



# Systematic investigation of keywords selection and processing strategy on search engine forecasting: a case of tourist volume in Beijing

Ziqi Yuan<sup>1</sup> · Guozhu Jia<sup>1</sup>

Received: 5 May 2022 / Revised: 26 September 2022 / Accepted: 27 September 2022 /  
Published online: 7 November 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

## Abstract

The timeliness, precision, and low cost of search data have great potential for projecting tourist volume. Obtaining valuable information for decision-making, particularly for predicting, is hampered by the vast amount of search data. A systematic investigation of keyword selection and processing has been conducted. Using Beijing tourist volume as an example, 11 different feature extraction algorithms were selected and combined with long short-term memory (LSTM), random forest (RF) and fuzzy time series (FTS) for forecasting tourist volume. A total of 1612 keywords were retrieved from Baidu Index demand mapping using the direct word extraction method, range word extraction method and empirical selection method. The remaining 813 keywords were subjected to feature extraction. Based on the forecasting results of medium and short-term (1-day, 7-days and 10-days), the forecasting results of Kernel principal component analysis (KPCA) and locally linear embedding (LLE) are relatively stable when the dimensionality is reduced to 5 dimensions. The forecasting results of t-stochastic neighbor embedding (t-SNE), isometric mapping (IsoMap) and locally linear embedding (LLE), locality preserving projections (LPP), independent component correlation (ICA) are relatively stable when the dimensionality is reduced to 10 dimensions. Accurately forecasting many factors (transportation, attraction, food, lodging, travel, tips, tickets, and weather) provides a solid foundation for tourism demand optimization and scientific management and a resource for tourists' holistic vacation planning.

**Keywords** Tourist volume forecasting · Search engine data · Feature extraction algorithm · LSTM · FTS · RF

---

✉ Guozhu Jia  
309739124@qq.com

<sup>1</sup> College of Physical and Electronics Engineering, Sichuan Normal University, Chengdu 610000, China

## 1 Introduction

Tourism forecasts based on online search data directly reflect tourists' worries about various parts of their travel itineraries enabling tourism resource allocation and service plans to be modified to comprehend better the tourism market (Li et al. 2018a, b, 2021a; Liu et al. 2018). All facets of tourism, such as cuisine, hotel, transport, and tips, are included in the Baidu search engine's tourism-related keywords. The search for keywords differs from person to person, resulting in numerous tourism-related keywords. Combining these keywords makes it possible to precisely describe tourist travel plans and increase the accuracy of projections, enabling the tourism management industry to deploy resources and make strategic decisions more efficiently (Li et al. 2020a, b). However, due to the different knowledge acquisition channels (content channel, primary knowledge channel and background channel) that influence visitors' keyword selection behaviour (Lu et al. 2020), keywords' selection and processing strategy must be considered. Gao and Sheng (2021) described the online search behaviour of travellers using tourism-related terms and subjectively picked four keywords for analysis. Lu and Liao (2021) screened out the initial keywords by analysing the main factors affecting the behaviour of tourism consumers and travel notes. In order to address the issue of data volume, the dimensionality reduction algorithm has garnered a great deal of interest as an efficient data extraction technique (Kuang et al. 2018).

The dimensionality reduction algorithm aims to address the intractable problem of keyword data dimension and enhance data quality by lowering data complexity. It is divided primarily into feature selection and feature extraction. Feature selection is used to pick a subset of associated features for effective data classification (Hoque et al. 2014). Feature extraction is to combine these features into fewer new features through algebraic transformation and retain the most pertinent information (Khalid et al. 2014; Zebari et al. 2020). Feature extraction algorithms (FEAs) are the focus of our research because they can better deal with issues in serial data sets, such as noise, complexity, and sparsity. This research will use several feature extraction algorithms to identify all 1612 tourism-related keywords and perform classification and dimensionality reduction.

This case study aims to present a keyword selection approach and evaluate practicable feature extraction algorithms that use search trend data to forecast tourist volume. The approach for the feature extraction method can handle a large number of search data sequences and includes sufficient keyword variables. Using a strategy similar to Li et al. (2017), we will collect tourism-related keywords according to various criteria. Due to the limits imposed on international tourists by the COVID-19 pandemic, this article focuses primarily on analysing many elements of Chinese tourists' tourism; hence the Baidu Index platform was selected for analysis. Through specific search keywords collected from the Baidu Index demand map, including tourism-related keywords ("transportation", "lodging", "travel guide", "travel", "attraction", "weather", "food" and "tickets"), this study empirically compared the different reduction results of feature extraction algorithms for forecasting Beijing tourist volume from January 2017 to September 2021.

Our research has the following structure: Sect. 2 briefly reviews the literature related to tourism through the network engine data and introduces the feature extraction algorithms. Section 3 introduces the methods. In Sect. 4, experiments were carried out, and the experimental results were analysed. Finally, Sect. 5 summarises the research.

## 2 Literature review

This section introduces the forecasting methods by tourism search data and methods for forecasting tourist volume in recent years, as well as the feature extraction algorithm and its application.

### 2.1 Search engine data and methods of tourist volume forecasting

With the expansion of the Internet, there is an increasing tendency to employ online data for digital monitoring (Huang et al. 2021). Large amounts of online data, in the form of consumers' online search behaviour, can enhance forecasts' accuracy significantly. Search engine data consists of the daily, weekly and monthly search queries that users input in real-time in search engines that provide various services, such as Google and Baidu, the two most popular internet search engines among the general public. They construct data platforms based on the data they look for on the platforms—Baidu Index and Google Trends—which are the most extensively utilized data for projecting tourist volume. Social data and user-generated content have become essential sources of timely and knowledge-based support for data-driven decision-making techniques to solve complicated tourism management issues (Cuomo et al. 2021). Search engines give researchers markers of tourist behaviour, such as current location and future trip plans.

Numerous studies have proved the usefulness of search engine data for travel forecasting. Several studies have demonstrated that Google Trends can enhance the accuracy of travel forecasting (Siliverstovs and Wochner 2018; Bokelmann and Lessmann 2019; Clark et al. 2019; Feng et al. 2019; Cevik 2020). Baidu Index is the most popular search engine in China; hence forecasting tourist volume in China is more accurate when utilising the Baidu Index (Yang et al. 2015). In addition, many studies have demonstrated that using the Baidu Index can considerably enhance the accuracy of Chinese tourist forecasting (Li et al. 2017; Huang et al. 2017; Kang et al. 2020; Ren et al. 2020a, b; Xie et al. 2021). Other research has investigated the combination of Google Trends and Baidu Index, particularly for improved forecasting of international passengers (Dergiades et al. 2018; Sun et al. 2018).

According to Li et al. (2021a, b, c), various models have been employed to anticipate tourist volume. Time series and econometric prediction models dominate, while artificial intelligence methods are still evolving. The predictive capability of the machine learning model substantially aids forecasters in estimating the optimal predictive performance and comprehending the impact of various data characteristics on forecasting performance (Zhang et al. 2021). Historically-based time series

models estimate visitor arrival patterns. Numerous research employs time series models to assess and anticipate tourism demand, with the autoregressive moving average models (ARIMA) (Huang et al. 2017; Li et al. 2017) and seasonal integrated autoregressive moving average models (SARIMA) being the most common (Bokelmann and Lessmann 2019; Sun 2021). Vector autoregressive (VAR) is a commonly employed model in econometrics (Padhi and Pati 2017; Ren et al. 2020a, b). In recent years, machine learning methods such as support vector regression (SVR), support vector machines (SVM), and random forests (RF) have been adopted by the majority of researchers (Feng et al. 2019; Li et al. 2020a, b; Yao et al. 2021). Artificial neural networks (ANN) (Law et al. 2019), Back Propagation Neural networks (BPNN) and Generalised Regression Neural networks (GRNN) (Xie et al. 2020) are examples of deep learning techniques. Long short-term Memory (LSTM) is the method of deep learning that is utilised most frequently (Li and Cao 2018; Bi et al. 2020; Li et al. 2020a, b; He et al. 2021; Kaya et al. 2022).

Throughout the selection and processing of the search data, Li et al. (2021a, b, c) examined feature selection approaches based on machine learning to pick the search data and determine the outcomes effectiveness. Li et al. (2018a; b) and Wang et al. (2021a, b) both employ principal component analysis (PCA) to minimise the dimension of data. Peng et al. (2017) select keywords by combining Hurst exponent (HE) and time difference correlation (TDC). Other studies utilise keyword correlations for selection (Wei and Cui 2018; Feng et al. 2019; Kang et al. 2020). As far as the author is aware, no systematic research has been conducted on selecting and processing tourism keyword phrases.

## 2.2 Feature extraction algorithm

A large number of tourism-related keywords cause the keyword dimensions to swell. Feature extraction algorithms aim to reduce data complexity to improve data quality and prevent dimensionality disasters (Anowar et al. 2021). In this study, 1612 keywords on various aspects of Beijing tourism were chosen, downscaled using feature extraction techniques, and the findings of various FEAs were compared in depth. As shown in Table 1.

**Table 1** Feature extraction algorithms (FEAs)

Linear FEAs	Supervised	Linear discriminant analysis (LDA) Independent component correlation (ICA) Locality preserving projections (LPP)
	Unsupervised	Principal component analysis (PCA) Truncated singular value decomposition (TSVD)
Nonlinear FEAs	Supervised	
	Unsupervised	Kernel PCA (KPCA) Multidimensional scaling (MDS) Isometric mapping (IsoMap) Laplacian eigenmaps (LE) Locally linear embedding (LLE) t-stochastic neighbor embedding (t-SNE)

Linear discriminant analysis (LDA) maximises the ratio between the inter-class variance and the within-class variance of any given dataset, hence providing the highest level of separability. LDA does not alter the original dataset's location but instead attempts to increase class separability and map a decision region between specified classes. This strategy also assists in comprehending the distribution of feature data (Balakrishnama and Ganapathiraju 1998). LDA is predominantly utilised in engineering (Li et al. 2021a, b, c), agriculture (Wei et al. 2021) and chemistry (Buzzini et al. 2021).

Independent Component Analysis (ICA) decomposes a random vector into linear components that are "as independent as feasible". The ICA method attempts to extract independent elements from a mixture of random variables or measurement data. ICA is often restricted to linear mixes, and the putative sources are considered independent (Westad and Kermit 2009). The majority of ICA's uses are in chemistry (Moreira de Oliveira et al. 2021), medicine ((Barborica et al. 2021) and materials (Mu et al. 2021).

