, (\mathrm{L}^4)QTR, (\mathrm{L}^1)QR, (\mathrm{L}^{10})QR \end{bmatrix}$$

Empirical results

We set 8 nodes in the hidden layer, the maximum training epochs to 5,000 and the error tolerance to 0.001. MSE is used as an error measurement during the training process. We train the model recursively to generate one-month-ahead forecasts. After training the ANN, the goodness of fit for Model 1 and Model 2 according to MSE is 8.71×10^8 and 1.70×10^9, respectively. The forecasting errors of Model 1 and Model 2 are measured by MAD, MAPE, MSE, RMSE and RMSPE. These error measures are presented in Table 2. The results show that Model 1 improves forecasting performance between 5.13% and 30.64% (MAD), 5.66% and 31.96% (MAPE), 7.90% and 50.09% (MSE), 4.03% and 29.35% (RMSE) and 4.59% and 31.75% (RMSPE) compared with Model 2. In particular, when tourist arrivals from January 2018 to March 2018 are taken into account, the performance of Model 1 increases to 30.64% (MAD), 31.96% (MAPE), 50.09% (MSE), 29.35% (RMSE) and 31.75% (RMSPE) compared with Model 2. With causal variables, performance during both training and testing is improved. This means that useful information is provided by causal variables. More accurate nonlinear relationships between inputs and tourism demand are established. These variables improve the forecasting ability of the neural network. This suggests that causal variables can help improve forecast accuracy when ANN-based models are used to predict tourism demand.

[Insert Table 2 Here]

Two questions still need to be addressed. The first is whether forecast error can be reduced when causal variables are introduced into econometric models with HTDS and SED. The second concerns the forecasting performance of ANN-based models relative to other forecasting models, including causal econometric models and time series models. To answer these questions, an ADL model with HTDS, CV and SED, an ADL model with HTDS and SED, and an ARIMA model with HTDS are estimated to generate monthly rolling forecasts for 2018. The goodness of fit for these three models is measured by MSE; the values are respectively 1.01×10^9, 1.41×10^9 and 3.71×10^9. Their forecasting performance is shown in Table 3. The performance of ADL (HTDS & SED) is 4.8768×10^4 (MAD), 8.5988×10^{-2} (MAPE), 4.0974×10^9 (MSE), 6.4011×10^4 (RMSE) and 1.0591×10^{-1} (RMSPE).

A comparison of ADL (HTDS & CV & SED) with ADL (HTDS & SED) shows that the accuracy of the ADL model improves by 5.35% (MAD), 3.50% (MAPE), 12.05% (MSE), 6.22% (RMSE) and 3.57% (RMSPE) after the causal variables are added to the model. With causal variables, both the fit and forecasting performance are improved. These improvements show that causal variables can provide useful information and enhance the explanatory power of econometric models of tourism demand.

A comparison of ANN_1 (HTDS & CV & SED) with ADL (HTDS & CV & SED) using the same data sources gives the same results, which confirm the findings of Song and Li (2008) and Wu et al. (2017) that the AI-based method is achieves excellent results when data for causal variables are lacking.

Lastly, a comparison of the performance of ANN_1 (HTDS & CV & SED) with ADL (HTDS & CV & SED), ADL (HTDS & SED) and ARIMA (HTDS), ANN_1 (HTDS & CV & SED) shows an improved accuracy of between 10.41% and 20.39%, 15.65% and 29.41%, 3.25% and 25.08%, respectively. These results show that the ANN_1 (HTDS & CV & SED) model outperforms the other models. There are two reasons for this. First, ANN_1 (HTDS & CV & SED) is an AI-based model, which Song and Li (2008) and Wu et al. (2017) have shown performs particularly well. Second, ANN_1 (HTDS & CV & SED) incorporates HTDS, CV and SED.

[Insert Table 3 Here]

Implications

This study has several implications for destination tourism management. To formulate effective tourism marketing and development strategies/policies, practitioners and policymakers should consider using three types of data in forecasting tourism demand. (1) Causal variables such as the income of tourists, prices at the destination, prices at the substitute destination, marketing expenditure, exchange rates and transportation costs are still very useful in explaining the determinants of tourism demand. (2) Search query data from sites such as Baidu and Google for keywords related to dining, lodging, shopping, transportation, tours and recreation also contain useful information for forecasting. (3) Lagged tourism demand data contain important

information on the dynamics of tourism demand from the historic perspective, which can enhance forecasting accuracy if used properly.

