Tourism Management 84 (2021) 104263

Contents lists available at ScienceDirect

Tourism Management

journal homepage: http://www.elsevier.com/locate/tourman

Hierarchical pattern recognition for tourism demand forecasting

Mingming Hu[a,b], Richard T.R. Qiu[c], Doris Chenguang Wu[d,*], Haiyan Song[b]

[a] Business School, Guangxi University Nanning, China
[b] Hospitality and Tourism Research Centre, School of Hotel and Tourism Management, The Hong Kong Polytechnic University, Kowloon, China
[c] Department of Integrated Resort and Tourism Management, Faculty of Business Administration, University of Macau, Taipa, China
[d] Business School, Sun Yat-sen University, Guangzhou, China

ARTICLE INFO

Keywords:
Tourism demand forecasting
Hierarchical pattern recognition
Calendar pattern
Tourism demand pattern
Floating holidays
Daily attraction visits

ABSTRACT

This study proposes a hierarchical pattern recognition method for tourism demand forecasting. The hierarchy consists of three tiers: the first tier recognizes the calendar pattern of tourism demand, identifying work days and holidays and integrating “floating holidays.” The second tier recognizes the tourism demand pattern in the data stream for different calendar pattern groups. The third tier generates forecasts of future tourism demand. Evidence from daily tourist visits to three attractions in China shows that the proposed method is effective in forecasting daily tourism demand. Moreover, the treatment of “floating holidays” turns out to be more effective and flexible than the commonly adopted dummy variable approach.

1. Introduction

In the past decade, the tourism industry has made a major contribution to the global economy. According to UNWTO (2019), the total volume of international tourism exports exceeded US$ 1.7 trillion in 2018, after nine consecutive years of sustained growth, accounting for 7% of global exports and 29% of global service exports. The significant role of the tourism industry in today’s economy suggests the need for comprehensive and accurate analysis of tourism demand. In recent years, tourism forecasting has received more attention from the industry (Li & Wu, 2019). From an aggregate perspective, accurate tourism forecasts can assist governments and destination management offices in strategic management, regulation, and infrastructure investment. At a disaggregate level, accurate forecast of visitor flows to a tourism site can provide tourism-related businesses with useful information for pricing, operation strategy, and crowd control (e.g. Chen, Bloomfield, & Fu, 2003).

Forecasts of tourism demand are primarily based on the analysis of consecutive measurements of tourism demand and other relevant evidence (see Song, Qiu, and Park (2019) for a comprehensive review). The frequency of tourism demand series can be as low as annually or as high as daily. Forecasting annual tourism demand provides interested parties with general trends and cycles, whereas the analysis of high-frequency data can be useful for precision marketing and management. Among the features investigated in the tourism demand forecast literature, seasonality stands out as a distinctive characteristic of tourism demand, with a major impact on tourism businesses’ planning and operation (Chen, Li, Wu, & Shen, 2019). Tourism demand tends to follow regular patterns, such as low and peak seasons at the quarterly/monthly level and weekends and holidays at the daily level.

The analysis of seasonality becomes more complicated when the frequency of data increases due to the interlocking of different seasonality patterns at various frequencies. In addition, as argued by Hyndman (2013), it is difficult to treat “moving events” such as Easter or Chinese New Year when working with daily data. These holidays do not occur at the same time every year and may include weekends. In other words, these holidays “float” on the calendar, leading to irregular seasonality patterns in daily tourism demand data. Since enjoying holidays is perceived as one of the major factors driving travelers to choose a destination (Heung, Qu, & Chu, 2001), the irregular seasonality patterns would have a significant impact on the accuracy of tourism demand forecasts. Although a dummy variable approach has been proposed to treat moving holidays (Hyndman, 2013), Enders and Li (2015) question its validity due to its subjectivity and inflexibility.

This study proposes an innovative hierarchical pattern recognition (HPR) approach for forecasting daily tourism demand for attractions. High frequency forecasting, such as daily, and high frequency seasonality identification are important for the regular and routine operation

* Corresponding author.
E-mail addresses: mingming.hu@gxu.edu.cn (M. Hu), richardqiu@um.edu.mo (R.T.R. Qiu), wucheng@mail.sysu.edu.cn (D.C. Wu), haiyan.song@polyu.edu.hk (H. Song).

https://doi.org/10.1016/j.tourman.2020.104263
Received 26 January 2020; Received in revised form 22 September 2020; Accepted 14 November 2020
Available online 14 December 2020

and management of attractions. For example, the short-run forecasting provides the attraction management office with relevant information to guide their human resource allocation and promotion strategies. The proposed method includes three tiers in which both floating holidays and the overlapping effect of different levels of seasonality are considered. The first tier recognizes the calendar patterns of the data stream and identifies work days and holidays (including weekends and floating holidays). The data stream is classified into different groups according to the calendar pattern. This tier provides the model with the flexibility to deal with floating holidays that occur irregularly in the calendar. The second tier recognizes the patterns of tourism demand trends at various seasonal levels in the historical pattern pool within the same calendar pattern group. This tier generates tourism demand forecasts at different seasonal levels and for different calendar patterns. The third tier collates the forecasts generated in the second hierarchy and produces the final forecast of tourism demand. The forecast generated in this tier integrates patterns from interlocking seasons and holidays. The proposed approach is tested with daily tourist arrival data from three famous attractions in China (Jiuzhaigou Valley, Kulangsu, and Siguniang Mountain). Six benchmark models (weekly naïve, simple *k*-NN, SARIMA, SARIMAX, ETS, and TBATS) are used for forecast performance evaluation and comparison. Forecasting horizons from one to 14 days ahead are evaluated separately, and MAPE and MASE are used to measure accuracy.

The rest of this study is structured as follows. Section 2 reviews the related literature on tourism demand forecasting and pattern recognition. Section 3 explains the proposed HPR approach in detail, followed by a brief introduction to the benchmark methods. Data and forecasting results are reported in Section 4, and the accuracy of the HPR approach is compared with the benchmarks. Section 5 concludes the study with a discussion of limitations and future directions.

2. Literature review

2.1. Tourism demand forecasting

Tourism products, such as unused hotel rooms and unoccupied airline seats, are impossible to stockpile. This means that tourism demand forecasting is essential for the government and industry. Indeed, an extensive tourism demand forecasting literature has developed in the past few decades. A literature review by Song et al. (2019) shows that from 1968 to 2018, over 600 studies investigated the modeling and forecasting of tourism demand, and their results have been adopted by governments, destination management organizations (DMOs), and industrial practitioners for managerial and strategic purposes. The majority of tourism demand forecast studies adopt either one or several methods for prediction, such as time series models, econometric methods, AI-based techniques, and subjective approaches.

Univariate time series methods extrapolate historical tourism demand series to generate forecasts (Chu, 2009; Wu, Song, & Shen, 2017). It is well known that tourism demand series can be associated with long and short memory series reflecting cyclical and seasonal changes in tourist behavior (Gil-Alana, 2005; Karlaftis & Vlahogianni, 2009; Gil-Alana, Mudida, & de Gracia, 2014; Sadaei, Guimarães, da Silva, Lee, & Eslami, 2017). Predictions of future tourism demand can be generated according to established memory patterns and data frequency in existing tourism demand series. The most frequently adopted univariate time series models are the no-change model (Naïve I), constant growth rate model (Naïve II), exponential smoothing (ES) model, Box-Jenkins models (autoregressive moving average (ARMA) family models), and structural time series (STS) models (Song et al., 2019; Wu et al., 2017). Univariate time series models have been found to provide reliable forecasts of tourism demand over the years. However, the impact of economic variables on tourism demand cannot be captured by univariate time series models.

Multivariate econometric methods complement univariate time series models by incorporating causal variables (Onafowora & Owoye, 2012; Song & Li, 2008; Wu et al., 2017). In addition to forecasting, econometric methods illustrate the relationship between tourism demand and causal variables, such as tourist income, tourism prices, substitute prices, exchange rates, transportation costs, marketing expenses, and climate change (Dritsakis, 2004; Goh, 2012; Law, 2000, 2001; Law & Au, 1999; Li, Goh, Hung, & Chen, 2018; Li, Song, & Li, 2017; Li, Song, & Witt, 2005; Lim, 1999; Lise & Tol, 2002; Song & Li, 2008; Wu, Cao, Wen, & Song, 2020). Due to their ability to develop econometric systems that link tourism demand and key economic factors, and to generate reliable forecasts, econometric methods have become popular in the tourism demand literature in recent decades (Song et al., 2019).

Due to the exponential development in computing technology, AI-based techniques have received substantial attention in various scientific disciplines. In the tourism demand forecast context, AI-based techniques aim to establish non-linear connections between tourism demand, its lagged values, and other explanatory variables (Claveria, Monte, & Torra, 2015; Kon & Turner, 2005; Law, 2000; Law & Au, 1999; Palmer, Montano, & Sesé, 2006). Artificial neural networks (Law, 2000; Law & Au, 1999), support vector regression models (Chen & Wang, 2007), and Gaussian> process regression (Wu, Law, & Xu, 2012) are typical AI-based techniques found in the tourism demand forecasting literature. AI-based techniques are sometimes referred to as a “black box” (Zhang, Patuwo, & Hu, 1998), due to the lack of theoretical foundation for the estimation process. Nevertheless, the demand for high-accuracy forecasting has made AI-based techniques very popular in the tourism demand forecast literature since 2000 (Song et al., 2019).

Subjective forecasts of tourism demand are usually generated on the basis of experts’ experience and opinions. This approach was adopted quite often before 1990 (Song et al. 2019), but has lost popularity due to the development of computational techniques and the evolution of time series and econometric models. More recently, it has been shown that judgmental approaches can be combined with other quantitative forecasting models to improve forecast accuracy (Lin, Goodwin, & Song, 2014; Tideswell, Mules, & Faulkner, 2001).