Locality preserving projections (LPP) creates a graph encapsulating the neighbourhood information of a given data collection, calculating the transformation matrix that translates the data points to a subspace using the graph Laplace operator idea. This linear translation preserves the ideal local neighbourhood information to some extent. This approach produces a discrete linear approximation of the continuous map provided naturally by the various geometry (He and Niyogi 2010). LPP is mainly utilised in agriculture (Shao 2019), chemistry (He et al. 2018), and imaging (Wei et al. 2020).

Principal Component Analysis (PCA) aims to transfer  $n$ -dimensional characteristics onto a  $k$ -dimensional space ( $k < n$ ). Instead of just deleting the leftover  $n-k$  dimensional features from the  $n$ -dimensional features, these  $k$ -dimensional features are brand-new orthogonal and reconstructed dimensional features. PCA seeks to summarise the structure of multivariate data and project it into fewer components, with each principal component accounting for a share of the original data set's variance (Wold et al. 1987). PCA is utilised in a variety of fields, including physics (Park et al. 2021), geography (Aidoo et al. 2021), medicine (Duarte and Riveros-Perez 2021) and tourism to some extent (Natalia et al. 2018).

Truncated singular value decomposition (TSVD) is a matrix factorisation algorithm. SVD decomposition is conducted on the data matrix, whereas PCA decomposition is performed on the covariance matrix. SVD is typically used to identify the primary components of a matrix. TSVD is a regularisation technique that compromises some precision for the stability of the solution, resulting in greater generalisability of the output (Chen et al. 2019). TSVD is mainly utilised in physics (Yuan et al. 2020), mathematics (Vu et al. 2021) and chemistry (Cattani et al. 2015).

Kernel Principal Component Analysis (KPCA) is a nonlinear variant of the PCA algorithm. PCA is linear and typically appears ineffective with nonlinear data. However, KPCA can access the nonlinear information contained within the dataset. Based on the kernel function principle, the input space is projected into the high-dimensional feature space using nonlinear mapping. Then the mapped data is then analysed using the principal component (Schölkopf et al. 1997). KPCA is

predominantly utilised in stock price forecasting (Nahil and Lyhyaoui 2018), physics (Cui and Shen 2021), and medicine (Wang et al. 2021a, b).

Multidimensional scaling (MDS) is a method for quantitative similarity determination. Formally, MDS accepts as input an estimate of the similarity between a group of elements. These stimuli may be explicit ratings or various 'indirect' measures (such as perceived confusion), and they may be perceived or conceptual, spatially conveying the relationship between items in which similar items are located near one another and different items are separated proportionally (Hout et al. 2013). MDA is primarily utilised in mathematics (Machado and Luchko 2021) and geology (Seraine et al. 2021).

Isometric Mapping (IsoMap) is a linear MDS version. It identifies a set of points that correspond to the provided distance matrix. In contrast to standard MDS, distances are defined differently. IsoMap uses geodesic distances rather than pairwise direct distances. The geodesic distances are approximated by the  $k$ -nearest neighbour graph's shortest path distance (Mousavi Nezhad et al. 2018). IsoMap is primarily utilised in manufacturing and chemical (Yang and Yan 2012; Bi et al. 2019).

Laplacian eigenmaps (LE) employ the weighted distance between two points as the loss function for dimension reduction. By building an adjacency matrix graph, LE reconstructs the local structural properties of the data manifold. If two data instances are comparable, they should be as close as possible in the dimensionality reduction target subspace (Belkin and Niyogi 2003). LE is primarily utilised in medical and engineering (Good et al. 2020; Arena et al. 2020).

Local linear embedding (LLE) approximates each data point to the neighbouring data point in the input space while capturing the local geometric properties of the complex embedded manifold with the best linear coefficients. LLE then locates a set of low-dimensional points, each of which can be linearly approximated by its neighbours with the same set of coefficients computed from high-dimensional data points in the input space while minimising the reconstruction cost (Roweis and Saul 2000). LLE is mainly used for measurement and modelling (Liu et al. 2021; Miao et al. 2022).

Stochastic neighbor embedding (SNE) is based on the concept that similar data points in high-dimensional space are translated to low-dimensional space with similar distances. To describe similarity, T-distributed stochastic neighbour embedding (t-SNE) converts this distance connection into a conditional probability. t-SNE has made two enhancements to SNE: one is to simplify the gradient formula by employing symmetric SNE, and the other is to replace Gaussian distribution with t-distribution in low-dimensional space (van der Maaten and Hinton 2008). t-SNE is mainly employed in energy (Han et al. 2021), food (Luo et al. 2021), and medicine (Cieslak et al. 2020).

### 3 Methodology

In this section, we first present the methodology's framework. He proposed forecasting of tourism demand is then detailed. The entire procedure is depicted in Fig. 1.

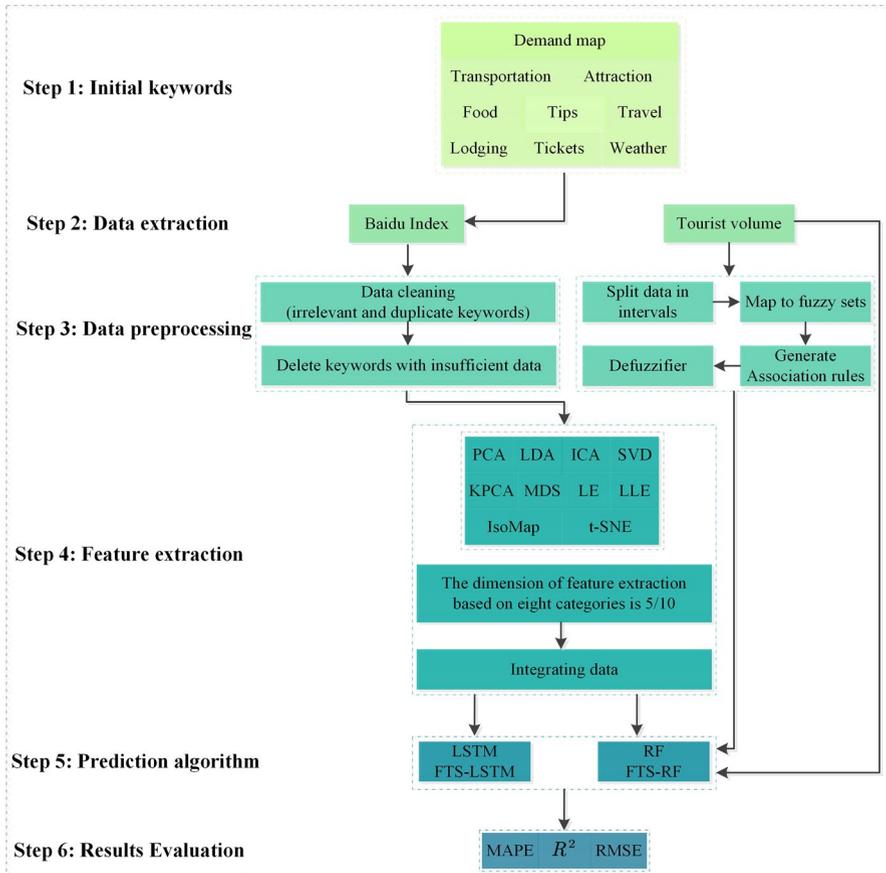


Fig. 1 Method framework

### 3.1 Selecting keyword indices

In the Internet era, they are searching for information before and exchanging it after travel has become an indispensable step. From their perspective, the use of Internet search engines for information retrieval has established avenues for tourists to prepare for the trip. Consequently, rising tourists are eager to utilise this information retrieval strategy. Whether true or false, Internet-posted comments may be accessible to other visitors. Therefore, the tourism search data might indicate tourism promotion and the travellers’ interest in tourist information.

According to the Chinese search engine product market development report by Big Data Research ([www.bigdata-research.cn](http://www.bigdata-research.cn)), Baidu’s total penetration rate ranks first at 70%. As the search engine most used by Chinese netizens, Baidu (<http://index.baidu.com/>) covers the great majority of search behaviours. Therefore, we will perform research for this paper using Baidu. Baidu Index is a platform for exchanging



- in the Baidu Index demand map. Then, using a range word extraction method based on eight distinct criteria, a total of 1122 keywords were eliminated.
- d) The time-series data for the sought-after keywords were retrieved from the Baidu index and then filtered using the empirical selection method. Typically, the empirical selection method (Li et al. 2017) was used to pick keywords, in which the frequency and quantity of keywords are weighted based on the cognitive level of the researcher. Following the removal of duplicate and inadequate data, 813 keywords remained.

This process aims to maximise the potential keywords symbolising Beijing's tourist attractions.

### 3.1.2 Selecting proper indices

The search volume of search words in various periods can be obtained on the Baidu Index platform, which can be regarded as a set of digital data or a time series. The Baidu index, the same as the date series of the number of tourists, is the benchmark index. There are many query words in the library of Baidu search queries. If all data from the Baidu index is used for forecasting, the complexity of generating the forecasting model and the pace of operations will increase significantly. In addition, an excessive amount of noise will be added to the model, which will have a detrimental effect on predictions. If Baidu index data for numerous query terms are taken at random, the forecasting model's accuracy is diminished. Consequently, it is crucial to analyze the keywords.

In order to further investigate whether the medium and low-frequency words that take into account a large number of meaningful relationships are more conducive to improving the results of predictive analysis, this paper takes into account potential keywords. It employs feature extraction algorithms to reduce the dimension. After accumulating relevant time series data, keywords with inadequate data are deleted. Figure 3 and Table 2 display the top ten keywords for each category.

## 3.2 Long short term memory network (LSTM)

In the early stages of prediction for the time series, the most recent samples and historical data are required. Thanks to the self-feedback mechanism of the hidden layer, the RNN model has some advantages in dealing with long-term dependence. However, there are still some difficulties in practical application, as shown in Fig. 4 (Bengio et al. 1994), the hidden layer information of the RNN at this moment is only derived from the current input and the hidden layer information of the previous moment, and there is no memory function. It is challenging to back-propagate the gradient at the end of a long sequence to the sequence before it when the sequence is lengthy. Hochreiter et al. (1997) proposed an LSTM model to address the gradient disappearance problem, also known as the long-term dependence of RNNs. It was later improved and popularized by Graves (2013).



