Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Forecast daily tourist volumes during the epidemic period using COVID-19 data, search engine data and weather data

Chuan Zhang *, Yu-Xin Tian

School of Business Administration, Northeastern University, Shenyang 110169, China

ARTICLE INFO

Keywords:
Tourist volume forecasting
COVID-19 data
Search traffic data
Variational mode decomposition
Gated recurrent unit network

ABSTRACT

The COVID-19 epidemic has brought a devastating blow to the tourism industry. Affected by the epidemic situation, the change of tourism volume of scenic spots is very unstable. Therefore, forecasting tourist volume in the context of COVID-19 epidemic is a new and challenging problem. In response, a novel multivariate time series forecasting framework based on variational mode decomposition (VMD) and gated recurrent unit network (GRU), i.e., VMD-GRU, is proposed to forecast daily tourist volumes during the epidemic. It takes the lead in using COVID-19 data, search traffic data and weather data. Through sufficient experiments and comparisons, the superiority of the approach is illustrated, and the predictive power of the above three types of data, especially the COVID-19 data, is revealed. Accurate forecast results from the method can help relevant government officials and tourism practitioners to better adjust tourism resources, cooperate with anti-epidemic work and reduce operational risks.

1. Introduction

At the beginning of 2020, the COVID-19 epidemic broke out and swept across the world, causing a huge blow to the tourism industry and a sharp drop in tourism demand (Li et al., 2021). According to the report “Tourism Economic Operation Analysis in 2020 and Development Forecast in 2021” released by China Tourism Academy (https://eng.ctaweb.org.cn/), the number of tourists in China in 2020 was 2.88 billion, representing a year-on-year decrease of 52.1%.

As COVID-19 virus is easy to spread and mutates continuously, the human struggle with COVID-19 virus may be long and lasting. People attach great importance to epidemic prevention, and travel cautiously. Even if there are a few new confirmed cases, people’s travel decisions will be greatly changed. Therefore, it is a new problem to forecast tourism volumes during the COVID-19 epidemic, and the epidemic data can be used as an important predictive variable.

In this paper, we study the forecasting of short-term (daily) tourist volume of a specific attraction during the period of epidemic. The short-term tourist volume forecast is often more important than the long-term forecast for a relatively small area (Bi et al., 2020). Because it can better help tourism operators to make contingency plans, allocate tourism resources, design tourism packages or set appropriate prices (Divino & McAleer, 2010), so as to cope with the drastic fluctuations of tourist volumes caused by epidemic situation and peak demand, better cooperate with epidemic prevention and control, and reduce operational risks.

Considering the volatility and complexity of the tourist volume time series, we propose a novel forecasting approach that is composed of variational mode decomposition (VMD) and gated recurrent unit networks (GRUs), namely the VMD-GRU model. In this framework, VMD is used to decompose the non-stationary historical tourist volume data and obtain some intrinsic mode functions (IMFs) with certain periodic laws, such as off-season, peak season and holiday information, etc. The GRU network, as an advanced recurrent neural network, can effectively predict time series. Specifically, the advantages of this forecast architecture are embodied in the following two aspects.

Firstly, it has been found that for the prediction of non-stationary time series data, the signal decomposition method can obtain better forecasting results (Chen et al., 2012). Empirical mode decomposition (EMD) and its improved method are widely used in existing studies (Zhang et al., 2008; Chen et al., 2012; Carmelossi Furlaneto et al., 2017), and it can decompose a complex sequence into a limited and usually small number of IMFs according to different frequencies (Huang et al., 1998). These IMFs have obvious regularity and stronger correlation, so they can make the results of a time series forecasting model more accurate (Chen et al., 2012). However, the EMD method has limitations:

* Corresponding author.
E-mail address: czhang@mail.neu.edu.cn (C. Zhang).

https://doi.org/10.1016/j.eswa.2022.118505
Received 27 April 2022; Received in revised form 5 August 2022; Accepted 8 August 2022
Available online 12 August 2022
0957-4174/© 2022 Elsevier Ltd. All rights reserved.
mode mixing is easy to occur and there is no strict mathematical theory support. The state-of-the-art signal decomposition method, VMD (Dragomiretskiy & Zosso, 2014), is a novel variational method, which can non-recursively decompose a non-stationary signal into a given number of mode functions, and can avoid the occurrence of mode mixing. It is supported by strict mathematical theory and has been applied in the field of forecasting (Huang & Deng, 2021; Guo et al., 2022; Qian Zhang et al., 2022). Therefore, we use the VMD as the first part of the proposed forecasting framework to decompose the original time series and obtain multiple stable time series to improve the forecast ability of the framework.

Secondly, we use the GRU network, an artificial intelligence model for processing serial series data, as the main body of the forecasting framework proposed in this paper. Compared with the traditional multivariate time series analysis model, such as ARIMAX, the GRU network model has the strong ability of nonlinear relationship description. Among them, the long short-term memory network (LSTM) is popular and widely used in the task of tourist volume forecasting (Law et al., 2019; Bi et al., 2020; Kulshrestha et al., 2020; Laaroussi et al., 2020; Bi et al., 2021; Binru Zhang et al., 2020), because it solves the problem of gradient vanishing and gradient explosion in the traditional recurrent neural network (RNN) when learning long-term dependence. However, the gating network structure of a LSTM unit is too complex and redundant, which affects the learning efficiency, and redundant information may affect the performance of the model. In response, we adopt the GRU network (Chung et al., 2014), as the main body of the method to obtain more effective and accurate forecasting results.

In addition to historical tourist volume data, we select COVID-19 data, search traffic data and weather data as variables used in the proposed method. The rationality is based on people’s travel decision-making process during the COVID-19 epidemic period, as shown in Fig. 1. A few days before the trip, people usually consult epidemic data of both whole nation and the destination from news to know how serious it is. If the epidemic in the whole nation and the destination area are not serious, and the risk of contracting COVID-19 to is within an acceptable range, then they may consider the further travel plans to the attraction; Otherwise, they will abandon the next travel plans because of their own worries about infection and government restrictions on travel to prevent and control the epidemic. Therefore, it is reasonable to forecast tourist volumes of attractions using the national and destination newly confirmed cases data of COVID-19.

Once the above conditions permit, people often search for information about the destination spots on the internet (Yang et al., 2015; Li et al., 2018), such as transportation, hotels and travel guide. These operations will leave traces of queries, and then the search engine platform builds corresponding search indices, such as Baidu index and Google trend, by counting the query times of each keyword every day. It can reflect the change of people’s attention to a certain tourist attraction, and they are forward-looking data, which can be well used for tourist volume forecasting as existing studies (Yang et al., 2015; Li et al., 2018; Sun et al., 2019; Bi et al., 2020; Li et al., 2020). Changes in the weather at tourist attractions can also directly affect people’s willingness to travel, especially for domestic tourism. Generally speaking, the tourist volume of an attraction will be significantly larger in normal weather (such as sunny and cloudy) than in extreme weather (such as heavy rain). Therefore, weather information is also an important data to forecast the tourist volume.

To verify the superiority of the method, taking Jiuzhaigou and Sigu'niang Mountain, two famous attractions in China, as examples, experiments are carried out to make all-around comparisons and discussions. Results show that, compared with other benchmark models, the proposed VMD-GRU model has better forecasting performance in tourist volume prediction problem. Further, by successively reducing the types of data used in the model, the strong predictive power of the COVID-19 data, search traffic data and weather data used is confirmed, respectively.