In the tourism demand forecasting literature, regardless of the method adopted, seasonality is acknowledged as a key feature of tourism demand series (Song & Li, 2008). Seasonality refers to the phenomenon in which tourist flows tend to have a similar pattern for the same period across different years. The description of seasonality relates to the frequency of the data stream. Low and peak seasons are usually captured at monthly or quarterly frequency, whereas the influence of weekends and specific holidays can only be revealed at daily frequency. The treatment of seasonality in daily data can be tricky, as a long enough daily data stream can incorporate not only popularity fluctuations from one season to another, but also the calendar patterns of weekends and holidays. Hyndman (2013) further highlights that floating holidays pose difficulties in treating daily data. Holidays such as Thanksgiving, Easter, and Chinese New Year appear at different times in the calendar each year, and actual vacations may also include weekends around these floating holidays. Such “floating holidays” exhibit more complicated and irregular seasonality patterns. In time series analysis, the dummy variable approach is the most common and effective way of dealing with floating holidays (Hyndman, 2013). However, Enders and Li (2015) specify three conditions for the validity of the dummy variable approach: (1) the date of the event is exogenous and known; (2) the impact of the event on the trend is immediate; (3) the event occurs at a single point in time and has a unique impact on the trend. In the context of tourism demand analysis, the validity of these three conditions is questionable. First, as stated earlier, although such holidays have known and exact dates on the calendar, the dates float around in the calendar. Second, the impact of holidays on tourist flows is not immediate. Quite often, a bell-shaped pattern emerges in tourism demand data around holidays (Kirillova & Lehto, 2015). Lastly, the impact of floating holidays interlocks with trends in low or peak seasons, which offsets or amplifies their influence. Therefore, the dummy variable approach may not be ideal for dealing with floating holidays in tourism forecasting.

2.2. Pattern recognition

Patterns are an important subject in many disciplines, such as biology, psychology, medicine, marketing, computer vision, artificial intelligence, and remote sensing. Watanabe (1985) defines a pattern as the "opposite of a chaos; it is an entity, vaguely defined, that could be given a name." A pattern can be a curve, a point in multidimensional space, a fingerprint image, a signature, or a human face. In past decades, pattern recognition techniques have advanced dramatically due to the availability of large databases, increased calculation speed and estimation accuracy, and reduced cost of data management (Jain, Duin, & Mao, 2000). This technique has been frequently adopted in numerous fields, including the analysis of DNA sequences (Brubaker, Bonham, Zanoni, & Kagan, 2015; Harding et al., 2017); the recognition of faces (Fagan, 2017), fingerprints (Jain, Arora, Cao, Best-Rowden, & Bhatnagar, 2016), and signatures (Galbally et al., 2015); the recognition of speech (Afouras, Chung, Senior, Vinyals, & Zisserman, 2018), images (Zoph, Vasudevan, Shlens, & Le, 2018), and characters and documents (Fujisawa, 2008); and the forecast of future values such as stock market trends (Liu & Kwong, 2007), electricity prices (Lora, Santos, Expósito, Ramos, & Santos, 2007), and solar radiation (Ghofrani, Azimi, & Youshi, 2019). The general process of pattern recognition involves three steps: data acquisition, pattern extraction, and pattern classification (Asht & Dass, 2012). The acquisition step retrieves the data and converts it into a structural format which is acceptable to computing devices. The pattern extraction step analyzes the converted structural data and extracts the potential patterns from the data. The pattern classification step categorizes the extracted patterns and applies the patterns in practice.

Throughout the applications of pattern recognition, various methods exist for the extraction and classification of patterns. These methods have advantages and disadvantages in specific contexts (see Asht and Dass (2012) for a methodological review). Despite the diversity of fields that use pattern recognition, there is a general trend for patterns to be represented as multiple features or measurements and as points in a multi-dimensional space (Jain et al., 2000). The essence of pattern recognition is classifying or categorizing points in this multi-dimensional space. There are two main types of classification in the literature: supervised classification, in which input patterns are identified as members of predefined classes, and unsupervised classification (e.g., clustering), in which patterns are assigned to hitherto unknown classes (Watanabe, 1985). The nearest neighbor (NN) algorithm (Cover & Hart, 1967) is a generic supervised classification method, which is used to determine the similarity between a target point and stored points in multi-dimensional space and to predict the characteristics of the target point according to several NN points (Huang, Lin, Huang, & Xing, 2017). The NN algorithm is frequently used in the forecasting literature, and multiple (k) neighbors are usually considered in the model. Lora et al. (2007) utilize a weighted-NN algorithm to forecast electricity prices in Spain. In this algorithm, the number of nearest neighbors and the window length of neighbors are determined by the minimization of training set forecast error, and forecasts of electricity prices are determined by the linear combination of the next-day prices of the chosen neighbors. Shelke and Thakare (2014) adopt a k-NN algorithm to classify daily electricity load data in India between 2012 and 2013 and generate load forecasts from the classified data using the Holt-Winters model. Cai et al. (2016) refine the k-NN algorithm by incorporating spatiotemporal correlations in a multistep model and generate forecasts of short-term traffic.

In the tourism demand forecasting context, Kamel, Atiya, El Gayar, and El-Shishiny (2008) investigate the accuracy of traditional forecast techniques and machine learning methods using annual tourist arrivals in Hong Kong, and reveal the reasonable predictive power of the k-NN algorithm, with a performance just below that of the generalized regression neural network model. Höpken, Ernesti, Fuchs, Kronenberg, and Lexhagen (2017) identify the superior performance of the k-NN model compared with linear regression in the forecast of monthly tourist arrivals at a Swedish mountain destination. Díaz and Mateu-Sbert (2011) provide point and sign forecasts of daily airport arrivals in Mallorca, using several types of k-NN model, and these algorithms demonstrate satisfactory forecasting ability. Following Díaz and Mateu-Sbert (2011), Olmedo (2016) confirms the nonlinear dynamics of daily airport arrivals in Mallorca, and validates the use of k-NN models in point and sign forecasts. Deviating from these preceding studies, where the dimension parameters and number of neighbors are optimized within the model, Rice, Park, Pan, and Newman (2019) set the embedding dimension according to the seasonality cycle (12 for monthly data) and the number of neighbors to three in their forecast of monthly campsite demand in US national parks. Four forecasting methods (moving average, Holt-Winters exponential smoothing, seasonal ARIMA, and neural network autoregression) are used for performance comparison. The results show, however, that the k-NN algorithm provides merely fair predictions, and is outperformed by seasonal ARIMA and exponential smoothing in three-, six-, and twelve-month-ahead forecasts (Rice et al., 2019). It is evident that the application of pattern recognition in tourism forecasting is quite rare and limited to monthly or annual in frequency. Furthermore, pattern recognition does not show obvious superior forecasting ability in tourism demand, in contrast to the reliable outstanding performance revealed in studies in other fields.

In the context of daily tourism demand analysis, floating holidays are an important factor in influencing the performance of the forecast. However, the investigation of floating holidays, especially at daily frequency, has thus far received little academic attention. This study therefore contributes to the tourism forecasting literature in four ways. First, an innovative pattern recognition method, hierarchical pattern recognition, is proposed for daily tourist arrival forecasting. Second, "floating holiday" patterns are captured using the k-NN algorithm to describe multiple and complicated seasonality characteristics in holiday-sensitive tourism demand series. Third, two ensemble steps are adopted in the forecasting process with one ensembles forecasts from the nearest patterns and the other ensembles forecasts from different time window lengths. Lastly, different forecasting horizons, from one to 14 days ahead, are examined separately to verify the forecasting ability of the proposed method.

3. Methodology

A hierarchical pattern recognition forecasting method is proposed in this study to forecast daily tourism demand. There are three hierarchies in this method (Fig. 1). In Tier 1, calendar pattern of the current date is recognized and compared with all historical data. Historical data with same the calendar pattern are pulled out in preparation for the next tier. In this tier, both the regular calendar patterns and the floating holiday patterns are integrated into the algorithm. In Tier 2, multiple time windows are utilized to construct tourism demand pattern of the current date. In each time window, the k-NN algorithm is used to identify two nearest neighbors from the pool prepared in Tier 1. In Tier 3, in each time window, the observations of the two nearest neighbors are linearly combined to generate one forecast value. The forecast values in all time windows are then ensembled to generate the final forecast.

In this section, the k-NN algorithm is firstly introduced as the basis of

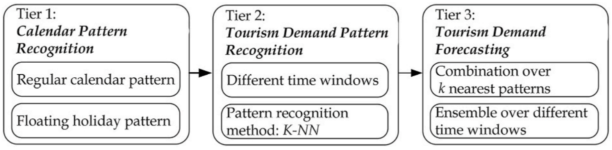
Tier 1:
Calendar Pattern Recognition
Regular calendar pattern
Floating holiday pattern
Tier 2:
Tourism Demand Pattern Recognition
Different time windows
Pattern recognition method: K-NN
Tier 3:
Tourism Demand Forecasting
Combination over k nearest patterns
Ensemble over different time windows

Fig. 1. Framework of hierarchical pattern recognition forecasting method.

the proposed method. The HPR forecasting method is then discussed. Benchmark models and forecasting accuracy evaluation methods are introduced at the end of the section.

3.1. *k*-Nearest neighbor algorithm

The *k*-NN algorithm is the proposed method for recognizing patterns. According to the *k*-NN algorithm, a tourist arrival series of time length *m* can be considered as a tourism demand pattern of time window *m* and be treated as one point in *m*-dimensional space. By rolling the time window on the tourist arrival series, multiple points can be extracted. Taking the point with the newest date as the current point and all other points as historical points, the distances between the current point and historical points in *m*-dimensional space can be measured to determine the similarity of the current tourism demand pattern to historical tourism demand patterns. The *k* neighbors with the highest similarity (the smallest distances) are selected and the future tourism demand values of these neighbors collectively generate the forecast of tourism demand. Fig. 2 illustrates the framework of the *k*-NN algorithm.