**Table 2** The top ten high-frequency words of different classes

Classes	Keywords
Transportation	Beijing Municipal Administration and Communications Card Beijing traffic Beijing transit Beijing bus line query Beijing subway Beijing Subway Map Beijing subway route map Beijing airport Beijing subway operating hours Beijing Real-time Public Transport
Attraction	Shichahai Badaling Great Wall National Museum of China The Old Summer Palace The Temple of Heaven Prince Gong's Mansion The Forbidden City Palace Museum Yonghe Palace The Summer Palace
Travel	Badaling Great Wall Day Tour One-day tour of Beijing Three-day tour of Beijing Two-day tour of Beijing Two-day tour around Beijing Travel around Beijing Beijing tourism Beijing Tour Visit Beijing Day trip to the Great Wall
Travel tips	Badaling Great Wall tour tips Beijing one day tour tips Fun places in Beijing Beijing tips Beijing Tourist Map Beijing travel tips Beijing Happy Valley Tips Beijing travel tips Beijing Self-Guided Tour Tips Summer Palace tips
Food	Quanjudu Roast Duck Beijing snacks Beijing specialties Special Food of Beijing Beijing cuisine Beijing hotel Huguo Temple snacks Wangfujing Food Alley Noodles with Soy Bean Paste Douzhir
Tickets	Beijing Zoo tickets Beijing Forbidden City tickets Beijing Happy Valley Tickets Beijing Wildlife Park tickets The Old Summer Palace tickets Prince Gong's Mansion tickets Forbidden City tickets Great Wall tickets The Summer Palace tickets How much is the entrance fee for the Forbidden City
Weather	The weather in Beijing Beijing weather forecast 15 days Tomorrow's weather in Beijing the temperature in Beijing Miyun weather Weather forecast for Yanqing Huairou weather Fangshan weather Changping weather Shunyi weather
Lodging	Gloucester Luk Kwok Beijing Marriott Hotel Beijing lodging tips Grand Millennium Beijing Beijing Hotel Xijiao Hotel Beijing Beijing guesthouse Beijing Hotel Reservation Beijing Youth Travel Service Diaoyutai State Guest House

$$C_t = f_t \cdot C_t + i_t \cdot \tilde{C}_t$$

### 3.3 Random forest (RF)

Random forest is an integrated method of decision-making based on classification or regressions trees proposed by Breiman (2001). It develops and merges hundreds of separate classification trees into a single classifier. Individual classifiers will use the majority rule to decide the final classification of a sample. Random forest features minimal configurable parameters, automatic generalisation error calculation, and good overfitting resistance. The illustration of RF is shown in Fig. 6.

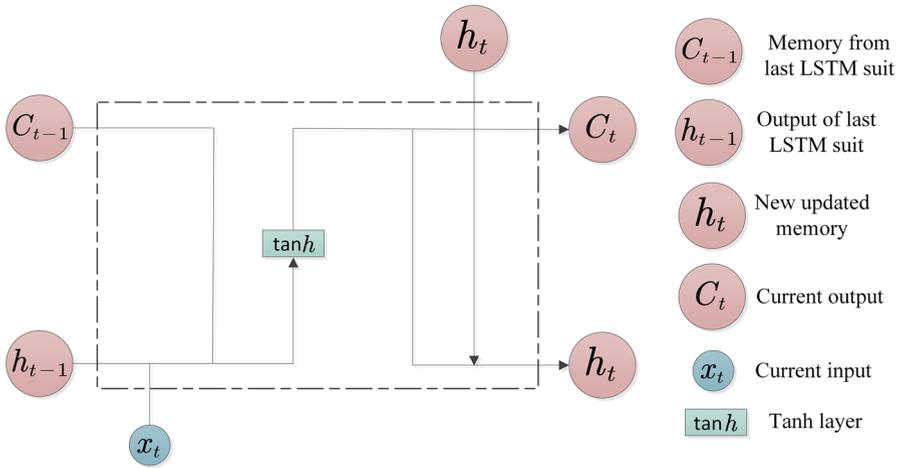


Fig. 4 The structure of the recurrent neural network (RNN) neural network

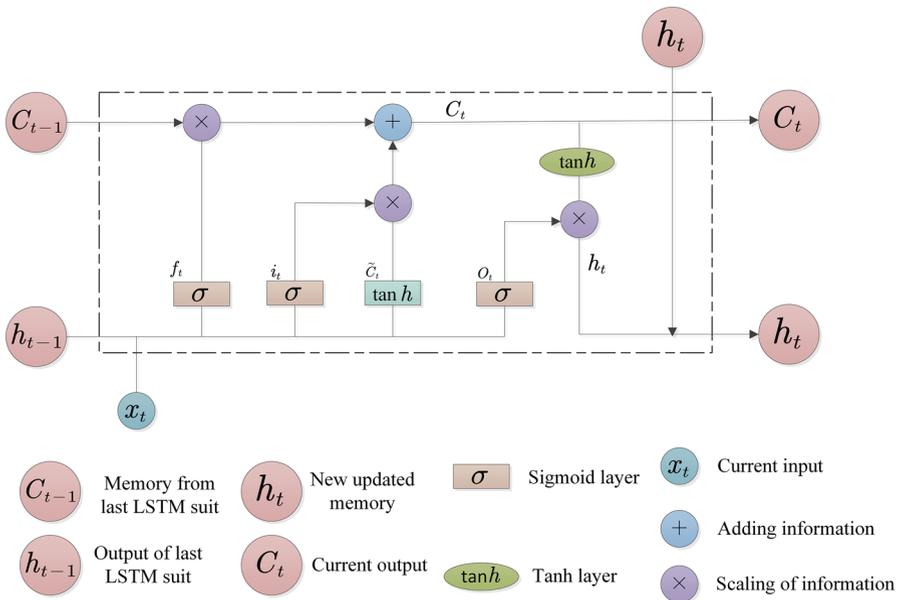


Fig. 5 The structure of the long short-term memory (LSTM) neural network

### 3.4 Fuzzy time series (FTS)

Song and Chissom (1993, 1994) first presented the concepts of fuzzy time series, where the values in a time series data are represented by fuzzy sets. FTS models have the advantage of using fuzzy sets when solving non-linear time series forecasts. For example, FTS can model non-linear and uncertain systems,

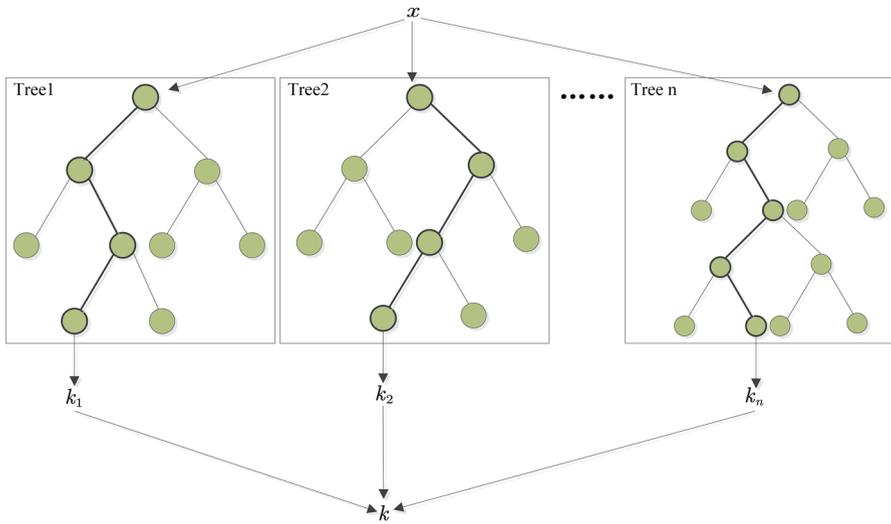


Fig. 6 A simple illustration of RF

incorporate expert opinions and experiences in the modelling process, deal with linguistic variables, and require no statistical assumptions. In general, the basic steps in designing an FTS model to produce forecasts are as follows.

Define the universe of discourse  $U$  and divide into equal intervals,  $U = \{u_1, u_2, \dots, u_n\}$ , the fuzzy set  $A_i$  defined in  $U$  can be expressed as:

$$A_i = \frac{f_{A_i}(u_1)}{u_1} + \frac{f_{A_i}(u_2)}{u_2} + \dots + \frac{f_{A_i}(u_n)}{u_n}$$

where  $f_{A_i}$  is membership function of  $A_i$ , and  $f_{A_i}(u_n) \in [0, 1]$ .

**Definition 1.** Let  $X(t)(t = 0, 1, 2, \dots, n)$  as a subset of real numbers. Give the universe of discourse  $U$ ,  $y_i(t)(i = 1, 2, \dots)$  is the fuzzy set defined on  $U$ .  $Y(t) = \{y_1(t), y_2(t), \dots\}$  is the fuzzy time series defined on  $X(t)$ .

**Definition 2** Assume that  $Y(t)$  is only affected by  $Y(t - 1)$  in the fuzzy time series relationship.  $Y(t)$  and  $Y(t - 1)$  are respectively the fuzzy sets at time  $t$  and  $t - 1$ .  $M(t - 1, t)$  is the fuzzy logic relationship between  $Y(t)$  and  $Y(t - 1)$ . Then the fuzzy relationship can be represented as  $Y(t) = Y(t - 1) * M(t - 1, t)$ .

**Definition 3** Set  $Y(t - 1) = A_i, Y(t) = A_j$ , the fuzzy logic relationship between  $Y(t)$  and  $Y(t - 1)$  can be denoted as  $A_i \rightarrow A_j$ . Where  $A_i$  and  $A_j$  correspond to left-hand side and right-hand side of fuzzy relationship, respectively.

### 3.5 Autoregressive integrated moving average model (ARIMA)

The ARIMA model is a popular and widely used statistical method for time series forecasting. AR is the autoregressive model, and MA is the moving average model,

Autoregressive models describe the relationship between current and historical values and use the historical time data of the variables themselves to predict themselves, but they have certain limitations. That is, the autoregressive model must meet the requirements of stationarity and must have autocorrelation.

The formula definition of the  $p$ -order autoregressive process is:

$$y_t = \mu + \sum_{i=1}^p r_i y_{t-i} + \varepsilon_t$$

$y_t$  is the current value,  $\mu$  is the constant term,  $p$  is the order,  $r_i$  is the autocorrelation coefficient, and  $\varepsilon_t$  is the error.

The model reflects that there is a linear relationship between the target value at  $t$ -time and the first  $t - 1 \sim p$  target values:

$$y_t \sim r_1 y_{t-1} + r_2 y_{t-2} + \dots + r_p y_{t-p}$$

The moving average model focuses on the accumulation of error terms in autoregressive models, and the expression is as follows:

$$y_t = \mu + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t$$

The model reflects that there is a linear relationship between the target value at  $t$ -time and the first  $t - 1 \sim p$  target values:

$$y_t \sim \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_p \varepsilon_{t-p}$$

ARIMA is a combination of autoregression and moving average. The specific mathematical model is as follows:

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

The basic principle of ARIMA is a model built by transforming data into smooth data by differencing and then regressing the dependent variable on its lagged values only and on the present and lagged values of the random error term.