Our contributions are summarized as follows. (1) A novel forecasting framework, VMD-GRU, is proposed. It gives full play to the advantages of VMD in time series decomposition and GRU in time series learning, and can well fit the non-stationary tourist time series to obtain more accurate forecasting results. (2) We are the first one that uses COVID-19 data, search traffic data, weather data and historical data simultaneously to forecast the tourist volume, which can obtain better performance. (3) For the selection of search traffic data, a time difference correlation analysis method based on distance correlation coefficient is proposed. Among them, the distance correlation can measure both linear correlation and nonlinear correlation between the index and tourism demand, so as to find more predictive search traffic indices. Last
Table 1

| Literature                          | Historical data | Search traffic data | Weather data | Economic data | Online reviews | Public holiday dates | COVID-19 Data |
|------------------------------------|-----------------|---------------------|--------------|---------------|----------------|----------------------|---------------|
| Alvarez-Díaz and Rosselló-Nadal (2010). | ✓               | ✓                   | ✓            | ✓             | ✓              | ✓                    | ✓             |
| Yang et al. (2015); Li et al. (2018); Law et al. (2019); Sun et al. (2019); Volchek et al. (2019); Höpken et al. (2021); Huang and Hao (2021); Tang et al. (2021). | ✓               | ✓                   | ✓            | ✓             | ✓              | ✓                    | ✓             |
| Bi et al. (2020).                  | ✓               | ✓                   | ✓            | ✓             | ✓              | ✓                    | ✓             |
| Li et al. (2020); Zhang et al. (2021). | ✓               | ✓                   | ✓            | ✓             | ✓              | ✓                    | ✓             |
| Bi, Li, Xu, et al. (2021).         | ✓               | ✓                   | ✓            | ✓             | ✓              | ✓                    | ✓             |
| Our study                          | ✓               | ✓                   | ✓            | ✓             | ✓              | ✓                    | ✓             |

but not least, this method expands studies of demand forecasting theory and has great practical application value.

The rest of this paper is structured as follows. Section 2 reviews the literature on the most recent forecasting model and variables used in tourism demand forecasting, respectively. Section 3 and Section 4 describe our method in detail. Section 5 illustrates the proposed method by using comparing experiments. Lastly, Section 6 presents the conclusions and outlines future prospects.

2. Literature review

2.1. Progress in tourist volume forecasting model

Models commonly used in the field of tourist volume forecasting include time series analysis model (Assaf et al., 2019), econometric model (Cao et al., 2017; Nicholas, 2021) and artificial intelligence (AI) model. In recent two years, the artificial intelligence model has become the mainstream in the forecast of tourist volume (Law et al., 2019; Sun et al., 2019; Bi et al., 2020; Kulshrestha et al., 2020; Laaroussi et al., 2020; Binru Zhang et al., 2020; Bi et al., 2021; Bi, Li, & Fan, 2021).

Benefiting from the advantages that no assumption on the data needs to be made, the adaptability and nonlinearity characteristics, AI approaches are suitable for non-linear prediction and show good performances in various fields of forecasting problems.

The most used artificial intelligence models are long short-term memory networks (LSTM) and artificial neural networks (ANN) in recent studies. For example, Bi et al. (2020) proposed a new method to forecast the daily tourist volume of tourist attractions based on LSTM network, and the experiment showed that LSTM has better forecasting performance than other machine learning methods such as support vector regression (SVR). Law et al. (2019) combined LSTM with attention mechanism to study the forecast framework of monthly tourist arrivals in Macau and the experimental results showed that their method was obviously superior to SVR and ANN. Binru Zhang et al. (2020) introduced the concept of LSTM network to deal with the complicated time series forecasting problem in tourism, and their study also showed the superiority of LSTM. Other similar studies can be found in Kulshrestha et al. (2020), Höpken et al. (2021) and their references.

In addition, the hybrid model has become the new study trend. For instance, Bi, Li, and Fan (2021) put forward a time series imaging model on the basis of deep learning. They encoded the time series data into the format of picture data first, and then extracted the image features by using the convolution neural network (CNN), and lastly inputted the extracted features into the LSTM network to achieve a more accurate forecasting of the tourist volume. He et al. (2021) introduced a SARIMA-CNN-LSTM model to forecast daily tourist demand data, and their approach adopted a SARIMA model and a deep neural network structure that combined CNN and LSTM layers to capture linear and nonlinear data features. Li et al. (2018) proposed the PCA-DEA-BPNN model. Before training the structure of the back-propagation neural network (BPNN), the principal component analysis (PCA) was used to decrease the dimension of the input data, and the adaptive differential evolution algorithm (ADE) was adopted to globally optimize the weights and thresholds of the BP network to improve the prediction performance of the BPNN.

A few studies combined signal processing methods with the AI forecasting model. For example, Chen et al. (2012) adopted empirical mode decomposition (EMD) to decompose the original data into a limited set of intrinsic mode functions (IMFs) and a residual firstly, and then used BPNN to model and forecast IMFs and the residual, and obtained the final forecast value through integrating these forecasting results. Xie et al. (2020) developed a decomposition-ensemble approach for tourism demand forecasting in areas with frequent irregular events. It is based on the complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN), data characteristic analysis, and the Elman’s neural network model. Liu et al. (2021) found that variational mode decomposition (VMD) method had a very good performance in forecasting influenza epidemics in Hong Kong. However, existing studies have not combined state-of-the-art variational mode decomposition (VMD) with deep learning networks for tourism forecasting. In this paper, we make up for this gap, and propose a novel VMD-GRU model. First, the historical data is processed by VMD, and several IMFs data are decomposed. Then, these IMFs data together with other predictive variable data are used as the inputs of GRU networks to realize the model training and forecasting task.

2.2. Progress in variables used in tourist volume forecasting

Studies have shown that integrating multi-source data instead of relying on a single source data has great advantages in forecasting tourism demand (Li et al., 2020). It can improve the forecasting accuracy and generalization ability of the model, and may produce more valuable forecasts. To date, the use of multi-source external data has become a recent study trend. Most recent studies on tourism forecast and the predictive variables they use are listed in Table 1.

From Table 1, the variables widely used in the latest tourism forecast literature are historical data, search engine data and weather data. For example, Bi et al. (2020) adopted search traffic data, weather data and historical tourist volume data to forecast daily tourist arrivals at Chinese attractions. Based on multiple sources of internet data, i.e., Baidu Index data and reviews on Ctrip and Quanar, Li et al. (2020) forecasted the number of tourists visiting Siguniang Mountain in China.

In this paper, on the basis of widely used historical data, search engine data and weather data, we consider the data of confirmed cases in COVID-19 as the prediction variable. The reasons why we do not use economic data, public holiday dates and so on are as follows. Firstly, the economic data reflects the long-term trend and their release frequency is low, generally monthly data, which is not applicable to the forecast of daily tourism demand. Secondly, our approach involves time series analysis, which itself can mine periodic laws. Therefore, there is no need to add calendar holiday data as a predictive variable. In addition, the historical time series contains information about the influence of online reviews on the tourist volume. To improve the training efficiency of the model and avoid multicollinearity, therefore, we do not introduce online
The impact of the epidemic on the tourism industry is obvious and foreseeable. Therefore, the COVID-19 data is an important variable for more accurate prediction of tourism demand. However, studies on forecasting tourist volume by using COVID-19 data are very rare. Pri-listya et al. (2021) used ARIMAX and SARIMAX models to analyze the influence of COVID-19 epidemic situation on foreign tourists of Indonesia and forecasted the tourist volume by history data, Google Trends data. They only treated the epidemic situation as a dummy variable. Different from existing studies, in this paper, we study the forecasting of daily tourist volumes with COVID-19 data, search traffic data and weather data. It enriches the research in this field and has good theoretical significance, and is applicable to the actual situation at present and has high practical significance.

