# Forecasting tourism demand with an improved mixed data sampling model

**Long Wen**

*School of Economics, University of Nottingham Ningbo China, Ningbo, 315100, PR China*
*Address: 199 Taikang East Road, Ningbo, China*
*Tel: +86-15967807611; Fax: +852-2362-9362;*
*Email: long.wen@nottingham.edu.cn*

**Chang Liu**

*School of Economics, University of Nottingham Ningbo China, Ningbo, 315100, PR China*
*Address: 199 Taikang East Road, Ningbo, China*
*Tel: +86-574-8818-0125; Fax: +86-574-8818-0125*
*Email: chang.liu@nottingham.edu.cn*

**Haiyan Song**

*Shenzhen Research Institute, School of Hotel and Tourism Management, The Hong Kong Polytechnic University, Hong Kong SAR*
*Adresss: 17 Science Museum Road, TST East, Kowloon, Hong Kong*
*Tel: +852-3400 2286; Fax: +852-2362-9362*
*Email: haiyan.song@polyu.edu.hk*

**Han Liu***

*Center for Quantitative Economics and Business School, Jilin University, Changchun, 130012, PR China*
*Acknowledgment: Qianjin Avenue, #2699, Gaoxin District, Changchun, 130012. China.*
*Email: hanliu@jlu.edu.cn*
*Tel: +86-13578656976; Fax: +86-431-85166347*
*Email:hanliu@jlu.edu.cn*

**Acknowledgement**

The authors would like to acknowledge the Natural Science Foundation of China for the financial support of the study (Grant No. NSFC71673233).

***Corresponding Author:**
Han Liu, Center for Quantitative Economics and Business School, Jilin University, Qianjin Avenue, #2699, Gaoxin District, Changchun, 130012. China.
Email: hanliu@jlu.edu.cn

# Forecasting tourism demand with an improved mixed data sampling model


## Abstract

Search query data reflect users' intentions, preferences and interests. The interest in using such data to forecast tourism demand has increased in recent years. The mixed data sampling (MIDAS) method is often used in such forecasting, but is not effective when moving average (MA) dynamics are involved. To investigate the relevance of the MA components in MIDAS models to tourism demand forecasting, an improved MIDAS model that integrates MIDAS and the seasonal autoregressive integrated moving average process is proposed. Its performance is tested by forecasting monthly tourist arrivals in Hong Kong from mainland China with daily composite indices constructed from a large number of search queries using the generalised dynamic factor model. The forecasting results suggest that this new model significantly outperforms the benchmark model. In addition, comparing the forecasts and nowcasts shows that the latter generally outperform the former.



**Keywords**: Tourism demand forecasting; MIDAS; Search query data; Generalised dynamic factor model; Nowcasts

# 1. Introduction

The perishable nature of the tourism industry makes accurately forecasting tourism demand an important task for tourism- and hotel-related decisionmakers. It is impossible to store unfilled airline seats and unsold hotel rooms. Therefore, accurate demand forecasts can help tourism practitioners make business decisions, such as those concerning scheduling, staffing and pricing. In addition, policymakers in tourist destinations need accurate forecasts to formulate tourism development policies, such as tourism infrastructure investments.

Traditional tourism demand forecasting studies have often used historical tourism demand and macroeconomic data. However, macroeconomic data, such as GDP and CPI, are usually delayed and may take several weeks or months to be published. The rapid development of information technology and the Internet has given rise to massive-scale and readily available data (Kambatla et al. 2014). Such data often reflect users' intentions and can serve as early indicators of various activities. For example, search queries have been used for various forecasting purposes, such as unemployment claims (Choi and Varian 2012), influenza epidemics (Ginsberg et al. 2009) and housing prices and sales (Wu and Brynjolfsson 2015).

Search query data have also gained popularity in forecasting tourism demand. Tourists use search engines to look for travel information on weather, transportation, hotels, attractions, travel guides and other tourists' opinions (Fesenmaier et al. 2011). These web search behaviours are recorded and reflect users' intentions, preferences and interests. Therefore, they can be valuable predictors of tourism demand. Although the use of search query data in tourism demand forecasting is relatively new, interest in this area has increased rapidly in recent years. Search query data have often been aggregated and converted into the same

frequency as tourism demand variables in previous studies because they are often sampled at a higher frequency (Choi and Varian 2012; Li et al. 2017; Pan et al. 2012; Rivera 2016; Yang et al. 2015). This can lead to information loss and poor forecasting performance because high frequency information is not used (Ghysels et al. 2007). Bangwayo-Skeete and Skeete (2015) were the first to introduce mixed data sampling (MIDAS) for tourism demand forecasting. They found that MIDAS performed better in forecasting monthly tourist arrivals using weekly Google Trends data in most forecasting exercises, whereas its performance was poor in other exercises. Compared with common benchmark models, such as the seasonal autoregressive integrated moving average (SARIMA) model, traditional MIDAS models often involve autoregressive (AR) components and are unable to incorporate moving average (MA) dynamics. Indeed, they are not effective when the underlying data include MA dynamics. In fact, in a recent study, Foroni et al. (2019) showed that MA components emerged in a MIDAS model in which the low frequency variable was the result of temporal aggregation. They investigated the effect of neglecting MA components in the forecasts and found that including MA components improved the forecasting performance of their Monte Carlo simulations and application to US macroeconomic variables. In this study, the same idea is introduced to tourism demand forecasting and the relevance of MA components is investigated in this context. In addition, Foroni et al. (2019) focused on forecasting macroeconomic variables and did not consider seasonal ARMA components. As tourism demand often exhibits strong seasonality, it is important to account for seasonality in the modelling process. Moreover, Foroni et al. (2019) arbitrarily determined the orders of AR and MA components in MIDAS models. Doing so may yield a higher probability of model misspecification. To overcome these problems, a new model

that integrates the MIDAS and SARIMA processes is proposed. This new model is an extension of traditional MIDAS models and is able to accommodate seasonal and non-seasonal ARMA components. The features of the MIDAS and SARIMA models are especially relevant in the tourism demand forecasting context. The mixed frequency aspect of the new model provides a more efficient way to utilise high frequency search query data. Furthermore, its seasonal and non-seasonal ARMA components capture important characteristics of tourism demand. In this study, the effectiveness of the model is investigated by forecasting monthly tourist arrivals in Hong Kong from mainland China, with daily composite indices constructed from a large number of search queries using the generalised dynamic factor model (GDFM). Previous studies have focused on forecasting or nowcasting tourism demand. In contrast, this study is the first to conduct a comparison analysis of forecasts and nowcasts. Such a comparison may be particularly useful for decisionmakers who need frequent updates to make more accurate forecasts. When forecasting tourism demand, traditional macroeconomic data, such as income level in the origin country and relative price level in the destination country, are often incomplete and subject to revision for the current and most recent periods. However, search query data are readily available on a daily and even hourly basis. They are especially useful in a nowcasting framework, which can enable more timely tourism demand forecast updates when new information becomes available. For example, timely and improved updates of nowcasts of demand are very valuable in hotel revenue management, which involves dynamic pricing.

The remainder of this paper is organised as follows. Section 2 reviews the relevant literature. Section 3 presents the data and the construction of the search query index. Section 4 discusses

the details of the models and their estimation results. Section 5 presents the forecasting and nowcasting results. Finally, Section 6 concludes.

# 2. Literature review

## 2.1. Tourism demand forecasting

Tourism demand forecasting is a well-established research area. The three main types of modelling techniques include non-causal time series, econometric and artificial intelligence (AI)-based methods.

Traditional time series models include Naïve 1 models (no change), Naïve 2 models (constant growth rate), exponential smoothing models and simple AR models (Song and Li 2008; Wu et al. 2017). They are often used as benchmarks in tourism forecasting studies. Autoregressive integrated moving average (ARIMA) models and SARIMA models are the most commonly used models, depending on the frequency of the time series. Various extensions of the ARIMA model have also been used in the literature. For example, Chu (2009) introduced an autoregressive ARMA (ARARMA) model and a fractionally integrated ARMA (ARFIMA) model to forecast tourist arrivals in nine destinations in the Asia-Pacific region and found that the ARFIMA model outperformed the SARIMA and ARARMA models. Similarly, Assaf et al. (2011) used several models based on fractional integration to forecast tourist arrivals in Australia, confirming that they outperformed the standard ARIMA and SARIMA models.

Structural time series (Turner and Witt 2001) and generalised autoregressive conditional heteroskedastic (Divino and McAleer 2010) models have also been widely used in the tourism literature. In recent years, more advanced time series models have been used to generate better forecasting performance than traditional time series models, such as innovations state

space models for exponential smoothing (ETS; Athanasopoulos et al. 2011), singular spectrum analysis (SSA) models (Hassani et al. 2017) and time-varying parameter structural time series models (Song et al. 2011). Decomposition methods, such as SSA, empirical mode decomposition (Yahya et al. 2017) and ensemble empirical mode decomposition (Zhang et al. 2017), have gained much popularity in recent years and have demonstrated good forecasting performance. These techniques have been used in univariate time series forecasting settings (Hassani et al. 2017; Hassani et al. 2015; Silva et al. 2019) and causal time series forecasting settings (Li and Law 2020).

Unlike non-causal time series models, econometric models can analyse the relationship between tourism demand and its key determinants, and the information can be used to provide policy recommendations. Several important factors affecting tourism demand have been identified in the literature, such as tourist income, tourism prices in a destination relative to those of the country of origin, tourism prices in competing destinations and real exchange rates (Song and Li 2008; Wu et al. 2017).

Spurious regression is often present in traditional regression analysis. Several modern econometric models have been introduced in tourism modelling and forecasting, such as the autoregressive distributed lag model (Song et al. 2012), the error correction model (Goh 2012), the vector autoregressive (VAR) model (Wong et al. 2006), the time-varying parameter model (Page et al. 2012), the almost ideal demand system model (Li et al. 2006) and the Bayesian VAR model (Gunter and Önder 2015; Wong et al. 2006). Numerous studies have concluded that econometric models perform better (Song et al. 2003), but some have confirmed that time

series models outperform econometric models in predicting tourism demand (Athanasopoulos et al. 2011).

In addition to time series and econometric methods, a variety of AI-based methods have been introduced in the tourism forecasting literature. The dominant model is the artificial neural network (ANN) model. It consists of several layers, each of which can contain multiple neurons. The ANN model is a nonparametric and data-driven method that can be used to model non-linear relationships. It is also the most frequently used AI-based method in tourism demand forecasting studies (Claveria et al. 2015; Law et al. 2019; Sun et al. 2019). Other AI-based methods used to forecast tourism demand include the support vector machine model (Chen et al. 2015; Hong et al. 2011), the fuzzy system model (Aladag et al. 2014), the rough set model (Goh et al. 2008) and grey theory (Sun et al. 2016).

Although various methods have been introduced and applied in the literature, there is a consensus that no model can outperform other models consistently under all conditions (Song and Li 2008). Using a meta-analysis, Peng et al. (2014) showed that their data characteristics and study features, such as demand measure, data frequency and origin/destination pairs, affected the forecasting accuracy of tourism demand.