### 3.6 Seasonal autoregressive integrated moving average (SARIMA)

Compared the ARIMA model, SARIMA adds four new parameters to define the feature of the seasonal component of the time series. Generally expressed as SARIMA( $p, d, q$ )( $P, D, Q$ ) $_s$ , where  $d$  is the differential value, which is determined by the unit root test,  $p$  is the autoregressive order, and  $q$  is the moving average order value, which is determined by AIC and BIC information criteria, respectively.  $P$  is

the periodic autoregressive order, and  $D$  is the periodic differential order,  $Q$  is the order of the periodic moving average,  $S$  is the cycle interval.

### 3.7 Evaluation criteria

For model training, we used 80% of the dataset as the training set and 20% of the dataset as the test set. To verify the forecasting accuracy of different models, we adopt three main evaluation criteria: mean absolute error (MAE), Mean Absolute Percentage Error (MAPE) and root mean squared error (RMSE).

$$MAE = \sum_{i=1}^n \frac{|\hat{y}_i - y_i|}{n}$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

where  $y_i$  indicates the actual tourist volumes, and  $\hat{y}_i$  is the predicted values of tourist volumes with learning model.

The average mean absolute error (MAE) and average root mean squared error (RMSE) from the time series cross-validation method were used to validate the prediction results.

## 4 Empirical study

### 4.1 Data collection

Beijing was chosen as the test destination because it is the capital of China. Beijing tourist volume is collected from the Beijing Municipal Bureau of Culture and Tourism. The daily frequency of these data ranges from January 1, 2017, to September 30, 2021, and Fig. 7 depicts the periodic swings of tourism data. Within the study, the number of tourists in Beijing displayed a constant temporal trend and, generally, a variable growth pattern. Before the emergence of COVID-19 (February 2020), there were four annual peaks: February, April (except for 2019), August, and October. In August 2018 and 2019, the number of tourists hits a plateau. The tourist volume showed a significant cyclical shift.

The search time series data was gathered from the Baidu index platform as an extra input component. After processing, 813 keyword data were obtained in total. The duration is between January 1, 2017, and September 30, 2021.

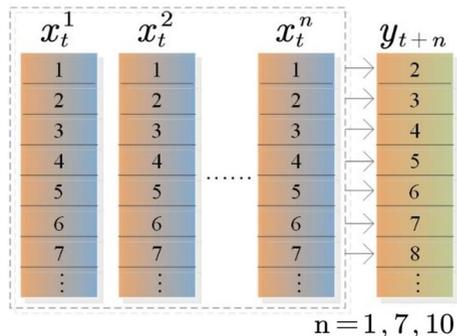


Fig. 7 Beijing tourist volume

### 4.2 Empirical results

In this section, the findings of feature extraction algorithms are used as inputs for LSTM and RF prediction methods. Due to many Beijing tourism-related keywords, multiple feature extraction methods were employed to downscale the data of different categories to 5 or 10 dimensions. Then the data, after being downscaled to 5 and 10 dimensions, were pooled to forecast the outcomes. Particularly during the feature extraction procedure with a dimension of 10, KPCA can extract a maximum of 9 dimensions. Therefore, in the feature extraction process with dimension 10, the experimental dimension of KPCA is 9. Similar to the short-, medium-, and long-term forecast forms of Lu et al. (2021), this paper inputs the independent variables in the form of Fig. 8. to forecast the 1-day, 7-days and 10-days periods. Our total dataset is 1734 data. For the prediction part, the training sets comprised 80 per cent of the overall dataset, while the test set comprised 20 per cent.

Fig. 8 Input–output feature structure



### 4.2.1 ARIMA algorithm

First, we conducted an ADF test on the data to test for stationarity, and the results are shown in Table 3. The original hypothesis of the ADF test is the existence of a unit root. As long as this statistic is less than a number at the 1% level, the original hypothesis can be rejected as highly significant, and the data is considered smooth. The test result of  $-3.638$  is less than  $-3.434$  at the 1% level, and the  $p$  value is less than  $0.05$ . Therefore, the time series is considered stationary.

We used the auto-ARIMA algorithm model to determine the parameters, which combines a unit root test, minimising AIC and MLE to obtain the ARIMA model. ARIMA models with different values of parameters were explored. The final model was determined to be an ARIMA (1, 1, 1) model, and the forecasting results are shown in Table 4.

### 4.2.2 SARIMA algorithm

Experiments are used to adjust the parameters of the final SARIMA model to (1, 1, 1) (2, 1, 0) [6]. The SARIMA forecasting results are displayed in Table 5. At the time of 1-day forecast, the MAE and RMSE are 0.183 and 0.255, At the time of forecasting 7-days, the MAE is 0.684 and the RMSE is 1.377. At the time of forecasting 10-days, the MAE is 0.578 and the RMSE is 1.166.

### 4.2.3 FTS algorithm

This work examines the isometric partitioning technique known as the grid based partitioning scheme, and employs a triangle affiliation function to map entities, as

**Table 3** ADF test

ADF	- 3.638
P-Value	0.005
Num of lags	10
Critical values	1%: - 3.434 5%: - 2.863 10%: - 2.568

**Table 4** Performance evaluation forecasting results using ARIMA algorithm

$n=1$		$n=7$		$n=10$	
MAE	RMSE	MAE	RMSE	MAE	RMSE
0.138	0.243	0.513	0.811	0.538	0.842

**Table 5** Performance evaluation forecasting results using SARIMA algorithm

$n=1$		$n=7$		$n=10$	
MAE	RMSE	MAE	RMSE	MAE	RMSE
0.183	0.255	0.684	1.377	0.578	1.166

depicted in Fig. 9. A splitting factor of 30 is used to map the number of visitors over the 30 days. The following formula represents the triangular membership scheme.

$$\omega(x)_i = \begin{cases} 0, & x \leq l \\ \frac{x-l}{m-l}, & l < x < m \\ \frac{h-x}{h-m}, & m < x < h \\ 0, & x \geq h \end{cases}$$

The notation  $\omega(x)_i$  determines the membership degree of element  $x$  to fuzzy set  $i$ , where  $l$  and  $h$  are lower and upper limit, and  $m$  is average of the two limits.

Relationship rules between the different categories are generated based on the dataset. For example,  $A_5 \rightarrow A_6$  means that the determinant of  $A_5$  values result in  $A_6$ . The forecasting uses these generated rules. These rules are determined by the relationships between the data points. There are two main factors that affect these forecasting results. These are the order of application and the method. According to the results in Table 5, for the time series analysis of this study, we used order 1 and probability weighted higher order fuzzy time series (PWHOFTS) for training. We fuzzed the prediction input section and compared the prediction results with those of the un-fuzzed data. Table 6 shows different metrics for different univariate methods such as RMSE (root mean square error), MAPE (mean absolute percentage error) and U (uncertainty coefficient). We fuzzed the number of tourists and then used machine learning methods to make predictions, comparing the prediction results with those of the un-fuzzed data.

#### 4.2.4 LSTM and FTS-LSTM algorithms

The forecasting results of LSTM and FTS-LSTM are displayed in Tables 7, 8, 9, and 10. When the dimensionality was 5, in Table 7, the average mean absolute error (MAE) for a 1-day forecast was approximately 0.1, except for t-SNE (0.237) and LE (0.389). The predictions were best for KPCA. The average MAE of the 7-days forecast was below 1.0, except for LE of 1.539 and LLE of 1.069. KPCA produced the best prediction results. The average MAE of the 10-days forecast was between 1.0 and 2.0, except for LE, which was 2.226. The best forecasts came from KPCA. In Table 9, the average MAE of 1-day forecasts were all around 0.1, except for ICA (0.219), t-SNE (0.247) and LE (0.337). The best forecasts were from KPCA and

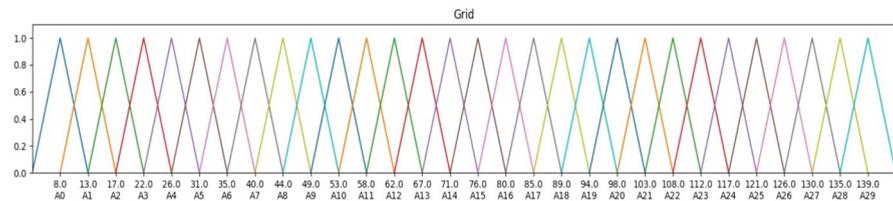


Fig. 9 Grid partition scheme with triangular membership function

**Table 6** Comparison between different methods and order

Method	Order	RMSE	MAPE	<i>U</i>
HOFTS	1	4.09	5.04	1.93
	2	3.94	4.75	1.86
	3	3.8	4.49	1.79
WHOFTS	1	2.5	2.53	1.18
	2	2.46	2.36	1.16
	3	2.44	2.3	1.15
PWHOFTS	<b>1</b>	<b>2.19</b>	<b>1.56</b>	<b>1.03</b>
	2	2.3	1.7	1.08
	3	2.37	1.87	1.12

Bold values indicate optimal results

IspMap. The expected 7-days average MAE was less than 1.0, except for t-SNE (1.022), LE (1.474) and LLE (1.114). The best forecasts were from KPCA and IspMap. The anticipated 10-days average MAE varied from 1.0 to 2.0 except for the LE of 2.240. The best forecasts were from KPCA and IspMap.

When the dimensionality is 10, the average MAE for the 1-day forecast in Table 8 was roughly 0.2, except for LE (0.513) and t-SNE (0.308). The forecasts for LPP were the best. The average MAE over the 7-days of forecasting was all close to 1.0. The best result performance was 0.695 for the LLE. The average MAE for the 10-days of forecasting was close to 1.0. The best result was 1.017 for IsoMap. In Table 10, the estimated average 1-day MAE was around 0.2. LLE and LPP had the best outcomes at 0.182. The projected 7-days mean MAE is approximately 1.0. The best performance was the LLE at 0.721. The anticipated 10-days average MAE was approximately 1.0, with IsoMap producing the best result at 1.031.

#### 4.2.5 RF and FTS-RF algorithms

The forecasting results of RF and FTS-RF are displayed in Tables 11, 12, 13 and 14. When the dimension is 5, the average MAE of the forecasted 1-day in Table 11 was approximately 0.1, with the best result for MDS at 0.184. The average MAE of the forecasted 7-days was approximately 1.0, while the best outcome was LLE at 0.996. The anticipated 10-days MAE averaged around 1.2, with TSVD achieving the most significant result at 1.257. In Table 13, the average MAE of the anticipated 1-day was approximately 0.1, with t-SNE producing the most significant result (0.187). The average MAE for the 7-days projected was approximately 1.0, while the best outcome was LLE at 0.997. The projected 10-days MAE averaged around 1.2, but the best outcome was LLE at 1.246.