3. Related methods and preliminaries

3.1. Variational mode decomposition (VMD)

Variational mode decomposition (VMD) is a complete and adaptive non-reursive signal processing method proposed by Dragomiretskiy and Zosso (2014). This method shows accuracy and robustness compared to traditional algorithms, e.g., empirical mode decomposition (EMD) method (Huang et al., 1998), because VMD solves the problems of end effects and mode mixing existing in EMD. The core principles of VMD are as follows.

Firstly, assuming that the original signal \( f(t) \) is decomposed into \( K \) components, the decomposed sequence is guaranteed to be mode components with limited bandwidth with central frequency, and the sum of estimated bandwidths of all modes is the smallest, i.e.,

\[
\begin{align*}
\min_{\{u_k\}, \{\omega_k\}} & \sum_{k=1}^{K} \left\| \delta(t) \otimes u_k(t) \right\|_2^2 \\
\text{s.t.} & \sum_{k=1}^{K} u_k = f(t)
\end{align*}
\]

where \( \{u_k\} = \{u_1, u_2, \cdots, u_K\} \) and \( \{\omega_k\} = \{\omega_1, \omega_2, \cdots, \omega_K\} \) are the \( k \)th decomposed signal and the center frequency of the signal, respectively. \( t \) denotes the time point, \( \delta(t) \) is the Dirac distribution, \( \otimes \) is the convolution operator and \( j^2 = -1 \). The expression \( \left( \delta(t) + \frac{j}{\omega} \right) \otimes u_k(t) \) is the Hilbert transform of \( u_k \).

Then, by solving equation (1) and introducing Lagrange multiplication operator \( \lambda \), the constrained variational problem is converted into unconstrained variational problem, and the augmented Lagrange expression is obtained as follows:

\[
\begin{align*}
L(u_k, \omega_k, \lambda) &= a \sum_{k=1}^{K} \left\| \left( \delta(t) + \frac{j}{\omega} \right) \otimes u_k(t) \right\|_2^2 + \\
&\left\| f(t) - \sum_{k=1}^{K} u_k \right\|_2^2 + \left\langle \lambda(t), f(t) - \sum_{k=1}^{K} u_k \right\rangle
\end{align*}
\]

where \( a \) is the quadratic penalty factor, which is adopted to diminish the Gaussian noise.

After optimizing the mode components and center frequencies through alternating direction multiplier iterative algorithm, the saddle point of augmented Lagrange function is searched. The expressions of \( u_k \), \( \omega_k \) and \( \lambda \) after optimization iteration are as follows, and the detailed process is shown in reference Dragomiretskiy and Zosso (2014).

\[
\begin{align*}
\hat{u}_k^{n+1}(\omega) &= f(\omega) - \sum_{i=1}^{n} \hat{u}_i(\omega) + \frac{\hat{\lambda}(\omega)}{2} \\
&\frac{1}{1 + 2\alpha(\omega - \omega_k)} \\
\hat{\omega}_k^{n+1} &= \int_{0}^{\pi} \omega \hat{u}_k(\omega) |^2 d\omega \\
\hat{u}_k^{n+1}(\omega) &= \int_{0}^{\pi} \hat{u}_k(\omega) |^2 d\omega \\
\hat{\lambda}_k^{n+1} &= \hat{\lambda}_k^n + \tau \left( f(\omega) - \sum_{i=1}^{n} \hat{u}_i^{n+1}(\omega) \right)
\end{align*}
\]

where \( \tau \) is noise tolerance, which is used to meet the fidelity requirement of signal decomposition. \( \hat{u}_k^{n+1}(\omega), \hat{\omega}_k(\omega), \hat{\lambda}(\omega) \) and \( \hat{\lambda}(\omega) \) are Fourier transforms corresponding to \( \hat{u}_k^{n+1}(\omega), \hat{u}_k(\omega), f(\omega) \) and \( \hat{\lambda}(\omega) \), respectively.

The main solving process of VMD is as follows:

Step 1: Initialize \( \{\hat{u}_k\}, \{\hat{\omega}_k\}, \tau \) and the max number of iterations \( N \). Let \( n = 0 \).

Step 2: Update \( u_k \) and \( \omega_k \) by Eq. (3) and Eq. (4).
The function is differentiable everywhere, thus ensuring that the gate is open or closed; Second, the output value of the function is between 0 and 1, which meets the physical definition of gating, i.e., sigmoid function.

Fig. 3 is an example of decomposition results of the historical tourist volumes of Jiuzhaigou by VMD when $K = 4$.

3.2. Gated recurrent unit network (GRU)

Traditional RNN has inherent defects: because gradients have plentiful multiplications in the process of back propagation, gradient vanishing and gradient explosion often occur when learning long-term dependency. Both cases will lead to the poor effect of the trained neural network. The neurons of LSTM networks add an input gate, a forgetting gate, an output gate and an internal memory unit, selectively retain important information, forget secondary information and abandon redundant memory, thus avoiding the defects of gradient vanishing and gradient explosion existing in the traditional RNN (Hochreiter & Schmidhuber, 1997).

Although LSTM improves the traditional RNN, its gating network structure is too complex and redundant, which affects the learning efficiency, and redundant information may affect the performance of the model. For this reason, Chung et al. (2014) proposed the gated recurrent unit network (GRU), an improved model of LSTM. Its structure is shown in Fig. 3. GRU combines the forget gate and the input gate into an update gate, and combines the memory cell and the hidden layer into a reset gate, thereby simplifying the operation of the whole structure and enhancing the performance.

Define $x_t$ as the input data of time point $t$, $r_t$ as the reset gate, $z_t$ as the update gate and $h_t$ as the candidate hidden state. The expressions of each part of GRU shown in Fig. 4 are as follows.

(1) Reset gate $r_t$ is.

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r)$$

(6)

where $\sigma$ means the “sigmoid” function, i.e., $\sigma(x) = \text{sigmoid}(x) = 1/(1 + e^{-x})$. The reasons why “sigmoid” is adopted as the activation function are as follows. First, the output value of the “sigmoid” function is between 0 and 1, which meets the physical definition of gating, i.e., when the input is large or small, the output will be very close to 1 or 0, thus ensuring that the gate is open or closed; Second, the “sigmoid” function is differentiable everywhere.

The reset gate controls how much information in the hidden state of the previous time period, i.e., $h_{t-1}$, will be retained in the candidate hidden state of the current time $h_t$. The larger $r_t$ indicates that the more state information was written at the previous time period.

(2) Update gate $z_t$ is.

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z)$$

(7)

It controls how much state information of the previous time period, i.e., $h_{t-1}$, is retained in the current state. Unlike the reset gate which focuses on the candidate hidden state at the current moment, the update gate focuses on controlling the influence of the hidden information in the past on the present.