In Fig. 2, $a(t)$ indicates the tourism demand at time *t*. The length of the time window is denoted by *m*, and $\{a(t\text{-}m+1), \ldots, a(t\text{-}1), a(t)\}$ represents the current pattern for time *t*, denoted as $A(t,m)$. All of the patterns with the same length in the data stream, $A(t\text{-}1,m)$ to $A(m,m)$, are considered as the historical patterns. Neighbors are searched among the historical patterns, and *k* patterns with highest similarity to the current pattern are chosen as the nearest neighbors of the current pattern $A(t,m)$. The future tourism demand at time $t+1$ can be derived by the linear combination of these nearest neighbors.

Similarity determination. With a given time window *m*, similarity between the current pattern, $A(t,m)$, and historical pattern, $A(h,m)$, can be determined by the Euclidean distance between the points in *m*-dimensional space (Díaz & Mateu-Sbert, 2011; Lora et al., 2007; Rice et al., 2019). Due to the temporal trend feature of tourism demand data, however, direct measurement of the Euclidean distance between the current and historical patterns could be misleading. Thus, relative Euclidean distance (RED) is adopted to diminish the effect of trend when measuring the similarity between patterns. In particular, the current pattern is detrended as follows:

$$\varphi(A(t,m)) \equiv \{[a(t-m+1) - L(A(t,m))], \ldots, [a(t) - L(A(t,m))]\}$$

where the trend term, $L(A(t,m))$, is defined by

$$L(A(t,m)) \equiv \frac{[a(t-m+1) + \cdots + a(t-1) + a(t)]}{m}$$

The RED between the current pattern $A(t,m)$ and historical pattern $A(h,m)$ is then calculated by:

$$RED(A(t,m), A(h,m)) \equiv \| \varphi(A(t,m)) - \varphi(A(h,m)) \|$$

and the similarity between the current pattern $A(t,m)$ and historical

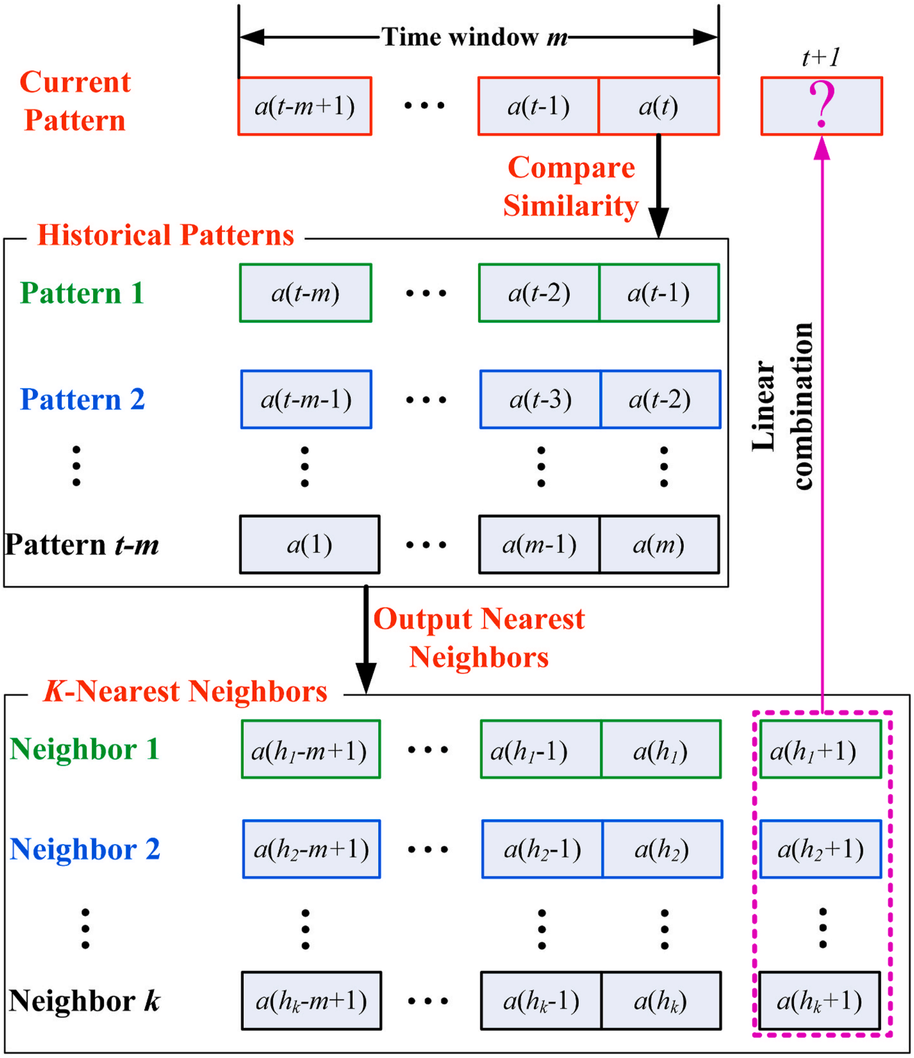

Fig. 2. The framework of the *k*-NN algorithm.

pattern $A(h,m)$ is computed as:

$$S(A(t,m),A(h,m)) = e^{-RED(A(t,m),A(h,m))}$$

Forecast generation. Neighbors of the current pattern can be chosen by evaluating the similarity, $S(A(t,m),A(h,m))$, between the current pattern and all historical patterns. The value of k is set to 2 in this study following Höpken et al. (2017); that is, the two nearest neighbors are selected and combined to generate the forecast of future tourism demand. In particular, the forecast of tourism demand at time $t+1$ is determined by the weighted arithmetic operator (Xu & Da, 2003), which is the weighted average of the future values of the two nearest neighbors adjusted by their trend values:

$$\hat{a}(t+1,m) = w(h_1,h_2)[a(h_1+1) + L(A(t,m)) - L(A(h_1,m))] + (1 - w(h_1,h_2))[a(h_2+1) + L(A(t,m)) - L(A(h_2,m))]$$

where $A(h_1,m)$ and $A(h_2,m)$ are the two nearest neighbor patterns of the current pattern $A(t,m)$. The weighting, $w(h_1,h_2)$, follows the formulation of relative weighting in information science and decision-making literature (e.g. Hu, Ren, Lan, Wang, & Zheng, 2014; Lan, Hu, Ye, & Sun, 2012; Lan, Zou, & Hu, 2020; Xu, 2015), and is calculated by the similarity:

$$w(h_1,h_2) = \frac{e^{-RED(A(t,m),A(h_1,m))}}{e^{-RED(A(t,m),A(h_1,m))} + e^{-RED(A(t,m),A(h_2,m))}}$$

3.2. Hierarchical pattern recognition forecasting method

The proposed HPR forecasting model has three tiers. In the first tier, the calendar patterns of the historical data are recognized and stored. This tier helps to identify work days and holidays in the data stream so that work days, regular weekends, and floating holidays can be treated accordingly in the forecasting process. The second tier recognizes the tourism demand pattern of work days and holidays (both weekends and floating holidays) in the data stream and generates forecasts of future tourism demand. The third tier integrates the future values of the nearest neighbors with different time windows and generates the forecast value.

Calendar pattern recognition. Daily tourism demand is significantly affected by holidays, including weekends and floating holidays. The patterns of tourism demand during holidays are different from those during weekdays. Therefore, identifying holidays (both weekends and floating holidays) is essential for the proposed HPR method. In the current method, work days are denoted as "1" and holidays (including weekends and floating holidays) are denoted as "0." Because a typical week consists of five work days and two weekend days, a consecutive five-day calendar pattern is always considered in the proposed method. Because travel planning relates to not only the previous and current situations, but also the upcoming schedule, the calendar pattern of the target date consists of the work day/holiday status for the two days before and two days after the target. A sample calendar pattern for forecasting the tourism demand of the first day of a three-day holiday is shown in Fig. 3.

Here in Fig. 3, to forecast the tourism demand at time $t+1$, the calendar pattern (1,1,0,0,0) is examined. The pattern recognition process is only conducted on historical tourism demand patterns with the same calendar pattern, such as (h_1-2, h_1-1, h_1, h_1+1, h_1+2) and (h_2-2, h_2-1, h_2, h_2+1, h_2+2) in Fig. 3.

The calendar pattern recognition tier incorporates the floating holiday feature into the proposed method. Since the calendar pattern of the current date is compared with those of all historical data, the patterns may not be selected at a regular time interval. This relaxes the assumption of seasonality in time-series analysis, where the seasons occur regularly throughout the year. In addition, together with the different time windows adopted in the later tiers of the hierarchy, the proposed method facilitates the analysis of the overlapping effect of different levels of seasonality. By recognizing calendar patterns, the proposed method explicitly takes the time constraints of the tourists into consideration and integrates this information into the forecasting process.

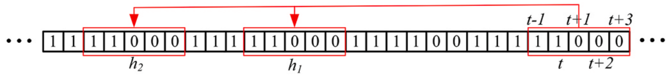

Fig. 3. Sample of calendar pattern recognition.