When the dimension is 10, in Table 12, the average MAE of the forecasted 1-day was around 0.2. The best result was ICA with 0.180. The average MAE for the 7-days forecast was almost 1.0, with IsoMap producing the best result at 0.948. The anticipated 10-days MAE averaged approximately 1.2, with t-SNE achieving the most significant result at 1.193. In Table 14, the average MAE of the forecasted

**Table 7** Performance evaluation of Beijing tourist volume forecasting results using LSTM algorithm (components = 5)

Feature extraction algorithms	$n = 1$			$n = 7$			$n = 10$					
	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE
PCA	0.091	0.167	0.162	0.238	0.600	0.739	0.756	0.943	0.919	1.094	1.070	1.325
MDS	0.131	0.162	0.184	0.245	0.524	0.665	0.743	0.874	0.735	1.002	0.984	1.289
LDA	0.124	0.155	0.190	0.232	0.743	0.873	0.915	1.085	1.120	1.337	1.366	1.680
ICA	0.111	0.156	0.185	0.234	0.669	0.928	0.856	1.162	0.974	1.567	1.217	1.870
t-SNE	0.136	0.237	0.212	0.311	0.661	0.955	0.855	1.154	1.008	1.308	1.190	1.513
LE	0.143	0.389	0.211	0.497	1.095	1.539	1.283	1.794	1.529	2.226	1.893	2.557
LLE	0.091	0.158	0.164	0.226	0.682	1.069	0.834	1.214	0.750	1.319	0.956	1.535
LPP	0.096	0.160	0.167	0.217	0.712	0.957	0.928	1.169	1.111	1.493	1.402	1.798
TSVD	0.091	0.140	0.166	0.208	0.492	0.659	0.675	0.872	0.696	0.907	0.905	1.156
IsoMap	0.132	0.161	0.199	0.219	0.596	0.656	0.773	0.848	0.774	0.913	0.869	1.150
KPCA	<b>0.093</b>	<b>0.115</b>	<b>0.166</b>	<b>0.187</b>	<b>0.466</b>	<b>0.500</b>	<b>0.635</b>	<b>0.678</b>	<b>0.596</b>	<b>0.652</b>	<b>0.815</b>	<b>0.868</b>

Bold values indicate optimal results

**Table 8** Performance evaluation of Beijing tourist volume forecasting results using LSTM algorithm (components = 10)

Feature extraction algorithms	n = 1			n = 7			n = 10					
	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE
PCA	0.138	0.195	0.215	0.271	0.606	0.725	0.799	0.928	0.979	1.148	1.252	1.433
MDS	0.130	0.214	0.200	0.311	0.791	1.068	1.034	1.401	1.178	1.311	1.254	1.655
LDA	0.150	0.186	0.220	0.266	0.621	0.907	0.820	1.146	0.932	1.130	1.166	1.408
ICA	0.125	0.167	0.193	0.235	0.718	0.881	0.957	1.121	0.723	1.033	0.927	1.305
t-SNE	0.212	0.308	0.282	0.373	0.764	1.020	0.947	1.215	0.928	1.307	1.124	1.561
LE	0.309	0.513	0.411	0.628	1.139	1.584	1.323	1.828	1.322	1.638	1.626	1.957
LLE	0.128	0.200	0.191	0.269	<b>0.482</b>	<b>0.695</b>	<b>0.625</b>	<b>0.858</b>	0.940	1.288	1.065	1.480
LPP	<b>0.113</b>	<b>0.138</b>	<b>0.175</b>	<b>0.203</b>	0.673	0.933	0.860	1.171	0.773	1.080	0.979	1.314
TSVD	0.150	0.199	0.219	0.267	0.596	0.867	0.761	1.102	1.040	1.203	1.298	1.507
IsoMap	0.141	0.190	0.225	0.260	0.719	0.925	0.904	1.181	<b>0.777</b>	<b>1.017</b>	<b>0.979</b>	<b>1.252</b>
KPCA (9)	0.122	0.142	0.191	0.217	0.540	0.588	0.708	0.768	0.705	0.772	0.957	0.995

Bold values indicate optimal results

**Table 9** Performance evaluation of Beijing tourist volume forecasting results using FTS-LSTM algorithm (components = 5)

Feature extraction algorithms	$n = 1$				$n = 7$				$n = 10$			
	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE
PCA	0.116	0.159	0.184	0.228	0.620	0.740	0.767	0.927	0.999	1.219	1.208	1.471
MDS	0.108	0.180	0.176	0.248	0.521	0.768	0.705	1.005	0.831	1.114	1.090	1.422
LDA	0.129	0.168	0.198	0.243	0.723	0.932	0.939	1.189	0.981	1.312	0.127	1.657
ICA	0.162	0.219	0.270	0.306	0.856	0.996	1.015	1.197	1.164	1.590	1.417	1.927
t-SNE	0.121	0.247	0.178	0.308	0.651	1.022	0.829	1.221	0.959	1.293	0.117	1.546
LE	0.195	0.337	0.250	0.468	0.999	1.474	0.137	1.700	1.566	2.240	1.731	2.579
LLE	0.114	0.167	0.189	0.240	0.587	1.114	0.733	1.281	1.056	1.507	1.228	1.720
LPP	0.097	0.162	0.160	0.220	0.774	0.965	0.939	1.171	1.163	1.456	1.357	1.803
TSVD	0.115	0.158	0.181	0.219	0.502	0.732	0.674	0.950	0.725	0.942	0.911	1.182
ISOMAP	0.100	0.153	0.166	0.220	0.487	0.653	0.652	0.835	0.698	0.803	0.909	1.017
KPCA	<b>0.100</b>	<b>0.121</b>	<b>0.175</b>	<b>0.193</b>	<b>0.440</b>	<b>0.532</b>	<b>0.612</b>	<b>0.721</b>	<b>0.633</b>	<b>0.745</b>	<b>0.855</b>	<b>0.970</b>

Bold values indicate optimal results

**Table 10** Performance evaluation of Beijing tourist volume forecasting results using FTS-LSTM algorithm (components = 10)

Feature extraction algorithms	<i>n</i> = 1			<i>n</i> = 7			<i>n</i> = 10					
	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE
	PCA	0.130	0.212	0.206	0.284	0.656	0.808	0.927	1.028	0.865	1.211	1.103
MDS	0.131	0.228	0.212	0.331	0.701	1.058	0.926	1.438	0.951	1.293	1.229	1.634
LDA	0.152	0.183	0.227	0.260	0.746	0.942	0.966	1.199	0.945	1.229	1.191	1.517
ICA	0.148	0.185	0.214	0.251	0.650	0.855	0.849	1.088	0.876	1.144	1.092	1.454
t-SNE	0.202	0.264	0.270	0.331	0.890	1.421	1.144	1.723	1.000	1.548	1.228	1.832
LE	0.370	0.797	0.453	0.985	0.980	1.395	1.174	1.645	1.435	1.854	1.748	2.158
LLE	<b>0.118</b>	<b>0.182</b>	<b>0.182</b>	<b>0.248</b>	<b>0.581</b>	<b>0.721</b>	<b>0.748</b>	<b>0.895</b>	0.819	1.390	1.017	1.589
LPP	<b>0.110</b>	<b>0.182</b>	<b>0.170</b>	<b>0.239</b>	0.727	0.818	0.900	1.020	1.019	1.189	1.253	1.437
TSVD	0.172	0.225	0.241	0.299	0.745	0.959	0.974	1.233	0.911	1.243	1.200	1.557
ISOMAP	0.156	0.195	0.228	0.255	0.686	0.803	0.792	0.990	<b>0.880</b>	<b>1.031</b>	<b>0.981</b>	<b>1.257</b>
KPCA (9)	0.126	0.144	0.198	0.219	0.481	0.567	0.648	0.751	0.652	0.740	0.861	0.969

Bold values indicate optimal results

**Table 11** Performance evaluation of Beijing tourist volume forecasting results using RF algorithm (components = 5)

Feature extraction algorithms	$n = 1$				$n = 7$				$n = 10$			
	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE
	PCA	0.005	0.191	0.005	0.550	0.522	1.023	0.645	1.358	0.595	1.301	0.819
MDS	<b>0.046</b>	<b>0.184</b>	<b>0.004</b>	<b>0.533</b>	0.471	1.016	0.681	1.364	0.647	1.287	0.877	1.643
LDA	0.051	0.189	0.005	0.558	0.492	1.011	0.665	1.361	0.598	1.284	0.867	1.643
ICA	0.053	0.187	0.004	0.555	0.440	1.014	0.626	1.342	0.599	1.264	0.863	1.633
t-SNE	0.049	0.186	0.004	0.531	0.469	1.011	0.588	1.367	0.604	1.265	0.805	1.615
LE	0.044	0.214	0.003	0.680	0.442	1.024	0.662	1.327	0.662	1.301	0.893	1.633
LLE	0.047	0.202	0.005	0.590	<b>0.495</b>	<b>0.996</b>	<b>0.607</b>	<b>1.258</b>	0.689	1.262	0.892	1.535
LPP	0.054	0.194	0.004	0.574	0.533	1.022	0.672	1.343	0.631	1.268	0.894	1.630
TSVD	0.047	0.202	0.004	0.578	0.502	1.006	0.619	1.327	<b>0.615</b>	<b>1.257</b>	<b>0.887</b>	<b>1.630</b>
ISOMAP	0.052	0.198	0.004	0.576	0.413	1.002	0.607	1.333	0.544	1.268	0.745	1.612
KPCA	0.052	0.194	0.006	0.565	0.425	1.018	0.635	1.346	0.607	1.274	0.868	1.655

Bold values indicate optimal results

**Table 12** Performance evaluation of Beijing tourist volume forecasting results using RF algorithm (components = 10)

Feature extraction algorithms	n = 1			n = 7			n = 10					
	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE
PCA	0.057	0.193	0.006	0.565	0.495	1.005	0.604	1.334	0.603	1.250	0.858	1.602
MDS	0.056	0.202	0.007	0.591	0.503	1.005	0.635	1.330	0.657	1.247	0.854	1.595
LDA	0.049	0.194	0.006	0.535	0.466	0.993	0.614	1.329	0.621	1.261	0.859	1.608
ICA	<b>0.059</b>	<b>0.180</b>	<b>0.005</b>	<b>0.533</b>	0.450	0.993	0.606	1.325	0.623	1.238	0.835	1.597
t-SNE	0.049	0.195	0.005	0.539	0.584	0.978	0.680	1.289	<b>0.629</b>	<b>1.193</b>	<b>0.778</b>	<b>1.526</b>
LE	0.049	0.215	0.005	0.611	0.447	0.957	0.610	1.265	0.570	1.263	0.762	1.584
LLE	0.052	0.201	0.005	0.589	0.392	1.111	0.564	1.478	0.570	1.264	0.819	1.575
LPP	0.055	0.200	0.007	0.569	0.447	0.996	0.629	1.315	0.606	1.253	0.843	1.587
TSVD	0.053	0.198	0.005	0.577	0.466	1.009	0.581	1.316	0.589	1.244	0.793	1.589
ISOMAP	0.052	0.202	0.006	0.581	<b>0.427</b>	<b>0.948</b>	<b>0.584</b>	<b>1.274</b>	0.650	1.204	0.817	1.540
KPCA (9)	0.051	0.192	0.005	0.561	0.434	1.001	0.565	1.314	0.561	1.265	0.819	1.605