## 2.2. Forecasting with search query data

People often search for information online and their search behaviour reflect their consumption preferences and decision-making processes (Du et al. 2014; Ghose et al. 2014). Search query data can serve as a powerful predictor to improve forecasting accuracy. Thus, forecasting using search query data has gained popularity in a number of research areas.

For example, Ginsberg et al. (2009) investigated a large number of Google search queries to track influenza-like illnesses; ultimately, their method improved early detection. Since then, researchers have explored the usefulness of search query data for forecasting unemployment rates (Askitas and Zimmermann 2009), consumer consumption (Vosen and Schmidt 2011), stock markets (Bordino et al. 2012; Da et al. 2011), automobile sales (Du and Kamakura 2012) and house prices and sales (Wu and Brynjolfsson 2015).

In recent years, forecasting tourism demand using search engine data has also attracted attention. For example, Choi and Varian (2012) used Google Trends data for the first 2 weeks of each month to predict the number of visits to Hong Kong in a given month. Pan et al. (2012) chose five related Google search queries to forecast demand for hotel rooms in Charleston, US, improving forecasting performance by including search query data. Pan and Yang (2017) used Google search engine queries and website traffic data to forecast hotel demand in Charleston and found that their forecasts were more accurate when they included both data sources. Rivera (2016) pointed out that Google Trends data differ each week because the data are constructed as a relative volume and come from a periodic sample of queries. Therefore, he proposed using a dynamic linear model and treated Google Trends data as a representation of an unobservable process. In addition, the association between hotel demand and Google Trends data can be better understood when the data are downloaded on multiple occasions. Yang et al. (2015) used Google Trends and the Baidu Index, which represents the absolute volume of the chosen search queries, to forecast the number of visitors to a province in China. They found that although the data from both search engines improved forecasting accuracy, the Baidu Index performed better. Li et al. (2017) used the GDFM to construct the composite

index from a large number of Baidu search queries to forecast tourist arrivals in Beijing. They showed improved forecasting performance using the GDFM compared with another dimension reduction method, principal component analysis (PCA). Recently, studies have also used ANNs to model the relationship between tourism demand and search query data. Sun et al. (2019) used Google and Baidu search data to forecast tourist arrivals in popular destinations in China, showing better forecasting performance when using the kernel extreme learning machine model. Similarly, Wen et al. (2019) used the Baidu Index to forecast tourist arrivals in Hong Kong from mainland China, using a newly proposed hybrid model integrating the ARIMA and ANN models. They found that the hybrid model outperformed component models. Law et al. (2019) applied a deep learning approach to forecast tourist arrivals in Macau using search query data. They showed that the deep learning approach significantly outperformed the support vector regression model and the traditional ANN model.

## 2.3. MIDAS regressions

Time series data are often collected at different frequencies, but most models require variables to be converted to the same low frequency. During this process, the potentially valuable information contained in high frequency variables is smoothed and lost. To tackle this problem, Ghysels et al. (2004) used MIDAS regressions to directly estimate equations with variables sampled at different frequencies.

The use of MIDAS regressions has proliferated in the macroeconomic literature. For example, Clements and Galvão (2008) used MIDAS to forecast quarterly output growth using monthly predictors and found significant improvement. Andreou et al. (2013) extracted a small set of daily financial data from a large panel of daily financial assets to predict quarterly real GDP

growth using MIDAS and elucidated the value of daily financial information. MIDAS has also been used to forecast inflation and oil prices. Monteforte and Moretti (2012) showed a reduction in inflation forecast errors in the euro area by including daily financial variables using MIDAS. Baumeister et al. (2015) investigated the predictive power of daily and weekly financial market data in forecasting monthly oil prices. They demonstrated that the preferred MIDAS model improved forecasting accuracy compared with no-change forecasts.

MIDAS regressions have also been widely used in the financial literature. Ghysels et al. (2009) compared several models generating multi-period ahead forecasts of stock return volatilities and found that MIDAS performed best for longer horizon forecasts. Gurgul et al. (2018) used MIDAS-based models for systemic risk assessment in the banking sector and found that the information contained in the macroeconomic variables helped predict short- and long-term risk components.

Bangwayo-Skeete and Skeete (2015) were the first to apply MIDAS with AR components in the tourism literature. Using weekly Google data to forecast monthly tourist arrivals in five Caribbean countries, they found that the MIDAS models generated better predictions than the baseline time series models for most of their experiments. However, MIDAS models can only accommodate AR dynamics and forecasting performance may deteriorate when MA dynamics are involved. They are not effective when the underlying data include MA dynamics. Foroni et al. (2019) showed that MA components in general emerged in a MIDAS model and improved the forecasting accuracy of US macroeconomic variables by including MA components. In this study, the relevance of MA components in tourism demand forecasting is investigated using an improved MIDAS model that incorporates seasonal ARMA components and automatically

selecting appropriate structures. This novel model combines the advantages of MIDAS and SARIMA and can offer desirable features for modelling tourism demand using search queries. In addition to accommodating the mixed frequency variables provided by MIDAS, it can also automatically choose appropriate seasonal and non-seasonal ARMA components, which are often present in tourism demand data.

## 3. Data and composite index

Hong Kong is one of the most popular tourist destinations in Asia. It is distinguished by its unique culture and often described as a place where 'East meets West'. Despite the modernised lifestyles of the people in Hong Kong, traditional Chinese practices and cultural events have been preserved, such as feng shui and the dragon boat festival. Tourism is one of the four pillar industries of the Hong Kong economy. In 2016, it contributed to approximately 5% of Hong Kong's GDP and 7% of total employment. After 2 years of decline in 2014 and 2015, the total number of arrivals reached a growth rate of 3.2% in 2017 with 58.5 million visitors (Tourism Commission—Tourism Fact Sheets 2018). Mainland China remains Hong Kong's largest source market, accounting for approximately 76% of all visitors. The increase in the number of visitors from mainland China in recent decades has been largely fuelled by visa liberalisation policies, such as the 2003 Individual Visit Scheme and Shenzhen residents' multiple-entry permits in 2009. The increased number of visitors from mainland has boosted tourism revenue and generated many job opportunities. However, it has also led to higher prices and a shortage of goods, causing tension between mainland visitors and Hong Kong residents. Thus, businesses and policymakers require accurate forecasts of tourist arrivals from mainland China to make informed decisions.

In this study, we used data on monthly tourist arrivals from mainland China to Hong Kong between January 2011 and February 2018. The data were collected from the Hong Kong Tourism Board's B2B website, PartnerNet (https://partnernet.hktb.com). The data were sampled from 2011 because the Baidu Index data are only available from 2011. Following previous studies, the log transformation was applied before starting the modelling process.

Although Google dominates the global market, it left the mainland China market in 2010 following a dispute with the Chinese government. Baidu has since become the most popular search engine in China, holding the largest market share (Yang et al. 2015). Given this study's interest in tourist arrivals in Hong Kong from mainland China, the Baidu Index was used.

To apply the search query data to tourism forecasting, keyword selection was conducted first. The most common method for selecting search query data is based on the researcher's intuition and prior knowledge (Brynjolfsson et al. 2014). This practice is common in the tourism field. For instance, Pan et al. (2012) chose five Google search queries to forecast hotel demand. Bangwayo-Skeete and Skeete (2015) also adopted this method and used 'hotels' and 'flights' as keywords to forecast tourist arrivals in the Caribbean. Although this method is easy to apply, it can omit important information by excluding relevant search queries. To mitigate this problem, the set of initial keywords can be extended by adding pertinent keywords using the functions of the search engine (Li et al. 2017; Yang et al. 2015). In this study, the initial set of keywords was thus extended and conducted according to the following steps to select the keywords in the Baidu Index:

1. Six aspects of tourism planning were specified: dining, shopping, transportation, tours, attractions and lodging. Several initial keywords were determined for each aspect.

2. Keywords highly correlated to the initial keywords were added using a demand map interface provided by Baidu. This step was iterated until convergence.
3. As Baidu does not provide the search query volume below a certain threshold, the availability of each search query was manually checked using the keywords.

Ultimately, 101 Baidu search queries were collected (the names of the translated search queries can be found in Appendix A).

With this large number of search queries, some AI models, such as the deep learning models used in Law et al. (2019), can directly incorporate these search queries and identify the most relevant ones. However, most econometric models, including the MIDAS models used in this study, cannot perform this task. As a result, the dimensionality of the search queries must be reduced before the modelling process. This can be done by extracting common components using various factor models, such as static and dynamic factor models. Static factor models express common components as a linear combination of a small number of unobserved static factors that are loaded simultaneously (Stock and Watson 2002). The GDFM proposed by Forni et al. (2000) encompasses the static factor model and its common components, $\chi_{it}$, are driven by $q$ unobservable common factors, $f_{jt}$, $j = 1, \dots, q$. For the observed variables $\{\mathrm{X}_{it}, i = 1, \dots, n, t = 1, \dots, T\}$, the model can be formulated as

$$\mathrm{X}_{it} = \chi_{it} + \varepsilon_{it}, \quad (1)$$

$$\chi_{it} = b_{i1}(L)f_{1t} + b_{i2}(L)f_{2t} + \cdots + b_{iq}(L)f_{qt}, \quad (2)$$

where $b_{ij}(L) = \sum_{k=1}^{\infty} b_{ijk}L^k$ is the factor loading, $L$ is the lag operator and $\varepsilon_{it}$ is the idiosyncratic component. The GDFM has two important characteristics: it is dynamic and allows for cross-correlation among idiosyncratic components. Unlike static factor models, in which

lagged factors are added as additional static factors, the common components of the GDFM can accommodate AR and MA responses. Distinguishing between leading and coincident variables a priori is not needed. The common components of the GDFM depend on cross-correlations at all leads and lags, so they can incorporate different lead and lag information from the variables (Forni et al. 2000). This is advantageous for constructing a composite index from a large number of search queries.

The GDFM has been adopted by several economic and financial institutions to analyse and predict economic activities. The Banca d'Italia published a real-time monthly coincident indicator of the euro area business cycle (Eurocoin) based on the GDFM (Altissimo et al. 2010). The Federal Reserve Bank of New York developed a similar index for estimating underlying inflation using these methods (Amstad and Potter 2009). In the tourism context, Li et al. (2017) were the first to use the GDFM to construct the composite index from Baidu search queries. They found that the GDFM-based index had better forecasting performance than PCA. Therefore, the GDFM was adopted in this study to construct the index.

Before estimating the GDFM, a number of common factors, $q$, must be determined. To this end, Forni et al. (2000) used the variance contribution rate, where $q$ is the number of factors whose variance contribution rates converge. However, this is a heuristic eye inspection rule. Nevertheless, Hallin and Liška (2007) proposed a formal test, using the log criterion of their study with the penalty function $p_1$ and lag window $\sqrt{T}$ to choose the number of factors, the maximum number of factors being set to 50. $c$ is the coefficient associated with the penalty function, $S_c$ is defined as the variability of the estimated $q$ when the size of the subsample increases and $q_{c;n}^{*T}$ is the estimated $q$ when the whole sample is used. The selection of $q$ can

be based on the plot of $S_c$ and $q_{c;n}^{*T}$ on $c$, where $c$ is based on the second stability interval (more details are provided by Hallin and Liška 2007). This method was applied to the search queries of this study. The plot is shown in Fig. 1.