Setting time windows. Typically, only one time window is selected in the k-NN algorithm to reflect the most influential seasonality cycle (e.g. Díaz & Mateu-Sbert, 2011; Olmedo, 2016; Rice et al., 2019). However, as mentioned earlier, daily tourism demand may be influenced by seasonality with different frequencies. In the proposed method, time windows from 2 days up to 28 days ($m = 2$ to 28) are jointly considered to account for interlocking seasonality at different frequencies. The forecasts of different time windows, $\hat{a}(t+1,m)$, are aggregated with reciprocal weightings to give the final forecast of tourism demand for day $t+1$:

$$\hat{a}(t+1) = \sum_{m=2}^{28} w(m)\hat{a}(t+1,m)$$

with

$$w(m) = \frac{\frac{1}{m}}{\sum_{s=2}^{28}\frac{1}{s}}, m = 2,3,...,28$$

The complete framework of the proposed HPR method is given in Fig. 4.

3.3. Benchmark models

Six models are used as benchmarks to evaluate the forecast performance of the proposed method. These include four time series models: the seasonal naïve model, seasonal autoregressive integrated moving average (SARIMA) model, exponential smoothing model (ETS), and exponential smoothing state space model with Box-Cox transformation, ARMA errors, trend, and seasonal components (TBATS). The other two are a time series model with intervention (SARIMA with explanatory variable, SARIMAX) and pattern recognition model (simple k-NN).

3.3.1. Seasonal naïve

In a seasonal naïve model, future forecasts are simply equal to the most recent available value in the corresponding season (Athanasopoulos, Hyndman, Song, & Wu, 2011). For weekly data, $\widehat{y}_t = y_{t-s}$, where y is the tourism demand, $\widehat{y}_t$ is its forecast, t refers to time, and s is 7.

3.3.2. Seasonal auto-regressive integrated moving average (SARIMA)

The SARIMA model belongs to the ARMA family originally proposed by Box and Jenkins (1970). It integrates an autoregressive (AR) component, a moving-average (MA) component, and seasonality into one model and has gained popularity in recent years (Song et al., 2019). A general SARIMA model is specified as $ARIMA(p,d,q)*(P,D,Q)_S$, where p is non-seasonal AR order, d is non-seasonal differencing, q is non-seasonal MA order, P is seasonal AR order, D is seasonal differencing, Q is seasonal MA order, and S is the time cycle of the seasonal pattern. For daily tourism demand forecasting, we set $S = 7$. The auto.arima() function of the "forecast" package (Hyndman et al., 2019) in the R software is utilized, which uses the Box-Jenkins approach to determine the optimized parameters.

3.3.3. SARIMAX model

As an extension of the SARIMA model, the SARIMAX model further includes exogenous variables in the modeling process. In this study, the exogenous variables included are dummy variables for holidays. The

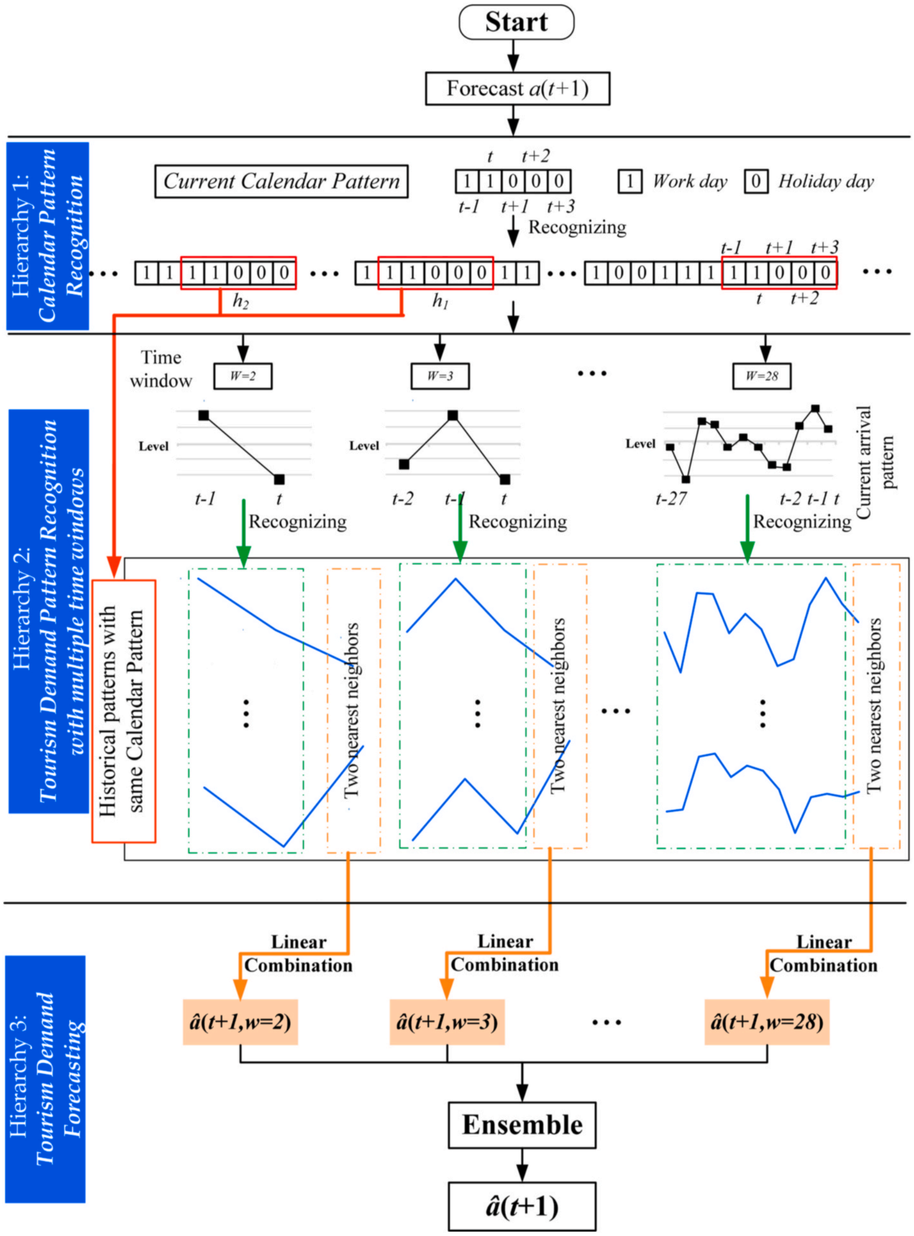

Fig. 4. The framework of the hierarchical pattern recognition method.

SARIMAX model is examined here because it has been noted that augmenting time series models with additional explanatory variables often improves forecasting accuracy (Wu et al., 2017).

3.3.4. *Exponential smoothing (ETS)*

ETS uses the average of past observations with exponentially decreased weights (Hyndman & Athanasopoulos, 2018). Holt (2004) first incorporated trend into simple exponential smoothing. The ETS method further develops this approach by including level, trend, seasonality, and smoothing (Hyndman & Athanasopoulos, 2018). The Holt-Winter method has two formats: additive and multiplicative. The additive model assumes a constant degree of seasonality; the multiplicative method assumes seasonality involves a multiplicative relationship between the trend, seasonality, and irregularity in the series. The general ETS model can be written as *ETS (e, t, s)*, where *e* denotes the error type ("A," "M," or "Z") with "A" for additive error, "M" for multiplicative error, and "Z" for automatically selecting the type of error. *t* denotes the trend type ("N," "A," "M," or "Z") with "N" for no trend, "A" for additive trend, "M" for multiplicative trend, and "Z" for automatically selecting the type of trend. *s* denotes the seasonality type ("N," "A," "M," or "Z") with "N" for no seasonality, "A" for additive seasonality, "M" for multiplicative seasonality, and "Z" for automatically selecting the type of seasonality. The ets() function of the "forecast" package (Hyndman et al., 2019) in the R software is utilized to determine the parameters automatically.

3.3.5. *Exponential smoothing state space model with Box-Cox transformation, ARMA errors, trend, and seasonal components (TBATS)*

De Livera, Hyndman, and Snyder (2011) extend the exponential smoothing method and develop an exponential smoothing state space model with Box-Cox transformation, ARMA errors, trends, and seasonal components (TBATS) to forecast high frequency time series with complex seasonal patterns, such as daily or higher frequencies. After setting the seasonality of time series with the msts() function in the "forecast" package (Hyndman et al., 2019) in the R software, the TBATS() function can be used to fit the figure automatically. Hyndman (2013) states that when the time series is long enough to take in more than a year, it is necessary to allow for annual and weekly seasonality. Thus, both 7 days and 365.25 days are taken as the periods of daily tourism demand.

3.3.6. *Simple k-NN*

To further examine the effectiveness of the calendar pattern recognition tier of the proposed method, a classical pattern recognition model, simple *k*-NN, is also included as a benchmark model. The parameter settings of the simple *k*-NN algorithm follow the same settings adopted in the proposed HPR method, with *k* set to 2 and the nearest neighbors linearly combined. The calculation of the simple *k*-NN algorithm follows the description in Section 3.1 and is done in the R software with kknn() in the Weighted k-Nearest Neighbors Classification and Clustering (kknn) package (Schliep, Hechenbichler, & Lizee, 2016). Table 1 summarizes the variables, optimization criteria, and estimation methods adopted in the benchmark models and the proposed HPR method.

3.4. *Forecasting accuracy assessment*

To evaluate the forecast accuracy of the proposed HPR method and the benchmark models, the mean absolute percentage error (MAPE) and mean absolute scaled error (MASE) are calculated:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|a_i - \widehat{a}_i\right|}{a_i}$$

$$MASE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|a_i - \widehat{a}_i\right|}{\frac{1}{T-1}\sum_{t=2}^{T}\left|a_t - a_{t-1}\right|}$$

where $\widehat{a}_i$ and a_i denote the forecast value and actual value of tourism demand, respectively. Smaller values of assessment measures indicate better forecasting performance of the associated models. Other forecast accuracy measurements, such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and root mean squared percentage error (RMSPE), are also available upon request.