Bold values indicate optimal results

**Table 13** Performance evaluation of Beijing tourist volume forecasting results using FTS-RF algorithm (components = 5)

Feature extraction algorithms	<i>n</i> = 1			<i>n</i> = 7			<i>n</i> = 10					
	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE
	PCA	0.050	0.198	0.018	0.574	0.507	0.644	1.355	0.612	1.277	0.871	1.653
MDS	0.056	0.193	0.005	0.567	0.490	0.669	1.363	0.649	1.279	0.867	1.632	
LDA	0.052	0.190	0.005	0.562	0.479	0.660	1.341	0.622	1.283	0.903	1.642	
ICA	0.050	0.197	0.006	0.561	0.434	0.641	1.324	0.575	1.270	0.841	1.630	
t-SNE	<b>0.052</b>	<b>0.187</b>	<b>0.005</b>	<b>0.569</b>	0.494	0.577	1.357	0.608	1.261	0.798	1.614	
LE	0.051	0.214	0.004	0.615	0.438	0.667	1.330	0.630	1.293	0.916	1.626	
LLE	0.046	0.205	0.004	0.585	<b>0.527</b>	<b>0.654</b>	<b>1.277</b>	<b>0.720</b>	<b>1.246</b>	<b>0.939</b>	<b>1.536</b>	
LPP	0.052	0.197	0.005	0.572	0.464	0.691	1.343	0.633	1.279	0.872	1.634	
TSVD	0.049	0.197	0.004	0.572	0.472	0.629	1.324	0.649	1.278	0.876	1.624	
IsoMap	0.045	0.201	0.004	0.574	0.424	0.601	1.327	0.526	1.260	0.763	1.618	
KPCA	0.056	0.197	0.006	0.557	0.438	0.614	1.352	0.635	1.275	0.841	1.654	

Bold values indicate optimal results

**Table 14** Performance evaluation of Beijing tourist volume forecasting results using FTS-RF algorithm (components = 10)

Feature extraction algorithms	<i>n</i> = 1			<i>n</i> = 7			<i>n</i> = 10					
	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE	Optimal MAE	Average MAE	Optimal RMSE	Average RMSE
	PCA	0.056	0.196	0.006	0.557	0.485	1.008	0.617	1.320	0.595	1.265	0.849
MDS	0.059	0.204	0.007	0.577	0.500	1.001	0.726	1.310	0.647	1.248	0.841	1.576
LDA	0.058	0.194	0.006	0.537	0.477	0.991	0.618	1.331	0.613	1.241	0.850	1.599
ICA	<b>0.052</b>	<b>0.193</b>	<b>0.004</b>	<b>0.548</b>	0.463	0.999	0.621	1.321	0.602	1.228	0.805	1.592
t-SNE	0.052	0.199	0.006	0.541	<b>0.531</b>	<b>0.962</b>	<b>0.660</b>	<b>1.277</b>	<b>0.660</b>	<b>1.187</b>	<b>0.836</b>	<b>1.543</b>
LE	0.051	0.222	0.004	0.620	0.443	0.973	0.609	1.265	0.553	1.275	0.754	1.569
LLE	0.050	0.204	0.005	0.594	0.395	1.095	0.600	1.505	0.588	1.239	0.778	1.574
LPP	0.057	0.201	0.007	0.567	0.429	1.002	0.638	1.321	0.610	1.251	0.836	1.601
TSVD	0.060	0.196	0.007	0.582	0.462	1.005	0.592	1.307	0.610	1.240	0.790	1.585
IsoMap	0.049	0.200	0.005	0.577	0.446	0.968	0.562	1.284	0.609	1.199	0.837	1.526
KPCA (9)	0.050	0.190	0.006	0.577	0.417	0.982	0.562	1.313	0.569	1.250	0.818	1.613

Bold values indicate optimal results

1-day was around 0.2. The best result was ICA with 0.193. The average MAE of the forecasted 7-days was almost 1.0, and the best result was t-SNE with 0.962. The anticipated 10-days MAE averaged approximately 1.2, with t-SNE achieving the most significant result at 1.187.

### 4.3 Discussion

#### 4.3.1 Results analysis

As the prediction period increases, the forecasting accuracy of all feature extraction algorithms gradually decreases. Comparing the prediction results for dimension 5 and dimension 10 demonstrates that the results for dimension 5 are better.

The machine learning model's 1-day forecasting for the search data was similar to those of the ARIMA and SARIMA models using tourism headcount data. The results of 7-days forecasts based on search data and the machine learning model mostly outperformed those of the ARIMA and SARIMA models. Forecasts using the search data with the machine learning model did not give as good a prediction as the ARIMA and SARIMA models when forecasting 10-days.

Machine learning models using search data for forecasting can outperform ARIMA and SARIMA models using tourism data for short-term (1-day, 7-days) forecasting but slightly underperform ARIMA and SARIMA models for medium-term (10-days) forecasting. This may be owing to the inherent complexity of the search data and the cyclical nature of passenger traffic.

#### 4.3.2 Methods analysis

Due to the speed and accuracy of web search data, its application to prediction problems in the tourism industry has become a popular research topic in recent years. Google, Baidu and other large search engines log users' search content, frequency, and location in real-time, creating structured data for Google Trends or Baidu Index. The quantity and kind of terms entered by users into search engines provide indirect information on their travel needs, interests, and level of concern for tourism destinations. Experiments have demonstrated that search engine keyword data can be utilised to forecast tourism demand.

Additionally, this paper collects keywords for various tourism-related characteristics and classifies them to include as many tourism-related aspects as feasible. The results demonstrate that our keyword selection and processing contribute to tourism forecasting. Therefore, it is possible to employ many facets of tourism as points of data collecting.

Given the sheer volume of search engine data and its wealth of information, researchers pick and minimise terms to retrieve relevant data for accurate prediction. In conventional linear techniques, however, nonlinear correlations and interactions between variables are frequently missed. The ARIMA and SARIMA models employed capture only linear and not nonlinear interactions. Establishing nonlinear

correlations between independent and dependent variables in a complex system environment allows for a more thorough examination of nonlinear impacts between variables through machine learning techniques.

In previous experiments, researchers have often reduced the dimensionality of the data using PCA alone. This study compares the effectiveness of 11 distinct feature extraction algorithms in conjunction with machine learning models for prediction. The results indicate that some feature extraction algorithms can predict the same or better outcomes than the PCA method. Moreover, when the dimensionality is reduced, the prediction result for dimension 5 is superior to dimension 10.

Lastly, comparable data for the work presented in this paper may have been acquired from other platforms, and the concerns addressed in the study vary by region. This paper's keyword data collection is based on the Baidu index platform in China, while data collection for other countries might be based on Google Trends. The analysis tools utilised in this paper are based on the Python programming language, and the experimental tool selected is Pycharm 2020. All of the utilised software libraries are Python's own.

## 5 Conclusions

This study proposed a systematic tourism keyword selection and processing strategy for the tourism industry, comprising several components (transportation, attraction, food, lodging, travel, tips, tickets, and weather). PCA, MDS, LDA, ICA, t-SNE, LE, LLE, LPP, TSVD, IsoMap, and KPCA extract keywords into 5 or 10 dimensions. At the same time, RF, LSTM, and FTS are used to compare the forecasting results of various feature extraction algorithms. Our experiment performs well in terms of short-term forecasting, although the results are slightly worse than those of the ARIMA and SARIMA models regarding medium-term forecasting. Overall, some feature extraction techniques can yield similar forecasting results, and some methods even achieve better forecast results than the prevalent PCA method.

Numerous tests have demonstrated that online big data is a significant and influential source for forecasting tourist volume, and this research provides additional evidence of the conclusion's validity. Based on a survey of the relevant literature, implementing various feature extraction algorithms in forecasting tourist volume remains uncommon. In conclusion, these findings indicate that selecting feature extraction algorithms carefully can enhance the accuracy of visitor volume forecasts. Future frameworks for the selection and processing of keywords will be more efficient.

Our work is not limitless. Feature selection and extraction methods are included in dimension reduction algorithms. In order to include as many data characteristics as feasible, this research focuses on studying and comparing feature extraction strategies rather than feature selection algorithms. In future research, the use of feature selection algorithms may be investigated.

This investigation is exploratory. Our research aims to demonstrate that different feature extraction methods can be considered in selecting and processing tourist

keywords to provide insight into the selection and processing of tourism keywords and serve as a foundation for future study. Additional research can be undertaken using several case studies to get more general conclusions.

**Acknowledgements** The authors would like to thank the anonymous reviewers and the editor for their helpful comments and suggestions. This work has been supported by the Major Cultivation Project of Education Department in Sichuan Province, China (18CZ0006).

**Data availability** The data supporting the results in this research and other findings of this study are available upon reasonable request from the corresponding author.