(3) Candidate hidden state $\tilde{h}_t$ is.

$$\tilde{h}_t = \tanh(W\cdot[h_{t-1}, x_t] + b_h)$$

(8)

where $\tanh$ means the “hyperbolic tangent” function, i.e., $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$. It includes the current input $x_t$ and the partial hidden state of the previous time period. The reason for using “hyperbolic tangent” as the activation function is that its output is between -1 and 1, which is consistent with the zero-centered feature distribution in most scenes. In addition, $\tanh$ function has a larger gradient than the “sigmoid function” when the input is close to 0, which usually makes the model converge faster.

(4) Hidden state $h_t$ is.

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t$$

(9)

where $(1 - z_t) \cdot h_{t-1}$ determines the information to be retained in the previous hidden state, and $z_t \cdot \tilde{h}_t$ determines the information to be retained in the candidate hidden state.

Generally speaking, the GRU unit first obtains two gating states: reset gate $r_t$ and update gate $z_t$ through the hidden state transmitted at the previous time period $h_{t-1}$ and the input at the current time period $x_t$. After the gating signal is obtained, reset gating is used firstly to obtain the reset data $r_t \cdot h_{t-1}$, which is spliced with the input information, and then scaled it to the range of $(1,1)$ by $\tanh$, thus obtaining the candidate hidden state $\tilde{h}_t$. Finally, the update gate is used for forgetting operation and memorizing operation at the same time, and the final hidden state $h_t$ is obtained.

3.3. Time difference correlation analysis for selecting important search index data

Although the GRU network can learn the time lags of input data
automatically for massive search traffic data, if all the collected search indices data are used for model training, the learning efficiency is reduced, and due to the possibility of introducing information with relatively weak correlation, the learning is interfered and the forecasting performance is negatively affected. Therefore, we need to use the time difference correlation analysis to filter the collected search index data, and eliminate the indices whose distance correlation with tourist volume data is less than a certain threshold within the specified lag period.

Pearson correlation coefficient is used in the time difference correlation analysis in existing studies. This coefficient can only calculate the linear correlation but cannot measure the nonlinear correlation. To eliminate this limitation, Székely et al. (2007) proposed the distance correlation which can measure degrees of linear correlation and nonlinear correlation. The calculation procedures of distance correlation are as follows.

1. Let $X$ and $Y$ be paired continuous variables of length $n$, and then we have.
\[ a_{ij} = \| X_i - X_j \| \quad j, k = 1, 2, \ldots, n \]  
(10)

\[ b_{ij} = \| Y_i - Y_j \| \quad j, k = 1, 2, \ldots, n \]  
(11)

where \( \{a_{ik}\} \) and \( \{b_{jk}\} \) represent the distance matrix between the respective elements of \( X \) and \( Y \), respectively.

(2) Calculate the center distance matrix, and the expressions are as follows.

\[ A_{k} = a_{ij} - \bar{a}_{r} + \bar{a} \]  
(12)

\[ B_{k} = b_{ij} - \bar{b}_{r} + \bar{b} \]  
(13)

where \( \bar{a}_{r} \) is the average of row \( r \) in \( \{a_{ik}\} \), \( \bar{a} \) is the average of column \( k \) in \( \{a_{ik}\} \), \( \bar{b}_{r} \) is the average of all elements in \( \{b_{jk}\} \) and the same applies to matrix \( \{b_{jk}\} \).

(3) Calculate the distance covariance as follows:

\[ \text{dCov}^2(X, Y) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{k} B_{k} \]  
(14)

(4) Calculate the distance variance of \( X \) and \( Y \), i.e.,

\[ \text{dVar}^2(X) = \text{dCov}^2(X, X) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{k}^2 \]  
(15)

\[ \text{dVar}^2(Y) = \text{dCov}^2(Y, Y) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} B_{k}^2 \]  
(16)

(5) Calculate the distance correlation as follows:

\[ \text{dCor}(X, Y) = \frac{\text{dCov}(X, Y)}{\sqrt{\text{dVar}(X) \cdot \text{dVar}(Y)}} \]  
(17)

Based on above, the process of selecting important search index data for tourist volume forecasting we proposed is as follows.

Let \( \Psi \) be the dependent variable data (e.g., historical tourist volume), \( \Omega \) be the search index data concerning the vth query word and \( M \) be the total number of alternative search indices concerning different query words. The set of alternative search indices data is \( \Omega = \{\Omega_1, \Omega_2, \ldots, \Omega_M\} \).

Denote \( \Omega_1 \) as the search index data concerning the vth query word with \( L (L > 0) \) period lagging behind the dependent variable. Then, the maximum absolute distance correlation between \( \Omega \) and \( \Psi \) at different lag orders is.

\[ \Gamma(\Omega, \Psi) = \max_{0 \leq l \leq L_\text{max}} d\text{Cor}(\Omega_l, \Psi) \]  
(18)

where \( L_{\text{max}} \) is the given maximum lag order.

Finally, the selection requirement for the search index data concerning the vth query word is \( \Gamma(\Omega, \Psi) > \xi \), where \( \xi \) is the threshold of the absolute distance correlation for search index data. This is the absolute value of the minimum distance correlation coefficient that should be met between the adopted search index with a certain lag order and the tourist volume demand. The threshold value needs to be determined through repeated experiments (for example, in the experiment of Jiuzhaigou, the threshold is preferably set to 0.8).

Generally speaking, for the correlation coefficient, if the absolute value is in \([0, 0.2]\), there is no correlation; if the absolute value is in \([0.2, 0.4]\), there is weak correlation; if the absolute value is in \([0.4, 0.6]\), there is moderate correlation; and if the absolute value is in \([0.6, 1.0]\), there is strong correlation (Chuan Zhang et al., 2020). To reduce the introduction of noise by eliminating indicators that do not contribute much to the forecasting, in this paper, the absolute distance correlation threshold \( \xi \) should be determined larger than 0.6 (Chuan Zhang et al., 2020; Chuan Zhang et al., 2022).

Using the proposed time difference correlation analysis method, indicators that have insignificant linear relationships but significant nonlinear relationships with tourism demand can also be selected. It is suitable for the following neural network model.

4. The method for daily tourist volume forecasting

Fig. 4 shows the framework of the novel method we proposed for forecasting daily tourist volumes during covid-19 epidemic. The method consists of four steps: step 1, collecting data; step 2, processing data; step 3, training GRU networks; step 4, forecasting tourist volumes. Descriptions of our method in detail are as follows.

4.1. Step 1. Collecting data

Four kinds of data need to be collected: historical data, COVID-19 data, search traffic data and weather data. Data collecting methods and mathematical expressions of these data are as follows.

(1) Historical data

Many tourist attractions have their own official website, on which the daily tourist volume is counted and disclosed to the public every day. Through simple web crawler tools, such as Houyi collector (https://www.houyicaiji.com/), the historical daily tourist volumes of attractions can be collected.

The tourist volume from the 1th day and the T-th day is known, and the tourist volume on the \((T + 1)\)-th day is unknown, i.e., we need to forecast it. The historical tourist volume data, \( H \), can be expressed as \( H = [h_1, h_2, \ldots, h_T] \), where \( h_t \) denotes the historical tourist volume in the t-th day, and \( t = 1, 2, \ldots, T \). For the convenience of understanding, taking the historical tourism volume data collected in Fig. 4 as an example, \( H = [h_1, h_2, \ldots, h_T] = [208, 295, \ldots, 59] \).