4. Data description

This study examines visitor arrivals for three famous tourism attractions in China: these are two 5A tourist attractions, Jiuzhaigou Valley in Sichuan province and Kulangsu in Fujian province, and one 4A tourist attraction, Siguniang Mountain in Sichuan province. Daily tourist arrival data are obtained from the management office of the attractions or their official websites. Public holidays and weekends are identified according to notices from the Chinese government.

Table 1
Models, variables, optimization criteria and estimation method.

Model	Variables	Optimization criteria	Estimation method
Seasonal Naïve	Daily arrivals	NA	Modeling weekly seasonality characteristics
SARIMA	Daily arrivals	Minimum AICc	Maximum likelihood
SARIMAX	Daily arrivals; Holiday dummy	Minimum AICc	Maximum likelihood
ETS	Daily arrivals	Minimum AICc	Maximum likelihood
TBATS	Daily arrivals	Minimum AIC	Maximum likelihood
k-NN	Daily arrivals	Minimum Distance	combination over nearest patterns
HPR	Daily arrivals; Holiday dummy	Minimum Distance	combination over nearest patterns; ensemble over time windows

Based on data availability, daily data for these attractions are collected for the periods of June 1, 2012–June 30, 2017; July 1, 2017–June 30, 2019; and September 25, 2015–November 30, 2018, respectively. Fig. 5 shows the time series plots for visitor arrivals at these three attractions, and significant seasonality characteristics can be observed in these time series.

For forecasting accuracy comparison, all data for each tourist attraction are divided into two parts: a training dataset and test dataset. The most recent half-year of data is used as a test dataset for forecasting accuracy evaluation, and the remainder is used as a training dataset for modeling. For each attraction, 14 forecasting horizons are examined and compared separately from one day ahead to two weeks ahead using increasing rolling windows.

5. Empirical results

5.1. *Tourism demand patterns of the three attractions*

Fig. 6 presents three major tourism demand patterns extracted from the data of each tourism attraction: weekend (two days holiday), long weekend (three days holiday), and the golden week (seven days holiday). The red round dots represent holidays and the blue square dots represent work days. It should be noted that the tourism demand patterns in Fig. 6 are based on detrended arrivals hence negative numbers exist.

The tourism demand patterns of weekends are quite similar in all three tourism attractions, with arrivals peaking on Saturday. This reflects the phenomenon that many tourists leave their homes on Friday after work and start the holiday right away on Saturday. On Sunday, however, due to work obligations on Monday, tourists may participate in less intense activities and prepare back home.

Regarding long weekends and golden weeks, similar patterns are observed in that the volume of tourist arrivals decreases as the end of the holiday draws near. However, in contrast to weekend patterns, the volume of tourist arrivals is not instantly high at the beginning of the

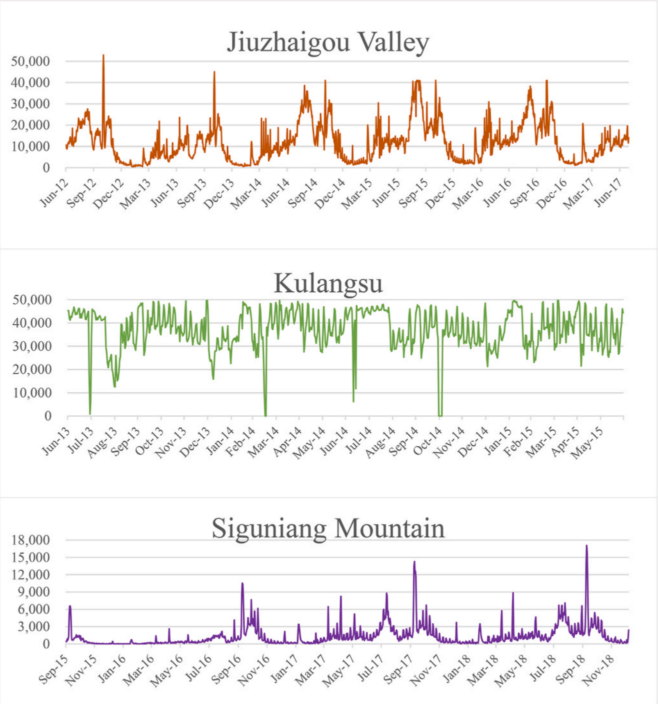

Fig. 5. Time series plots of daily visitor arrivals.

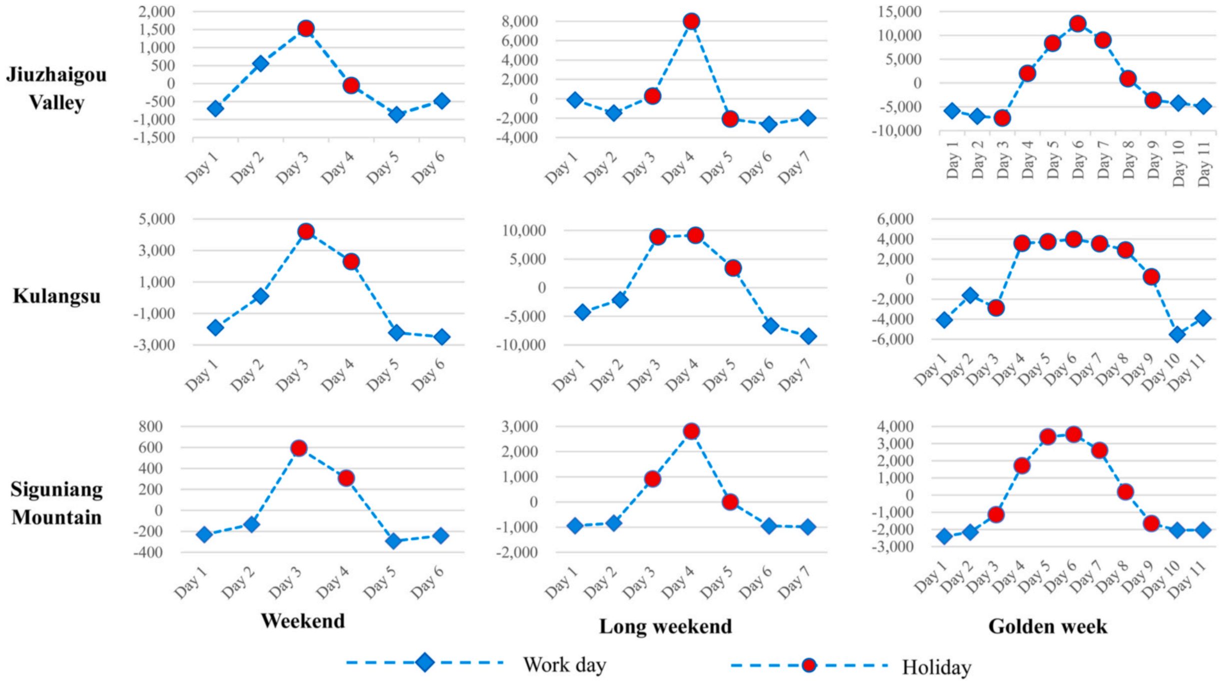

Fig. 6. Tourism demand pattern in different kinds of holidays.

holiday. For Jiuzhaigou Valley and Siguniang Mountain, it takes around two days for the tourist arrivals to reach their peaks, whereas the increase takes around one day for Kulangsu. This may be due to the fact that as the length of the holiday increases, tourists can enjoy a more relaxed journey to the attractions instead of rushing to them. Nonetheless, some differences are identified in the tourism demand patterns of Kulangsu in comparison with those in Jiuzhaigou Valley and Siguniang Mountain. In Kulangsu, the volume of tourism arrivals reaches the peak level on the second day and remains relatively constant throughout the remainder of the holiday. In contrast, the tourism demand patterns in Jiuzhaigou Valley and Siguniang Mountain fit more of a bell-shaped curve. This observation can be explained by two features of Kulangsu in terms of transportation and hotel arrangements. Kulangsu has easier access by airplane or high-speed rail, whereas Jiuzhaigou Valley and Siguniang Mountain require long distance coach travel after a flight or rail journey. In addition, Kulangsu has hotels within the attraction, whereas in Jiuzhaigou Valley and Siguniang Mountain, tourists have to stay outside of the attraction areas.

The above tourism demand patterns extracted from the data can assist the understanding of characteristics of the tourism attractions. The attraction management offices can utilize these patterns to generate accurate forecasts on tourist arrivals and, together with other social and economic concerns, to establish effective strategies for operations and marketing.

5.2. Daily tourism demand forecast of three attractions

Table 2 shows the point forecasting performance for the Jiuzhaigou Valley tourism attraction. Both MAPEs and MASEs indicate consistent conclusions. Among the six benchmarks, it can be generally observed that the seasonal naïve model and *k*-NN algorithm give the poorest forecasts with the average MAPEs of 0.4540 and 0.4837 respectively and average MASEs of 1.7411 and 2.8689 respectively. In addition, as we expected, SARIMA with holiday dummies consistently outperforms SARIMA without holiday dummies, with average MAPEs of 0.3914 and 0.4119 respectively and average MASEs of 1.4423 and 1.5064 respectively. TBATS is the most accurate among the six benchmark models, with average MAPEs of 0.3265 and average MASEs of 1.2588. This may

Table 2
Forecasting performance for Jiuzhaigou Valley.

