

# A hybrid MIDAS approach for forecasting hotel demand using large panels of search data

Tourism Economics

1–25

© The Author(s) 2021



Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/13548166211015515

[journals.sagepub.com/home/teu](https://journals.sagepub.com/home/teu)**Binru Zhang** 

Yangtze Normal University, China

**Nao Li**

Beijing Technology and Business University, China; State Key Laboratory of Resources and Environmental Information System, China

**Rob Law** 

Hong Kong Polytechnic University, Hong Kong

**Heng Liu**

University of International Business and Economics, China

## Abstract

The large amounts of hospitality and tourism-related search data sampled at different frequencies have long presented a challenge for hospitality and tourism demand forecasting. This study aims to evaluate the applicability of large panels of search series sampled at daily frequencies to improve the forecast precision of monthly hotel demand. In particular, a hybrid mixed-data sampling regression approach integrating a dynamic factor model and forecast combinations is the first reported method to incorporate mixed-frequency data while remaining parsimonious and flexible. A case study is undertaken by investigating Sanya, the southernmost city in Hainan province, as a tourist destination using 9 years of the experimental data set. Dynamic factor analysis is used to extract the information from large panels of web search series, and forecast combinations are attempted to obtain the final prediction results of the individual forecasts to enhance the prediction accuracy further. The empirical analysis results suggest that the developed hybrid forecast approach leads to improvements in monthly forecasts of hotel occupancy over its competitors.

## Corresponding author:

Nao Li, School of International Economics and Management, Beijing Technology and Business University, No. 33, Fucheng Road, Haidian District, Beijing, 100048, China; State Key Laboratory of Resources and Environmental Information System, China.

Email: [linao@btbu.edu.cn](mailto:linao@btbu.edu.cn)

**Keywords**

Dynamic factor model, forecast combinations, hotel demand, hybrid MIDAS approach, mixed-frequency data, search engine data

**Introduction**

Given the perishability of hospitality and tourism products, the real-time and accurate forecasting of hospitality and tourism demand is vital in revenue management to guide enterprises in long-term investment planning and strategic decisions (Chu, 2011; Li and Law 2020; Weatherford and Kimes, 2003). Inaccurate forecasts mislead them to make erroneous development plans and decisions, thus resulting in the unnecessary wastage of resources and even leading to the bankruptcy of the business (Norsworthy and Tsai, 1998). The high-frequency search data generated by Internet users can objectively reflect tourists' attention and the potential hospitality and tourism demand, they can also help to enhance the forecast performance of the model (Yang et al., 2015; Zhang et al., 2020). However, the current prediction technology can only deal with the prediction task of the same frequency data, which, to a certain extent, limits the application of the model and the predictive capability (Kim and Swanson, 2017). Furthermore, when the web search data are used for prediction, a large number of key word explanatory variables bring new challenges to the model. This study addresses these two problems in the prediction of hotel demand.

With the comprehensive popularization of Internet technology, Internet information search is closely related to people's daily lives and has become an important tool, particularly for tourists, to make decisions and conduct online transactions (Fesenmaier et al., 2009, 2010; TIA, 2008; Vila et al., 2018). The search record generated by Internet information exploration reflects the public opinions and attention of the tourists and essentially gives an early signal of tourist flow (Li and Law, 2020; Zhang et al., 2017). However, in practical applications, a dilemma usually faced by researchers is that the hospitality and tourism-related data are sampled at different frequencies. On the one hand, hospitality and tourism demand variables are generally monthly frequency data, while the online search data are generally daily or weekly high-frequency data containing potentially valuable feature information (Ghysels et al., 2007). On the other hand, researchers cannot directly incorporate mixed-frequency queries into a prediction model to forecast hospitality and tourism demand. High-frequency variables must be converted to low frequency by simple weighted average or summation method to meet the requirements of the consistent frequency of the variables in the current prediction technology, including time series models, econometric models, and artificial intelligence technologies. However, this solution may not be optimal, and it will lose potential useful information, thus resulting in progressively ineffective and inconsistent estimates and affecting the prediction accuracy (Andreou et al., 2010; Ghysels et al., 2007; Kim and Swanson, 2017; Owyang et al., 2010). Therefore, a forecasting scheme must be developed to enhance the predictability of the model. The mixed-data sampling (MIDAS) model developed by Ghysels et al. (2004) can solve the prediction problem of the data sampled at different frequencies. It allows the predicted variable to have different sampling frequencies, thereby avoiding the information loss caused by simple weighted average or summation scheme. This method has been widely used in the prediction of various macroeconomic indicators (Pan et al., 2018; Yuan and Lee, 2019).

The second challenge is incorporating large data sets into the MIDAS model while still being parsimonious. This incorporation involves the aggregation of multiple indicators and attracts the attention of academics (Brynjolfsson et al., 2016). Existing research mainly focuses on three methods to integrate web search data into the prediction model: first, by using statistical methods, such as Pearson correlation coefficient, to select key word variables with predictive ability and directly bringing the selected variables into the model (Choi and Varian, 2012; Zhang et al., 2017); second, by using a simple weighted average or summation, where multiple key word variables are aggregated into a comprehensive indicator (Yang et al., 2015); and third, by extracting the composite search index using factor analysis (Li et al., 2015).

Although these methods improve the prediction ability of the model under certain conditions, they are no longer suitable for a large number of key word variables. Directly bringing the key word series into the model likely results in numerous information duplication and overfitting problems (Varian, 2014), and the parameter proliferation problems caused by too many explanatory variables may be more evident (Owyang et al., 2010). The second method may lead to the information loss of key word variables and affect the prediction accuracy (Li et al., 2017). Factor analysis can extract common factors from large data sets (Stock and Watson, 2002). Using these factors can enhance the forecast precision of the model (Kim and Swanson, 2017). However, factor analysis mainly deals with sectional data, and the hospitality and tourism demand forecast uses time series in a large sample. In addition, the key assumption of the classical factor model is that the heterogeneity parts are orthogonal to one another, which means that all the correlations among the data are attributed to static common factors. Thus, the prediction results are not optimal. The dynamic factor model relaxes these assumptions, and it can extract information from high-dimensional time series data. Studies have shown that the introduction of dynamic factor models is conducive to the prediction of macroeconomic variables (Forni et al., 2003; Li et al., 2017). Forecast combinations can address structure breaks and model instability under certain circumstances (Timmermann, 2006). A large number of prior studies have shown that forecast combinations can improve prediction performance by using information from all models rather than relying on a single specific model (Aiolfi et al., 2010; Stock and Watson, 2004). Following the recent work by Andreou et al. (2013) and Gomez Zamudio and Ibarra (2017), and so on, this study introduces dynamic factor model and forecast combinations as two complementary approaches to deal with large-scale data that cannot be handled in one go.

The application of a large amount of hospitality and tourism-related search data sampled at different frequencies in demand forecasting needs further exploration. This study proposes a hybrid MIDAS approach integrating a dynamic factor model and forecast combinations to predict hotel occupancy. To evaluate the prediction power of the constructed approach, Sanya, a tourist city in China's Hainan province, is used as an experimental case by collecting the key word variables related to the destination from January 2011 to December 2019. To ensure the parsimony of the model, the dynamic factor method is used to obtain the factor variables before forecasting, and each common factor is taken into the MIDAS model to predict the hotel occupancy. Then, the prediction results of the individual forecasts of the single-factor MIDAS model are combined to obtain the final predicted value. We conclude that the methodologies described herein can lead to improvements in prediction accuracy over its competitors.

The rest of the article is structured as follows: an extensive literature review is undertaken in the second section. Then, the conceptual analysis is presented in the third section. The fourth section constructs prediction methods used in this work. Next, the fifth section elaborates on the empirical

analysis, results and discussion, and robustness check. Finally, the article concludes and offers a future course of action in this area.

## Literature review

### *Hospitality and tourism demand forecasting*

Various demand forecasting methods reported so far in hospitality and tourism mainly include time series models, econometric models, and artificial intelligence technologies (Song and Li, 2008; Song et al., 2019).

The time series methods represented by the autoregressive (AR) integrated moving average and the exponential smoothing model are used to predict demand based on the historical data derived from systematic observations (Chu, 2009; Guizzardi and Stacchini, 2015; Li et al., 2017; Song and Li, 2008; Stock and Watson, 2002). Econometric techniques, such as vector autoregression, construct models on the basis of statistical theory to detect the causal relationship between tourism demand variables and their influencing factors (Smeral, 2019; Song and Witt, 2006; Song and Li, 2008; Wong et al., 2007). Artificial intelligence methods include support vector regression (SVR), artificial neural network, deep learning method, and so on. These methods mainly model the nonlinear nature of tourism demand. The research results show that when the data possess nonlinear characteristics, artificial intelligent methods can effectively improve the model prediction capabilities compared with the benchmark models (Hassani et al., 2015; Law and Au, 1999; Law et al., 2019; Pai and Hong, 2005; Palmer et al., 2006; Zhang et al., 2020). However, no fixed prediction technology has excelled in all circumstances (Song and Li, 2008).

The aforementioned demand prediction models require that the predicted variables and the prediction variables have an equal sampling frequency, which, however, limits the application of the model to a certain extent and affects prediction accuracy (Kim and Swanson, 2017).