(Insert Fig. 1 about here)

The second stability interval appeared at the interval between 0.31 and 0.34 (with $S_c$ equal to 0) and the estimated $q$ was equal to 4. Therefore, the number of common factors was set to 4 for the search queries.

The common components were then calculated using standardised search query data with a mean of 0 and a standard deviation of 1. The coincidental index at time *t* was constructed using the common components, $z_t = \sum_{i=1}^{n} \chi_{it}$. As the search queries were collected daily, this index also had a daily frequency. The relationship between the daily index and the log transformation of monthly tourist arrivals is plotted in Fig. 2.

(Insert Fig. 2 about here)

The close relationship between the daily index and monthly tourist arrivals is clearly illustrated.

# 4. Research methods

In this section, the specifications and estimation procedure of the following competing models are presented: the SARIMA model, the SARIMA model with an exogenous variable (SARIMAX) and the traditional and improved MIDAS models. Data up to February 2017 were used for the

estimation procedure and the remaining data were used to evaluate the forecasting performance.

## 4.1. SARIMA and SARIMAX

The SARIMA model can account for seasonality, which is a common feature of tourism demand. It is the most commonly used time series model in the tourism demand forecasting literature and is often used as a benchmark (Song and Li 2008; Wu et al. 2017). A SARIMA (*p, d , q*)(*P, D , Q*) model with seasonal frequency *m* can be specified as follows:

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d y_t = c + \Theta(B^m)\theta(B)\epsilon_t, \quad (3)$$

where $y_t$ is the log of tourist arrivals, $B$ is the backshift operator, $\Phi(x)$ and $\Theta(x)$ represent the seasonal AR and MA components (which are polynomials of order *P* and *Q*), respectively, $\phi(x)$ and $\theta(x)$ represent the non-seasonal AR and MA components (which are polynomials of order *p* and *q*), respectively, and $\epsilon_t$ is a white noise process. The *forecast* package in the R program (R Core Team 2016) was used to automatically select the orders and estimate the coefficients (Hyndman and Khandakar 2008) as follows:

1. The order of seasonal differencing $D$ was chosen using a test suggested by Wang et al. (2006), which is based on a measure of seasonal strength.
2. The order of non-seasonal differencing $d$ was chosen using the KPSS unit-root test (Kwiatkowski et al. 1992).
3. A stepwise procedure was used to traverse the model space and the orders and *p, q, P* and *Q* were chosen based on the corrected Akaike information criterion (AIC).

The SARIMAX model simply adds an exogenous variable to SARIMA so that it becomes a regression model with SARIMA errors. Therefore, the estimation procedure of the SARIMAX

model is almost identical to that of the SARIMA model, except that the regression is conducted first. It can be formulated as follows:

$$y_t = \beta_0 + \beta_1 x_t + n_t \quad (4)$$

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d n_t = \Theta(B^m)\theta(B)\epsilon_t, \quad (5)$$

where $x_t$ is the exogenous variable (which may include lagged variables) and $n_t$ is the error from the regression model. It is equivalent to substituting the differencing terms in the following regression equation:

$$(1-B^m)^D(1-B)^d y_t = (1-B^m)^D(1-B)^d(\beta_0 + \beta_1 x_t) + n_t' \quad (6)$$

$$\Phi(B^m)\phi(B)n_t' = \Theta(B^m)\theta(B)\epsilon_t, \quad (7)$$

where $n_t'$ is $(1-B^m)^D(1-B)^d n_t$. Furthermore, this is equivalent to differencing $y_t$ and $x_t$ before fitting the model with ARMA errors. As non-stationary errors suggest the existence of spurious regression, it is necessary to difference the variables first.

The SARIMAX model can be rewritten as follows:

$$(1-B^m)^D(1-B)^d\Phi(B^m)\phi(B)y_t = (1-B^m)^D(1-B)^d\Phi(B^m)\phi(B)(\beta_0 + \beta_1 x_t) + \Theta(B^m)\theta(B)\epsilon_t \quad (8)$$

It can be seen that the same AR terms are applied to $y_t$ and $x_t$.

The exogenous variable used in this study was the composite index constructed from the search queries using the GDFM. As it was available daily, temporal aggregation was conducted by averaging the daily index for each month. However, the number of days varies in different months. To enable a direct comparison between the SARIMAX and MIDAS models, the 30 days preceding the first day of each month were considered to be a full last month. The monthly composite index at time *t* is denoted as $index_t$.

The monthly index with at least one lag was added to the SARIMAX model and the lag length was determined based on the AIC and the Bayesian information criterion (BIC). A monthly index with one lag was found to generate the smallest AIC and BIC.

After estimation, the fitted SARIMA and SARIMAX models can be written as

$$SARIMA: \quad (1 + 0.2651B - 0.5819B^2)(1 - B^{12})(1 - B)y_t = (1 - 0.6750B^{12})(1 - 1.0745B + 1.2188B^2 - 0.4372B^3)\epsilon_t \tag{9}$$

$$SARIMAX: \quad \begin{cases} y_t = 0.2122index_{t-1} + n_t \\ (1 - 0.8780B - 0.5241B^2)(1 - B^{12})(1 - B)n_t = (1 - 0.5468B^{12})\epsilon_t \end{cases} \tag{10}$$

The details of the estimation results are summarised in Table 1.

(Insert Table 1 about here)

The $index_{t-1}$ coefficient was positive and significant at the 1% level, indicating that an increase in search queries leads to an increase in tourist arrivals the following month. The smaller AIC and BIC values of the SARIMAX model suggest that including the search query index fitted the model better. A Ljung-Box test was conducted to check the residuals of the fitted models and the *p* values were reported. The large *p* values indicate that the residuals were independently distributed and that the models were properly specified.

## 4.2. MIDAS models

Search query data are available at a higher frequency than tourist arrival data. They contain potentially valuable information, and temporal aggregation can lead to information loss (Ghysels et al. 2007). Most time series regressions involve data sampled at the same frequency, so high frequency information cannot be used directly. As an alternative to the common

solution of converting all data to the same low frequency, MIDAS can directly accommodate variables sampled at different frequencies. MIDAS models can be applied in cases where high frequency variables are used to forecast a low frequency variable. In addition, they may have more salient advantages when the frequencies of the variables are significantly different, as using traditional methods can lead to greater information loss during temporal aggregation. Therefore, MIDAS models are well suited to this study using monthly tourist arrivals and daily search queries.

The basic MIDAS model for a single explanatory variable can be written as

$$y_t = \beta_0 + \beta_1 \sum_{i=1}^{l} \omega(i;\Theta) L_{HF}^{i} z_t + \epsilon_t, \quad (11)$$

where $y_t$ is the log of tourist arrivals, $L_{HF}$ is the high frequency lag operator, $\omega(i;\Theta)$ is a polynomial that assigns the weights to the high frequency variable $z_t$ at lag *i*, $l$ is the maximum lag on the high frequency variable and $\epsilon_t$ is a white noise process.

Different weighting schemes can be used as functional constraints. A weighting scheme defined by the vector of parameters $\Theta = (\theta_1, \theta_2, \dots, \theta_p)$ can be written as

$$\omega(i;\Theta) = \frac{f(i,\Theta)}{\sum_{j=1}^{l} f(j,\Theta)} \quad (12)$$

The most popular specifications for $f(i,\Theta)$ include the exponential Almon function and the beta function (Ghysels et al. 2007). Ghysels et al. (2007) argued that the beta function was flexible enough to accommodate different weighting shapes with only two parameters. For comparison purposes, the exponential Almon specification used in this study also used two parameters. In addition, a Gompertz function was added as an additional comparison. The specifications of these three functions with two parameters ($\theta_1, \theta_2$) are expressed below:

Exponential Almon: $f(i,\Theta) = \exp(\theta_1 i + \theta_2 i^2)$.

Beta: $f(i,\Theta) = \frac{(k)^{\theta_1-1}(1-k)^{\theta_2-1}\Gamma(\theta_1+\theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)}$, where $k = \frac{i}{l}$ and $\Gamma$ is the standard gamma function.

Gompertz: $f(i,\Theta) = \exp(\theta_2 i)\exp(-\theta_1 \exp(\theta_2 i))$.

MIDAS models can be expanded to include AR dynamics. However, this process is not straightforward, as noted by Ghysels et al. (2007). Consider a MIDAS-AR model with one lag of $y_t$:

$$y_t = \beta_0 + \lambda y_{t-1} + \beta_1 \sum_{i=1}^{l} \omega(i;\Theta) L_{HF}^{i} z_t + \epsilon_t \quad (13)$$

It can be rewritten as

$$y_t = \beta_0(1-\lambda)^{-1} + \beta_1(1-\lambda L_{LF})^{-1} \sum_{i=1}^{l} \omega(i;\Theta) L_{HF}^{i} z_t + \tilde{\epsilon_t}, \quad (14)$$

where $L_{LF}$ is the low frequency lag operator and $\tilde{\epsilon_t} = (1-\lambda L_{LF})^{-1}\epsilon_t$. The polynomial on $z_t$ is a combination of $L_{LF}$ and $L_{HF}$. This generates a seasonal response of $y_t$ to $z_t$, whether $z_t$ demonstrates seasonal patterns. This strategy is generally considered inappropriate. Therefore, Clements and Galvão (2008) suggested introducing AR dynamics in $y_t$ as a common factor,

$$y_t = \beta_0 + \lambda y_{t-1} + \beta_1 \sum_{i=1}^{l} \omega(i;\Theta) L_{HF}^{i}(1-\lambda L_{LF}) z_t + \epsilon_t, \quad (15)$$

where the same AR dynamics are applied to $y_t$ and $z_t$ so that the response of $y_t$ to $z_t$ is non-seasonal. This model was adopted in this study as MIDAS-AR. Before estimating MIDAS-AR, the amount of seasonal and non-seasonal differencing for tourist arrivals was determined using the same tests as the SARIMA model (Kwiatkowski et al. 1992; Wang et al. 2006). The same orders of differencing were applied to the daily index for interpretability. In addition, the AR orders were chosen based on the AIC and BIC.

MIDAS-AR models are less effective when MA dynamics are involved, therefore it is desirable to include MA components in MIDAS. Foroni et al. (2019) proved the usefulness of incorporating MA components into MIDAS models to predict US macroeconomic variables. However, they did

not consider seasonal components and the orders of ARMA components were determined arbitrarily. To remedy these shortcomings, an improved MIDAS model integrating MIDAS and SARIMA (MIDAS-SARIMA) was proposed. This model can be written as:

$$y_t = \beta_0 + \beta_1 \sum_{i=1}^{l} \omega(i;\Theta) L_{HF}^{i} z_t + n_t \quad (16)$$

The difference between this new model and a standard MIDAS is that $n_t$ is a SARIMA process. Thus, the MIDAS-SARIMA model can accommodate seasonal and non-seasonal ARMA dynamics. This model is distinguished from standard MIDAS models by its seasonal structure. In addition, it integrates the automatic order selection procedure of SARIMA models to determine the best structure of the MIDAS-SARIMA model. As a result, this model combines the advantages of the MIDAS and SARIMA models and offers considerable potential to improve forecasting accuracy. The MIDAS-SARIMA model applies the same AR dynamics to $y_t$ and $z_t$. The estimation procedure of the MIDAS-SARIMA model is the same as that of the SARIMAX model, except that the first step is a MIDAS regression instead of a linear regression.