	MAPE							MASE						
Forecast horizon	Weekly Naïve	*k*-NN	SARIMA	SARIMAX	ETS	TBATS	HPR	Weekly Naïve	*k*-NN	SARIMA	SARIMAX	ETS	TBATS	HPR
1	0.4003	0.2141	0.2065	0.1972	0.1784	0.1729	**0.1482**	1.5559	1.5922	0.7966	0.7705	0.7351	0.7134	**0.5096**
2	0.3975	0.2641	0.2642	0.2423	0.2445	0.2184	**0.1880**	1.5519	1.9757	0.9856	0.9251	0.9944	0.8608	**0.6600**
3	0.3975	0.2783	0.3114	0.2868	0.2906	0.2618	**0.2000**	1.5539	2.1052	1.1496	1.0860	1.1550	1.0056	**0.7335**
4	0.3931	0.2891	0.3407	0.3206	0.3084	0.2871	**0.2321**	1.5507	2.2140	1.2496	1.2037	1.2493	1.1004	**0.8252**
5	0.3883	0.3552	0.3629	0.3482	0.3281	0.3106	**0.2532**	1.5482	2.5177	1.3546	1.3041	1.3507	1.1882	**0.8862**
6	0.3834	0.3722	0.3836	0.3736	0.3592	0.3246	**0.2704**	1.5465	2.5890	1.4442	1.3964	1.4688	1.2453	**0.9690**
7	0.3808	0.4091	0.4124	0.3997	0.4052	0.3369	**0.2934**	1.5455	2.6991	1.5329	1.4822	1.5685	1.2918	**1.0387**
8	0.5220	0.4768	0.4421	0.4287	0.4360	0.3549	**0.3087**	1.9042	2.9008	1.6052	1.5515	1.6272	1.3408	**1.1111**
9	0.5209	0.5293	0.4705	0.4471	0.4724	0.3693	**0.3139**	1.9199	3.1571	1.7001	1.6170	1.7315	1.4107	**1.1415**
10	0.5205	0.5989	0.4852	0.4657	0.4861	0.3800	**0.3269**	1.9403	3.3705	1.7440	1.6862	1.7499	1.4645	**1.1586**
11	0.5157	0.6602	0.5067	0.4822	0.5276	0.3848	**0.3344**	1.9379	3.5410	1.8215	1.7459	1.8901	1.4816	**1.1992**
12	0.5135	0.7106	0.5191	0.4928	0.5460	0.3889	**0.3537**	1.9377	3.7343	1.8563	1.7794	1.9379	1.5007	**1.2598**
13	0.5122	0.7856	0.5259	0.4977	0.5563	0.3907	**0.3533**	1.9415	3.8157	1.9053	1.8156	1.9790	1.5043	**1.2945**
14	0.5098	0.8277	0.5355	0.4976	0.5546	0.3908	**0.3656**	1.9411	3.9520	1.9441	1.8289	1.9949	1.5152	**1.3429**
Average	*0.4540*	*0.4837*	*0.4119*	*0.3914*	*0.4067*	*0.3265*	***0.2815***	*1.7411*	*2.8689*	*1.5064*	*1.4423*	*1.5309*	*1.2588*	***1.0093***

be attributed to its incorporation of ARMA errors, trends, and seasonal components in one model.

More importantly, it can be observed that the proposed pattern recognition method consistently outperforms all six benchmark models for all forecasting horizons from one day ahead to 14 days ahead when both MAPE and MASE are considered. The average MAPE over all horizons of the proposed pattern recognition method is 0.2815, whereas the average MAPEs of the six benchmark models vary from 0.3265 to 0.4837. Similarly, the average MASE over all horizons of the proposed pattern recognition method is 1.0093, whereas the average MASEs of the six benchmark models vary from 1.2588 to 2.8689. This provides strong evidence that the proposed method is superior to the six benchmark models and is an effective method of high frequency tourism demand forecasting.

Tables 3 and 4 report the forecasting performance for the other two tourism attractions, Kulangsu and Siguniang Mountain. Generally all models perform best for Kulangsu, and perform worst for Siguniang Mountain. Taking one-day-ahead forecasting as an example, the MAPEs vary from 0.0914 to 0.1787 for Kulangsu, from 0.1482 to 0.4003 for Jiuzhaigou Valley, and from 0.2468 to 0.5618 for Siguniang Mountain.

When comparing models for all three attractions, the empirical results indicate consistent findings, which confirms the robustness of our findings for daily tourism attraction demand forecasting. Our general observations for the three attractions are as follows: on average over all considered horizons, the seasonal naïve model and simple *k*-NN algorithm perform the worst among all models; the SARIMA model augmented by holiday dummies outperforms the original SARIMA model; the proposed HPR method outperforms the six benchmark models, indicating the effectiveness of this method for high frequency tourism demand forecasting.

In the comparisons between SARIMAX and SARIMA, and between HPR and *k*-NN, the methods including the considerations of holidays (SARIMAX and HPR) outperform those models without such considerations. This observation further confirms the important role of holidays in daily tourism demand forecasting and validates the use of holiday patterns in the proposed HPR.

When different forecasting horizons are further examined and compared, it is interesting to note that though our proposed method beats all benchmark models for all horizons in the case of Jiuzhaigou Valley, it does not perform the best consistently over all horizons in the cases of Kulangsu and Siguniang Mountain. Particularly, in the case of Kulangsu (in Table 3), for horizons of 1–9, the proposed HPR method is the most accurate model according to both MAPE and MASE. However, for horizons of 10–14, the MAPEs show that SARIMAX outperforms the proposed HPR method and is the most accurate model. The MASEs also show that for horizons of 10–14, the most accurate model is SARIMAX, followed by TBATS, with HPR ranking third. For Siguniang Mountain in Table 4, according to MASE, the proposed HPR method is the most accurate for all horizons. However, according to MAPE, the proposed HPR method is the most accurate for horizons of 1–11, and is the second best for horizons 12–14. In particular, it is outperformed by ETS for the horizons of 12 and 13, and outperformed by seasonal naïve for the horizon of 14.

We therefore conclude our observations as follows: firstly, the proposed HPR method performs the best on average over all forecasting horizons of 1–14; secondly, when each horizon is examined separately, the proposed HPR method is very promising and beats all benchmark models for three tourist attractions for the short-term forecasting from one-day-ahead to nine-day-ahead; Thirdly, when longer horizons from 10-day-ahead to 14-day-ahead are considered, the proposed HPR method still produces the best forecasts outperforming other models in 17 out of 30 cases, followed by SARIMAX (10 cases), ETS (2 cases), and weekly Naïve model (one case). These observations indicate that (1) recognizing holiday patterns plays an important role in improving daily tourism demand forecasting; (2) the proposed HPR method provides robust and promising forecasts on high frequency tourism demand data, in which the calendar patterns of the holidays can be reflected in the time-series; (3) our proposed method is especially effective for short-run forecasting practice. The above results are validated in terms of both construct validity and predictive validity (Armstrong, 2001). The construct validity refers to the validation of input variables, such as the calendar patterns of the holidays and the lagged values of tourism demand. In the present study, the calendar patterns of the holidays are verified by comparing the specifications of SARIMAX, SARIMA, HPR and *k*-NN, and the inclusion of the lagged values of tourism demand is the basic norm of time series analysis. The predictive validity refers to the accuracy of the forecasts, which is verified by the forecast performance comparisons between the proposed method and the benchmark models over 14 forecast horizons and across three tourist attractions.

6. Conclusion

Due to floating holidays and the multiple seasonality of daily tourism demand, daily tourism forecasting remains challenging. This study proposes an innovative pattern recognition method for daily tourism demand forecasting. The daily tourist arrivals from three tourism attractions in China are used separately for empirical validation. Forecasting horizons from one to 14 days ahead are examined. The empirical results show that the proposed technique is consistently superior to the six benchmark methods of seasonal naïve, conventional *k*-NN algorithm, SARIMA, SARIMAX, ETS, and TBATS in short-run forecasting. Therefore, the proposed hierarchical pattern recognition method is a

Table 3
Forecasting performance for Kulangsu.

	MAPE							MASE						
Forecast horizon	Weekly Naïve	*k*-NN	SARIMA	SARIMAX	ETS	TBATS	HPR	Weekly Naïve	k-NN	SARIMA	SARIMAX	ETS	TBATS	HPR
1	0.1787	0.1112	0.1052	0.0983	0.1003	0.1029	**0.0914**	1.6450	0.8603	0.9805	0.9020	0.9335	0.9512	**0.8486**
2	0.1789	0.1408	0.1494	0.1305	0.1433	0.1388	**0.1091**	1.6448	1.0951	1.3765	1.1924	1.3195	1.2635	**1.0102**
3	0.1785	0.1538	0.1615	0.1424	0.1590	0.1521	**0.1212**	1.6454	1.1861	1.4890	1.3041	1.4612	1.3755	**1.1260**
4	0.1762	0.1735	0.1671	0.1482	0.1741	0.1592	**0.1317**	1.6343	1.3490	1.5579	1.3720	1.6295	1.4565	**1.2343**
5	0.1756	0.1755	0.1667	0.1483	0.1844	0.1589	**0.1328**	1.6307	1.3629	1.5624	1.3775	1.7261	1.4604	**1.2515**
6	0.1755	0.1871	0.1657	0.1487	0.1825	0.1560	**0.1349**	1.6316	1.4514	1.5525	1.3797	1.7009	1.4278	**1.2795**
7	0.1732	0.2037	0.1633	0.1461	0.1740	0.1539	**0.1369**	1.6169	1.5931	1.5322	1.3602	1.6205	1.4179	**1.3109**
8	0.2198	0.2035	0.1688	0.1499	0.1871	0.1619	**0.1442**	2.0795	1.5879	1.5828	1.3949	1.7419	1.4927	**1.3766**
9	0.2187	0.2082	0.1737	0.1531	0.1978	0.1637	**0.1483**	2.0716	1.6191	1.6321	1.4288	1.8453	1.5041	**1.4148**
10	0.2185	0.2309	0.1739	**0.1528**	0.2020	0.1611	0.1577	2.0708	1.7922	1.6395	**1.4302**	1.9023	1.4868	1.5020
11	0.2193	0.2291	0.1734	**0.1528**	0.2095	0.1645	0.1670	2.0806	1.7966	1.6398	**1.4347**	1.9856	1.5333	1.5836
12	0.2186	0.2227	0.1724	**0.1521**	0.2180	0.1680	0.1666	2.0757	1.7368	1.6332	**1.4314**	2.0721	1.5574	1.5961
13	0.2177	0.2309	0.1716	**0.1507**	0.2143	0.1650	0.1700	2.0683	1.8011	1.6303	**1.4223**	2.0384	1.5273	1.6357
14	0.2156	0.2192	0.1713	**0.1499**	0.2184	0.1644	0.1771	2.0500	1.7113	1.6291	**1.4165**	2.0669	1.5214	1.7086
Average	*0.1975*	*0.1921*	*0.1631*	*0.1446*	*0.1832*	*0.1550*	***0.1421***	*1.8532*	*1.4959*	*1.5313*	***1.3462***	*1.7174*	*1.4268*	*1.3485*

Table 4
Forecasting performance for Siguniang Mountain.