(2) COVID-19 data

COVID-19 data is an important index for predicting tourist volumes. It can be downloaded directly through the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19).

The data of daily new confirmed cases in COVID-19, \( C \), can be represented as 
\[ C = \begin{bmatrix} c_1^0 & c_2^0 & \cdots & c_T^0 \\ c_1^1 & c_2^1 & \cdots & c_T^1 \\ \vdots & \vdots & \ddots & \vdots \\ c_1^L & c_2^L & \cdots & c_T^L \end{bmatrix} \]
where \( c_i^L \) represents the number of newly confirmed cases in whole country and the destination area respectively in the t-th day, and \( t = 1, 2, \ldots, T \). Taking the collected data in Fig. 4 as an example, \( C = \begin{bmatrix} 31 & 34 & \cdots & 23 \\ 2 & 2 & \cdots & 0 \end{bmatrix} \).

(3) Search traffic data

Taking Baidu Index (https://index.baidu.com/) as an example, we use Python programming to obtain Baidu search index data.

The most critical step before collecting Baidu search data is to identify key words. The way to generate a candidate word set is using a central word related to the tourist attraction to elicit other key related words or phrases by adopting the “related searches” function provided by Baidu index platform (Li et al., 2018; Sun et al., 2019; Li et al., 2020; Chuan Zhang et al., 2022). For example, to collect Baidu search indices related to Jiuzhaigou scenic spot, firstly the core search word “Jiuzhaigou” should be determined, and then enter this word into the query box of Baidu Index website. The box will automatically list many related search words, such as “Jiuzhaigou tickets” and “Jiuzhaigou accommodation”.

Assuming that there are \( Q \) Baidu search key words collected, the collected search traffic data \( S \) can be expressed as:

\[ S = \begin{bmatrix} s_1^0 & s_2^0 & \cdots & s_T^0 \\ s_1^1 & s_2^1 & \cdots & s_T^1 \\ \vdots & \vdots & \ddots & \vdots \\ s_1^Q & s_2^Q & \cdots & s_T^Q \end{bmatrix} \]
and then the decomposition process can be expressed as. Fig. 4 as an example, have

\[ W^k = \begin{bmatrix} w^k_{1,1} & w^k_{1,2} & \cdots & w^k_{1,T} \\ w^k_{2,1} & w^k_{2,2} & \cdots & w^k_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ w^k_{1,1} & w^k_{1,2} & \cdots & w^k_{1,T} \end{bmatrix}, \]

(23)

where predictor \( IMF^k \) represents the \( k \)-th IMF decomposed from historical tourist volume data by VMD, \( \xi \) represents the value of the \( k \)-th IMF in the \( t \)-th day, and \( t = 1, 2, \ldots, T \).

(2) Processing search traffic data using time difference correlation analysis.

As mentioned before, the total number of search indices concerning candidate words collected is \( Q \). Assuming that there are \( Q \) search indices with different lag orders selected as predictors, the set of them is \( KW = \{ kw_1, kw_2, \ldots, kw_Q \} \) (\( Q \leq Q \)). The corresponding search traffic data is.

\[ S^q = \begin{bmatrix} s_{1,1}^q & s_{1,2}^q & \cdots & s_{1,T}^q \\ s_{2,1}^q & s_{2,2}^q & \cdots & s_{2,T}^q \\ \vdots & \vdots & \ddots & \vdots \\ s_{d,1}^q & s_{d,2}^q & \cdots & s_{d,T}^q \end{bmatrix}, \]

(24)

where \( s_{q,t}^q \) is the Baidu search data of the \( q \)-th selected predictor, \( s_{q,t}^q \) is the search index value of the predictor \( s_{q,t}^q \) in the \( t \)-th day, and \( t = 1, 2, \ldots, T \).

(3) Processing weather data.

Notice that the data of weather conditions are non-digital, which cannot be directly used for mathematical calculating. Hence, we need to convert this kind of data into dummy variables based on their suitability for travel. Reference to Bi et al. (2020), we divide the weather conditions into three predictors: “Good”, “Moderate” and “Bad”, and convert them into one-hot encodings. Table 2 shows the conversion rules in detail. According to the rules, all the three predictors are dummy variables with values of 0 or 1, where 0 indicates “False” and 1 indicates “True”.

Then, the weather data will be processed as.

\[ W^w = \begin{bmatrix} w_{1,1}^w & w_{1,2}^w & \cdots & w_{1,T}^w \\ w_{2,1}^w & w_{2,2}^w & \cdots & w_{2,T}^w \\ \vdots & \vdots & \ddots & \vdots \\ w_{1,1}^w & w_{1,2}^w & \cdots & w_{1,T}^w \end{bmatrix}, \]

(25)

where \( w_{w,1}^w = \text{good} \), \( w_{w,2}^w = \text{moderate} \) and \( w_{w,3}^w = \text{bad} \) are predictors indicating whether the weather conditions are “good”, “moderate” and “bad” respectively. The values of \( w_{w,1}^w \), \( w_{w,2}^w \) and \( w_{w,3}^w \) are either 0 or 1, which respectively indicates whether the weather condition is “good”, “moderate” or “bad” in the \( t \)-th day, \( t = 2, \ldots, T + 1 \).

(4) Combining the above processed data into one matrix \( D_{t+1} \), i.e.,

\[ D_{t+1} = \begin{bmatrix} d_1 \\ \vdots \\ d_{J-1} \end{bmatrix}, \]

(26)

where \( d_j \) represents the number of forecasting predictors, \( k = 2, \ldots, K + \#Q \).
In the matrix $\mathbf{D}_{j-T}$, the historical data, IMFs obtained by decomposing historical data with VMD, COVID-19 data and search traffic data are from the 1-th day to the T-th day, while the weather data are from the 2-th day to the (T + 1)th day.

(5) Data normalization.

Different features have different orders of magnitude and dimensions. In other words, they are incommensurate. If they are directly used, the modeling accuracy will be affected. Therefore, to improve the training efficiency and forecast accuracy, they should be standardized using Max-Min normalization method as follows.

$$\tilde{d}_j = \frac{d_j - \min(d_j)}{\max(d_j) - \min(d_j)}, \quad j = 1, 2, ..., J; t = 1, 2, ..., T, \quad \text{(27)}$$

in which $d_j$ denotes the $t$-th value of $d$ and $\tilde{d}_j$ denotes the normalization result of $d_j$. Then, $\mathbf{D}_{j-T}$ can be processed as $\mathbf{D}_{\tilde{j-T}}$ based on the obtained $\tilde{d}_j$, i.e.,

$$\mathbf{D}_{\tilde{j-T}} = \begin{bmatrix} \tilde{d}_1 & \tilde{d}_2 & ... & \tilde{d}_T \\ \tilde{d}_j & \tilde{d}_j & ... & \tilde{d}_j \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{d}_{T-t} & \tilde{d}_{T-t} & ... & \tilde{d}_{T-t} \end{bmatrix}$$

For the convenience of processing, we standardized all the predictors in the matrix $\mathbf{D}_{j-T}$, including the dummy variable of weather conditions. Since the values of weather condition features are only 0 or 1 after the process of one-hot encoding, they remain unchanged after normalization by Eq. (27).