Unlike the SARIMAX model, which assigns the same weights to the high frequency variable after temporal aggregation, the MIDAS-SARIMA model relaxes this restriction, which is useful because the search query data for different days are likely to have different effects on monthly tourist arrivals. This may lead to different weights of the daily composite index. The flexibility provided by the MIDAS-SARIMA model probably improves the forecasting accuracy of monthly tourist arrivals.

The number of lags of the daily composite index was set to 30 for MIDAS models in accordance with the SARIMAX model. That is, the 30 daily indices preceding the first day of the month were used to forecast tourist arrivals for the following month. This setting enabled a direct

comparison between the SARIMAX and MIDAS models. The estimation results of the MIDAS-AR and MIDAS-SARIMA models are summarised in Table 2. MIDAS-AR-Almon, MIDAS-AR-Beta and MIDAS-AR-Gom denote MIDAS-AR models with the exponential Almon, beta and Gompertz weighting schemes, respectively. MIDAS-SARIMA-Almon, MIDAS-SARIMA-Beta and MIDAS-SARIMA-Gom denote MIDAS-SARIMA models with the exponential Almon, beta and Gompertz weighting schemes, respectively.

(Insert Table 2 about here)

Seasonal and non-seasonal differencing were performed for all MIDAS models. The MIDAS-AR models gave the same structures, with two lags of the AR dynamics, and the estimated coefficients of the two AR components were close for different weighting schemes. The same was observed for the MIDAS-SARIMA models, which had the same MA(1) and SMA(1) structures and similar estimated coefficients. This suggests that the different weighting schemes made little difference in the estimation of the MIDAS models. This result is consistent with the results of Bangwayo-Skeete and Skeete (2015). The AIC and BIC values suggest that the MIDAS-SARIMA models had a better fit and were more appropriate than the MIDAS-AR models. The weights of the daily indices can be visualised. For example, the weights of the 30 daily indices for the MIDAS-SARIMA models are plotted in Fig. 3.

(Insert Fig. 3 about here)

All three weighting schemes weighted the most recent indices more heavily. Most weights were put on the last 15 days, whereas the earlier days had almost 0 weight. Furthermore, the beta weighting scheme put the highest weight on Day 2, whereas Day 1 was given very little weight. The exponential Almon and Gompertz weighting schemes generated similar patterns to that of the beta weighting scheme. However, their weighting curves were much smoother than that of the beta weighting scheme. The close estimates of $\beta_1$ shown in Table 2 suggest that the total weights of the daily indices were similar.

# 5. Result evaluation

## 5.1. Forecasting

In this subsection, the forecasting performance of the models using data from March 2017 to February 2018 is evaluated. Search query data with one lag were used in the modelling process, with the tourist arrivals and search query data available up to time t. Thus, the ARIMAX and MIDAS models had to first be estimated using tourist arrivals data up to time t and search query data up to time t-1. The results were then used to generate the forecasts at time t+1, with search query data at time t. Therefore, only one-step-ahead forecasts could be generated in this study. Longer-term forecasts may be further investigated in a future study with a different estimation procedure that uses search query data of lags longer than one but not conducted here. An expanding window approach was used to generate the one-step-ahead dynamic forecasts. For example, the data on tourist arrivals up to February 2017 and the search query data up to January 2017 were first used to estimate the models, then the search query data for February 2017 were used to forecast tourist arrivals in March 2017. The estimation period was then extended by 1 month and the models were re-estimated using the same procedure. The

forecasts were generated at each round until all 12 one-step-ahead forecasts were calculated for the period from March 2017 to February 2018.

Forecast accuracy was evaluated using five commonly used forecast error measures, including the mean absolute deviation (MAD), the mean squared error (MSE), the mean absolute percentage error (MAPE), the root mean square percentage error (RMSPE) and Theil's U statistic (Goh and Law 2002; Law et al. 2019). The MAD and MSE are absolute error measures. In contrast, the MAPE and RMSPE are relative error measures. Finally, Theil's U was constructed based on the error ratio of the underlying model to the seasonal naïve model. A seasonal naïve model basically predicts that monthly tourist arrivals for the following year will be the same as this year for the same month. A value less than 1 indicates that the performance of the model is superior to that of the naïve model. Their specifications are as follows:

$$\text{MAD} = \frac{1}{n}\sum_{t=1}^{n} |A_t - F_t|$$

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} (A_t - F_t)^2$$

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \frac{|A_t - F_t|}{A_t}$$

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \left(\frac{A_t - F_t}{A_t}\right)^2}$$

$$\text{U} = \frac{\sqrt{\sum_{t=1}^{n}(A_t - F_t)^2}}{\sqrt{\sum_{t=1}^{n}(A_t - A_{t-12})^2}},$$

where $A_t$ is the actual value, $F_t$ is the forecast value at time *t* and *n* is the length of the forecast period (n = 12 in this study). Two extra benchmark models have been added to the forecasting practice: the ETS and the seasonal naïve model. Athanasopoulos et al. (2011) found that the ETS performed particularly well for monthly data in their tourism forecasting competition. The seasonal naïve model is also widely used as a benchmark model in forecasting seasonal tourism demand (Athanasopoulos et al. 2011). Table 3 presents the results of the models' forecasting performance.

(Insert Table 3 about here)

The rankings were mostly consistent based on the error measures. All of the models except for the seasonal naïve model had Theil's U values of less than 1, suggesting that all of the models outperformed the seasonal naïve model in terms of the squared error (SE).
Overall, the seasonal naïve model performed the worst, especially in terms of the MSE and RMSPE. In contrast, the ETS performed well in terms of the MSE and RMSPE. The SARIMAX model performed better than the SARIMA model based on the error measures and showed that search queries improved the forecasting accuracy of tourist arrivals. This result is consistent with the findings of previous studies (Li et al. 2017; Pan et al. 2012; Pan and Yang 2017; Yang et al. 2015). The MIDAS-AR models generated forecasts comparable to those of SARIMA and only outperformed SARIMA in terms of the MAPE and MAD. This result contrasts the findings of Bangwayo-Skeete and Skeete (2015). The MIDAS-AR models were also outperformed by the SARIMAX and MIDAS-SARIMA models, indicating that traditional MIDAS models have

limitations when using the information provided by high frequency search query data. The SARIMAX model performed well due to its flexibility to incorporate seasonal and non-seasonal ARMA components. Finally, the MIDAS-SARIMA models demonstrated the best performance in terms of all measures and remarkable improvements compared with the MIDAS-AR models. Thus, this improved MIDAS model combining the strengths of the MIDAS and SARIMAX models improved the forecasting performance. The results also suggest that forecasting performance was improved by integrating MA components into MIDAS models, which is consistent with the findings of Foroni et al. (2019).

MIDAS-SARIMA-Almon demonstrated the best performance based on the MAD, MAPE and RMSPE, whereas MIDAS-SARIMA-Gom demonstrated the best performance based on the MSE and Theil's U. Overall, different weighting schemes generated comparable forecasting performance, which is consistent with the findings of Bangwayo-Skeete and Skeete (2015). The results also indicate that the most recent search query data, which were assigned most weights in the MIDAS-SARIMA models, were the most valuable in predicting tourist arrivals.

To further test the significance of forecasting differences between the two better benchmark models (SARIMA and ETS) and the models using search query data, a Diebold-Mariano (DM) test was conducted (Diebold and Mariano 1995). The test was based on the forecasting differences of four measures, namely the absolute deviation, the SE, the absolute percentage error (APE) and the squared percentage error (SPE), which were used to calculate the MAD, MSE, MAPE and RMSPE, respectively. As Theil's U had the same denominator derived from the seasonal naïve model and its numerator was calculated from the MSE, the corresponding DM test largely depended on the MSE and was therefore omitted. Tables 4 and 5 present the

results of the DM tests for SARIMA and the ETS, respectively. The null hypothesis of the DM test is that the accuracy of the forecasts generated by the benchmark and alternative models does not differ.

(Insert Table 4 about here)

(Insert Table 5 about here)

As indicated by the DM test statistics, although the SARIMAX model outperformed the SARIMA model, the difference was not significant. Only the MIDAS-SARIMA models performed significantly better than SARIMA based on the error measures (at least at the 10% significance level), confirming the superiority of the proposed model and highlighting the importance of added flexibility to accommodate MA dynamics in MIDAS models. In the case of the ETS, only MIDAS-SARIMA-Almon generated significantly better results in terms of all measures. MIDAS-SARIMA-Gom significantly outperformed the ETS in terms of the AE and APE.

## 5.2. Nowcasting

The traditional models used in this study, such as the benchmark models and ARIMAX, are unable to update the forecasts until a full month of search query data are available, as they cannot incorporate high frequency search query data that offer a new daily index after each day. However, daily nowcasts can be generated using MIDAS models as new search query data become available every day. For example, when the search query data for the first day of the month are available, they can be added to the MIDAS models and used to predict tourist

arrivals for that month (nowcasting). This can be repeated every day until the end of the month. Again, due to the variable number of days in each month, nowcasts with search query data for 30 days starting from the first day of the same month were produced. As the MIDAS-SARIMA models had the best forecasting performance, their nowcasting performance was further investigated. Nowcasting is conducted in a similar way to forecasting. Nowcasting models must be refitted when new daily search query data become available, using the monthly tourist arrivals data until the end of the last month and the daily search query data until the end of that day. Then the nowcasts of the monthly tourist arrivals in the current month can be generated. This process can be repeated over an entire month to update the nowcasts of tourist arrivals in that month. The accuracy of these nowcasts can be investigated to determine whether updated nowcasts with more daily search query data perform better.

Nowcasting performance is plotted against the number of days of search query data added in Fig. 4-8. The X axis represents the number of days of search query data added to generate the nowcasts. Day 0 denotes the forecasting performance of the same model. The horizontal dotted line drawn on the forecasting performance level facilitates the comparison between forecasts and nowcasts.

(Insert Fig. 4 about here)

(Insert Fig. 5 about here)

(Insert Fig. 6 about here)

(Insert Fig. 7 about here)

(Insert Fig. 8 about here)

The nowcasts showed a certain level of fluctuation for all models. Overall, the exponential Almon weighting scheme gave the best results. In addition, most of the points were below the dotted forecasting line of the corresponding colour, which became more apparent as the number of days increased. This suggests that nowcasting generally outperforms forecasting using the MIDAS-SARIMA models, especially when more data become available. The percentage of the nowcasts outperforming the forecasts were calculated for each model, as shown in Table 6. The exponential Almon and beta weighting schemes had more nowcasts that outperformed the forecasts based on all measures. However, the Gompertz weighting scheme had better nowcasts only with respect to the MAPE. Nevertheless, a downward trend was still visible for the Gompertz weighting scheme (as shown in Fig. 4-8), indicating that the nowcasts generally improved as more search query data became available.