However, including too many data series could lead to model overfitting. Thus, the AIC or Bayesian Information Criterion (BIC) index should be used to decide upon the inclusion of the variables and their lags.

Using selected variables and lags, AI-based models or econometric models can be utilised to forecast tourism demand. The results of this paper and those of Song and Li (2008) and Wu et al. (2017) show that an AI-based model can outperform other models if used properly, especially in situations where the sample size is small.

The AI-based model is also more flexible than econometric models in forecasting tourism demand in different frequencies. This is useful given that mixed frequency data can provide useful information for decision making at different time intervals.

Conclusion

The ability to accurately forecast tourism demand is important for practitioners and policymakers. With the growing use of the Internet and smartphones, search engines have become a globally important platform for users searching for information. Because search queries indicate the interests of users, search query data can contribute to tourism demand forecasting and improve the accuracy of pure time series models. We also investigate whether causal variables can improve accuracy. Theoretically, a conceptual framework for tourism demand realisation is proposed to support the

integration of historical tourism demand, causal variables and search queries. Based on this conceptual framework, we find that causal variables help the decision-making process and improve performance. To quantitatively analyse the role of causal variables, we specify two competing models. Model 1 contains the causal variables as input for the ANN model, whereas Model 2 does not include any causal variables. The training sample is 156 observations of tourism demand for short-haul trips from Hong Kong to Macau between January 2005 and December 2017 (Twelve more observations are used for testing). We train the models and generate one-month-ahead forecasts recursively from January 2018 to December 2018. Comparison of the models proves that causal variables improve the forecasting accuracy of the ANN and ADL models. To test the performance of the ANN model with causal variables, we compare it with two ADL models and one ARIMA model. The comparison confirms that the ANN model with causal variables outperforms these benchmark models. The reasons are in two aspects. The one comes from the model. AI-based model can generate particular accurate forecast. It is proved both in this paper and the existing literature (Song and Li, 2008; Wu et al., 2017). The second reason is the multiple data sources. ANN model with causal variables incorporates more information to interpret the tourism demand. The role of search query data is proved by Bangwayo-Skeete and Skeete (2015) that search query data help to improve the accuracy of forecasting for tourism demand. That of historical series is shown by Odaki (1993) and Gil-Alana (2005) that time series own short- or long- memories. That of the causal variables are quantitively proved in this paper that causal variables help to improve the performance.

In the context of research combining historical tourism demand series, causal variables and search query data to predict tourism demand, the main contribution of this study is that we propose a conceptual framework supporting the integration of causal variables, search queries and historical data. In addition, we quantitatively prove the superiority of the AI-based model with causal variables in tourism demand forecasting.

This study has several limitations, some of which could be addressed in future research. First, tourist income, tourism prices at the destination and substitute prices are used as causal economic variables; however, the populations, exchange rates, transportation costs, marketing promotion and climates of various markets may also affect tourist arrivals. Future researchers could include these variables in forecasting models to further improve their accuracy. Second, this study focuses on short-haul travel from Hong Kong to Macau. Future researchers could further explore the performance of the models in long-haul tourism demand forecasting.

References

Bangwayo-Skeete PF and Skeete RW (2015) Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. Tourism Management 46: 454-464.

Box GEP and Jenkins GM (1970) Time series analysis, forecasting and control. San Francisco: Holden Day.

Burger CJSC, Dohnal M, Kathrada M and Law R (2001) A practitioner's guide to time-series methods for tourism demand forecasting – A case study of Durban, South Africa. Tourism Management 22(4): 403-409.

Chang YW and Liao MY (2010) A seasonal ARIMA model of tourism forecasting: The case of Taiwan. Asia Pacific Journal of Tourism Research 15(2): 215-221.

Chen KY and Wang CH (2007) Support vector regression with genetic algorithms in forecasting tourism demand. Tourism Management 28(1): 215-226.

Chu FL (2004) Forecasting tourism demand: A cubic polynomial approach. Tourism Management 25(2): 209-218.

Chu FL (2009) Forecasting tourism demand with ARMA-based methods. Tourism Management 30(5): 740-751.

Claveria O, Monte E and Torra S (2015) Tourism demand forecasting with neural network models: Different ways of treating information. International Journal of Tourism Research 17(5): 492-500.