### *Forecasting with search engine data and dimension reduction*

When consumers' travel motivation is stimulated, tourists tend to explore and understand the related information of various holiday-making destinations to arrive at decisions (Gnoth, 1997; Uysal and Jurowski, 1994). With the popularity of the Internet, online information query has become an important and essential tool for consumer decision-making (Fesenmaier et al., 2009, 2010; Ho and Liu, 2005). Compared with traditional information search, web information search not only saves consumers' time and cost but also provides more credible information (O'Connor, 1999). Moreover, information search behavior usually precedes tourism behavior (Pan et al., 2012). Tourist information search reveals different aspects of the travel process (Kim et al., 2007). These information searches are recorded by search engines, which objectively reflect the attention of the tourists. Therefore, key word variables can be used as the leading indicators of hospitality and tourism demand. Compared with statistical data, this kind of online search data is timely, easily accessible, and sensitive to tourists' behavior. Many prior studies have reported in prediction research on social and economic activities based on the search engine data (Li et al., 2017).

In the field of hospitality and tourism, researchers mainly use Google and Baidu search queries to predict demand. To select effective key word variables for prediction, data dimensionality reduction methods mainly include using statistical techniques to filter key words, weighted sum scheme to aggregate key word variables into a comprehensive index, and factor analysis method (Choi and Varian, 2012; Li et al., 2017; Loureno et al., 2020; Yang et al., 2015; Zhang et al., 2017,

2019). The empirical outcomes suggest that the addition of search engine data can further enhance the forecast accuracy. For example, the ARMA with weekly Google trend data was used to forecast the monthly hotel room demand in Charleston City in the southeastern United States. In the empirical application, five key words are directly brought into the model for prediction, and the results indicate that the addition of Google data can improve the predictability of the model (Pan et al., 2012). For dimensionality reduction, Yang et al. (2015) transformed the key word variables into a composite index by a weighted average scheme and verified the cointegration relationship between the key word composite index and the tourist volume of China's Hainan. Then, they incorporated the composite variable into the ARMA to predict Chinese tourist flow. Considering the nonlinear fluctuation characteristics of tourism demand, Zhang et al. (2017) used the Pearson cross-correlation analysis method to retain the four key word prediction variables with a higher correlation with the predicted variables. The selected variables were incorporated into the SVR algorithm to forecast the tourist flow of China. The results show that the addition of search data can enhance forecast power. Sun et al. (2019) constructed a single-layer artificial neural network algorithm and used Google and Baidu composite search index to forecast the tourist volume of Beijing; the composite index was constructed by a shift and sum approach. Other studies have also been reported to corroborate similar conclusions (Choi and Varian, 2012; Huang et al., 2017; Li et al., 2018; Park et al., 2016; Sun et al., 2020; Wei and Cui, 2018; Zhang et al., 2020). With the advance of academic research, online big data from multiple sources such as search engines and social media platforms are used to predict tourism demand (Li et al., 2020).

The existing literature mainly uses the time series model and artificial intelligence method when using high-frequency web online queries to forecast trends. However, none of these prediction techniques can directly model the mixed-frequency data set. The solution is to convert the high-frequency variable into a low-frequency one by a simple weighted average, thereby losing the characteristic information of high-frequency data (Ghysels et al., 2007; Owyang et al., 2010). In addition, the prediction accuracy will be affected by directly bringing the key word variables into the model and aggregating them into a comprehensive index (Li et al., 2017). Moreover, the classical factor model can show good dimensionality reduction ability in processing cross-sectional data. However, with the enlargement of time series, classic factor analysis is no longer the optimal scheme. The application of Internet search data in the hotel demand forecast needs further exploration. Furthermore, Baidu high-frequency web online data has not been applied in the hotel demand forecast, and the specific prediction effect needs to be tested.

### *Forecasting with MIDAS and forecast combinations*

To avoid information loss caused by transforming mixed-frequency data into the same frequency by simple averaging and summing, the MIDAS regression developed by Ghysels et al. (2004) can be used to construct a forecasting technique with different frequencies while still being parsimonious. However, the MIDAS model was originally proposed for mixed-frequency data to forecast the volatility of the stock market (Ghysels et al., 2005; Leon et al., 2007). Later, the MIDAS and its variants have been successfully deployed in the field of macroeconomic forecasting (Li et al., 2015; Xu et al., 2020).

The application of MIDAS in hotel demand forecasting remains unexplored. Bangwayo-Skeete and Skeete (2015) directly introduced data from Google Trends into the MIDAS model to predict the tourist flows to five tourist destinations in the Caribbean. Meanwhile, Qin and Liu (2019) used weekly key word variables into the multivariate MIDAS approach to predict the tourist flows in

China. Wen et al. (2020) proposed an improved MIDAS approach to forecast tourist volumes in Hong Kong from Mainland China with search index. The results suggest that the introduction of MIDAS can improve the forecasting accuracy of the model. Other similar studies have also been reported similar findings (Havranek and Zeynalov, 2019; Volchek et al., 2019; Wen et al., 2019).

Since the original study conducted by Bates and Granger (1969), forecast combinations have attracted extensive attention. This method combines the prediction results of individual models by using predetermined weighting schemes, thereby improving the forecasting performance over that offered by individual models (Gomez-Zamudio and Ibarra, 2017). Timmermann (2006) and Kim and Swanson (2014) have discussed this topic. Timmermann (2006) provided an excellent survey on forecast combinations. One of the justified reasons for using forecast combinations is that due to the various uncertainties in the modeling process, we regard the results of individual models as approximations, and we can combine the information sets of the prediction results of different models to generate a better prediction (Aastveit et al., 2014; Timmermann, 2006). A second reason is that forecast combinations can, to a certain extent, address the instability and structural mutation of the model under certain conditions, whereas a single model cannot (Stock and Watson, 2004). Moreover, a single model may suffer from an unknown form of error setting bias (Stock and Watson, 2004). A large and growing body of research shows that forecast combinations can improve prediction accuracy and robustness by using the information from all models rather than relying on a single specific model (Aiolfi et al., 2010; Timmermann, 2006). Forecast combinations have been widely used in the field of macroeconomic forecasting (Timmermann, 2006).

To address large panels of high-frequency data, Andreou et al. (2013) proposed two complementary prediction methods: (1) using data dimension reduction methods such as principal component analysis and factor analysis to extract a few common factors from a large number of time series and (2) combining the results of individual MIDAS models with a single factor, more accurate, and robust results are obtained. This approach has been applied to a large number of empirical studies. For example, to forecast Mexico's GDP, Gomez Zamudio and Ibarra (2017) used factor analysis to extract a few common factors from 392 daily financial series and combined the forecast results of the single-factor MIDAS models with forecast combinations. The results show that this approach significantly improves prediction accuracy relative to the single-factor MIDAS model. Sen Dogan and Midilic (2019) successfully predicted Turkey's economic growth by using the two complementary methods of principal component analysis and forecast combinations. Similar studies have also been conducted by other researchers (Andreou et al., 2013; Kim and Swanson, 2017; Stock and Watson, 2004). This study is encountering the same problem, we followed the method proposed by Andreou et al. (2013) to deal with large-scale data that cannot be handled in one go.

In contrast to the existing research, this study constructs a hybrid MIDAS approach integrating a dynamic factor model and forecast combinations to predict the hotel occupancy in the spirit of their work and to evaluate the role of high-frequency search data in hotel demand forecasting. The dynamic factor model is used to extract feature information from large panels of key word variables, which reduces the dimension of prediction variables and avoids information redundancy and multicollinearity. To ensure the parsimony of modeling and the stability of prediction, the prediction results of the single-factor MIDAS models are combined with forecast combinations (Sen Dogan and Midilic, 2019). The results show that compared with the benchmark model, the introduction of MIDAS can improve the prediction ability of the model.

## Methodology

### MIDAS regression approach

According to the theory of the distributed lag model, the MIDAS method developed by Ghysels et al. (2004) can directly model mixed-frequency data in a parsimonious, simple, and flexible way. Given the possible autocorrelation of tourism demand, the lag term of explained variables is introduced into the model. We consider the following MIDAS regression technique:

$$y_t = \alpha + \sum_{i=1}^p \gamma_i L^i y_t + \beta \sum_{k=1}^m \Theta(k; \theta) L_{HF}^k F_t + \mu_t \quad (1)$$

where the function  $\Theta(k; \theta)$  is the weighting polynomial determining the time summation,  $k$  is the lag number, and  $\theta$  is the hyperparametric vector that determines the shape of the weight function.  $F$  is the high-frequency prediction variable, which is obtained by dynamic factor analysis, and  $y$  is the hotel occupancy.  $L$  denotes a polynomial lag operator, and  $m$  denotes the sampling rate.  $\alpha$ ,  $\beta$ , and  $\gamma$  are the parameters to be determined.  $\mu$  represents the disturbance term.