The labels we use in model training, i.e., the actual values, have also been standardized, as shown in Fig. 4. The purpose is to accelerate the convergence of weight parameters in the process of training the neural network, which is helpful for fast training.

(5) Construction of training samples.

Denote as the time step of the GRU networks. $T-t$s training samples, i.e., $\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_{T-t}$, are constructed on the basis of the obtained $\mathbf{D}_{\tilde{j-T}}$, where $\mathbf{T}_g$ denotes the $g$-th training sample, $g = 1, 2, ..., T-t$. $\mathbf{T}_g$ is a matrix, i.e.,

$$\mathbf{T}_g = \begin{bmatrix} \tilde{d}_1^g, \tilde{d}_2^g, ..., \tilde{d}_T^g \\ \vdots \\ \tilde{d}_1^g, \tilde{d}_2^g, ..., \tilde{d}_T^g \\ \vdots \\ \tilde{d}_1^g, \tilde{d}_2^g, ..., \tilde{d}_T^g \\ \tilde{d}_1^{g+1}, \tilde{d}_2^{g+1}, ..., \tilde{d}_T^{g+1} \end{bmatrix}$$

where data in the 1-th to ts-th column are inputs of GRU networks and the data in the (ts + 1)th column represent the GRU output.

4.3. Step 3. Training GRU networks.

GRU networks is trained by using the above obtained $T-t$s constructed training data, i.e., $\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_{T-t}$, and then, the GRU networks with practical forecasting power can be gained. For the complete process and principle of training GRU, please refer to Cho et al. (2014).

4.4. Step 4. Forecasting tourist volumes.

In Step 2 of Fig. 5, we have obtained $T-t$s training samples and the input data $\mathbf{ID}$, which is used for the forecast of the $(T + 1)$th day. The $\mathbf{ID}$ has the same form as $\mathbf{T}_g$, except that the $(ts + 1)$th column data in the $\mathbf{ID}$ is unsuspected and it is prepared to be predicted.

Inputting the data $\mathbf{ID}$ into the trained GRU networks in Step 3, then the tourist volume on the $(T + 1)$th day, $\tilde{d}_{T-t+1}$, can be obtained. Through the inverse normalization, it can be converted into the forecasted tourist volume of the $(T + 1)$th day, and the formula is as follows:

$$h_{T+1} = \frac{1}{T} \sum_{t=1}^{T} \left[ \max(d) - \min(d) \right] + \min(d), \quad j = 1, 2, ..., J; t = 1, 2, ..., T, \quad \text{(30)}$$

where $h_{T+1}$ is the tourist volume of the $(T + 1)$th day.

5. Experimental study.

In this section, we take two well-known attractions in China: Jiuzhaigou and Siguniang Mountain, as examples, to conduct comprehensive experimental comparisons and discussions.

---

**Table 3**

| Data name          | Description                                      | Source                             | Date range   |
|--------------------|--------------------------------------------------|------------------------------------|--------------|
| Historical data    | Historical tourist volume of Jiuzhaigou and Siguniang Mountain. | Official websites of these tourist attractions | 2020.4.1~2021.11.20 |
| COVID-19 data      | New confirmed COVID-19 cases daily in Chinese mainland and Sichuan province. | The Center for Systems Science and Engineering (CSSE) at Johns Hopkins University | 2020.4.1~2021.11.20 |
| Search traffic data| Relevant Baidu search indices concerning these tourist attractions. Take Jiuzhaigou as an example, the related query words are “Jiuzhaigou”, “Jiuzhaigou Tickets”, “Jiuzhaigou pictures”, “Jiuzhaigou weather”, “Jiuzhaigou location”, “Jiuzhaigou tourism strategy”, “Jiuzhaigou accommodation”, “Jiuzhaigou tourism” and “Jiuzhaigou scenic spot”. | Baidu Index | 2020.3.1~2021.11.20 |
| Weather data       | The temperature and weather condition data concerning these tourist attractions. | Tianqi website | 2020.4.2~2021.11.21 |
concerning keywords listed in Table A2 will be used. It is necessary to select search words that have good predictive ability. For example, temperature, the lowest temperature, and the weather conditions. Higher than that in Jiuzhaigou indicates that, it does not affect the forecasting, because the low search frequency of keywords related to Siguniang Mountain is relatively small. Although there are many zero or very low frequencies of keywords related to the two scenic spots are consistent with the trend of the tourist indices are shown in Table 3.

From Table A2, we can find that the search frequency of keywords related to the two scenic spots are consistent with the trend of the tourist volume in the two scenic spots as a whole, i.e., the average search frequency of keywords related to Jiuzhaigou is obviously more than that of Siguniang Mountain, and the number of zero or very low frequencies of keywords related to Jiuzhaigou is relatively small. Although there are many zero or very low frequencies of keywords related to Siguniang Mountain, it does not affect the forecasting, because the low search frequency reflects the low attention, and the corresponding number of tourists may also be small. Most importantly, not all the search indices concerning keywords listed in Table A2 will be used. It is necessary to use time difference correlation analysis method to eliminate useless search keywords, so as to ensure that the search frequency data of the selected search words have good predictive ability.

As can be seen from Table A3, the weather data includes the highest temperature, the lowest temperature and the weather conditions. Among them, the weather conditions are processed into three predictors (or features): “Good”, “Moderate” and “Bad”, according to the rules provided in Table 2 of Section 4.2. All three predictors are one-hot encodings with values of 0 or 1, where 0 indicates “False” and 1 indicates “True”. Obviously, the temperature in Jiuzhaigou is generally higher than that in Siguniang Mountain. The ratios of “Good” and “Bad” weather conditions (i.e., the mean values of the “Good” and “Bad” predictors) in Jiuzhaigou are 0.8 and 0.04, respectively, while the ratios of “Good” and “Bad” weather conditions in Siguniang Mountain are 0.78 and 0.06, respectively. This indicates that there are more tourists in scenic spots with better weather conditions, which further indicates that the weather data can forecast the tourism demand.

Table A4 presents the data of confirmed cases in COVID-19 throughout China and Sichuan Province. Although the overall value is not particularly large, in view of the infectivity of COVID-19, the panic and influence caused by the increase of confirmed cases are very great. National and local epidemic data should be considered separately for the following reasons. If the epidemic situation in the whole country is not severe, but only the epidemic situation in the destination area is seriously rebounding, many travelers will give up traveling to this area for fear of contracting the virus, and choose other places. If there is no epidemic situation in the area where the tourist attraction is located, but the national epidemic situation is serious, it will only affect the number of tourists from other areas to the tourist attraction, but will not affect the tourists from the place where the attraction is located. It can be seen that for a specific attraction, the impact of the national epidemic and the local epidemic on the number of tourists to the attraction is different. Therefore, if the data of confirmed cases in the whole country and the destination area are combined, the information will be lost and the forecast performance will be affected.

Taking Jiuzhaigou as an example, Fig. 5 shows curves of Baidu search indices screened out by the time difference correlation analysis method, and the intermediate process is shown in Table 4.

### 5.2. Experimental design

1. **Forecasting scheme and the split of the data set.**

   We use the fixed forecasting scheme to conduct experiments. The data set is divided into the training set and the test set. We use the data from October 21, 2021 to November 20, 2021, totaling 30 days, as the test set to evaluate forecast performance. The training set consists of the data from April 1, 2020 to October 20, 2021, totaling 568 days. To facilitate the determination of hyperparameters, such as the number of IMFs, unit number of GRU network and time steps and so on, we take 20% samples from the training set as the validating set.