Dergiades T, Mavragani E and Pan B (2018) Google Trends and tourists' arrivals: Emerging biases and proposed corrections. Tourism Management 66: 108-120.

Dormann CF, Elith J, Bacher S, et al. (2013) Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography 36(1): 27-46.

Dritsakis N (2004) Cointegration analysis of German and British tourism demand for Greece. Tourism Management 25(1): 111-119.

Fong DKC (2017) Macau visitor profile study 2017. Institute for the Study of Commercial Gaming, University of Macau (2017).

Friedman JH (1997) On bias, variance, 0/1 – Loss, and the curse-of-dimensionality. Data Mining and Knowledge Discovery 1(1): 55-77.

Gil-Alana LA (2005) Modelling international monthly arrivals using seasonal univariate long-memory processes. Tourism Management 26(6): 867-878.

Ginsberg J, Mohebbi MH, Patel RS, et al. (2009) Detecting influenza epidemics using search engine query data. Nature 457(7232): 1012.

Goh C (2012) Exploring impact of climate on tourism demand. Annals of Tourism Research 39(4): 1859-1883.

Goh C and Law R (2011) The methodological progress of tourism demand forecasting: A review of related literature. Journal of Travel & Tourism Marketing 28(3): 296-317.

Gunter U and Önder I (2015) Forecasting international city tourism demand for Paris: Accuracy of uni- and multivariate models employing monthly data. Tourism Management 46: 123-135.

Heung VCS, Qu H and Chu R (2001) The relationship between vacation factors and socio-demographic characteristics: The case of Japanese leisure travelers. Tourism Management 22(3): 259-269.

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

Joseph K, Wintoki MB and Zhang Z (2011) Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. International Journal of Forecasting 27(4): 1116-1127.

Kon SC and Turner LW (2005) Neural network forecasting of tourism demand. Tourism Economics 11(3): 301-328.

Kulendran N and Wilson K (2000) Modelling business travel. Tourism Economics 6(1): 47-59.

Law R (2000) Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. Tourism Management 21(4): 331-340.

Law R (2001) The impact of the Asian financial crisis on Japanese demand for travel to Hong Kong: A study of various forecasting techniques. Journal of Travel & Tourism Marketing 10(2-3): 47-65.

Law R and Au N (1999) A neural network model to forecast Japanese demand for travel to Hong Kong. Tourism Management 20(1): 89-97.

Law R, Li G, Fong DKC and Han X (2019) Tourism demand forecasting: A deep learning approach. Annals of Tourism Research 75, 410-423.

Li G, Song H and Witt SF (2005) Recent developments in econometric modelling and forecasting. Journal of Travel Research 44(1): 82-99.

Li G, Song H and Witt SF (2006) Time varying parameter and fixed parameter linear AIDS: An application to tourism demand forecasting. International Journal of Forecasting 22(1): 57-71.

Li H, Goh C, Hung K and Chen JL (2018) Relative climate index and its effect on seasonal tourism demand. Journal of Travel Research 57(2): 178-192.

Li H, Song H and Li L (2017) A dynamic panel data analysis of climate and tourism demand: Additional evidence. Journal of Travel Research 56(2): 158-171.

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

Lim C (1999) A meta-analytic review of international tourism demand. Journal of Travel Research 37(3): 273-284.

Lise W and Tol RS (2002) Impact of climate on tourist demand. Climatic Change 55(4): 429-449.

Lu D (2011) Emic interpretations of global gaming destinations: Travel blog stories about experiencing Macau, Las Vegas, and Monaco. In Tourism sensemaking: Strategies to give meaning to experience (pp. 111-152). Emerald Group Publishing Limited.

Mangion ML, Durbarry R and Sinclair MT (2005) Tourism competitiveness: Price and quality. Tourism Economics 11(1): 45-68.

Morley CL (2009) Dynamics in the specification of tourism demand models. Tourism Economics 15(1): 23-39.

Odaki M (1993) On the invertibility of fractionally differenced ARIMA processes. Biometrika 80(3): 703-709.

Onafowora OA and Owoye O (2012) Modelling international tourism demand for the Caribbean. Tourism Economics 18(1): 159-180.

Önder I and Gunter U (2016) Forecasting tourism demand with Google Trends for a major European city destination. Tourism Analysis 21(2-3): 203-220.

Pai P and Hong W (2005) An improved neural network model in forecasting arrivals. Annals of Tourism Research 32(4): 1138-1141.