Weighting function  $\Theta(k; \theta)$  has several function forms, and its basic purpose is to maintain parsimony and flexibility. According to Ghysels et al. (2004), the beta probability density function is used as the weighting function. The literature shows that the lowest prediction error rate can be obtained in most cases. In addition, the weighting function can depict various distribution shapes with only few parameters. Given its parsimonious representation and flexible shape, the weighting function has been successfully applied in various prediction tasks (Ghysels et al., 2007; Li et al., 2017). The form of the beta weighting function with two parameters is given as:

$$\Theta(k; \theta_1, \theta_2) = \frac{f\left(\frac{k}{m}, \theta_1, \theta_2\right)}{\sum_{j=1}^m f\left(\frac{j}{m}, \theta_1, \theta_2\right)} \quad (2)$$

where

$$f(i, \theta_1, \theta_2) = \frac{i^{\theta_1-1} (1-i)^{\theta_2-1} \Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1) \Gamma(\theta_2)} \quad (3)$$

$\theta_1$  and  $\theta_2$  are the hyperparameters that determine the shape of the function, and

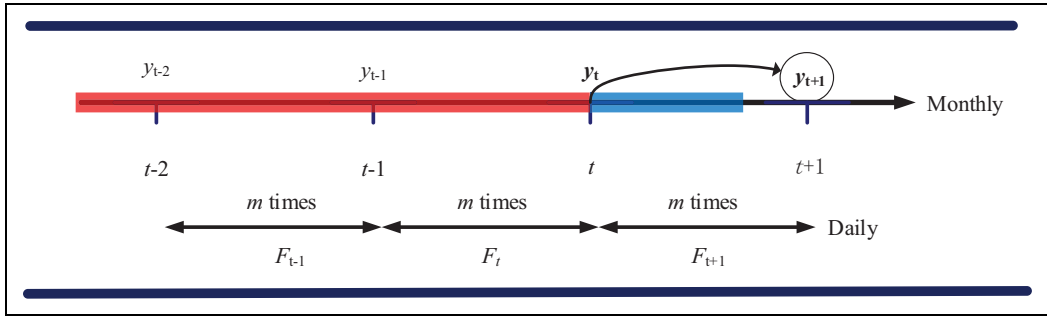
$$\Gamma(\theta_q) = \int_0^{\infty} e^{-i} i^{\theta_q-1} di \quad (4)$$

is the gamma function in standard form. The equal weighting scheme can be obtained when  $\theta_1 = \theta_2 = 1$ .

The MIDAS model can also be extended to the multiple-lag form of predictor  $F$  as

$$y_t = \alpha + \sum_{i=1}^p \gamma_i L^i y_t + \beta \sum_{k=1}^M \Theta(k; \theta) L_{HF}^k F_t + \mu_t \quad (5)$$

where  $M = m \times n$  denotes the lag order. If the influence of lagged prediction variables is monotonically decreasing after a certain period, this expression is a useful parsimony representation (Bangwayo-Skeete and Skeete, 2015). The parameters of MIDAS are estimated by the



**Figure 1.** Forecast timeline.

nonlinear least square method, and the lag order of predicted variables and high-frequency variables are determined by the AIC and BIC criteria.

In the MIDAS model, in addition to introducing the lags of the predicted variable  $y$ , we focus on the information of the high-frequency factor  $F$ , which was sampled  $m$  times between two consecutive sample periods of  $y$  (e.g. between  $t - 1$  and  $t$ ). Figure 1 shows the forecast timeline, which presents the data information available for predicting hotel occupancy at time  $t + 1$ . The figure intuitively and simply presents one-step forward forecasts. Moreover, generalization to multistep forward forecasts is evident. We assume that at time  $t$ , we are interested in forecasting the hotel occupancy  $y_{t+1}$ , where the observation of  $y$  at time  $t + 1$  is circled on the timeline. The forecasting experiment uses all available information up to time  $t$ , as shown in the red part of the timeline. The blue part of the timeline shows the new information available during the period of  $t + 1$  (i.e. leads); this information would be related to the hotel occupancy at time  $t + 1$ , which can be used for nowcasting (Raul and Luis, 2017). This study uses the constructed model and the data specified in the red section on the timeline to perform out-of-sample forecasts. In addition, the nowcasting experiment will not be explored here.

### Dynamic factor model

In this work, the dynamic factor model is introduced to obtain comprehensive variables from large panels of search data. The basic purpose of the dynamic factor model is to obtain a few common factors through dimension reduction. These few factors can explain the main part of the original variable information and avoid information redundancy, overfitting, and multicollinearity (Li et al., 2017). Research shows that the introduction of the dynamic factor model is beneficial to the prediction of macroeconomic variables (Forni et al., 2003; Li et al., 2017).

Supposedly, a large data set  $X$  with  $p$  variables will be used for prediction. Each variable has  $n$  observations and assumes that  $n > p$ . The goal is to find the common factor vector  $F = (F_1, F_2, \dots, F_m)$  ( $m < p$ ) and parameter set  $\Lambda$  to explain the original variable to the greatest extent. The dynamic factor model can be expressed as follows:

$$\begin{cases} X_t = \Lambda F_t + u_t \\ F_t = \Phi F_{t-1} + \eta_t \\ u_{it} = \alpha_{it}(L)u_{it-1} + \varepsilon_{it} \\ i = 1, 2, \dots, n \end{cases} \quad (6)$$



The number of factors is determined by the suggestion of Bai and Ng (2002), and the parameters are determined by the maximum likelihood method. Once the dynamic factors are obtained, they are taken into the MIDAS model to carry out a single-factor model prediction experiment. The final forecasts are obtained through the forecast combinations.

### Forecast combinations

To make full use of the information of each common factor, the forecast combinations are introduced to combine the prediction results of individual MIDAS with a single factor without increasing the estimated parameters of the model and maintaining its parsimony. Previous studies have agreed that forecast combinations with preset weighting scheme can improve the prediction accuracy of a single model (Timmermann, 2006) and can deal with the instability and structural mutation of the univariate model (Watson and Stock, 2004).  $\hat{y}_{i,t+h|t}$  represents the  $i$ th individual out-of-sample forecasted value of  $y_{i,t+h|t}$  estimated at time  $t$ . Thus, a combination forecast model is the weighted average of the predicted values of  $l$  individual MIDAS models:

$$f_{t+h|t} = \sum_{i=1}^l \omega_{i,t} \hat{y}_{i,t+h|t} \quad (7)$$

where  $\omega$  denotes the weighting scheme. Three weighting schemes are often used in previous studies viz. equally weighted weights, BIC-weighted forecast, and mean squared error (MSE) weighting. In this study, the MSE-related model averaging scheme is used, which can significantly enhance the forecast power of the model (Andreou et al., 2013). The MSE-related weighting scheme is as follows:

$$\omega_{i,t} = \frac{m_{i,t}^{-1}}{\sum_{i=1}^l m_{i,t}^{-1}} \quad (8)$$

$$m_{i,t} = \sum_{i=T_0}^t \left( y_{s+h}^h - \hat{y}_{i,s+h|s}^h \right)^2 \quad (9)$$

where  $T_0$  is the time point of the first out-of-sample observed value, and  $\hat{y}_{i,s+h|s}^h$  represents the out-of-sample forecasted value. Evidently, greater is the weight assigned when the MSE value of the  $i$ th model is smaller.

### Forecast evaluation

To evaluate the forecast power of the constructed hybrid forecast approach, the AR model and autoregressive distributed lag (ADL) model are introduced as the benchmark models, which can be expressed as follows:

$$y_t = \alpha + \sum_{i=1}^p \beta_i y_{t-i} + \varepsilon_t \quad (10)$$

$$y_t = \alpha + \sum_{i=1}^p \beta_i y_{t-i} + \sum_{i=0}^q \gamma_i X_{t-i} + \varepsilon_t \quad (11)$$

AR uses historical observations of hotel occupancy for prediction, and the lag order  $p$  is determined by using AIC and BIC criteria. ADL introduces public factor exogenous vector  $X$  and their lag terms based on AR. To ensure frequency consistency, the high-frequency factor obtained is converted into a low-frequency factor variable by using the equal weighting scheme, and the lag number is determined according to the AIC and BIC criteria. The parameters of AR and ADL are estimated by the least square method. In addition, the individual MIDAS model with a single factor is also used as a benchmark model to examine the effectiveness of the combination forecasts. Root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination ( $R$ ) are used as the standards to measure prediction performance. The expressions of the three indicators are as follows:

$$\text{RMSE}(y_t, \hat{y}_t) = \sqrt{\frac{1}{\tau} \sum_{t=1}^{\tau} (y_t - \hat{y}_t)^2} \quad (12)$$

$$\text{MAPE}(y_t, \hat{y}_t) = \frac{1}{\tau} \sum_{t=1}^{\tau} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \times 100\% \quad (13)$$

$$R(y_t, \hat{y}_t) = 1 - \frac{\sum_{t=1}^{\tau} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{\tau} (y_t - \bar{y}_t)^2} \quad (14)$$

where  $y_t$  and  $\hat{y}_t$  represent the observed values and the forecasted values, respectively.  $\tau$  denotes the testing size, and  $\bar{y}_t$  is the mean value of the observed values in the testing set.

RMSE is an absolute index, and its value ranges from zero to infinity. When the predicted value and the real value are completely consistent, it is equal to 0; greater error entails greater value. MAPE is a relative index with a range of  $[0, \infty]$ . When MAPE is 0%, it is a perfect model; if MAPE is greater than 100%, it is a poor model. The value range of  $R$  is  $[0, 1]$ . If the result is 0, the fitting effect of the model is poor; if the result is 1, the model fits the data perfectly. Generally speaking, a larger value entails better model fitting.