(Insert Table 6 about here)

# 6. Conclusion

Search query data are increasingly used to improve the accuracy of tourism demand forecasting. The aim of this study was to investigate the performance of an improved MIDAS model (the

MIDAS-SARIMA model) with the flexibility to accommodate seasonal and non-seasonal ARMA dynamics in predicting monthly tourist arrivals in Hong Kong from mainland China. The results confirmed the superiority of the proposed MIDAS-SARIMA model compared with traditional MIDAS and other benchmark models.

Traditional MIDAS-AR models are ineffective when MA dynamics are involved. The MIDAS-AR models produced similar results to those of the benchmark model and were outperformed by the SARIMAX model. Although MIDAS-AR could better use the valuable information contained in the high frequency data, this advantage was outweighed by the limitation of its structure. The improved forecasting performance of the proposed MIDAS-SARIMA model compared with the MIDAS-AR models is consistent with the findings of Foroni et al. (2019), who demonstrated the relevance of MA components in MIDAS models in forecasting macroeconomic variables. In addition to accommodating MA components, the MIDAS-SARIMA model proposed in this study could incorporate seasonal ARMA components and automatically choose appropriate structures. As a result, the MIDAS-SARIMA model produced the best forecasts and was the only model that could significantly outperform the SARIMA benchmark model, as indicated by the DM tests.

A comparison of forecasts and nowcasts was also conducted. As new search query data became available, their information could be incorporated into the MIDAS models using the mixed frequency structure and daily nowcasts could be generated. The nowcasts outperformed the forecasts most of the time for the exponential Almon and beta weighting schemes. Although the forecasts of the Gompertz weighting scheme outperformed most nowcasts, the scheme

overall showed a downward trend in error measures. Thus, the nowcasts were generally more accurate as more search query data became available.

The results of this study have important implications for research in this area. Search query data have received considerable attention in forecasting tourism demand in recent years. However, using the valuable information contained in these data is problematic and requires appropriate methods. Bangwayo-Skeete and Skeete (2015) were the first to use MIDAS models, which were found to have better forecasting performance than benchmark time series models. However, they did not compare these MIDAS models with models that could also include search query information, such as the SARIMAX model. Therefore, whether the benefits of MIDAS can outweigh the cost of the limitations of its structure is unclear. Indeed, some studies have found no evidence supporting the use of mixed frequency methods (Rivera 2016). The forecasting performance of MIDAS models is likely to be hindered by their inability to accommodate MA dynamics. The improved MIDAS model proposed in this study not only overcame this shortcoming, but also accommodated seasonal dynamics. In addition, the automatic structure determination reduced the risk of misspecification. Overall, the forecasting results confirm the merits of this new model.

The results of this study also have implications for decisionmakers in the tourism sector. Specifically, they confirm the value of search query data in forecasting tourism demand, showing that forecasting accuracy can be further improved using the improved MIDAS model. The benefits of this improved forecasting accuracy are significant, as indicated by the DM tests, whereas the cost is minimal, as the search engine data can often be retrieved for free. Once the model is developed, updating the forecasts and nowcasts is easy. As nowcasts can be generated

daily, they are particularly important for those who need frequent updates of tourism demand forecasts in their day-to-day business operations.

Various possibilities exist for future research. First, only search query data were used in this study. Other forms of big data, such as social media and device data, may be valuable predictors that can improve forecasting accuracy. In addition, the results suggest that the most recent search query data have more forecasting power. Therefore, improved forecasting accuracy may gradually disappear as the forecasting horizon increases. As only short-term forecasts were generated in this study, it would be interesting to see whether the usefulness of search query data wanes as the forecast horizon increases. Finally, more origin and destination pairs should be used to further generalise the results of this study.

# Appendix A

| No. | Search query name | No. | Search query name | No. | Search query name |
|---|---|---|---|---|---|
| | **Dining** | 33 | Hong Kong travel map | 66 | Hong Kong Ocean Park |
| 1 | **Hong Kong food** | 34 | Hong Kong subway price | 67 | Hong Kong Times Square |
| 2 | **Hong Kong snack** | 35 | Hong Kong subway schedule | 68 | Hong Kong Mong Kok |
| 3 | **Hong Kong food tips** | 36 | **Hong Kong airport** | 69 | Hong Kong Causeway Bay |
| 4 | **Hong Kong food recommendation** | 37 | Hong Kong airport express | 70 | Hong Kong Avenue of Stars |
| 5 | Hong Kong specialty | 38 | Hong Kong airport duty free shop | 71 | Hong Kong Victoria Harbour |
| 6 | What are Hong Kong specialties | 39 | Octopus card | 72 | Hong Kong attractions |
| 7 | Hong Kong specialty food | 40 | Citybus | 73 | Madame Tussauds Hong Kong |
| 8 | Macau food | 41 | Futian Port | 74 | Hong Kong Ocean Park tips |
| 9 | Tsui Wah Restaurant | 42 | Huanggang Port | 75 | Hong Kong Ocean Park ticket |
| 10 | Taiwan food | 43 | Kowloon bus | 76 | Hong Kong Disneyland |
| 11 | Hong Kong restaurants | 44 | Luohu Port | 77 | Hong Kong Disneyland Resort |
| | | 45 | Customs clearance time of Luohu Port | 78 | Hong Kong Ocean Park ticket price |
| | **Shopping** | 46 | Shenzhen Bay Port | 79 | Hong Kong Disneyland ticket price |
| 12 | **Hong Kong shopping** | 47 | Hong Kong International Airport | 80 | Hong Kong Disneyland tips |
| 13 | **Hong Kong shopping list** | 48 | Hong Kong Cross-Harbour Tunnel | 81 | Macau tourist attractions |
| 14 | **Hong Kong shopping tips** | 49 | Hong Kong airlines | 82 | Wong Tai Sin |
| 15 | Hong Kong Ladies Market | | | 83 | Hong Kong Wax Museum |
| 16 | What is worth buying in Hong Kong | | **Tours** | 84 | Hong Kong Museum of History |
| 17 | Hong Kong shopping guide | 50 | **Hong Kong travel tips** | 85 | Hong Kong tourist attractions encyclopedia |
| 18 | Hong Kong shopping map | 51 | **Hong Kong self-guided tour tips** | 86 | Hong Kong tourist attractions pictures |
| 19 | Hong Kong travel shopping guide | 52 | **Hong Kong travel guide** | 87 | Hong Kong Jockey Club |
| 20 | Go to Hong Kong shopping tips | 53 | **Hong Kong tourism tips** | 88 | Hong Kong Victoria Harbour night view |
| 21 | Hong Kong Mong Kok shopping tips | 54 | **Hong Kong tourism self-guided tour** | | **Lodging** |
| 22 | Exchange rate of Hong Kong dollar to Chinese yuan | 55 | Hong Kong weather | 89 | **Hong Kong hotels** |
| 23 | Exchange rate of Hong Kong dollar | 56 | Hong Kong weather forecast | 90 | Peninsula Hotel Hong Kong |
| 24 | Hong Kong shopping centers | 57 | Hong Kong one-day trip | 91 | **Hong Kong hotels booking** |
| 25 | Hong Kong cosmetics | 58 | Hong Kong one-day trip tips | 92 | Hong Kong accommodation |
| 26 | Hong Kong airport duty free shops | 59 | Hong Kong tips | 93 | Four Seasons Hotel Hong Kong |
| 27 | Hong Kong duty free shops | 60 | Hong Kong travel agencies | 94 | **Hong Kong hotels recommendation** |
| | | 61 | Hong Kong Observatory | 95 | Hong Kong hotels booking website |
| | **Transportation** | 62 | Hong Kong self-help tour | 96 | Hong Kong hotels group-booking |
| 28 | **Hong Kong map** | 63 | Hong Kong self-guided tour | 97 | L'Hotel Nina et Convention Centre |
| 29 | **Hong Kong subway** | | | 98 | Hong Kong hotels map |
| 30 | Hong Kong subway circuit map | | **Attractions** | 99 | Hong Kong hotels reservation |
| 31 | Hong Kong whole map HD | 64 | **Hong Kong tourist attractions** | 100 | Hong Kong hostels |
| 32 | Hong Kong subway map | 65 | **Hong Kong tourist attractions introduction** | 101 | Hong Kong accommodation guide |

Note: Keywords in bold indicate the initial keywords specified.

# References

Aladag, C. H., E. Egrioglu, U. Yolcu, and V. R. Uslu. 2014. “A high order seasonal fuzzy time series model and application to international tourism demand of Turkey.” Journal of Intelligent and Fuzzy Systems, 26(1): 295–302.

Altissimo, F., R. Cristadoro, M. Forni, M. Lippi, and G. Veronese. 2010. “New Eurocoin: Tracking economic growth in real time.” The Review of Economics and Statistics, 92(4): 1024–34.

Amstad, M., and S. Potter. 2009. “Real time underlying inflation gauges for monetary policymakers.” FRB of New York Staff Report, 420.

Andreou, E., E. Ghysels, and A. Kourtellos. 2013. “Should macroeconomic forecasters use daily financial data and how?” Journal of Business & Economic Statistics, 31(2): 240–51.

Askitas, N., and K. F. Zimmermann. 2009. “Google econometrics and unemployment forecasting.” Applied Economics Quarterly, 55(2): 107–20.

Assaf, A. G., C. P. Barros, and L. A. Gil-Alana. 2011. “Persistence in the short- and long-term tourist arrivals to Australia.” Journal of Travel Research, 50(2): 213–29.

Athanasopoulos, G., R. J. Hyndman, H. Song, and D. C. Wu. 2011. “The tourism forecasting competition.” International Journal of Forecasting, 27(3): 822–44.

Bangwayo-Skeete, P. F., and R. W. Skeete. 2015. “Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach.” Tourism Management, 46: 454–64.

Baumeister, C., P. Guérin, and L. Kilian. 2015. “Do high-frequency financial data help forecast oil prices? The MIDAS touch at work.” International Journal of Forecasting, 31(2): 238–52.

Bordino, I., S. Battiston, G. Caldarelli, M. Cristelli, A. Ukkonen, and I. Weber. 2012. “Web search queries can predict stock market volumes.” PLOS ONE, 7(7): e40014.

Brynjolfsson, E., T. Geva, and S. Reichman. 2014. “Using crowd-based data selection to improve the predictive power of search trend data.” The International Conference on Information Systems (ICIS 2014), Auckland, New Zealand.

Chen, R., C. Y. Liang, W. C. Hong, and D. X. Gu. 2015. “Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm.” Applied Soft Computing, 26: 435–43.

Choi, H., and H. Varian. 2012. “Predicting the present with Google Trends.” Economic Record, 88(1): 2–9.

Chu, F. L. (2009). “Forecasting tourism demand with ARMA-based methods.” Tourism Management, 30(5): 740–51.

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

Clements, M. P., and A. B. Galvão. 2008. “Macroeconomic forecasting with mixed-frequency data.” Journal of Business & Economic Statistics, 26(4): 546–54.

Da, Z., J. Engelberg, and P. Gao. 2011. “In search of attention.” The Journal of Finance, 66(5): 1461–99.

Diebold, F. X., and R. S. Mariano. 1995. “Comparing predictive accuracy.” Journal of Business & Economic Statistics, 13(3): 253–63.