Palmer A, Montano JJ and Sesé A (2006) Designing an artificial neural network for forecasting tourism time series. Tourism Management 27(5): 781-790.

Pan B, Wu CD and Song H (2012) Forecasting hotel room demand using search engine data. Journal of Hospitality and Tourism Technology 3(3): 196-210.

Pesaran MH and Shin Y (1998) An autoregressive distributed-lag modelling approach to cointegration analysis. Econometric Society Monographs 31: 371-413.

Pesaran MH, Shin Y and Smith RJ (1996) Testing for the existence of a long-run relationship. DAE Working Paper No. 9622. Department of Applied Economics, Cambridge University, Cambridge.

Ramos, CMQ and Rodrigues PMM (2014). Tourism demand modelling and forecasting: An overview. Revista de Turismo Contemporâneo – RTC, Natal 2: 323-340.

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

Sakamoto Y, Ishiguro M and Kitagawa G (1986) Akaike information criterion statistics. Dordrecht: D. Reidel.

Song H and Li G (2008) Tourism demand modelling and forecasting – A review of recent research. Tourism Management 29(2): 203-220.

Song H, Li G, Witt SF and Athanasopoulos G (2011) Forecasting tourist arrivals using time-varying parameter structural time series models. International Journal of Forecasting 27(3): 855-869.

Song H, Qu RTR and Park J (2019) A review on tourism demand forecasting. Annals of Tourism Research 75: 338-362.

Song H and Witt SF (2006) Forecasting international tourist flows to Macau. Tourism management 27(2): 214-224.

Song H and Witt SF (2012) Tourism demand modelling and forecasting. New York: Routledge.

Song H, Witt SF and Jensen TC (2003) Tourism forecasting: Accuracy of alternative econometric models. International Journal of Forecasting 19(1): 123-141.

Song H, Witt SF and Li G (2003) Modelling and forecasting the demand for Thai tourism. Tourism Economics 9(4): 363-387.

Song H, Wong KK and Chon KK (2003) Modelling and forecasting the demand for Hong Kong tourism. International Journal of Hospitality Management 22(4): 435-451.

Tsui WHK, Balli HO, Gilbey A and Gow H (2014) Forecasting of Hong Kong airport's passenger throughput. Tourism Management 42: 62-76.

Uysal M and El Roubi MS (1999) Artificial neural networks versus multiple regression in tourism demand analysis. Journal of Travel Research 38(2): 111-118.

Volchek K, Liu A, Song H and Buhalis D (2018) Forecasting tourist arrivals at attractions: Search engine empowered methodologies. Tourism Economics. doi: https://doi.org/10.1177/1354816618811558

Wanas N, Auda G, Kamel MS and Karray FAKF (1998) On the optimal number of hidden nodes in a neural network. Canadian Conference on Electrical and Computer Engineering 2: 918-921.

Wang L, Zeng Y and Chen T (2015) Back propagation neural network with adaptive differential evolution algorithm for time series forecasting. Expert Systems with Applications 42(2): 855-863.

Wu DC, Li G and Song H (2012) Economic analysis of tourism consumption dynamics: A time-varying parameter demand system approach. Annals of Tourism Research 39(2): 667-685.

Wu DC, Song H and Shen S (2017) New developments in tourism and hotel demand modelling and forecasting. International Journal of Contemporary Hospitality Management 29(1): 507-529.

Wu L and Brynjolfsson E (2015) The future of prediction: How Google searches foreshadow housing prices and sales. In A Goldfarb, SM Greenstein and CE Tucker (Eds.), Economic analysis of the digital economy (pp. 89-118). Chicago: University of Chicago Press.

Wu Q, Law R and Xu X (2012) A sparse Gaussian process regression model for tourism demand forecasting in Hong Kong. Expert Systems with Applications 39(5): 4769-4774.

Yang X, Pan B, Evans JA and Lv B (2015) Forecasting Chinese tourist volume with search engine data. Tourism Management 46: 386-397.