To evaluate the predictive power of non-nested models, the Diebold and Mariano (DM) method was used to verify whether there exists a significant difference between two non-nested forecast models (Diebold and Mariano, 1995). To achieve this goal, the following statistical hypotheses were put forward:

**H<sub>0</sub>:** Differences between the constructed model and its competitors are not statistically significant;

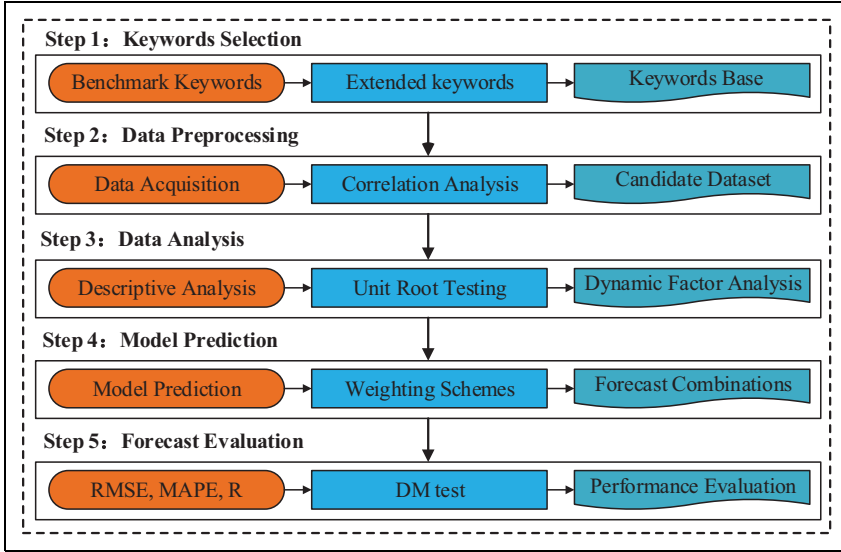
**H<sub>1</sub>:** Differences between the constructed model and its competitors are statistically significant.

A loss function  $L(e_{i,t})$  (such as squared error or absolute error) is selected to measure the prediction accuracy.  $e_{i,t}$  is the prediction error calculated by model  $i$  at time  $t$  based on the predicted values and the actual values. Under the null hypothesis, the constructed model and its alternative models have the same prediction ability, that is:

$$H_0 : L(e_{1,t}) = L(e_{2,t}) \quad (15)$$

When comparing predictive capabilities, the DM test introduces the loss differential:

$$d_{12t} = L(e_{1,t}) - L(e_{2,t}) \quad (16)$$



**Figure 2.** Flow diagram of the hybrid mixed-data sampling approach.

Equation (16) is used to construct the DM test statistics (Diebold and Mariano 1995):

$$DM = \frac{\bar{d}_{12}}{\bar{\sigma}_{d_{12}}} \quad (17)$$

where  $\bar{d}_{12}$  and  $\bar{\sigma}_{d_{12}}$  are the sample mean and the variance of the  $d_{12t}$ , respectively. When the DM test value is negative, the prediction performance of the first model is better on average and vice versa.

### Hybrid MIDAS forecasting approach

As shown in Figure 2, the hybrid MIDAS forecasting approach integrating a dynamic factor model and forecast combinations is constructed through five steps for empirical analysis. The specific steps are as follows:

*Step 1: Key word selection.* On the basis of Yang et al. (2015) and Zhang et al. (2020), we use the search engine automatic recommendation technology provided by Baidu to obtain the benchmark key words. These benchmark key words include all aspects of information queries related to the six elements of a tourist destination. Then, the other key words related to the benchmark terms are searched circularly to form a database of key words.

*Step 2: Data acquisition and preprocessing.* First, the observations of all key words in the key word database are obtained. To obtain the key word variables with prediction ability before dynamic factor analysis, irrelevant queries must be eliminated. To achieve this goal, Pearson cross-correlation analysis is carried out to identify the search data related to the predicted variables (Zhang et al., 2017) and form the initial experimental data set. In this study, the key word variables are converted using an equal weighting scheme to conduct the Pearson cross-correlation analysis.

*Step 3: Data analysis.* To explore the basic characteristics of each variable, descriptive statistical analysis is conducted. Considering the requirements of the model for the stationarity of time series, the Augmented Dickey–Fuller (ADF) unit root test is carried out for all variables. Meanwhile, unstable variables are transformed by the difference method. The dynamic factor model is used to obtain factor variables from all the search series. Given the requirement of the equal sampling rate of high-frequency variables in each month, the observation value of each factor in each month is adjusted to 30 days. When a month has 31 days, the observation of the 30th day is replaced by the average value of the last 2 days. The interpolation method is used to supplement the data every February. The adjusted public factors and predicted variables are used as experimental data sets.

*Step 4: Prediction experiment.* The experimental data set is divided into two parts viz. the training section and the test section, which are used for the model estimation and prediction tests, respectively. A single common factor is taken into the MIDAS for prediction experiments, and the prediction results of each single-factor MIDAS model are combined to generate the final forecasts.

*Step 5: Forecast evaluation.* The prediction accuracy between the proposed prediction method and the benchmark models is evaluated by using predictive performance metrics and DM significance tests.

## Empirical results

### *Data acquisition and preprocessing*

To evaluate the prediction performance of the constructed hybrid prediction method, this study takes Sanya, a famous tourist destination in China's Hainan province, as a case study. The city received 23.9633 million overnight tourists in 2019, which was an increase of 10.0% over the previous year. The average occupancy of the tourist hotels was 71.81%—an increase of 0.36% over the previous year. The monthly data of hotel occupancy is collected from the official website of Sanya (<http://lwj.sanya.gov.cn/>), covering January 2011 to December 2019. For convenience, the values of hotel occupancy are expressed in decimals with values range from 0 to 1.

As the econometric model should be driven by demand theory, price and income should be used as the forecast variables in the model. However, the official database does not directly provide historical data on average room rates. The consumer price index (CPI) can reflect changes in price levels, but the selected empirical case only provides CPI data after January 2013 (the time range of experimental data was from January 2011 to December 2019), and there is a serious lack of data every month. In addition, the disposable income data collected only include the annual data for nine years, and no suitable method has been found to forecast monthly tourism demand by using the annual data as the prediction variable. Therefore, following the forecasting approach of Bangwayo-Skeete and Skeete (2015) and Volchek et al. (2019), this study only uses the search data for prediction research.

The Internet search data are collected from Baidu (<http://index.baidu.com/>), which provides daily and weekly search data. These search queries are based on the search volume of netizens on Baidu. According to step 1 of the prediction approach, we search for key words related to Hainan and Sanya tourism given that Sanya is the core tourist city of Hainan. These key words include food and accommodation, transportation, scenic spots, shopping, and entertainment. Finally, 58 key words are collected to form a key word database.

**Table 1.** Pearson correlation analysis between key word variables and predicted variable.

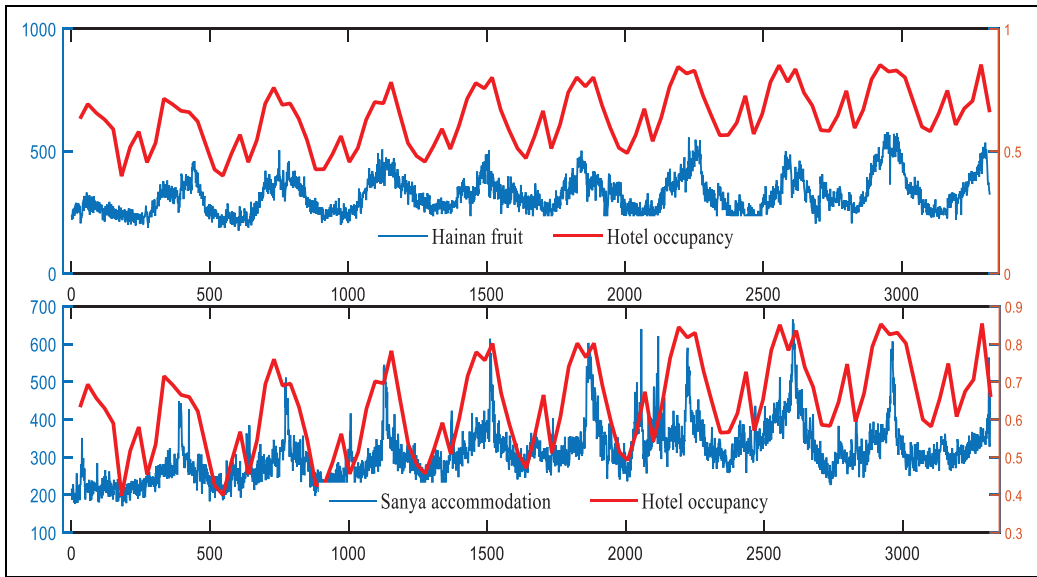
Key words	C	Lag order	Key words	C	Lag order
Haikou specialty	0.737	2	Nantian hot spring	0.773	3
Hainan scenic spot	0.734	2	Sanya tourist map	0.794	7
Hainan tourist map	0.719	1	Sanya cuisine	0.722	3
Hainan fruit	0.743	2	Sanya Nanshan Temple	0.755	2
Hainan specialty	0.670	2	Sanya specialty	0.764	2
Hainan characteristic fruit	0.787	2	Sanya weather forecast	0.703	1
Hainan self-driving travel	0.640	3	Sanya accommodation	0.647	1

Note: C represents the correlation coefficient between key word variables and the predicted variable.