Divino, J. A., and M. McAleer. 2010. “Modelling and forecasting daily international mass tourism to Peru.” Tourism Management, 31(6): 846–54.

Du, R. Y., Y. Hu, and S. Damangir. 2014. “Leveraging trends in online searches for product features in market response modeling.” Journal of Marketing, 79(1): 29–43.

Du, R. Y., and W. A. Kamakura. 2012. “Quantitative trendspotting.” Journal of Marketing Research, 49(4): 514–36.

Fesenmaier, D. R., Z. Xiang, B. Pan, and R. Law. 2011. “A framework of search engine use for travel planning.” Journal of Travel Research, 50(6): 587–601.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin. 2000. “The generalized dynamic-factor model: Identification and estimation.” Review of Economics and Statistics, 82(4): 540–54.

Foroni, C., M. Marcellino, and D. Stevanovic. 2019. “Mixed-frequency models with moving-average components.” Journal of Applied Econometrics, 34(5): 688–706.

Ghose, A., P. G. Ipeirotis, and B. Li. 2014. “Examining the impact of ranking on consumer behavior and search engine revenue.” Management Science, 60(7): 1632–54.

Ghysels, E., P. Santa-Clara, and R. Valkanov. 2004. “The MIDAS touch: Mixed data sampling regression models” (CIRANO Working Paper). CIRANO.

Ghysels, E., A. Sinko, and R. Valkanov. 2007. “MIDAS regressions: Further results and new directions.” Econometric Reviews, 26(1): 53–90.

Ghysels, E., R. I. Valkanov, and A. R. Serrano. 2009. “Multi-period forecasts of volatility: Direct, iterated, and mixed-data approaches.” In EFA 2009 Bergen Meetings Paper.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. 2009. "Detecting influenza epidemics using search engine query data." Nature, 457(7232): 1012–14.

Goh, C. 2012. "Exploring impact of climate on tourism demand." Annals of Tourism Research, 39(4): 1859–83.

Goh, C., and R. Law. 2002. "Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention." Tourism Management, 23(5): 499–510.

Goh, C., R. Law, and H. M. Mok. 2008. "Analyzing and forecasting tourism demand: A rough sets approach." Journal of Travel Research, 46(3): 327–38.

Gunter, U., and I. Önder. 2015. "Forecasting international city tourism demand for Paris: Accuracy of uni- and multivariate models employing monthly data." Tourism Management, 46: 123–35.

Gurgul, H., R. Mestel, and R. Syrek. 2018. "MIDAS models in banking sector—Systemic risk comparison." Managerial Economics, 18(2): 165-81.

Hallin, M., and R. Liška. 2007. "Determining the number of factors in the general dynamic factor model." Journal of the American Statistical Association, 102(478): 603–17.

Hassani, H., E. S. Silva, N. Antonakakis, G. Filis, and R. Gupta. 2017. "Forecasting accuracy evaluation of tourist arrivals." Annals of Tourism Research, 63: 112–27.

Hassani, H., A. Webster, E. S. Silva, and S. Heravi. 2015. "Forecasting U.S. tourist arrivals using optimal singular spectrum analysis." Tourism Management, 46: 322–35.

Hong, W.-C., Y. Dong, L. -Y. Chen, and S.-Y. Wei. 2011. “SVR with hybrid chaotic genetic algorithms for tourism demand forecasting.” Applied Soft Computing, 11(2): 1881–90.

Hyndman, R. J., and Y. Khandakar. 2008. “Automatic time series forecasting: The forecast package for R.” Journal of Statistical Software, 27(3): 1–22.

Kambatla, K., G. Kollias, V. Kumar, and A. Grama. 2014. “Trends in big data analytics.” Journal of Parallel and Distributed Computing, 74(7): 2561–73.

Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, and Y. Shin. 1992. “Testing the null hypothesis of stationarity against the alternative of a unit root.” Journal of Econometrics, 54(1): 159–78.

Law, R., G. Li, D. K. C. Fong, and X. Han. 2019. “Tourism demand forecasting: A deep learning approach.” Annals of Tourism Research, 75: 410–23.

Li, X., and R. Law. 2020. “Forecasting tourism demand with decomposed search cycles.” Journal of Travel Research, 59(1): 52–68.

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

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

Monteforte, L., and G. Moretti. 2012. “Real-time forecasts of inflation: The role of financial variables.” Journal of Forecasting, 32(1): 51–61.

Page, S., H. Song, and D. C. Wu. 2012. “Assessing the impacts of the global economic crisis and swine flu on inbound tourism demand in the United Kingdom.” Journal of Travel Research, 51(2): 142–53.

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

Pan, B., and Y. Yang. 2017. “Forecasting destination weekly hotel occupancy with big data.” Journal of Travel Research, 56(7): 957–70.

Peng, B., H. Song, and G. I. Crouch. 2014. “A meta-analysis of international tourism demand forecasting and implications for practice.” Tourism Management, 45: 181–93.

Rivera, R. 2016. “A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data.” Tourism Management, 57: 12–20.

Silva, E. S., H. Hassani, S. Heravi, and X. Huang. 2019. “Forecasting tourism demand with denoised neural networks.” Annals of Tourism Research, 74: 134–54.

Song, H., W. C. Gartner, and A. D. A. Tasci. 2012. “Visa restrictions and their adverse economic and marketing implications—Evidence from China.” Tourism Management, 33(2): 397–412.

Song, H., and G. Li. 2008. “Tourism demand modelling and forecasting—A review of recent research.” Tourism Management, 29(2): 203–20.

Song, H., G. Li, S. F. Witt, and G. Athanasopoulos. 2011. “Forecasting tourist arrivals using time-varying parameter structural time series models.” International Journal of Forecasting, 27(3): 855–69.

Song, H., S. F. Witt, and T. C. Jensen. 2003. “Tourism forecasting: Accuracy of alternative econometric models.” International Journal of Forecasting, 19(1): 123–41.

Stock, J. H., and M. W. Watson. 2002. “Forecasting using principal components from a large number of predictors.” Journal of the American Statistical Association, 97(460): 1167–79.

Sun, X., W. Sun, J. Wang, Y. Zhang, and Y. Gao. 2016. “Using a Grey–Markov model optimized by Cuckoo search algorithm to forecast the annual foreign tourist arrivals to China.” Tourism Management, 52: 369–79.

Sun, S., Y. Wei, K.-L. Tsui, and S. Wang. 2019. “Forecasting tourist arrivals with machine learning and Internet search index.” Tourism Management, 70: 1–10.

Tourism Commission—Tourism Fact Sheets. 2018. Retrieved August 12, 2019, from https://www.tourism.gov.hk/english/papers/papers_fact_sheets_2018.html.

Turner, L. W., and S. F. Witt. 2001. “Forecasting tourism using univariate and multivariate structural time series models.” Tourism Economics, 7(2): 135–47.

Varian, H. R. 2014. “Big data: New tricks for econometrics.” Journal of Economic Perspectives, 28(2): 3–28.

Vosen, S., and T. Schmidt. 2011. “Forecasting private consumption: Survey-based indicators vs. Google trends.” Journal of Forecasting, 30(6): 565–78.

Wang, X., K. Smith, and R. Hyndman. 2006. “Characteristic-based clustering for time series data.” Data Mining and Knowledge Discovery, 13(3): 335–64.

Wen, L., C. Liu, and H. Song, 2019. “Forecasting tourism demand using search query data: A hybrid modelling approach.” Tourism Economics, 25(3): 309–29.

Wong, K. K. F., H. Song, and K. S. Chon. 2006. “Bayesian models for tourism demand forecasting.” Tourism Management, 27(5): 773–80.

Wu, L., and E. Brynjolfsson. 2015. “The future of prediction: How Google searches foreshadow housing prices and sales.” In Economic Analysis of the Digital Economy, 89–118. University of Chicago Press.

Wu, D. C., H. Song, and S. Shen. 2017. “New developments in tourism and hotel demand modeling and forecasting.” International Journal of Contemporary Hospitality Management, 29(1): 507–29.

Yahya, N. A., R. Samsudin, and A. Shabri. 2017. “Tourism forecasting using hybrid modified empirical mode decomposition and neural network.” International Journal of Advances in Soft Computing and its Applications, 9(1): 14–31.

Yang, X., B. Pan, J. A. Evans, and B. Lv. 2015. “Forecasting Chinese tourist volume with search engine data.” Tourism Management, 46: 386–97.

Zhang, G., J. Wu, B. Pan, J. Li, M. Ma, M. Zhang, and J. Wang. 2017. “Improving daily occupancy forecasting accuracy for hotels based on EEMD-ARIMA model.” Tourism Economics, 23(7): 1496–514.

**Table 1.** Results of the SARIMA and SARIMAX models.

| SARIMA | | SARIMAX | |
|---|---|---|---|
| Non-seasonal differencing | Yes | Non-seasonal differencing | Yes |
| Seasonal differencing | Yes | Seasonal differencing | Yes |
| AR(1) | 0.2651 | AR(1) | -0.8780*** |
| AR(2) | -0.5819*** | AR(2) | -0.5241*** |
| MA(1) | -1.0745*** | SMA(1) | -0.5468*** |
| MA(2) | 1.2188*** | $index_{t-1}$ | 0.2122*** |
| MA(3) | -0.4372** | | |
| SMA(1) | -0.6750*** | | |
| Residual variance | 0.00387 | Residual variance | 0.00356 |
| AIC | -150.22 | AIC | -157.74 |
| BIC | -135.45 | BIC | -147.27 |
| Ljung-Box test | 0.35 | Ljung-Box test | 0.70 |

Note: ***, ** and * indicate that the estimates are significant at the 1%, 5% and 10% levels, respectively.

**Table 2.** Results of the MIDAS-AR and MIDAS-SARIMA models.

| MIDAS-AR-Almon | | MIDAS-AR-Beta | | MIDAS-AR-Gom | | MIDAS-SARIMA-Almon | | MIDAS-SARIMA-Beta | | MIDAS-SARIMA-Gom | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Non-seasonal differencing | Yes | Non-seasonal differencing | Yes | Non-seasonal differencing | Yes | Non-seasonal differencing | Yes | Non-seasonal differencing | Yes | Non-seasonal differencing | Yes |
| Seasonal differencing | Yes | Seasonal differencing | Yes | Seasonal differencing | Yes | Seasonal differencing | Yes | Seasonal differencing | Yes | Seasonal differencing | Yes |
| $\beta_1$ | 0.2342** | $\beta_1$ | 0.2349** | $\beta_1$ | 0.2376** | $\beta_1$ | 0.2932** | $\beta_1$ | 0.2788** | $\beta_1$ | 0.2881** |
| $\theta_1$ | -0.1196 | $\theta_1$ | 0.9607*** | $\theta_1$ | 1.9965 | $\theta_1$ | 0.2809 | $\theta_1$ | 1.1262 | $\theta_1$ | 0.7226 |
| $\theta_2$ | 0.0026 | $\theta_2$ | 1.0779 | $\theta_2$ | 0.5920 | $\theta_2$ | -0.0281 | $\theta_2$ | 8.1216 | $\theta_2$ | 3.9867 |
| AR(1) | -0.9096*** | AR(1) | -0.9274*** | AR(1) | -0.9141*** | MA(1) | -0.6464*** | MA(1) | -0.6559*** | MA(1) | -0.6465*** |
| AR(2) | -0.5265*** | AR(2) | -0.5348*** | AR(2) | -0.5353*** | SMA(1) | -0.5540*** | SMA(1) | -0.5077*** | SMA(1) | -0.5437*** |
| AIC | -139.10 | AIC | -138.91 | AIC | -138.65 | AIC | -145.96 | AIC | -145.86 | AIC | -146.21 |
| BIC | -128.80 | BIC | -128.60 | BIC | -128.35 | BIC | -133.39 | BIC | -133.30 | BIC | -133.64 |
| Ljung-Box test | 0.29 | Ljung-Box test | 0.31 | Ljung-Box test | 0.27 | Ljung-Box test | 0.38 | Ljung-Box test | 0.53 | Ljung-Box test | 0.48 |

Note: ***, ** and * indicate that the estimates are significant at the 1%, 5% and 10% levels, respectively.