## Declarations

**Conflict of interest** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

- Aidoo EN, Appiah SK, Awashie GE, Boateng A, Darko G (2021) Geographically weighted principal component analysis for characterising the spatial heterogeneity and connectivity of soil heavy metals in Kumasi Ghana. *Heliyon* 7(9):e08039. <https://doi.org/10.1016/j.heliyon.2021.e08039>
- Anowar F, Sadaoui S, Selim B (2021) Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Comput Sci Rev* 40:100378. <https://doi.org/10.1016/j.cosrev.2021.100378>
- Arena P, Patanè L, Spinosa AG (2020) Robust modelling of binary decisions in Laplacian Eigenmaps-based Echo State Networks. *Eng Appl Artif Intell* 95(July):103828. <https://doi.org/10.1016/j.engpai.2020.103828>
- Balakrishnama S, Ganapathiraju A (1998) Linear discriminant analysis—a brief tutorial. *Inst Signal Inform Process* 18(1998):1–8
- Barborica A, Mindruta I, Sheybani L, Spinelli L, Oane I, Pistol C, Donos C, López-Madrona VJ, Vulliemzo S, Bénar CG (2021) Extracting seizure onset from surface EEG with independent component analysis: insights from simultaneous scalp and intracerebral EEG. *NeuroImage: Clin.* <https://doi.org/10.1016/j.nicl.2021.102838>
- Belkin M, Niyogi P (2003) Laplacian Eigenmaps for dimensionality reduction and data representation. *Neural Comput* 15(6):1373–1396. <https://doi.org/10.1162/089976603321780317>
- Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. *IEEE Trans Neural Netw* 5(2):157–166
- Bi Q, Huang N, Zhang S, Shuai C, Wang Y (2019) Adaptive machining for curved contour on deformed large skin based on on-machine measurement and isometric mapping. *Int J Mach Tools Manuf* 136:34–44. <https://doi.org/10.1016/j.ijmachtools.2018.09.001>
- Bi JW, Liu Y, Li H (2020) Daily tourism volume forecasting for tourist attractions. *Ann Tour Res* 83:102923. <https://doi.org/10.1016/j.annals.2020.102923>
- Bokelmann B, Lessmann S (2019) Spurious patterns in Google Trends data—an analysis of the effects on tourism demand forecasting in Germany. *Tour Manag* 75(February):1–12. <https://doi.org/10.1016/j.tourman.2019.04.015>
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Buzzini P, Curran J, Polston C (2021) Comparison between visual assessments and different variants of linear discriminant analysis to the classification of Raman patterns of inkjet printer inks. *Forensic Chem* 24(March):100336. <https://doi.org/10.1016/j.forc.2021.100336>
- Cattani L, Maillet D, Bozzoli F, Rainieri S (2015) Estimation of the local convective heat transfer coefficient in pipe flow using a 2D thermal quadrupole model and truncated singular value decomposition. *Int J Heat Mass Transf* 91:1034–1045. <https://doi.org/10.1016/j.ijheatmasstransfer.2015.08.016>

- Cevik S (2020) Where should we go? Internet searches and tourist arrivals. *Int J Financ Econ*. <https://doi.org/10.1002/ijfe.2358>
- Chen Z, Qin L, Zhao S, Chan THT, Nguyen A (2019) Toward efficacy of piecewise polynomial truncated singular value decomposition algorithm in moving force identification. *Adv Struct Eng* 22(12):2687–2698. <https://doi.org/10.1177/1369433219849817>
- Cieslak MC, Castelfranco AM, Roncalli V, Lenz PH, Hartline DK (2020) t-distributed stochastic neighbor embedding (t-SNE): a tool for eco-physiological transcriptomic analysis. *Mar Genomics* 51(September):100723. <https://doi.org/10.1016/j.margen.2019.100723>
- Clark M, Wilkins EJ, Dagan DT, Powell R, Sharp RL, Hillis V (2019) Bringing forecasting into the future: using Google to predict visitation in US national parks. *J Environ Manag* 243(February):88–94. <https://doi.org/10.1016/j.jenvman.2019.05.006>
- Cui J, Shen BW (2021) A kernel principal component analysis of coexisting attractors within a generalized Lorenz model. *Chaos, Solitons Fractals* 146:110865. <https://doi.org/10.1016/j.chaos.2021.110865>
- Cuomo MT, Tortora D, Foroudi P, Giordano A, Festa G, Metallo G (2021) Digital transformation and tourist experience co-design: big social data for planning cultural tourism. *Technol Forecast Soc Change* 162(June):120345. <https://doi.org/10.1016/j.techfore.2020.120345>
- Dergiades T, Mavragani E, Pan B (2018) Google Trends and tourists' arrivals: emerging biases and proposed corrections. *Tour Manage* 66:108–120. <https://doi.org/10.1016/j.tourman.2017.10.014>
- Duarte P, Riveros-Perez E (2021) Understanding the cycles of COVID-19 incidence: principal component analysis and interaction of biological and socio-economic factors. *Ann Med Surg* 66(June):102437. <https://doi.org/10.1016/j.amsu.2021.102437>
- Feng Y, Li G, Sun X, Li J (2019) Forecasting the number of inbound tourists with Google Trends. *Procedia Comput Sci* 162(Ictqm):628–633. <https://doi.org/10.1016/j.procs.2019.12.032>
- Gao S, Sheng Y (2021) Research on Kaifeng Tourism demand modeling and forecasting based on Baidu Index. *Stat Theory Pract* (11):44–49. <https://kns-cnki-net-443.vpn.sicnu.edu.cn/kcms/detail/detail.aspx?FileName=TJLS202111004&DbName=CJFQ2021>
- Good WW, Erem B, Zenger B, Coll-Font J, Bergquist JA, Brooks DH, MacLeod RS (2020) Characterizing the transient electrocardiographic signature of ischemic stress using Laplacian Eigenmaps for dimensionality reduction. *Comput Biol Med* 127:104059. <https://doi.org/10.1016/j.combiomed.2020.104059>
- Graves A (2013) Generating sequences with recurrent neural networks. 1–43
- Greff K, Srivastava RK, Koutnik J, Steunebrink BR, Schmidhuber J (2017) LSTM: a search space odyssey. *IEEE Trans Neural Netw Learn Syst* 28(10):222–2232. <https://doi.org/10.1109/TNNLS.2016.2582924>
- Han Y, Liu S, Cong D, Geng Z, Fan J, Gao J, Pan T (2021) Resource optimization model using novel extreme learning machine with t-distributed stochastic neighbor embedding: application to complex industrial processes. *Energy* 225:120255. <https://doi.org/10.1016/j.energy.2021.120255>
- He X, Niyogi P (2010) Locality preserving projections. *Neural Inform Process Syst* 16:153
- He F, Wang C, Fan SKS (2018) Nonlinear fault detection of batch processes based on functional kernel locality preserving projections. *Chemom Intell Lab Syst* 183(May):79–89. <https://doi.org/10.1016/j.chemolab.2018.10.010>
- He K, Ji L, Wu CWD, Tso KFG (2021) Using SARIMA–CNN–LSTM approach to forecast daily tourism demand. *J Hosp Tour Manag* 49(September):25–33. <https://doi.org/10.1016/j.jhtm.2021.08.022>
- Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780
- Hoque N, Bhattacharyya DK, Kalita JK (2014) MIFS-ND: a mutual information-based feature selection method. *Expert Syst Appl* 41(14):6371–6385. <https://doi.org/10.1016/j.eswa.2014.04.019>
- Hout MC, Papeš MH, Goldinger SD (2013) Multidimensional scaling. *Wiley Interdiscip Rev Cogn Sci* 4(1):93–103. <https://doi.org/10.1002/wcs.1203>
- Huang X, Zhang L, Ding Y (2017) The Baidu Index: uses in predicting tourism flows—a case study of the forbidden city. *Tour Manag* 58:301–306. <https://doi.org/10.1016/j.tourman.2016.03.015>
- Huang W, Cao B, Yang G, Luo N, Chao N (2021) Turn to the internet first? Using online medical behavioral data to forecast COVID-19 Epidemic trend. *Inf Process Manag* 58(3):102486. <https://doi.org/10.1016/j.ipm.2020.102486>
- Kang J-f, Guo X-Y, Fang L (2020) Tourism trend forecasting based on Baidu index spatial and temporal distribution. *J Southwest Normal Univ (nat Sci Edition)* 45(10):72–81. <https://doi.org/10.13718/j.cnki.xsxb.2020.10.012>

- Kaya K, Yılmaz Y, Yaslan Y, Ögüdücü ŞG, Çıngır F (2022) Demand forecasting model using hotel clustering findings for hospitality industry. *Inf Process Manag* 59(1):102816. <https://doi.org/10.1016/j.ipm.2021.102816>
- Khalid S, Khalil T, Nasreen S (2014) A survey of feature selection and feature extraction techniques in machine learning. In: *Proceedings of 2014 Science and Information Conference, SAI 2014*, October 2016, pp 372–378. <https://doi.org/10.1109/SAI.2014.6918213>
- Kuang L, Yang LT, Chen J, Hao F, Luo C (2018) A holistic approach for distributed dimensionality reduction of big data. *IEEE Trans Cloud Comput* 6(2):506–518. <https://doi.org/10.1109/TCC.2015.2449855>
- Law R, Li G, Fong DKC, Han X (2019) Tourism demand forecasting: a deep learning approach. *Ann Tour Res* 75(January):410–423. <https://doi.org/10.1016/j.annals.2019.01.014>
- Li Y, Cao H (2018) Prediction for tourism flow based on LSTM neural network. *Procedia Comput Sci* 129:277–283. <https://doi.org/10.1016/j.procs.2018.03.076>
- Li X, Pan B, Law R, Huang X (2017) Forecasting tourism demand with composite search index. *Tour Manag* 59:57–66. <https://doi.org/10.1016/j.tourman.2016.07.005>
- Li J, Xu L, Tang L, Wang S, Li L (2018a) Big data in tourism research: a literature review. *Tour Manag* 68:301–323. <https://doi.org/10.1016/j.tourman.2018.03.009>
- Li S, Chen T, Wang L, Ming C (2018b) Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu Index. *Tour Manag* 68:116–126. <https://doi.org/10.1016/j.tourman.2018.03.006>
- Li H, Hu M, Li G (2020a) Forecasting tourism demand with multisource big data. *Ann Tour Res* 83(March):102912. <https://doi.org/10.1016/j.annals.2020.102912>
- Li Y, Zenggang X, Ailing C (2020b) A method of predicting tourist flow based on multi-scale combination. *Statistics and Decision-making*, 36(22):177–180. <https://kns-cnki-net-443.vpn.sicnu.edu.cn/kcms/detail/detail.aspx?FileName=TJJC2020b22041&DbName=DKFX2020b>
- Li C-N, Qi Y-F, Shao Y-H, Guo Y-R, Ye Y-F (2021a) Robust two-dimensional capped l2, 1-norm linear discriminant analysis with regularization and its applications on image recognition. *Eng Appl Artif Intell* 104(June):104367. <https://doi.org/10.1016/j.engappai.2021.104367>
- Li X, Law R, Xie G, Wang S (2021b) Review of tourist volume forecasting research with internet data. *Tour Manag* 83(August):104245. <https://doi.org/10.1016/j.tourman.2020.104245>
- Li X, Li H, Pan B, Law R (2021c) Machine learning in internet search query selection for tourist volume forecasting. *J Travel Res* 60(6):1213–1231. <https://doi.org/10.1177/0047287520934871>
- Liu YY, Tseng FM, Tseng YH (2018) Big data analytics for forecasting tourism destination arrivals with the applied vector autoregression model. *Technol Forecast Soc Change* 130(December):123–134. <https://doi.org/10.1016/j.techfore.2018.01.018>
- Liu Q, He H, Liu Y, Qu X (2021) Local linear embedding algorithm of mutual neighborhood based on multi-information fusion metric. *Measurement* 186(May):110239. <https://doi.org/10.1016/j.measurement.2021.110239>
- Lu L, Liao X (2021) Construction of tourism search index from the perspective of selection domain and analysis of its forecasting effect: a case study of Mt.Siguniang. *J Central South Univ for Technol Soc Sci* 15(2):100–110. <https://doi.org/10.14067/j.cnki.1673-9272.2021.02.013>
- Lu W, Liu Z, Huang Y, Bu Y, Li X, Cheng Q (2020) How do authors select keywords? A preliminary study of author keyword selection behavior. *J Inform* 14(4):1–17. <https://doi.org/10.1016/j.joi.2020.101066>
- Lu W, Huang S, Yang J, Bu Y, Cheng Q, Huang Y (2021) Detecting research topic trends by author-defined keyword frequency. *Inform Process Manag*. <https://doi.org/10.1016/j.ipm.2021.102594>
- Luo N, Yang X, Sun C, Xing B, Han J, Zhao C (2021) Visualization of vibrational spectroscopy for agro-food samples using t-distributed stochastic neighbor embedding. *Food Control* 126:107812. <https://doi.org/10.1016/j.foodcont.2020.107812>
- Machado JT, Luchko Y (2021) Multidimensional scaling and visualization of patterns in distribution of nontrivial zeros of the zeta-function. *Commun Nonlinear Sci Numer Simul* 102:105924. <https://doi.org/10.1016/j.cnsns.2021.105924>
- Miao J, Yang T, Sun L, Fei X, Niu L, Shi Y (2022) Graph regularized locally linear embedding for unsupervised feature selection. *Pattern Recogn* 122:108299. <https://doi.org/10.1016/j.patcog.2021.108299>
- Moreira de Oliveira A, Alberto Teixeira C, Wang Hantao L (2021) Evaluation of the retention profile in flow-modulated comprehensive two-dimensional gas chromatography and independent component