According to step 2, the daily search volume time series of 58 key words over the period under consideration is collected. To exclude the information that is not related to the predicted variables before performing dynamic factor analysis, Pearson cross-correlation analysis is employed to calculate the maximum correlation coefficient between the hotel occupancy and the 0–12 order lag variables of each key word variable. The selection of the maximum lag order is 12 considering the periodicity of the monthly data. To reduce the information loss, the threshold of the correlation coefficient is set to 0.6, and 14 key words are reserved. There are three reasons to set the threshold to 0.6. First, the correlation coefficient between key word variables and forecasted variable is statistically significant at least at 10% level (the original assumption is that the correlation coefficient is equal to 0) when the correlation coefficient is greater than 0.6. However, the correlation coefficient between key word variables and forecasted variable is not statistically significant when the correlation coefficient is less than 0.6, even many key word variables are negatively correlated with the forecasted variable. Second, the key word variables with correlation coefficient below 0.6 do not show the same seasonal characteristics and periodic fluctuations as the forecasted variables, and these key words contain much noise, which can not provide useful information for forecasting. Finally, if the threshold is set to 0.7, the three key word variables of “Hainan specialty,” “Hainan self-driving travel,” and “Sanya accommodation” that have significant correlation with the predicted variable may be excluded from the experimental data set. Through preliminary experiments, we found that excluding any one of these three key words will reduce the correlation between the common factors and the predicted variable, which further leads to information loss of key word variables and affects the prediction accuracy. Therefore, to keep the important characteristics of these three key words as much as possible, the threshold is not greater than 0.6.

The outcomes of the cross-correlation analysis are displayed in Table 1. The table shows a strong correlation between information search terms and hotel occupancy, and the optimal lag order of most variables is 1 to 3. The maximum lag order is 7, indicating that most information search behavior occurs approximately two months before trip. These key words are the leading indicators of hotel occupancy.

Figure 3 further intuitively presents the trend between the hotel occupancy and the two key word variables of “Hainan fruit” and “Sanya accommodation” among the 14 key variables. It presents consistent fluctuation characteristics and a close correlation between the daily key word variables and the monthly hotel occupancy. Given the seasonality, the tourist flow is at its peak from January to August every year, and the information search also exhibits similar traits.



**Figure 3.** Fluctuation trend between “Hainan fruit” as well as “Sanya accommodation” and hotel occupancy.

However, the fluctuation characteristics of different key words in a local area show heterogeneity, which means that different key words reflect potential tourists’ demand from various aspects and provide different characteristic information for hotel occupancy prediction. This observation prompts us to explore the validity of these high-frequency search data in the hotel occupancy forecast further.

### Data analysis

According to step 3, the descriptive statistical analysis and stable analysis are carried out for 14 key words and predicted variables. In this work, the stability of variables is tested by using the ADF test. The results of the descriptive statistical analysis and the ADF test are presented in Table 2. Test results (the last two columns of Table 2) show that except for “Sanya specialty” rejecting the original hypothesis at the 5% level, all other time series reject the original hypothesis at the 1% level, thus indicating that all variables meet the stationarity.

For dimensionality reduction, the dynamic factor model is used to extract threshold from 14 key word variables. Finally, two common factors denoted by  $F_1$  and  $F_2$  are obtained, which explain more than 80% of the original variable information. Moreover, according to the load matrix of dynamic factor analysis results, the correlation coefficients between  $F_1$  and the key word variables of “Hainan tourist map,” “Hainan fruit,” “Hainan characteristic fruit,” “Hainan self driving travel,” “Sanya tourist map,” “Sanya weather forecast,” and “Sanya accommodation” all exceed 0.67. This means that  $F_1$  mainly explains the information about eating, accommodation, and transportation related to the destination. The correlation coefficients between  $F_2$  and the key word variables of “Haikou specialty,” “Hainan specialty,” “Nantian hot

**Table 2.** Descriptive statistical analysis and ADF test of all variables.

Variable	Observation	Mean	Standard deviation	Min	Max	ADF test	Conclusion
Hotel occupancy	108	0.65	0.12	0.40	0.85	−4.324***	Stable
Haikou specialty	3287	177.47	25.76	120	283	−3.764***	Stable
Hainan scenic spot	3287	270.93	43.57	165	579	−5.336***	Stable
Hainan tourist map	3287	468.62	130.38	229	1080	−4.549***	Stable
Hainan fruit	3287	320.00	72.91	175	577	−4.050***	Stable
Hainan specialty	3287	691.02	141.12	359	1145	−3.546***	Stable
Hainan characteristic fruit	3287	2773.44	1551.15	490	9844	−4.416***	Stable
Hainan self-driving travel	3287	189.17	58.55	70	553	−6.542***	Stable
Nantian hot spring	3287	231.27	56.04	135	550	−4.765***	Stable
Sanya tourist map	3287	241.82	39.41	120	418	−4.671***	Stable
Sanya cuisine	3287	289.91	39.07	180	474	−4.687***	Stable
Sanya Nanshan Temple	3287	310.07	70.29	152	665	−5.445***	Stable
Sanya specialty	3287	327.93	69.49	158	544	−3.147**	Stable
Sanya weather forecast	3287	2211.02	878.84	840	5025	−4.751***	Stable
Sanya accommodation	3287	266.18	76.54	149	597	−4.227***	Stable

Note: ADF: Augmented Dickey–Fuller. The values of hotel occupancy are expressed in decimals with values range from 0 to 1.

\*\*\* $p < 0.01$ .

\*\* $p < 0.05$ .

\* $p < 0.1$ .

spring,” “Sanya cuisine,” “Sanya Nanshan Temple,” and “Sanya specialty” all exceed 0.71, which suggests that  $F_2$  mainly captures the information about travelling, shopping, and entertainment related to the destination.

Given that the sampling rate of common factor variables in each month is inconsistent, the common factor variable is adjusted according to step 3. In addition, the sampling rate of each month after adjustment is 30 days. The adjusted common factors and the predicted variables constitute the experimental data set. To carry out the prediction experiment, the experimental data set is divided into two sections: the initial data sample from January 2011 to December 2018 as the training section for model estimation and the data from January 2019 to December 2019 as the testing section for forecasting.

## Results and discussions

For convenience, MIDAS-1 refers to the MIDAS model with the common factor  $F_1$  as a high-frequency prediction variable, and MIDAS-2 refers to the MIDAS model with the common factor  $F_2$  as a high-frequency prediction variable. The constructed hybrid forecasting approach is represented as MIDAS-C, which is the weighted average of MIDAS-1 and MIDAS-2. In the process of prediction experiments, a rolling window scheme is applied for all the models. AIC and BIC criteria are used to determine the lag order of the hotel occupancy and the high-frequency factor variables. In the weighting scheme, only two parameters in the Beta function are estimated, and the parameters of all MIDAS models are estimated by the nonlinear least square method.

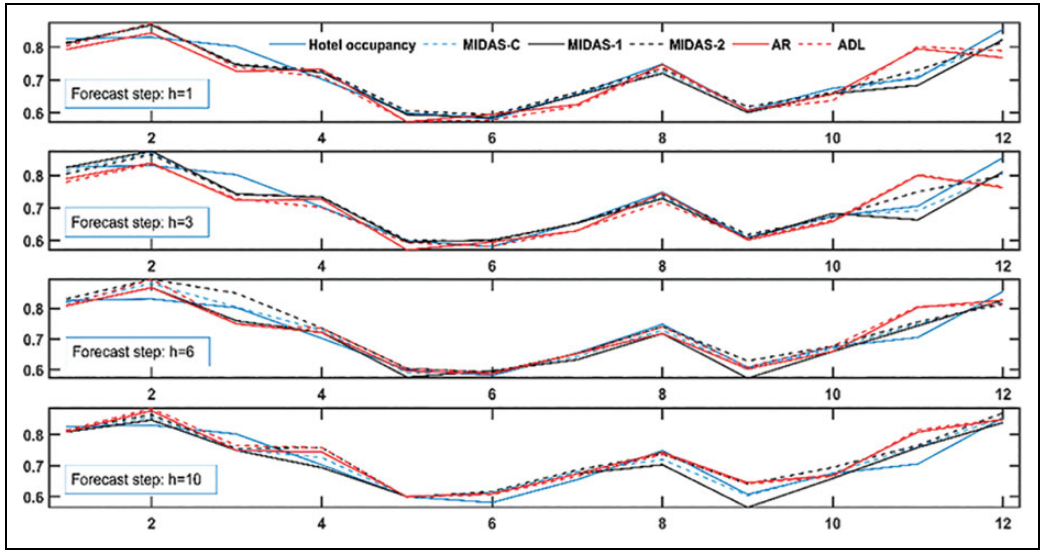


Figure 4. Comparison of one step and multiple steps ahead forecasts of forecast models.

Table 3. Comparison of RMSE values of each model for the Sanya data set.

Model	<i>h</i> = 1		<i>h</i> = 3		<i>h</i> = 6		<i>h</i> = 10	
	RMSE	IR (%)	RMSE	IR (%)	RMSE	IR (%)	RMSE	IR (%)
AR	0.0461	46.64	0.0478	48.54	0.0439	41.91	0.0470	40.00
ADL	0.0442	44.34	0.0471	47.77	0.0398	35.93	0.0425	33.65
MIDAS-1	0.0306	19.61	0.0344	28.49	0.0284	10.21	0.0312	9.62
MIDAS-2	0.0312	21.15	0.0357	31.09	0.0319	20.06	0.0345	18.26
MIDAS-C	0.0246		0.0279		0.0255		0.0282	

Note: IR: improvement rate; RMSE: root mean square error; AR: autoregressive; ADL: autoregressive distributed lag; MIDAS: mixed-data sampling. The values highlighted in italics denote the best statistical indices.