	MAPE							MASE						
Forecast horizon	Weekly Naïve	*k*-NN	SARIMA	SARIMAX	ETS	TBATS	HPR	Weekly Naïve	*k*-NN	SARIMA	SARIMAX	ETS	TBATS	HPR
1	0.5618	0.4748	0.3553	0.3156	0.2565	0.2818	**0.2468**	2.9904	1.8492	1.4042	1.2832	1.2511	1.2956	**0.9586**
2	0.5597	0.6575	0.6393	0.5470	0.4075	0.4436	**0.3184**	2.9863	2.3293	2.2450	2.0497	1.9499	1.8623	**1.2316**
3	0.5596	0.7188	0.8658	0.7221	0.5342	0.5374	**0.3761**	2.9876	2.5070	2.8212	2.5841	2.5149	2.2944	**1.4535**
4	0.5750	0.7136	1.0322	0.8747	0.6429	0.5758	**0.4348**	2.9968	2.6861	3.1295	2.9054	2.8993	2.4271	**1.6149**
5	0.5741	0.6938	1.1215	0.9648	0.6568	0.5898	**0.4717**	3.0003	2.8506	3.2672	3.0657	3.0121	2.5077	**1.7229**
6	0.5739	0.7508	1.1600	1.0236	0.6598	0.5993	**0.5170**	3.0015	2.7784	3.3138	3.1549	3.1885	2.5244	**1.8202**
7	0.5738	0.6569	1.1779	1.0516	0.6109	0.6125	**0.5620**	3.0049	2.7420	3.3398	3.1863	3.1756	2.5478	**1.8759**
8	0.8038	0.7538	1.2464	1.1161	0.6542	0.6824	**0.6244**	3.8491	2.9359	3.4564	3.2796	3.3709	2.7078	**2.0557**
9	0.8029	1.0346	1.3501	1.2016	0.6994	0.7493	**0.6672**	3.8458	3.3255	3.5794	3.4080	3.4348	2.7948	**2.2665**
10	0.8038	1.1589	1.4038	1.2571	0.7187	0.7962	**0.6851**	3.8508	3.5799	3.6543	3.4939	3.6221	2.9505	**2.4117**
11	0.8155	1.3064	1.4672	1.3302	0.7539	0.8226	**0.7365**	3.8629	3.6867	3.7091	3.5804	3.6608	3.0604	**2.5224**
12	0.8161	1.4194	1.5153	1.3852	**0.7770**	0.8431	0.7930	3.8676	3.9818	3.7458	3.6593	3.7102	3.0811	**2.6740**
13	0.8178	1.2543	1.5187	1.4172	**0.7854**	0.8672	0.8174	3.8805	3.9569	3.7900	3.7071	3.7623	3.1802	**2.7818**
14	**0.8126**	1.1923	1.5524	1.4598	0.8260	0.8840	0.8447	3.9036	3.9880	3.8464	3.7859	3.9058	3.2222	**2.8381**
Average	*0.6893*	*0.9133*	*1.1719*	*1.0476*	*0.6416*	*0.6632*	***0.5782***	*3.4306*	*3.0855*	*3.2359*	*3.0817*	*3.1042*	*2.6040*	***2.0163***

promising method of high frequency tourism demand forecasting.

This study makes three methodological contributions. First, it introduces an innovative pattern recognition method, hierarchical pattern recognition, to daily tourist arrival forecasting. The empirical results demonstrate its superior forecasting ability compared with six conventional time series models. Second, this study provides new insights for dealing with “floating holidays,” and capturing the multiple and complicated seasonality characteristics of holiday-sensitive tourism demand series. Third, two ensembles are adopted in the forecasting process with one ensemble based on the nearest patterns and the other ensemble based on different time window lengths.

The findings of this study also provide useful practical implications. First, considering the superior forecasting performance of the proposed method, practitioners at the tourist attractions can use it for short-run forecasting, which is the foundation for planning and marketing strategy formulation. For example, when extremely high tourist arrivals are predicted, the attraction managers should adopt relevant strategies such as crowd control or arranging more staff and resources in order to guarantee the safety of tourists and maintain visitor satisfaction levels. When very low tourist arrivals are predicted, price adjustment and online marketing strategies may be good options. Second, the special treatment of floating holidays and multiple seasonality is a useful tool that can help governments and DMOs understand and accommodate the fluctuation of demand. This will benefit their sustainable tourism and holiday arrangement policies. It is true that a high volume of tourist arrivals during holidays brings economic benefits to the attractions as well as the destinations where it is located. However, the excessive demand could also damage the sustainable development of natural tourism attractions. Therefore, when big fluctuations of tourism arrivals and excessive demand of attractions in holidays are predicted, the governments could develop polices to protect natural attractions. Such measures may include informing people of busy and less busy days, limiting the entry quota for admissions, and promoting some less popular sites as substitutes for over-crowed attractions. Third, the proposed HPR method could facilitate a better understanding of newly emerged tourist arrival patterns. Among others, the COVID-19 pandemic is having a severe impact on tourism. Arguably, that the tourism industry may be dramatically changed after the pandemic, and the proposed HPR method could help the governments and DMOs to quickly identify the post-COVID-19 tourism patterns and draft strategies and policies accordingly.

The findings of this study suggest a few future research directions. Firstly, this study proposes a hierarchical pattern recognition method for high frequency tourism demand forecasting with *k*-NN as the recognition algorithm. Future research could explore the effectiveness of different distance measurement scales when applying pattern recognition techniques. Secondly, in addition to holidays, other variables such as weather conditions, search engine query, and social media data could be incorporated into the forecasting process for the extension of the proposed method. Thirdly, though the predictive superiority of the proposed method has been verified compared with the benchmark models for three tourism attractions, the performance of the proposed method could be further tested in more diverse contexts, such as hotels, destinations, or different data frequencies. Lastly, interval forecasts can effectively supplement point forecasts by providing further information on their variability and uncertainty (Li, Wu, Zhou, & Liu, 2019). Therefore, in future research, instead of producing point forecasting, it is also valuable to explore using pattern recognition techniques to produce accurate tourism intervals.

CRediT author statement

Mingming Hu: Conceptualization, Methodology, Investigation, Data curation, Software, Formal analysis, Writing - original draft, Writing - review and editing. **Richard T.R. Qiu**: Conceptualization, Methodology, Investigation, Formal analysis, Writing - original draft; Writing - review and editing. **Doris Chenguang Wu**: Conceptualization, Formal analysis, Writing - original draft, Writing - review and editing. **Haiyan Song**: Conceptualization, Methodology, Writing - original draft, Writing - review and editing.

Impact statement

This study proposes an innovative hierarchical pattern recognition technique for forecasting the demand for tourism. The proposed method is capable of capturing the multiple and complicated seasonality characteristics in the holiday-sensitive tourism demand series, and is particularly effective for high frequency and short-run tourism demand forecast. This method could be used by the destination governments as well as tourism practitioners related to tourist attractions, airlines, hotels and online travel agencies to generate accurate daily forecasts. These forecasts can provide the destination governments and practitioners with useful information for visitor flow control, pricing, stuff arrangement, revenue management and other operational strategy formulations.

Declarations of competing interest

None.

References

Afouras, T., Chung, J., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–11.

Armstrong, J. S. (2001). Evaluating forecasting methods. In *Principles of forecasting* (pp. 443–472). Boston, MA: Springer.

Asht, S., & Dass, R. (2012). Pattern recognition techniques: A review. *International Journal of Computer Science and Telecommunications, 3*(8), 25–29.

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting, 27*(3), 822–844.

Box, G., & Jenkins, G. (1970). *Time series analysis, forecasting and control*. San Francisco: Holden Day.

Brubaker, S., Bonham, K., Zanoni, I., & Kagan, J. (2015). Innate immune pattern recognition: A cell biological perspective. *Annual Review of Immunology, 33*, 257–290.

Cai, P., Wang, Y., Lu, G., Chen, P., Ding, C., & Sun, J. (2016). A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting. *Transportation Research Part C: Emerging Technologies, 62*, 21–34.

Chen, R., Bloomfield, P., & Fu, J. (2003). An evaluation of alternative forecasting methods to recreation visitation. *Journal of Leisure Research, 35*(4), 441–454.

Chen, J. L., Li, G., Wu, D. C., & Shen, S. (2019). Forecasting seasonal tourism demand using a multiseries structural time series method. *Journal of Travel Research, 58*(1), 92–103.

Chen, K., & Wang, C. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management, 28*(1), 215–226.

Chu, F. (2009). Forecasting tourism demand with ARMA-based methods. *Tourism Management, 30*(5), 740–751.

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

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27.

De Livera, A. M., Hyndman, R., & Snyder, R. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association, 106*(496), 1513–1527.