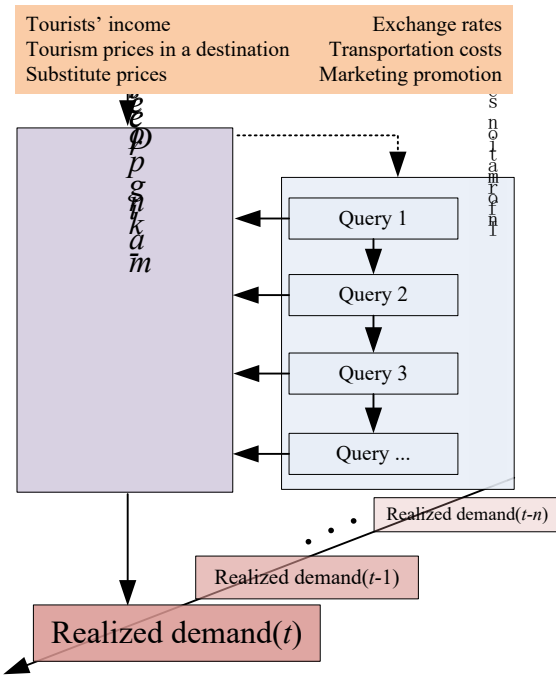

Figure 1. Tourism demand realisation - A conceptual framework.

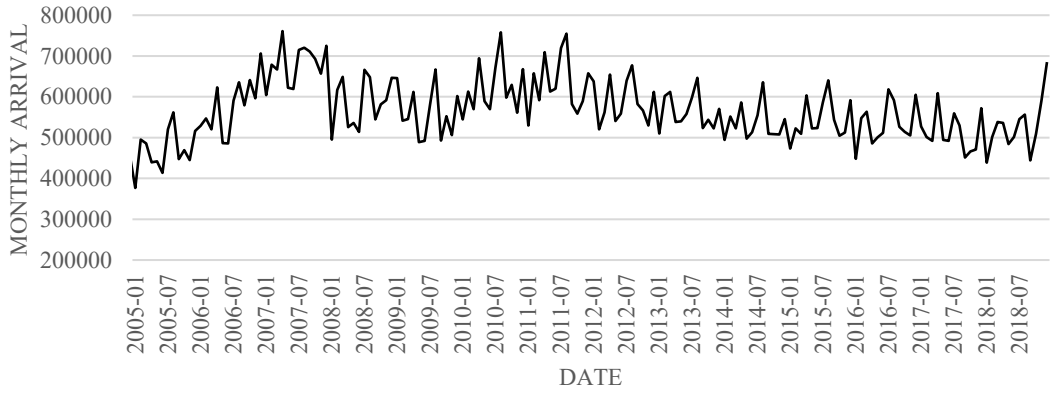

Figure 2. Tourist arrivals from Hong Kong to Macau.

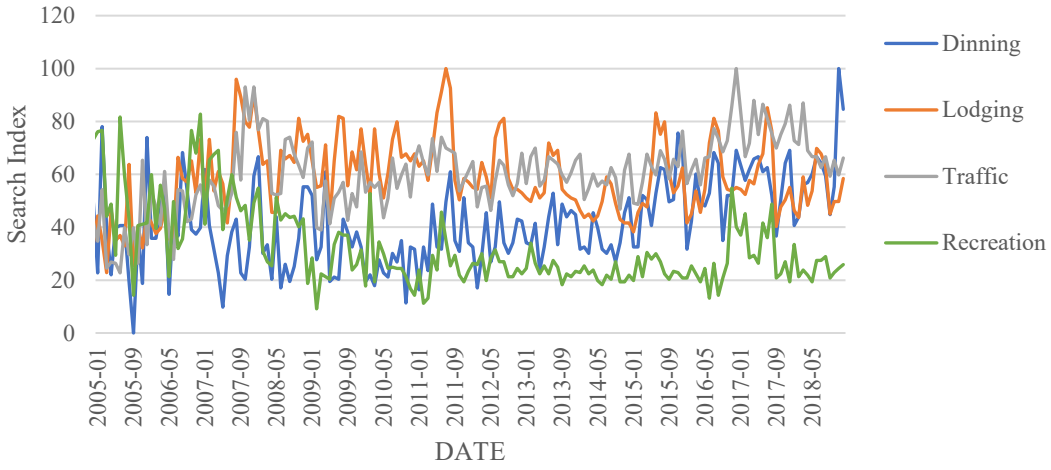

Figure 3. Search query data.