Figure 4 shows the prediction curves of each model on the test set by one step and multiple steps ( $h = 1, 3, 6, 10$ ) ahead horizons. The figure shows that the fitting power of each forecasting technique on the test section is preferable. Among them, the deviation of the MIDAS-C prediction curve from the expected value is smaller, which shows that the prediction method constructed in this study has a good fitting ability. The fitting effect of AR amongst the benchmark model is the worst. To illustrate the predictive ability of each model on the test set further, the predictive performance index values of each model must be compared.

Tables 3 and 4 show the comparison of the one step and multiple steps ahead statistical index (RMSE and MAPE) values of each model on the test set. As shown in Tables 3 and 4, MIDAS-C performs best on the two indicators, which means that the prediction power of the proposed approach is better than that of its competitors. Specifically, compared with the AR and ADL models, the improvement rate (IR) of MIDAS-C on RMSE and MAPE by all forecast horizons



**Table 4.** Comparison of MAPE values of each model for the Sanya data set.

Model	<i>h</i> = 1		<i>h</i> = 3		<i>h</i> = 6		<i>h</i> = 10	
	MAPE (%)	IR (%)	MAPE (%)	IR (%)	MAPE (%)	IR (%)	MAPE (%)	IR (%)
AR	4.7176	48.76	4.7974	49.61	3.4938	25.16	4.2964	29.10
ADL	4.6260	47.75	4.2219	42.74	3.6879	29.13	4.2623	28.53
MIDAS-1	3.1687	23.71	3.4092	29.09	3.7682	30.61	3.7123	17.95
MIDAS-2	3.3746	28.37	3.5129	31.19	3.2091	18.52	4.2257	27.91
MIDAS-C	2.4173		2.6859		2.6148		3.0461	

Note: IR: improvement rate; MAPE: mean absolute percentage error; AR: autoregressive; ADL: autoregressive distributed lag; MIDAS: mixed-data sampling. The values highlighted in italics denote the best statistical indices. The MAPE values are presented as percentages.

**Table 5.** Comparison of *R* values of each model for the Sanya data set.

Model	<i>h</i> = 1		<i>h</i> = 3		<i>h</i> = 6		<i>h</i> = 10	
	<i>R</i>	IR (%)	<i>R</i>	IR (%)	<i>R</i>	IR (%)	<i>R</i>	IR (%)
AR	0.8755	10.77	0.8669	11.87	0.9205	0.95	0.9127	−0.71
ADL	0.8959	8.25	0.8753	10.80	0.9065	2.50	0.9121	−0.65
MIDAS-1	0.9595	1.07	0.9329	3.96	0.9049	2.69	0.8847	2.43
MIDAS-2	0.9474	2.36	0.9217	5.22	0.8799	5.60	0.8591	5.48
MIDAS-C	0.9698		0.9548		0.9292		0.9062	

Note: IR: improvement rate; AR: autoregressive; ADL: autoregressive distributed lag; MIDAS: mixed-data sampling. The values highlighted in italics denote the best statistical indices.

exceeds 33% and 25%, respectively. Compared with the univariate MIDAS model, the prediction accuracy of MIDAS-C on RMSE and MAPE by all forecast horizons increased by approximately 9% and 17%, respectively.

Table 5 shows the comparison of the one step and multiple steps ahead *R* values of each model on the test set. As shown in Table 5, the fitting ability of the prediction model constructed in this research has varying degrees of improvement on the other forecast horizons except the 10-step ahead forecasts compared with its competitors, it is approximately 0.9% higher than AR and ADL and at least 1% higher than the single-factor MIDAS model.

To verify further whether the improvement of prediction accuracy is significant, we carried out a DM test on each forecast step, and the outcomes are shown in Table 6. The test statistics and *p*-value in Table 6 show that on one step and multiple steps ahead prediction results reject the null hypothesis at different significance levels (1% statistical significance level in most cases). This outcome indicates that the forecast accuracy of MIDAS-C is significantly different from that of other models.

In summary, the aforementioned analysis reveals that the prediction accuracy of the proposed hybrid MIDAS approach integrating a dynamic factor model and forecast combinations is significantly improved compared with its competitors. Moreover, the results confirm that daily search data is helpful to the hotel occupancy prediction, thus adding evidence to the conclusions of

**Table 6.** Tests of equal predictive ability with the DM test for the Sanya data set.

Forecast Step: h	DM statistics (MSE)			
	MIDAS-C vs. AR	MIDAS-C vs. ADL	MIDAS-C vs. MIDAS-I	MIDAS-C vs. MIDAS-2
$h = 1$	−4.493 (0.000)***	−5.663 (0.000)***	−2.779 (0.006)***	−2.433 (0.015)**
$h = 3$	−3.998 (0.000)***	−3.325 (0.001)***	−2.452 (0.014)**	−2.814 (0.005)***
$h = 6$	−3.254 (0.001)***	−5.020 (0.000)***	−4.014 (0.000)***	−4.383 (0.000)***
$h = 10$	−4.065 (0.000)**	−3.856 (0.000)**	−2.328 (0.020)**	−9.115 (0.000)**

Forecast Step: h	DM statistics (MAPE)			
	MIDAS-C vs. AR	MIDAS-C vs. ADL	MIDAS-C vs. MIDAS-I	MIDAS-C vs. MIDAS-2
$h = 1$	−7.083 (0.000)***	−7.321 (0.000)***	−2.443 (0.015)**	−5.826 (0.000)***
$h = 3$	−6.434 (0.000)***	−2.978 (0.003)***	−2.036 (0.042)**	−4.250 (0.000)***
$h = 6$	−3.682 (0.000)***	−4.006 (0.000)***	−8.337 (0.000)***	−2.939 (0.003)***
$h = 10$	−4.907 (0.000)**	−4.229 (0.000)**	−2.022 (0.043)**	−8.064 (0.000)**

Note: MIDAS: mixed-data sampling; AR: autoregressive; ADL: autoregressive distributed lag; MAPE: mean absolute percentage error; DM: Diebold and Mariano; MSE: mean squared error. The number in parentheses indicates the  $p$ -value.  
 \*\*\* $p < 0.01$ .  
 \*\* $p < 0.05$ .  
 \* $p < 0.1$ .

Bangwayo-Skeete and Skeete (2015) and Qin and Liu (2019). This confirmation is attributed to the following facts. First, the AR model uses only the historical information of the predicted variable and does not add the key word variables with predictive ability. Its prediction ability is significantly lower than that of the MIDAS-C. Second, the ADL model converts two high-frequency common factor variables into monthly variables by equal weighting scheme, thus losing the dynamic characteristic information of the original high-frequency data (Kim and Swanson, 2017) and resulting in its prediction ability to be significantly lower than that of the proposed MIDAS-C. Finally, MIDAS-C has stronger predictive ability than the two univariate MIDAS models, thus indicating that the forecast combinations can significantly improve the forecasting accuracy. This outcome is consistent with the conclusions of Timmermann (2006), as the forecast combinations are more stable than the univariate MIDAS model (Watson and Stock, 2004). Additionally, the constructed method integrates the main feature information of the original key word variables.

### Another case study

To enhance the generalization of this study, we take Beijing as another numerical example. The monthly data of hotel occupancy is collected from the Wind database, a leading financial database curated in China, and the search engine data are collected from Baidu. We used the same detailed forecasting procedure described in the “Hybrid MIDAS forecasting approach” subsection for the forecasting experiment over the same time span. Through the first three steps of the prediction framework, 10 key word variables with prediction ability are obtained, and two common factors are extracted by using the dynamic factor model. The forecasting results are shown in Tables 7, 8, and 9.

**Table 7.** Comparison of RMSE values of each model for the Beijing data set.

Model	$h = 1$		$h = 3$		$h = 6$		$h = 10$	
	RMSE	IR (%)	RMSE	IR (%)	RMSE	IR (%)	RMSE	IR (%)
AR	0.0348	29.02	0.0316	15.19	0.0319	11.91	0.0308	27.27
ADL	0.0298	17.11	0.0300	10.67	0.0308	8.77	0.0315	28.89
MIDAS-1	0.0254	2.76	0.0307	12.70	0.0307	8.47	0.0254	11.81
MIDAS-2	0.0252	1.98	0.0277	3.25	0.0294	4.42	0.0244	8.20
MIDAS-C	0.0247		0.0268		0.0281		0.0224	

Note: IR: improvement rate; AR: autoregressive; ADL: autoregressive distributed lag; MIDAS: mixed-data sampling; RMSE: root mean square error. The values highlighted in italics denote the best statistical indices.

**Table 8.** Comparison of MAPE values of each model for the Beijing data set.

Model	$h = 1$		$h = 3$		$h = 6$		$h = 10$	
	MAPE (%)	IR (%)	MAPE (%)	IR (%)	MAPE (%)	IR (%)	MAPE (%)	IR (%)
AR	3.8350	35.83	3.7950	41.86	3.8034	35.16	3.1104	28.12
ADL	3.5322	30.33	3.4450	35.95	3.5735	30.98	3.4286	34.79
MIDAS-1	2.6753	8.01	2.8524	22.64	2.7833	11.39	3.1129	28.18
MIDAS-2	2.7889	11.76	3.1944	30.93	3.3573	26.54	2.7713	19.32
MIDAS-C	2.4610		2.2065		2.4663		2.2358	

Note: IR: improvement rate; MAPE: mean absolute percentage error; AR: autoregressive; ADL: autoregressive distributed lag; MIDAS: mixed-data sampling. The values highlighted in italics denote the best statistical indices. The MAPE values are presented as percentages.