2. **Assessment metrics.**

   In this paper, two metrics are used to calculate the forecasting error of a forecasting model: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Their calculation formula is as follows:

   \[
   MAE = \frac{1}{N} \sum_{t=1}^{N} |x_t - \hat{x}_t|, \tag{31}
   \]

   \[
   RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (x_t - \hat{x}_t)^2}, \tag{32}
   \]

   where \(x_t\) and \(\hat{x}_t\) represent the actual value and the forecast value in the \(t\)-th period, respectively. Obviously, the lower the value of MAE and RMSE, the more reliable and accurate will be the forecasting.

   To measure whether the forecasting accuracy of the proposed approach is significantly better than a benchmark model from a statistical perspective. We adopt Diebold-Mariano (DM) test (Diebold & Mariano, 1995), and the null and alternative hypotheses are.

   \[
   H_0: E[g(e_1)] = E[g(e_2)]. \tag{33}
   \]

   \[
   H_1: E[g(e_1)] \neq E(g(e_2)). \tag{34}
   \]

   where \(g(\cdot)\) is loss function of forecasting, \(e_1\) and \(e_2\) are the forecasting errors of the two models. The DM test is calculated by.

   \[
   DM = \frac{\sum_{t=1}^{N} [g(e_1) - g(e_2)]^2}{\sqrt{s^2/N}}, \tag{35}
   \]

   where \(s^2\) is an estimator of the variance \([g(e_1) - g(e_2)]\). If \(DM > 0\), the model 2 is superior to the model 1; otherwise, the model 1 is superior to the model 2.

   (3) **Selection of hyperparameters.**

   When training the proposed model, several hyperparameters should be determined firstly, e.g., IMF number, time step, unit number, batch number and training numbers. Among them, IMF number, time step and unit number have greater influence on the performance of forecasting. To find the best values of the three hyperparameters, we changed the

| Key word | Maximum absolute distance correlation | The lag order of the maximum absolute distance correlation | Is selected? |
|----------|--------------------------------------|--------------------------------------------------------|-------------|
| Jiuzhaigou Tickets | 0.811 | 3 | Yes |
| Jiuzhaigou pictures | 0.667 | 3 | No |
| Jiuzhaigou weather | 0.842 | 3 | Yes |
| Jiuzhaigou location | 0.631 | 1 | No |
| Jiuzhaigou tourism strategy | 0.670 | 10 | No |
| Jiuzhaigou accommodation | 0.588 | 4 | No |
| Jiuzhaigou tourism | 0.826 | 3 | Yes |
| Jiuzhaigou scenic spot | 0.789 | 5 | No |
| Jiuzhaigou scenic spot | 0.535 | 2 | No |
Expert Systems With Applications 210 (2022) 118505

C. Zhang and Y.-X. Tian

hyperparameters to be adjusted within a certain range while keeping the values of other hyperparameters unchanged. By comparing the average RMSE values of verification sets, we chose the appropriate hyperparameters based on the principle of minimum RMSE. In the following, Jiuzhaigou is taken as an example to explain the detailed process.

First, we keep time step and unit number in GRU network constant to find the optimal of IMF number $K$. Every time the value of $K$ is changed, we repeat the same experiment for 5 times, and then record the average RMSE of these 5 verifications. Finally, the average RMSE of the

Fig. 6. The average values of RMSE under different $K$ (Take Jiuzhaigou as an example).

Fig. 7. The average values of RMSE under different hyperparameter combinations (Take Jiuzhaigou as an example).

Fig. 8. The actual and forecasted tourist volume of Jiuzhaigou with different models.

Fig. 9. The actual and forecasted tourist volume of Siguniang Mountain with different models.

Fig. 10. The MAE of each model.

Fig. 11. The RMSE of each model.

Fig. 12. The actual and forecasted tourist volume of Jiuzhaigou using different kinds of variables.
verification corresponding to each $K$ value is compared, and the $K$ corresponding to the lowest RMSE is selected. To visualize the results, we plotted the curves for the validation set RMSE values under different $K$ values as shown in Fig. 6. Obviously, when $K = 17$, the average RMSE is the lowest, and 17 is the best value of IMF number of Jiuzhaigou tourists.

Both time step and unit number are important hyperparameters of GRU network. By comparing the forecasting performance of verification set under different combinations of time step and unit number, the most suitable combination has been found. In the same way, keeping $K = 17$, change the combinations of time step and unit number to repeat experiments, and calculate the average RMSE of each attempt under different combinations. The results are shown in Fig. 7, and the closer the color of the region is to light yellow, the smaller the RMSE of the model trained by the hyperparameters combination, and the better the forecasting result. From Fig. 7, the model performs best when time step is 9 and unit number is 40.

Similarly, the above approach is also adopted for determining model hyperparameters in the experiment of forecasting tourist volume of Siguniang Mountain and the later experimental comparisons.

(4) Selection of comparative benchmarks.

We set up two groups of comparative experiments to prove the superiority of the VMD-GRU model and the remarkable predictive power of COVID-19 data, search traffic data and weather data.

To illustrate the superiority of the VMD-GRU model proposed in this paper, we set some benchmark models as follows.

- RM: Simple random walk. Also known as Naïve Forecast. The idea is to assume that the future demand completely replicates the past model. In this model, the forecasted value of the next period is equal to the actual demand of the current period (Hein & Spudeck, 1988).
- ARIMAX: Autoregressive integrated moving average model with exogenous input variables (Pan et al., 2012).
- SVR: Support vector regression (Huang & Hao, 2021).
- ANN: Artificial neural networks (Palmer et al., 2006; Höpken et al., 2021).
- LSTM: Long short-term memory networks (Law et al., 2019; Bi et al., 2020; Laaroussi et al., 2020).
- GRU: Gated recurrent unit networks (Laaroussi et al., 2020).
- EMD-GRU: A hybrid model composed of empirical mode decomposition and gated recurrent unit networks.
- EEMD-GRU: A hybrid model composed of ensemble empirical mode decomposition and gated recurrent unit networks.
- VMD-LSTM: A hybrid model composed of variational mode decomposition and long short-term memory networks.
- VMD-GRU: A hybrid model composed of variational mode decomposition and gated recurrent unit networks (our model).

To verify whether COVID-19 data, search traffic data and weather data can make progress in improving predictive ability of VMD-GRU model, we successively reduced the data variables used in the model and generated the following comparative experiments.

- H: Only historical data (H).
5.4. Comparisons and discussions

Firstly, according to Fig. 8 and Fig. 9, the forecasted values of our model, VMD-GRU, are the closest to the actual values, followed by VMD-LSTM model. From Table 5, the proposed model is superior to the benchmarks in MAE and RMSE, which proves the validity of the method proposed in this paper.

Secondly, compared with the time series analysis model ARIMAX, the machine learning models SVR, ANN, LSTM and GRU all have better forecasting performance. The traditional machine learning models, SVR and ANN, are inferior to deep learning models, LSTM and GRU. Even the simplest model, i.e., RM, is better than SVR and ANN in this study. GRU performs better than LSTM network, and its average MAE and RMSE are 9.44 % and 4.62 % lower than LSTM, respectively. Similarly, the average MAE and RMSE of VMD-GRU model are respectively 30.22 % and 20.53 % lower than VMD-LSTM model, which also shows that GRU is superior to LSTM.