- analysis of weathered heavy oils. *Microchem J* 172(PB):106978. <https://doi.org/10.1016/j.microc.2021.106978>
- Mousavi Nezhad M, Gironacci E, Rezanian M, Khalili N (2018) Stochastic modelling of crack propagation in materials with random properties using isometric mapping for dimensionality reduction of nonlinear data sets. *Int J Numer Methods Eng* 113(4):656–680. <https://doi.org/10.1002/nme.5630>
- Mu X, Chen L, Mikut R, Hahn H, Kübel C (2021) Unveiling local atomic bonding and packing of amorphous nanophases via independent component analysis facilitated pair distribution function. *Acta Mater* 212:116932. <https://doi.org/10.1016/j.actamat.2021.116932>
- Nahil A, Lyhyaoui A (2018) Short-Term stock price forecasting using kernel principal component analysis and support vector machines: the case of Casablanca stock Exchange. *Procedia Comput Sci* 127:161–169. <https://doi.org/10.1016/j.procs.2018.01.111>
- Natalia P, Clara RA, Simon D, Noelia G, Barbara A (2019) Critical elements in accessible tourism for destination competitiveness and comparison: principal component analysis from Oceania and South America. *Tour Manag* 75:169–185. <https://doi.org/10.1016/j.tourman.2019.04.012>
- Padhi SS, Pati RK (2017) Quantifying potential tourist behavior in choice of destination using Google Trends. *Tour Manag Perspect* 24:34–47. <https://doi.org/10.1016/j.tmp.2017.07.001>
- Park CW, Lee I, Kwon S-H, Son S-J, Ko D-K (2021) Classification of CARS spectral phase retrieval combined with principal component analysis. *Vib Spectrosc* 117(June):103314. <https://doi.org/10.1016/j.vibspec.2021.103314>
- Peng G, Liu Y, Wang J, Gu J (2017) Analysis of the forecasting capability of web search data based on the HE-TDC method—forecasting of the volume of daily tourism visitors. *J Syst Sci Syst Eng* 26(2):163–182. <https://doi.org/10.1007/s11518-016-5311-7>
- Ren H, Liu T, Kang J, Pan N, Li M-L, Ai S (2020a) A forecasting method of urban daily tourist scale based on Baidu Index. *J Zhejiang Univ (Nat Sci)*, 47(06):753–761. <https://kns-cnki-net-443.vpn.sicnu.edu.cn/kcms/detail/detail.aspx?FileName=HZDX2020a06014&DbName=CJFQ2020a>
- Ren H, Liu T, Kang J, P Ning, Li M-L, Ai S (2020b) A forecasting method of urban daily tourist scale based on Baidu Index. *J Zhejiang Univ (Nat Sci)*, 47(06):753–761. <https://kns-cnki-net-443.vpn.sicnu.edu.cn/kcms/detail/detail.aspx?FileName=HZDX2020b06014&DbName=CJFQ2020b>
- Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326. <https://doi.org/10.1126/science.290.5500.2323>
- Schölkopf B, Smola A, Müller KR (1997) Kernel principal component analysis. In: *International conference on artificial neural networks*. Springer, Berlin, Heidelberg, pp 583–588
- Seraine M, Campos JEG, Martins-Ferreira MAC, de Alvarenga CJS, Chemale F, Angelo TV, Spencer C (2021) Multi-dimensional scaling of detrital zircon geochronology constrains basin evolution of the late Mesoproterozoic Paranoá Group, central Brazil. *Precambrian Res*. <https://doi.org/10.1016/j.precamres.2021.106381>
- Shao Y (2019) Supervised global-locality preserving projection for plant leaf recognition. *Comput Electron Agric* 158(January):102–108. <https://doi.org/10.1016/j.compag.2019.01.022>
- Silverstovs B, Wochner DS (2018) Google Trends and reality: do the proportions match? Appraising the informational value of online search behavior: evidence from Swiss tourism regions. *J Econ Behav Organ* 145:1–23. <https://doi.org/10.1016/j.jebo.2017.10.011>
- Song Q, Chissom BS (1993) Forecasting enrollments with fuzzy time series—part I. *Fuzzy Sets Syst* 54(1):1–9
- Song Q, Chissom BS (1994) Forecasting enrollments with fuzzy time series—part II. *Fuzzy Sets Syst* 54(1):1–9
- Sun Y (2021) Forecast model construction of inbound tourism market based on seasonal ARIMA model. *J Nat Sci, Harbin Normal Univ* 37(04):56–60. <https://kns-cnki-net-443.vpn.sicnu.edu.cn/kcms/detail/detail.aspx?FileName=HEBY202104008&DbName=CJFQ2021>
- Sun S, Wei Y, Tsui KL, Wang S (2019) Forecasting tourist arrivals with machine learning and internet search index. *Tour Manag* 70:1–10. <https://doi.org/10.1016/j.tourman.2018.07.010>
- Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. *Adv Neural Inf Process Syst* 4(January):3104–3112
- Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9(11)
- Vu T, Chunikhina E, Raich R (2021) Perturbation expansions and error bounds for the truncated singular value decomposition. *Linear Algebra Appl* 627:94–139. <https://doi.org/10.1016/j.laa.2021.05.020>
- Wang L, Wang S, Yuan Z, Peng L (2021a) Analyzing potential tourist behavior using PCA and modified affinity propagation clustering based on Baidu Index: taking Beijing city as an example. *Data Sci Manag* 2:12–19. <https://doi.org/10.1016/j.dsm.2021.05.001>

- Wang X, Zhang Y, Yu B, Salhi A, Chen R, Wang L, Liu Z (2021b) Prediction of protein-protein interaction sites through eXtreme gradient boosting with kernel principal component analysis. *Comput Biol Med* 134(June):104516. <https://doi.org/10.1016/j.combiomed.2021.104516>
- Wei J, Cui H (2018) The construction of regional tourism index and its micro-dynamics characteristics: a case study of Xi'an. *J Syst Sci Math* 38(02):177–194. <https://kns-cnki-net-443.vpn.sicnu.edu.cn/kcms/detail/detail.aspx?FileName=STYS201802004&DbName=DKFX2018>
- Wei W, Dai H, Liang W (2020) Regularized least squares locality preserving projections with applications to image recognition. *Neural Netw* 128:322–330. <https://doi.org/10.1016/j.neunet.2020.05.023>
- Wei Y, Gu K, Tan L (2021) A positioning method for maize seed laser-cutting slice using linear discriminant analysis based on isometric distance measurement. *Inform Process Agric*. <https://doi.org/10.1016/j.inpa.2021.05.002>
- Westad F, Kermit M (2009) Independent component analysis. *Comprehensive chemometrics* (Second Edition, vol 2). Elsevier. <https://doi.org/10.1016/b978-0-444-64165-6.02006-1>
- Wold S, Esbensen K, Geladi P (1987) Principal component analysis. *Chemom Intell Lab Syst* 2(1–3):37–52
- Xie G, Qian Y, Wang S (2020) A decomposition-ensemble approach for tourist volume forecasting. *Ann Tour Res* 81:102891. <https://doi.org/10.1016/j.annals.2020.102891>
- Xie G, Qian Y, Wang S (2021) Forecasting Chinese cruise tourism demand with big data: an optimized machine learning approach. *Tour Manag* 82:104208. <https://doi.org/10.1016/j.tourman.2020.104208>
- Yang X, Yan H (2012) Analysis of DNA deformation patterns in nucleosome core particles based on isometric feature mapping and continuous wavelet transform. *Chem Phys Lett* 547:73–81. <https://doi.org/10.1016/j.cplett.2012.08.001>
- Yang X, Pan B, Evans JA, Lv B (2015) Forecasting Chinese tourist volume with search engine data. *Tour Manag* 46:386–397. <https://doi.org/10.1016/j.tourman.2014.07.019>
- Yao L, Ma R, Wang H (2021) Baidu Index-based forecast of daily tourist arrivals through rescaled range analysis, support vector regression, and autoregressive integrated moving average. *Alex Eng J* 60(1):365–372. <https://doi.org/10.1016/j.aej.2020.08.037>
- Yuan X, Liu Z, Wang Y, Xu Y, Zhang W, Mu T (2020) The non-negative truncated singular value decomposition for adaptive sampling of particle size distribution in dynamic light scattering inversion. *J Quant Spectrosc Radiat Transf* 246:106917. <https://doi.org/10.1016/j.jqsrt.2020.106917>
- Zebari R, Abdulazeez A, Zeebaree D, Zebari D, Saeed J (2020) A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *J Appl Sci Technol Trends* 1(2):56–70. <https://doi.org/10.38094/jastt1224>
- Zhang Y, Li G, Muskat B, Vu HQ, Law R (2021) Predictivity of tourism demand data. *Ann Tour Res* 89:103234. <https://doi.org/10.1016/j.annals.2021.103234>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.