**Table 9.** Comparison of  $R$  values of each model for the Beijing data set.

Model	$h = 1$		$h = 3$		$h = 6$		$h = 10$	
	$R$	IR (%)	$R$	IR (%)	$R$	IR (%)	$R$	IR (%)
AR	0.9221	0.79	0.9107	0.68	0.9046	0.50	0.9406	0.14
ADL	0.9144	1.64	0.9137	0.35	0.9149	-0.63	0.9132	3.14
MIDAS-1	0.9253	0.44	0.8915	2.85	0.8910	2.03	0.9255	1.77
MIDAS-2	0.9267	0.29	0.9115	0.59	0.9003	0.98	0.9310	1.17
MIDAS-C	0.9294		0.9169		0.9091		0.9419	

Note: IR: improvement rate; AR: autoregressive; ADL: autoregressive distributed lag; MIDAS: mixed-data sampling. The values highlighted in italics denote the best statistical indices.

The results show that MIDAS-C performs better than the other indicators and forecasting steps except the ADL model in the  $R$  index of the 6-step ahead horizon.

The significance test results in Table 10 show that, in terms of the MSE index, the forecasting accuracy of MIDAS-C is significantly different from that of other models except MIDAS-2 on

**Table 10.** Tests of equal predictive ability with the DM test for the Beijing data set.

Forecast Step: h	DM statistics (MSE)			
	MIDAS-C vs. AR	MIDAS-C vs. ADL	MIDAS-C vs. MIDAS-I	MIDAS-C vs. MIDAS-2
<i>h</i> = 1	−4.616 (0.000)***	−3.249 (0.001)***	−3.833 (0.000)***	−3.344 (0.001)***
<i>h</i> = 3	−5.733 (0.000)***	−3.942 (0.000)***	−3.153 (0.002)***	−1.627 (0.104)
<i>h</i> = 6	−2.906 (0.004)***	−3.888 (0.000)***	−2.856 (0.004)***	−2.347 (0.019)**
<i>h</i> = 10	−5.472 (0.000)**	−3.588 (0.000)**	−6.753 (0.000)**	−8.241 (0.000)**

Forecast Step: h	DM statistics (MAPE)			
	MIDAS-C vs. AR	MIDAS-C vs. ADL	MIDAS-C vs. MIDAS-I	MIDAS-C vs. MIDAS-2
<i>h</i> = 1	−5.903 (0.000)***	−5.387 (0.000)***	−4.724 (0.000)***	−4.951 (0.000)***
<i>h</i> = 3	−9.081 (0.000)***	−6.665 (0.000)***	−9.474 (0.000)***	−3.770 (0.000)***
<i>h</i> = 6	−5.580 (0.000)***	−6.627 (0.000)***	−2.310 (0.021)**	−1.767 (0.077)*
<i>h</i> = 10	−3.907 (0.000)**	−5.506 (0.000)**	−5.174 (0.000)**	−5.476 (0.000)**

Note: MIDAS: mixed-data sampling; AR: autoregressive; ADL: autoregressive distributed lag; MAPE: mean absolute percentage error; DM: Diebold and Mariano; MSE: mean squared error. The number in parentheses indicates the *p*-value.  
\*\*\**p* < 0.01.  
\*\**p* < 0.05.  
\**p* < 0.1.

three-step ahead forecasts. In terms of the MAPE indicator, the results suggest that the forecasting accuracy of MIDAS-C is significantly different from that of other models on all forecast horizons. Overall, the empirical analysis once again suggests that the developed approach can produce preferable and robust forecasting results.

Conclusions and implications

Accurate demand forecasting is of utmost help in scientific decision-making to hospitality and tourism-related industries. With the popularity of the Internet, web information search objectively reflects consumers’ potential demand and provides a reliable source of data for demand forecasting. However, forecasting accuracy is affected by the inconsistency of data frequency and increasing key word variables. In this study, we use the dynamic factor model to obtain two factors and directly model the monthly hotel occupancy and daily key word variables in the hybrid MIDAS framework. Then, we construct two single-factor MIDAS models for the two common factors. Then, the final forecasts are obtained by forecast combinations. The results show that compared with the baseline models, the proposed hybrid approach can significantly improve accuracy while maintaining parsimony and flexibility, and the forecast results are more robust. Moreover, MIDAS-C alleviates the fitting problem to a certain extent. We conclude that incorporating daily search data into the constructed predictive model can lead to improvements in the monthly forecasts of hotel occupancy over its baseline models.

The contribution of this work to the existing research is mainly reflected in three aspects. First, the current hotel demand prediction method deals with the problem of equal frequency data prediction. Hence, the hybrid MIDAS approach integrating a dynamic factor model and forecast

combinations in this study is reported to model the mixed-frequency data directly in a parsimonious and flexible way to reduce the prediction error rate of the model. Second, the effectiveness of the daily search data in the prediction of hotel occupancy is verified under the hybrid MIDAS approach for the first time. Lastly, to use the feature information of key word variables effectively, the dynamic factor model is used to obtain factor variables from large panels of daily search data. Moreover, the forecast combinations are used to obtain the final forecast values of the individual MIDAS model to improve the prediction ability further.

The research in this article has important theoretical implications. First, given the inconsistent frequency of the predicted variables and the prediction variables, the hybrid MIDAS approach is introduced to model the mixed-frequency data directly, which ensures the parsimony and flexibility of the modeling. Second, the dynamic factor model is introduced to extract the feature information from large amounts of search engine data. The original variable feature information is retained, while the dimension reduction is achieved. Thus, the model estimation is parsimonious. Finally, the forecast combinations are constructed to avoid the instability of the single-factor MIDAS model and improve the prediction ability.

This research has practical implications for policy makers, researchers, and practitioners. In particular, the results show that the daily key word variables contain useful feature information. The data is updated daily, while the frequency of hotel occupancy is updated monthly. Generally, the hotel occupancy observations of the current month are released in the middle and late periods of the following month. Therefore, before the official data are released, the current month's daily search data is available and can be used to predict the current month's hotel occupancy rate (nowcasting) in advance. From the policy point of view, forecast results can provide the necessary information support to policy makers and business managers in various aspects, such as planning, marketing, income management, investment, and annual budget. For example, in the off-season and peak season, hotels use forecast results to implement additional scientific promotion plans, ensure a reasonable allocation of hotel resources, and avoid unnecessary wastage. With regard to the prediction method, the prediction approach constructed in this work can be generalized to other demand forecasts, such as tourist flow and revenue, to solve the frequency inconsistency and the large amounts of key word variables to reduce the prediction error rate.

However, given the data collection restrictions, the experimental data set does not include search data from other search engines and other hospitality and tourism-related series such as macroeconomic data and various online review records. Integrating other data sources to verify and compare the performance of demand forecasting would be a direction of further research.

## **Acknowledgment**

The authors are grateful to the editors and the anonymous reviewers for their valuable comments and suggestions.

## **Declaration of conflicting interests**


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

## **Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the Humanities and Social Science Fund of Ministry of

Education of China, grant number 20YJC630202 and supported by a grant from State Key Laboratory of Resources and Environmental Information System.

## ORCID iDs

Binru Zhang  <https://orcid.org/0000-0003-4495-4060>

Rob Law  <https://orcid.org/0000-0001-7199-3757>

## References

- Aastveit KA, Gerdrup KR, Jor AS, et al. (2014) Nowcasting GDP in real time: a density combination approach. *Journal of Business and Economic Statistics* 32: 48–68.
- Aiolfi M, Capistrán C and Timmermann AG (2010) Forecast combinations. *Working Paper* 29: 135–196.
- Andreou E, Ghysels E and Kourtellis A (2010) Regression models with mixed sampling frequencies. *Journal of Econometrics* 158: 246–261.
- Andreou E, Ghysels E and Kourtellis A (2013) Should macroeconomic forecasters use daily financial data and how? *Journal of Business & Economic Statistics* 31: 240–251.
- Bai J and Ng S (2002) Determining the number of factors in approximate factor models. *Econometrica* 70: 191–221.
- 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.
- Bates JM and Granger CWJ (1969) The combination of forecasts. *Journal of the Operational Research Society* 20: 451–468.
- Brynjolfsson E, Geva T and Reichman S (2016) Crowd-squared: amplifying the predictive power of search trend data. *MIS Quarterly* 40: 941–961.
- Choi H and Varian H (2012) Predicting the present with google trends. *Economic Record* 88: 2–9.
- Chu FL (2009) Forecasting tourism demand with ARMA-based methods. *Tourism Management* 30: 740–751.
- Chu FL (2011) A piecewise linear approach to modeling and forecasting demand for Macau tourism. *Tourism Management* 32: 1414–1420.
- Diebold FX and Mariano RS (1995) Comparing predictive accuracy. *Journal of Business & Economic Statistics* 13: 253–263.
- Fesenmaier DR, Cook SD, Zach F, et al. (2009) *Travelers' Use of the Internet*. Washington, DC: Travel Industry Association of America.
- Fesenmaier DR, Xiang Z, Pan B, et al. (2010) A framework of search engine use for travel planning. *Journal of Travel Research* 50: 587–601.
- Forni M, Hallin M, Lippi M, et al. (2003) Do financial variables help forecasting inflation and real activity in the euro area? *Journal of Monetary Economics* 50: 1243–1255.
- Ghysels E, Santa-Clara P and Valkanov R (2004) The MIDAS touch: mixed data sampling regression models. *Working Papers* Retrieved from <https://www.researchgate.net/publication/5004749> (accessed 13 December 2016).
- Ghysels E, Santa-Clara P and Valkanov R (2005) There is a risk-return trade-off after all. *Journal of Financial Economics* 76: 509–548.
- Ghysels E, Sinko A and Valkanov R (2007) Midas regressions: further results and new directions. *Econometric Reviews* 26: 53–90.
- Gnoth J (1997) Tourism motivation and expectation formation. *Annals of Tourism Research* 24: 283–304.
- Gomez-Zamudio LM and Ibarra R (2017) Are daily financial data useful for forecasting GDP? Evidence from Mexico. *Economía* 17: 173–203.
- Guizzardi A and Stacchini A (2015) Real-time forecasting regional tourism with business sentiment surveys. *Tourism Management* 47: 213–223.
- Hassani H, Webster A, Silva ES, et al. (2015) Forecasting U.S. tourist arrivals using optimal singular spectrum analysis. *Tourism Management* 46: 322–335.