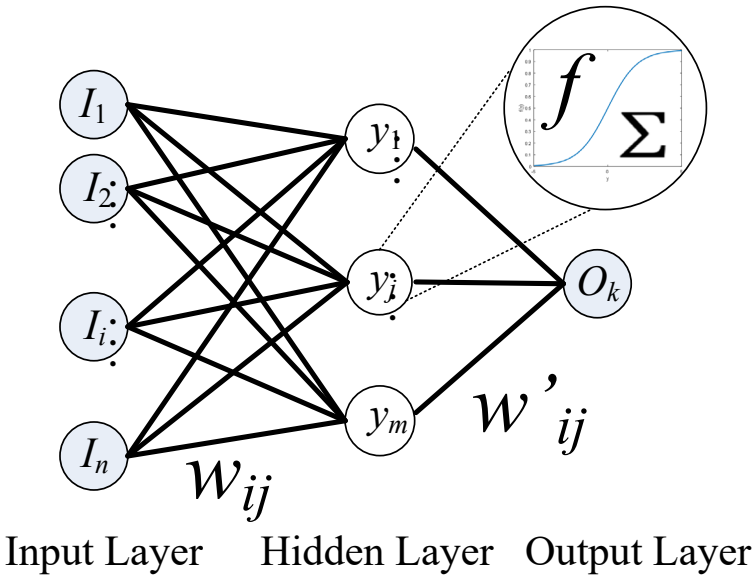

Figure 4. Back-propagation neural network.

Table 1 Keywords of search queries

Query	Keywords	Min	Max	Mean	Variance
Dining (QD)	Macau restaurant (A) Macau delicious food (B)	6	62	25.54	101.02
Lodging (QL)	Macau hotel (A&B)	17	72	43.64	115.70
Transportation (QTR)	Macau ferry (A&B) Macau steamer ticket (B)	9	62	43.34	99.30
Recreation (QR)	Macau entertainment (B) Macau casino (A&B)	6	66	21.04	93.91

Note: (A) In English; (B) in traditional Chinese

Table 2 Comparison of average performance

Data Sources	Accuracy model	Testing period			
		Jan.-Mar.	Jan.-Jun.	Jan.-Sep.	Jan.-Dec.
Model 1	MAD	**2.4470E+04** (30.64%)	**2.1833E+04** (11.87%)	**2.2893E+04** (5.46%)	**3.7005E+04** (5.13%)
	MAPE	**4.8634E-02** (31.96%)	**4.3612E-02** (12.35%)	**4.6609E-02** (5.78%)	**6.6062E-02** (5.66%)
	MSE	**8.0056E+08** (50.09%)	**6.2511E+08** (36.77%)	**7.2389E+08** (24.02%)	**2.8925E+09** (7.90%)
	RMSE	**2.8294E+04** (29.35%)	**2.5002E+04** (20.48%)	**2.6905E+04** (12.83%)	**5.3782E+04** (4.03%)
	RMSPE	**5.4749E-02** (31.75%)	**4.9183E-02** (21.75%)	**5.5705E-02** (12.36%)	**8.9338E-02** (4.59%)
Model 2	MAD	3.5278E+04	2.4774E+04	2.4215E+04	3.9006E+04
	MAPE	7.1482E-02	4.9757E-02	4.9468E-02	7.0029E-02
	MSE	1.6040E+09	9.8862E+08	9.5270E+08	3.1405E+09
	RMSE	4.0050E+04	3.1442E+04	3.0866E+04	5.6040E+04
	RMSPE	8.0215E-02	6.2854E-02	6.3559E-02	9.3635E-02

*Note: (**%) indicates the percentage of accuracy improvement compared with Model 2.*

Table 3 Forecasting values and accuracy among benchmark models

	ANN_1(HTDS & CV & SED)	ADL (HTDS & CV & SED)	ADL (HTDS & SED)	ARIMA (HTDS)
MAD	3.7005E+04	4.6160E+04 (19.83%)	4.8768E+04 (24.12%)	4.7389E+04 (21.91%)
MAPE	6.6062E-02	8.2981E-02 (20.39%)	8.5988E-02 (23.17%)	8.8172E-02 (25.08%)
MSE	2.8925E+09	3.6037E+09 (19.74%)	4.0974E+09 (29.41%)	3.0902E+09 (6.40%)
RMSE	5.3782E+04	6.0031E+04 (10.41%)	6.4011E+04 (15.98%)	5.5590E+04 (3.25%)
RMSPE	8.9338E-02	1.0213E-01 (12.53%)	1.0591E-01 (15.65%)	1.0165E-01 (12.11%)

*Note: (**%) indicates the percentage of accuracy improvement of ANN_1 (HTDS & CV & SED).*