**Table 3.** Evaluation of the one-step-ahead dynamic forecasts.

| Measure | MAD | MSE | MAPE | RMSPE | Theil's U |
|---|---|---|---|---|---|
| Seasonal Naïve | 288,795 | 1.77E+11 | 7.30% | 10.01% | 1 |
| ETS | 286,297 | 1.24E+11 | 7.34% | 8.70% | 0.837 |
| SARIMA | 290,781 | 1.45E+11 | 7.46% | 9.33% | 0.904 |
| SARIMAX | 268,737 | 1.26E+11 | 6.94% | 8.83% | 0.844 |
| MIDAS-AR-Almon | 285,371 | 1.54E+11 | 7.31% | 9.63% | 0.932 |
| MIDAS-AR-Beta | 285,183 | 1.54E+11 | 7.30% | 9.62% | 0.933 |
| MIDAS-AR-Gom | 280,302 | 1.52E+11 | 7.19% | 9.58% | 0.927 |
| MIDAS-SARIMA-Almon | **248,653** | 1.08E+11 | **6.34%** | **8.08%** | 0.78 |
| MIDAS-SARIMA-Beta | 255,198 | 1.17E+11 | 6.58% | 8.52% | 0.814 |
| MIDAS-SARIMA-Gom | 250,081 | **1.06E+11** | 6.47% | 8.16% | **0.775** |

Note: Figures in bold indicate the best forecasting performance for each measure.

**Table 4.** DM test statistics for SARIMA.

| Measure | AE | SE | APE | SPE |
|---|---|---|---|---|
| SARIMAX | -0.935 | -1.083 | -0.804 | -0.818 |
| MIDAS-AR-Almon | -0.179 | 0.370 | -0.182 | 0.359 |
| MIDAS-AR-Beta | -0.186 | 0.384 | -0.201 | 0.358 |
| MIDAS-AR-Gom | -0.334 | 0.296 | -0.317 | 0.301 |
| MIDAS-SARIMA-Almon | -2.107** | -1.484* | -2.260** | -1.669** |
| MIDAS-SARIMA-Beta | -2.101** | -1.431* | -1.902** | -1.334* |
| MIDAS-SARIMA-Gom | -2.081** | -1.328* | -2.084** | -1.338* |

Note: ***, ** and * indicate that the estimates are significant at the 1%, 5% and 10% levels, respectively.

**Table 5.** DM test statistics for ETS.

| Measure | AE | SE | APE | SPE |
|---|---|---|---|---|
| SARIMAX | -0.757 | 0.189 | -0.629 | 0.271 |
| MIDAS-AR-Almon | -0.023 | 0.712 | -0.025 | 0.693 |
| MIDAS-AR-Beta | -0.029 | 0.719 | -0.041 | 0.695 |
| MIDAS-AR-Gom | -0.148 | 0.669 | -0.138 | 0.655 |
| MIDAS-SARIMA-Almon | -2.213** | -1.578* | -2.209** | -1.656** |
| MIDAS-SARIMA-Beta | -1.323 | -0.510 | -1.193 | -0.325 |
| MIDAS-SARIMA-Gom | -1.812** | -1.219 | -1.628* | -1.012 |

Note: ***, ** and * indicate that the estimates are significant at the 1%, 5% and 10% levels, respectively.

**Table 6.** Percentage of nowcasts outperforming forecasts.

| Measure | MAD | MSE | MAPE | RMSPE | Theil's U |
|---|---|---|---|---|---|
| MIDAS-SARIMA-Almon | 0.73 | 0.83 | 0.60 | 0.90 | 0.83 |
| MIDAS-SARIMA-Beta | 0.63 | 0.87 | 0.53 | 0.80 | 0.87 |
| MIDAS-SARIMA-Gom | 0.50 | 0.33 | 0.63 | 0.37 | 0.33 |

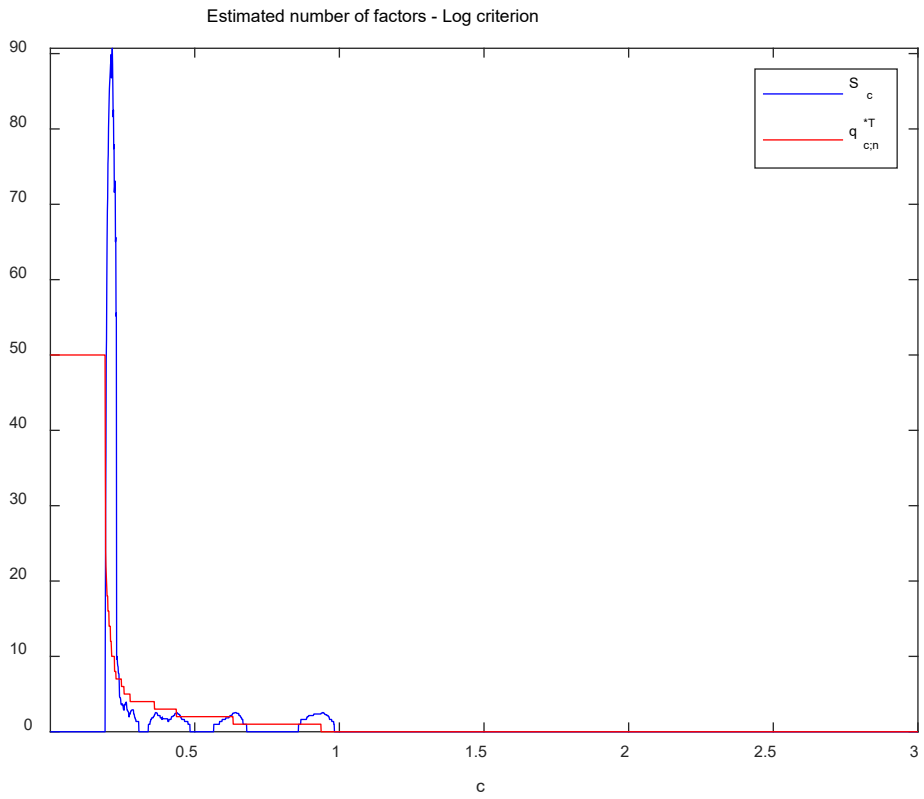


**Fig. 1.** Log criterion for factor selection.

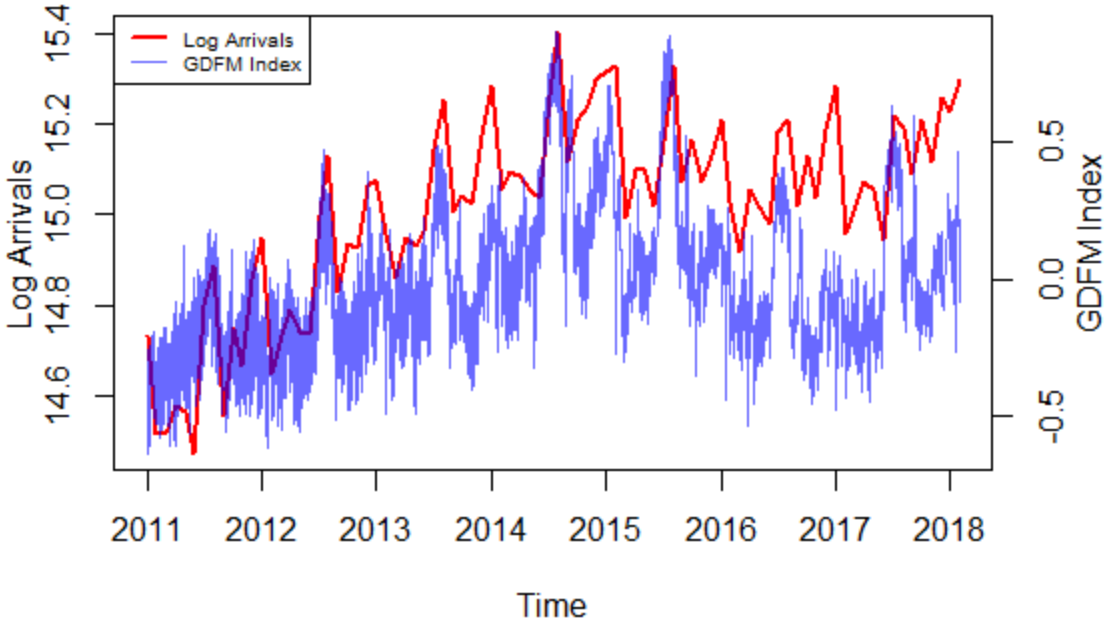


**Fig. 2.** Daily index and log of tourist arrivals.

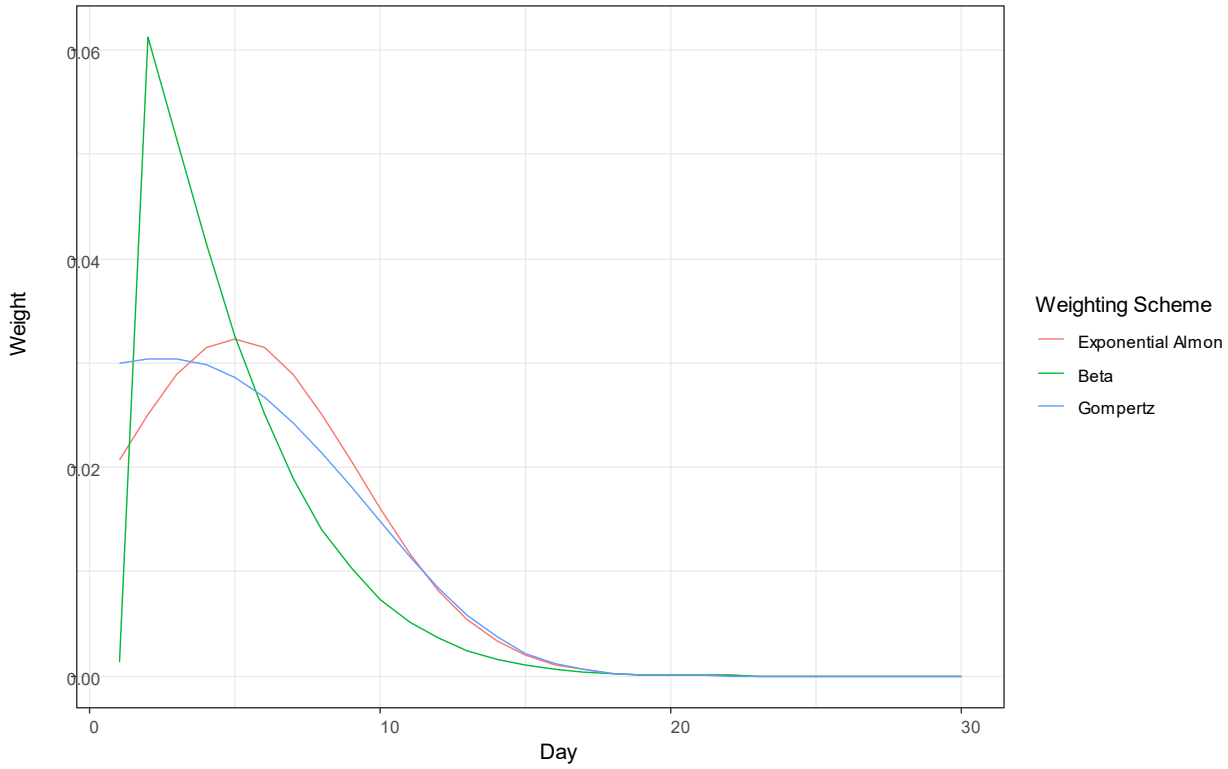


**Fig. 3.** Different weighting schemes of the daily indices for the MIDAS-SARIMA models. The x-axis represents the number of days preceding the first day of the following month.