- Havranek T and Zeynalov A (2019) Forecasting tourist arrivals: google trends meets mixed-frequency data. *Tourism Economics* 27: 129–148.
- Ho CI and Liu YP (2005) An exploratory investigation of web-based tourist information search behavior. *Asia Pacific Journal of Tourism Research* 10: 351–360.
- 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.
- Kim DY, Lehto XY and Morrison AM (2007) Gender differences in online travel information search: implications for marketing communications on the internet. *Tourism Management* 28: 423–433.
- Kim HH and Swanson NR (2014) Forecasting financial and macroeconomic variables using data reduction methods: new empirical evidence. *Journal of Econometrics* 178: 352–367.
- Kim HH and Swanson NR (2017) Methods for backcasting, nowcasting and forecasting using factor- MIDAS: with an application to Korean GDP. *Journal of Forecasting* 37: 281–302.
- Law R and Au N (1999) A neural network model to forecast Japanese demand for travel to Hong Kong. *Tourism Management* 20: 89–97.
- Law R, Li G, Fong DKC, et al. (2019) Tourism demand forecasting: a deep learning approach. *Annals of Tourism Research* 75: 410–423.
- Leon A, Nave JM and Rubio G (2007) The relationship between risk and expected return in Europe. *Journal of Banking & Finance* 31: 495–512.
- Li H, Hu M and Li G (2020) Forecasting tourism demand with multisource big data. *Annals of Tourism Research* 83: 102912.
- Li S, Chen T, Wang L, et al. (2018) Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management* 68: 116–126.
- Li X and Law R (2020) Forecasting tourism demand with decomposed search cycles. *Journal of Travel Research* 59: 52–68.
- Li X, Pan B, Law R, et al. (2017) Forecasting tourism demand with composite search index. *Tourism Management* 59: 57–66.
- Li X, Shang W, Wang S, et al. (2015) A MIDAS modelling framework for Chinese inflation index forecast incorporating Google search data. *Electronic Commerce Research & Applications* 14: 112–125.
- Loureno N, Gouveia CM and António R (2020) Forecasting tourism with targeted predictors in a data-rich environment. *Economic Modelling* article in press.
- Norsworthy JR and Tsai DH (1998) *Macroeconomic Policy as Implicit Industrial Policy: Its Industry and Enterprise Effects*. Berlin, Germany: Springer.
- O'Connor P (1999) *Electronic Information Distribution in Tourism and Hospitality*. Wallingford: CAB International.
- Owyang MT, Armesto MT and Engemann KM (2010) *Forecasting with Mixed Frequencies*. Louis, MO: Federal Reserve Bank of St. Louis Review.
- Pai PF and Hong WC (2005) An improved neural network model in forecasting arrivals. *Annals of Tourism Research* 32: 1138–1141.
- Palmer A, Montano JJ and Sese A (2006) Designing an artificial neural network for forecasting tourism time series. *Tourism Management* 27: 781–790.
- Pan B, Wu CD and Song H (2012) Forecasting hotel room demand using search engine data. *Journal of Hospitality & Tourism Technology* 3: 196–210.
- Pan Z, Wang Q, Wang Y, et al. (2018) Forecasting U.S. real GDP using oil prices: a time-varying parameter MIDAS model. *Energy Economics* 72: 177–187.
- Park S, Lee J and Song W (2016) Short-term forecasting of Japanese tourist inflow to south Korea using google trends data. *Journal of Travel & Tourism Marketing* 34: 1–12.
- Qin M and Liu H (2019) Baidu Index, mixed frequency model and Sanya tourism demand. *Tourism Tribune* 34: 116–126.

- Raul I and Luis MG (2017) Are daily financial data useful for forecasting GDP? Evidence from Mexico. *Working paper* Available at: <https://ideas.repec.org/p/bdm/wpaper/2017-17.html> (accessed 16 August 2018).
- Sen Dogan B and Midilic M (2019) Forecasting Turkish real GDP growth in a data-rich environment. *Empirical Economics* 56: 367–395.
- Smeral E (2019) Seasonal forecasting performance considering varying income elasticities in tourism demand. *Tourism Economics* 25: 355–374.
- Song H and Li G (2008) Tourism demand modelling and forecasting—a review of recent research. *Tourism Management Analysis Behaviour & Strategy* 29: 203–220.
- Song H and Witt SF (2006) Forecasting international tourist flows to Macau. *Tourism Management* 27: 214–224.
- Song H, Qiu RTR and Park J (2019) A review of research on tourism demand forecasting: launching the annals of tourism research curated collection on tourism demand forecasting. *Annals of Tourism Research* 75: 338–362.
- Stock JH and Watson MW (2002) Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97: 1167–1179.
- Stock JH and Watson MW (2004) Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting* 23: 405–430.
- Sun S, Li Y, Wang S, et al. (2020) Tourism demand forecasting with tourist attention: an ensemble deep learning approach. Working paper 2002.07964, arXiv.org, revised Mar 2020.
- Sun S, Wei Y, Tsui KL, et al. (2019) Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management* 70: 1–10.
- TIA (2008) *Travelers' Use of the Internet* (2008 edn). Washington, DC: Travel Industry Association of America.
- Timmermann AG (2006) Forecast combinations. In: Elliott G, Granger C and Timmermann A (eds) *Handbook of Economic Forecasting*. North Holland: Amsterdam: Elsevier, 1, pp. 135–196.
- Uysal M and Jurowski C (1994) Testing the push and pull factors. *Annals of Tourism Research* 21: 844–846.
- Varian HR (2014) Big data: new tricks for econometrics. *Journal of Economic Perspectives* 28: 3–27.
- Vila TD, Vila NA, Alén González E, et al. (2018) The role of the internet as a tool to search for tourist information. *Journal of Global Information Management* 26: 58–84.
- Volchek K, Liu A, Song H, et al. (2019) Forecasting tourist arrivals at attractions: search engine empowered methodologies. *Tourism Economics* 25: 425–447.
- Weatherford LR and Kimes SE (2003) A comparison of forecasting methods for hotel revenue management. *International Journal of Forecasting* 19: 401–415.
- Wei JR and Cui HM (2018) The construction of regional tourism index and its micro-dynamic characteristics: a case study of Xi'an. *Journal of Systems Science and Mathematical Sciences* 38: 177–194.
- Wen L, Liu C and Song H (2019) Forecasting tourism demand using search query data: a hybrid modelling approach. *Tourism Economics* 25: 309–329.
- Wen L, Liu C, Song H, et al. (2020) Forecasting tourism demand with an improved mixed data sampling model. *Journal of Travel Research* 60: 336–353.
- Wong KKF, Song H, Witt SF, et al. (2007) Tourism forecasting: to combine or not to combine? *Tourism Management* 28: 1068–1078.
- Xu Q, Wang L, Jiang C, et al. (2020) A novel (U) MIDAS-SVR model with multi-source market sentiment for forecasting stock returns. *Neural Computing & Applications* 32: 5875–5888.
- Yang X, Pan B, Evans JA, et al. (2015) Forecasting Chinese tourist volume with search engine data. *Tourism Management* 46: 386–397.
- Yuan FC and Lee CH (2019) Intelligent sales volume forecasting using google search engine data. *Soft Computing* 24: 2033–2047.
- Zhang B, Huang X, Li N, et al. (2017) A novel hybrid model for tourist volume forecasting incorporating search engine data. *Asia Pacific Journal of Tourism Research* 22: 245–254.



Zhang B, Li N, Shi F, et al. (2020) A deep learning approach for daily tourist flow forecasting with consumer search data. *Asia Pacific Journal of Tourism Research* 25: 323–339.

Zhang B, Pu Y, Wang Y, et al. (2019) Forecasting hotel accommodation demand based on LSTM model incorporating internet search index. *Sustainability* 11: 4708.

### Author biographies

**Binru Zhang**, Doctor of Economics, is an associate professor at the School of Finance and Economics in the Yangtze Normal University. His research interests are tourism policies and strategies and tourism economics. He hosted one Humanities and Social Science Fund of Ministry of Education of China and one Chongqing Social Science Planning of China. He has over 10 academic theses on Inquiry into Economic Issues, Asia Pacific Journal of Tourism Research, and so on and compiled one book.

**Nao Li**, PhD, is a Professor at School of International Economics and Management of Beijing Technology and Business University. Her research interests include big data analysis and mining, computer simulation, and technology applications in tourism.

**Rob Law**, PhD, is a Professor at the School of Hotel and Tourism Management, the Hong Kong Polytechnic University. His research interests are information management and technology applications.

**Heng Liu**, PhD, is a student at the School of International Trade and Economics in the University of International Business and Economics. His research interest is Financial Technology.