Díaz, Á., & Mateu-Sbert, J. (2011). Forecasting daily air arrivals in Mallorca Island using nearest neighbour methods. *Tourism Economics, 17*(1), 191–208.

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

Enders, W., & Li, J. (2015). Trend-cycle decomposition allowing for multiple smooth structural changes in the trend of US real GDP. *Journal of Macroeconomics, 44*, 71–81.

Fagan, J. (2017). The origins of facial pattern recognition. In *Psychological development from infancy* (pp. 83–113). New York: Routledge.

Fujisawa, H. (2008). Forty years of research in character and document recognition—an industrial perspective. *Pattern Recognition, 41*(8), 2435–2446.

Galbally, J., Diaz-Cabrera, M., Ferrer, M., Gomez-Barrero, M., Morales, A., & Fierrez, J. (2015). Online signature recognition through the combination of real dynamic data and synthetically generated static data. *Pattern Recognition, 48*(9), 2921–2934.

Ghofrani, M., Azimi, R., & Youshi, M. (2019). Pattern recognition and its application in solar radiation forecasting. In *Pattern recognition—selected methods and applications*. London: IntechOpen. https://www.intechopen.com/books/pattern-recognition-selected-methods-and-applications/pattern-recognition-and-its-application-in-solar-radiation-forecasting.

Gil-Alana, L. (2005). Modelling international monthly arrivals using seasonal univariate long-memory processes. *Tourism Management, 26*(6), 867–878.

Gil-Alana, L., Mudida, R., & de Gracia, F. (2014). Persistence, long memory and seasonality in Kenyan tourism series. *Annals of Tourism Research, 46*, 89–101.

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

Harding, S., Benci, J., Irianto, J., Discher, D., Minn, A., & Greenberg, R. (2017). Mitotic progression following DNA damage enables pattern recognition within micronuclei. *Nature, 548*(7668), 466.

Heung, V. C., Qu, H., & Chu, R. (2001). The relationship between vacation factors and socio-demographic and travelling characteristics: The case of Japanese leisure travellers. *Tourism Management, 22*(3), 259–269.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting, 20*(1), 5–10.

Höpken, W., Ernesti, D., Fuchs, M., Kronenberg, K., & Lexhagen, M. (2017). Big data as input for predicting tourist arrivals. In *Information and communication technologies in tourism 2017* (pp. 187–199). Berlin: Springer.

Huang, M., Lin, R., Huang, S., & Xing, T. (2017). A novel approach for precipitation forecast via improved K-nearest neighbor algorithm. *Advanced Engineering Informatics, 33*, 89–95.

Hu, M., Ren, P., Lan, J., Wang, J., & Zheng, W. (2014). Note on "Some models for deriving the priority weights from interval fuzzy preference relations". *European Journal of Operational Research, 237(2)*, 771–773.

Hyndman, R. (2013). *Forecasting with daily data*. https://robjhyndman.com/hyndsight/dailydata/. (Accessed 6 August 2019).

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.) https://otexts.org/fpp2/.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., et al. (2019). *Package "forecast*. https://cran.r-project.org/web/packages/forecast/forecast.pdf.

Jain, A., Arora, S., Cao, K., Best-Rowden, L., & Bhatnagar, A. (2016). Fingerprint recognition of young children. *IEEE Transactions on Information Forensics and Security, 12*(7), 1501–1514.

Jain, A., Duin, R., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(1), 4–37.

Kamel, N., Atiya, A., El Gayar, N., & El-Shishiny, H. (2008). Tourism demand forecasting using machine learning methods. *ICGST International Journal on Artificial Intelligence and Machine Learning, 8*, 1–7.

Karlaftis, M., & Vlahogianni, E. (2009). Memory properties and fractional integration in transportation time-series. *Transportation Research Part C: Emerging Technologies, 17* (4), 444–453.

Kirillova, K., & Lehto, X. (2015). An existential conceptualization of the vacation cycle. *Annals of Tourism Research, 55*, 110–123.

Kon, S., & Turner, L. (2005). Neural network forecasting of tourism demand. *Tourism Economics, 11*(3), 301–328.

Lan, J., Hu, M., Ye, X., & Sun, S. (2012). Deriving interval weights from an interval multiplicative consistent fuzzy preference relation. *Knowledge-Based Systems, 26*, 128–134.

Lan, J., Zou, H., & Hu, M. (2020). Dominance degrees for intervals and their application in multiple attribute decision-making. *Fuzzy Sets and Systems, 383*, 146–164.

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

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

Law, R., & Au, N. (1999). A neural network model to forecast Japanese demand for travel to Hong Kong. *Tourism Management, 20*(1), 89–97.

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

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

Lin, V. S., Goodwin, P., & Song, H. (2014). Accuracy and bias of experts' adjusted forecasts. *Annals of Tourism Research, 48*, 156–174.

Lise, W., & Tol, R. (2002). Impact of climate on tourist demand. *Climatic Change, 55*(4), 429–449.

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

Li, G., Song, H., & Witt, S. (2005). Recent developments in econometric modelling and forecasting. *Journal of Travel Research, 44*(1), 82–99.

Liu, J., & Kwong, R. (2007). Automatic extraction and identification of chart patterns towards financial forecast. *Applied Soft Computing, 7*(4), 1197–1208.

Li, G., & Wu, D. C. (2019). Introduction to the special focus: Tourism forecasting—new trends and issues. *Tourism Economics, 25*(3), 305–308.

Li, G., Wu, D. C., Zhou, M., & Liu, A. (2019). The combination of interval forecasts in tourism. *Annals of Tourism Research, 75*, 363–378.

Lora, A., Santos, J., Expósito, A., Ramos, J., & Santos, J. (2007). Electricity market price forecasting based on weighted nearest neighbors techniques. *IEEE Transactions on Power Systems, 22*(3), 1294–1301.

Olmedo, E. (2016). Comparison of near neighbor and neural network in travel forecasting. *Journal of Forecasting, 35*(3), 217–223.

Onafowora, O., & Owoye, O. (2012). Modelling international tourism demand for the Caribbean. *Tourism Economics, 18*(1), 159–180.

Palmer, A., Montano, J., & Sesé, A. (2006). Designing an artificial neural network for forecasting tourism time series. *Tourism Management, 27*(5), 781–790.

Rice, W., Park, S., Pan, B., & Newman, P. (2019). Forecasting campground demand in US national parks. *Annals of Tourism Research, 75*, 424–438.

Sadaei, H., Guimarães, F., da Silva, C., Lee, M., & Eslami, T. (2017). Short-term load forecasting method based on fuzzy time series, seasonality and long memory process. *International Journal of Approximate Reasoning, 83*, 196–217.

Schliep, K., Hechenbichler, K., & Lizee, A. (2016). *Package "kknn*. https://cran.r-project.org/web/packages/kknn/kknn.pdf.

Shelke, M., & Thakare, P. (2014). Short term load forecasting by using data mining techniques. *International Journal of Science and Research, 3*, 1363–1367.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting—a review of recent research. *Tourism Management, 29*(2), 203–220.

Song, H., Qiu, R., & Park, J. (2019). A review on tourism demand forecasting: Launching the *Annals of Tourism Research* Curated Collection on tourism demand forecasting. *Annals of Tourism Research, 75*, 338.

Tideswell, C., Mules, T., & Faulkner, B. (2001). An integrative approach to tourism forecasting: A glance in the rearview mirror. *Journal of Travel Research, 40*(2), 162–171.

UNWTO. (2019). *UNWTO international tourism highlights 2019 edition*. Madrid, Spain: World Tourism Organization (UNWTO).

Watanabe, S. (1985). *Pattern recognition: Human and mechanical*. New York: Wiley.

Wu, D. C., Cao, Z., Wen, L., & Song, H. (2020). Scenario forecasting for global tourism. *Journal of Hospitality & Tourism Research*. https://doi.org/10.1177/1096348020919990

Wu, D. C., Song, H., & Shen, S. (2017). New developments in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management, 29*(1), 507–529.

Xu, Z. (2015). *Uncertain multi-attribute decision making: Methods and applications*. Springer.

Xu, Z., & Da, Q. L. (2003). An overview of operators for aggregating information. *International Journal of Intelligent Systems, 18*(9), 953–969.

Zhang, G., Patuwo, B., & Hu, M. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting, 14*(1), 35–62.

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8697–8710).

Wu, Q., Law, R., & Xu, X. (2012). A sparse Gaussian process regression model for tourism demand forecasting in Hong Kong. *Expert Systems with Applications, 39*(5), 4769–4774. https://doi.org/10.1016/j.eswa.2011.09.159

Mingming Hu , PhD, is an assistant professor in the Business School of Guangxi University and postdoctoral fellow in the School of Hotel and Tourism Management at The Hong Kong Polytechnic University. His researcher interests are in the areas of tourism demand forecasting, tourist behavior and decision-making analysis. Email: mingming.hu@gxu.edu.cn

Richard T. R. Qiu , PhD, is an assistant professor in the Department of Integrated Resort and Tourism Management, Faculty of Business Administration at the University of Macau. His research interests are tourist choice behavior, tourism economics and tourism demand forecasting. Email: richardqiu@um.edu.mo

Doris Chenguang Wu , PhD, is professor of tourism forecasting in the Business School at the Sun Yat-sen University, China. Her research interests include tourism demand forecasting and tourism impact analysis. Email: wucheng@mail.sysu.edu.cn

Haiyan Song, PhD, is Chan Chak Fu Professor in International Tourism in the School of Hotel and Tourism Management at The Hong Kong Polytechnic University. His research interests are in tourism demand modeling and forecasting, tourism supply chain management and wine economics. Email: haiyan.song@polyu.edu.hk