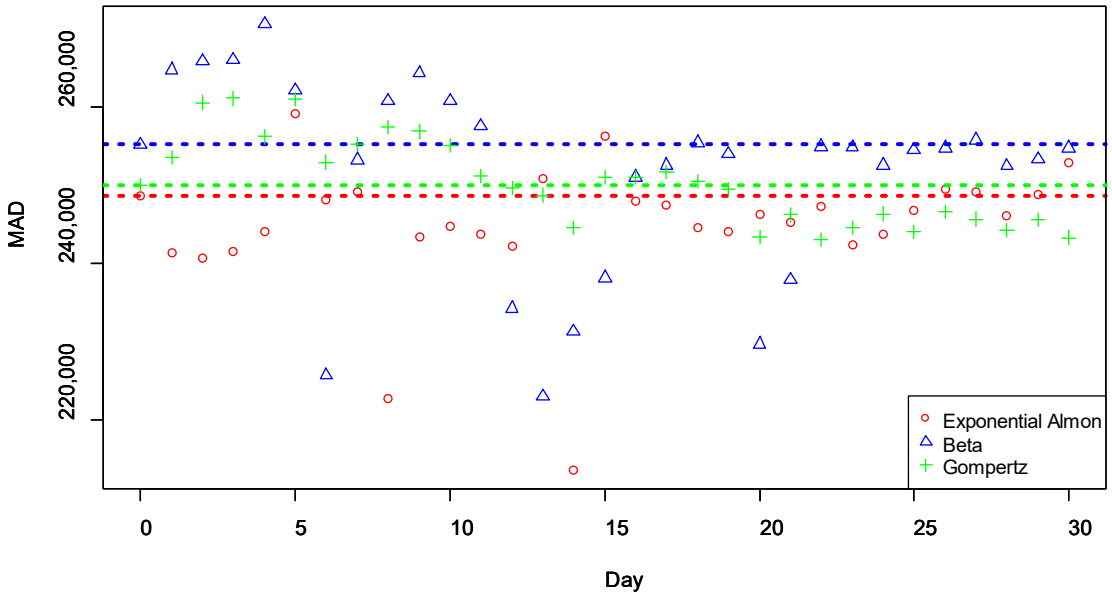


**Fig. 4.** MAD of nowcasting for the MIDAS-SARIMA models with different weighting schemes.

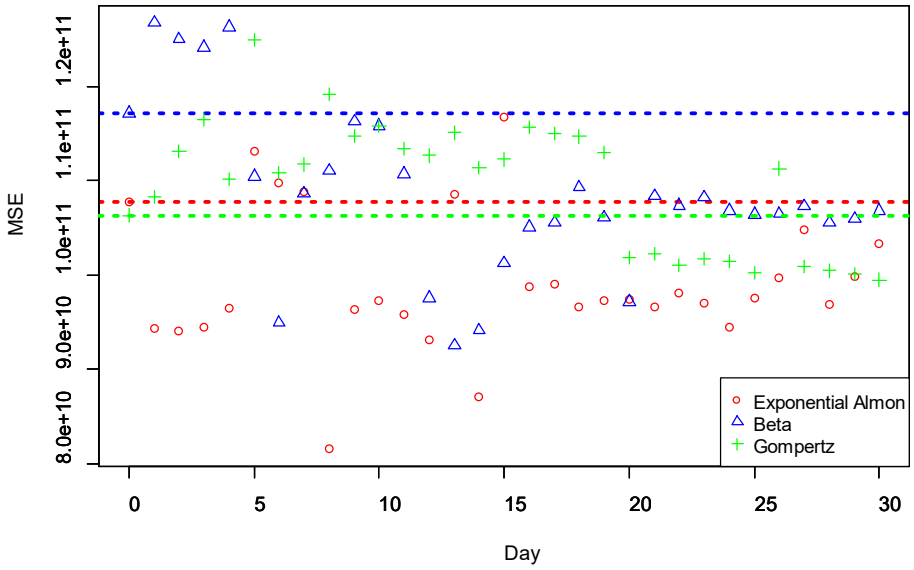


**Fig. 5.** MSE of nowcasting for the MIDAS-SARIMA models with different weighting schemes.

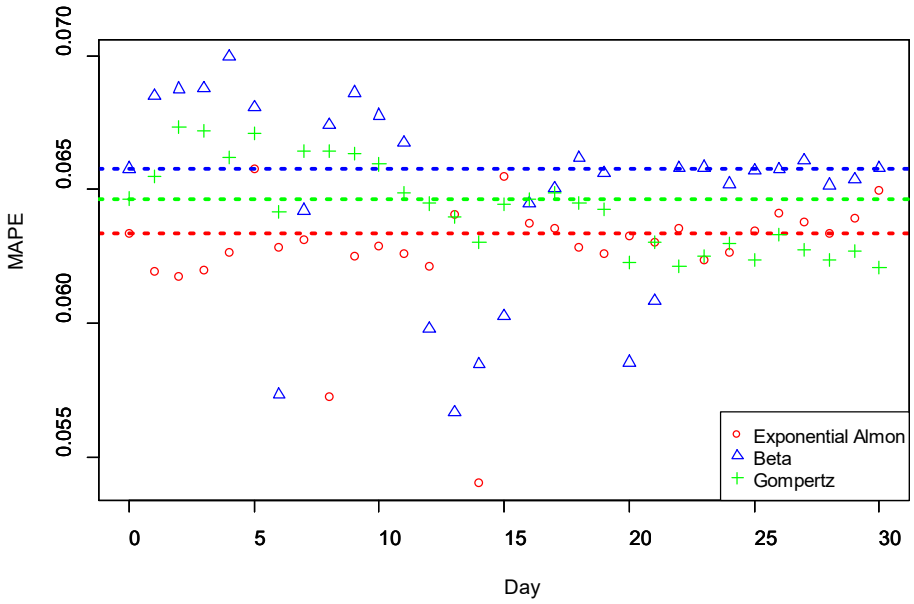


**Fig. 6.** MAPE of nowcasting for the MIDAS-SARIMA models with different weighting schemes.

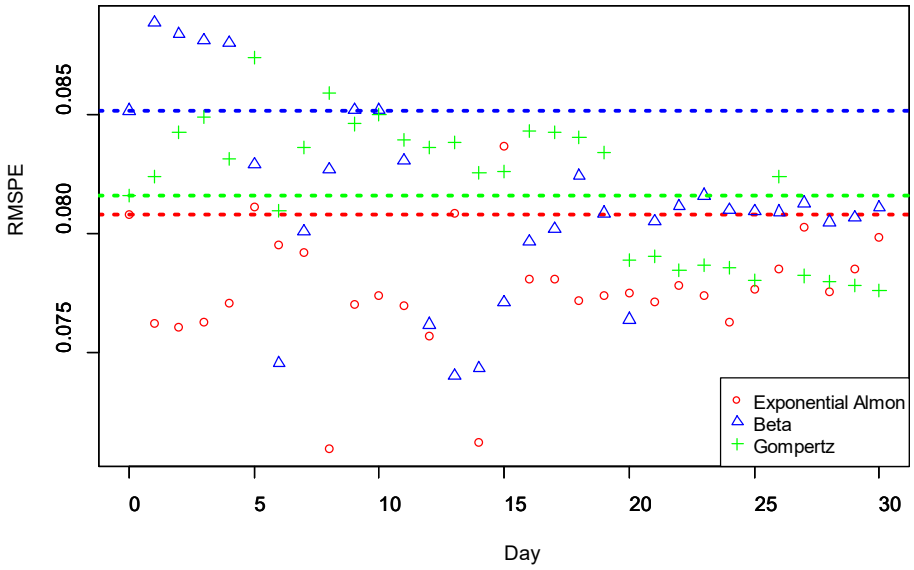


**Fig. 7.** RMSPE of nowcasting for the MIDAS-SARIMA models with different weighting schemes.

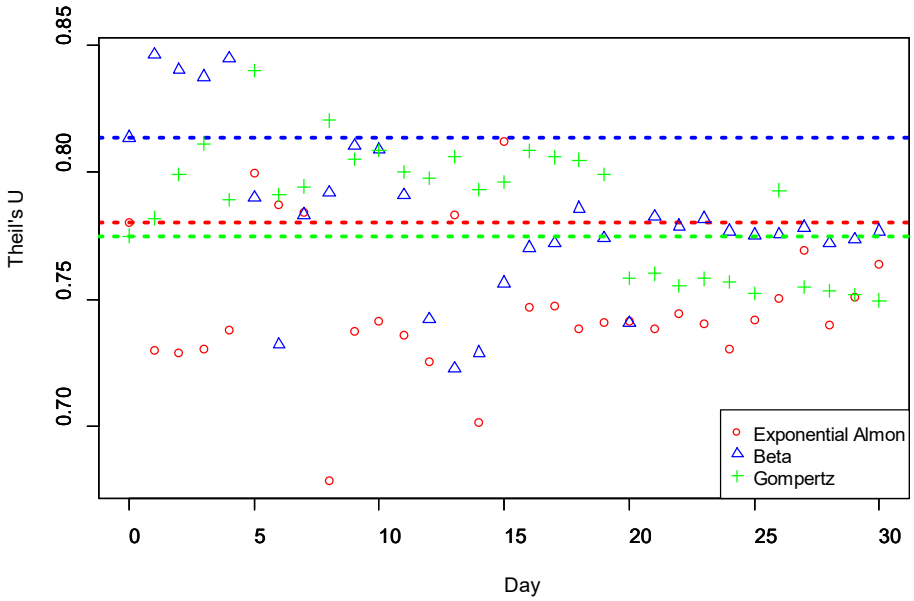


**Fig. 8.** Theil's U of nowcasting for the MIDAS-SARIMA models with different weighting schemes.