Finally, compared with GRU model, the average MAE and RMSE of EMD-GRU model are 9.44 % and 4.62 % lower than GRU model respectively, but the average MAE and RMSE of EEMD-GRU model are 18.03 % and 7.64 % higher than GRU model respectively, and the average MAE and RMSE of VMD-GRU model are 64.32 % and 61.70 % lower than GRU model respectively. These illustrate that the empirical mode decomposition plays a minor role in this task, and even has side effects (e.g., EEMAD), while the variational mode decomposition can significantly improve the forecasting performance of GRU model. In a word, our VMD-GRU model is superior and reasonable.

(2) Discussion of the comparison results of different forecasting variables.

As can be seen from Fig. 12 and Fig. 13, VMD-GRU model which simultaneously uses historical data, COVID-19 data, search traffic data and weather data has forecasted values that are closest to the actual values.

To explain in detail how these three kinds of data can enhance the forecasting performance, we make the following comparative analysis in combination with Table 6.

Firstly, the forecasting performance of “HW”, “HS” and “HC” are all better than “H”. Specifically, the average MAE and RMSE of “HW” are 16.8 % and 20.66 % lower than “H”, respectively; The average MAE and RMSE of “HS” are 25.14 % and 32.45 % lower than “H”, respectively; The average MAE and RMSE of “HC” are 32.56 % and 36.83 % lower than “H”, respectively. It can be analyzed that, COVID-19 data has the greatest contribution on improving the forecast performance, followed by search traffic data and finally weather data.

Secondly, the average MAE and RMSE of “HWS” is 13.9 % and 21.71 % lower than “HW”, respectively; The average MAE and RMSE of “HSC” are 29.88 % and 24.51 % lower than “HS”, respectively; The average MAE and RMSE of “HWC” are 64.32 % and 61.70 % lower than “HC”. This further illustrates that the contribution of COVID-19 data on forecasting performance is greater than search traffic data and weather data.

Finally, “HWSC” has the best forecasting performance. Specifically,
the average MAE and RMSE of “HWSC” are 53.55 % and 69.21 % lower than “HWS”, 47.37 % and 66.67 % lower than “HWC” and, 36.60 % and 62.50 % lower than “HSC”. These also indicate that the improvement effect of COVID-19 data on forecasting performance of tourist volumes is greater than search traffic data and the improvement effect of search traffic data is greater than weather data. Moreover, considering the three types of data simultaneously, the model has the best performance.

In conclusion, using historical data to predict the tourist volume recognizes the ductility of the changing trends, but when unexpected events occur, such as the epidemic situation, only considering historical data will limit the accuracy of prediction (Johann du Preez & Stephen, 2003). Therefore, we conclude that it is necessary to consider a variety of external data, especially the information related to the events that directly affect people’s travel decisions (such as the epidemic) in the forecast of travel volume in special periods. COVID-19 data, search traffic data and weather data here are important explanatory variables, and it is reasonable and effective to consider these three types of data to forecast tourist volumes.

(3) Diebold-Mariano Test.

We conduct DM tests under different forecast models and different predictive variables to illustrate whether the forecasting accuracy of the proposed model is significantly superior to benchmarks (Zhao et al., 2022). Table 7 and Table 8 show the results of the DM test of different forecasting models and different forecasting variables, respectively.

It can be seen from the results that the proposed approach is better than the benchmark models at different significance levels, both in the model and the predictive variables used, indicating that our forecasting approach is obviously effective.

6. Conclusions

This paper studies daily tourist volumes forecasting during the period of epidemic, and proposes an innovative VMD-GRU method, which can use historical data, COVID-19 data, search traffic data and weather data concurrently. Taking two well-known attractions: Jiuzhaigou and Siguniang Mountain, as examples, experiments are carried out. The results show that: 1) VMD is very suitable for the prediction of non-stationary tourist volumes and can greatly improve the forecasting performance; 2) GRU network has the best performance in the task of forecasting tourist volumes; 3) The introduction of COVID-19 data, search traffic data and weather data can enhance the predictive power of the forecasting model, and COVID-19 data has the biggest improvement on the forecast performance, followed by search traffic data and finally weather data.

The main innovations of this study are as the following: 1) This is the first study that applies VMD to the forecasting of tourist volumes, and combined with GRU, proposes a novel VMD-GRU model to fulfill multivariate time series forecasting based on multiple data sources. 2) This is the first work to integrate historical data, COVID-19 data, search traffic data and weather data into the construction of daily tourist volumes forecast method. 3) The distance correlation is adopted to improve the time difference correlation analysis for selecting significant search key words.

It is important to highlight that, this background is novel and different from existing studies, which provides a better choice for the operational decision-making of tourism industry in epidemic situation, and this study has laid a good foundation for the application of artificial intelligence model in the task of forecasting tourist volumes. Accurate forecasts of daily tourist volumes are sound decision guidelines for tourism operators. On the basis of the forecasting results, tourism operators can develop pricing strategies or tourism packages to raise the number of tourists if the tourism demand is low. Meanwhile, tourism operators can draw up crisis plans to prevent the risk of the epidemic spreading due to the detention of tourists or excessive tourists in periods of high demand.

The limitations of this paper are as follows. First, in view of the different epidemic prevention policies in different countries, whether the proposed method can be applied to the forecast of tourist volumes in other countries or other epidemic prevention policies needs further verification, and future research can turn the focus to other countries and regions. Second, more types of predictive variables can be tried in our method framework, such as online review data. In addition, low-frequency variables, such as economic data, can be used to investigate the long-term trend of tourism volume. Third, it is also a good attempt to construct a composite indicator from various variables through certain methods, which can reduce the dimension of the input of the forecast model. Finally, a support system can be developed based on the proposed method.

CRediT authorship contribution statement

Chuan Zhang: Supervision, Conceptualization, Funding acquisition, Writing – review & editing. Yu-Xin Tian: Conceptualization, Methodology, Software, Visualization, Writing – original draft, Validation, Writing – review & editing.

Declaration of Competing 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.

Data availability

Data will be made available on request.

Acknowledgement

This work was supported by the National Social Science Fund of China (19BGL229).

Appendix

Table A1

| Attraction name      | Count | Mean   | Standard deviation | Median | Min  | Max   |
|----------------------|-------|--------|--------------------|--------|------|-------|
| Jiuzhaigou           | 599   | 6183.85| 5956.71            | 4747   | 30   | 40,645|
| Siguniang Mountain   | 599   | 1552.10| 2494.34            | 825    | 1    | 20,000|
Table A2
Descriptive statistics of search traffic data.

| Attraction name | Search query | Count | Mean | Standard deviation | Median | Min | Max |
|-----------------|--------------|-------|------|--------------------|--------|-----|-----|
| Jiuzhaigou      | Tickets      | 630   | 452.08 | 238.05          | 386    | 109 | 1265 |
|                 | pictures     | 630   | 248.18 | 44.78           | 253    | 100 | 400 |
|                 | weather      | 1303.50 | 1057.97 | 263.5            | 79    | 908 |      |
|                 | location     | 266.15 | 80.82  | 273.93          | 475    | 87  | 1205 |
|                 | tourism strategy | 499.67 | 209.67 | 65.66           | 66     | 0   | 202  |
|                 | accommodation| 75.16  | 53.66  | 375             | 375    | 66  | 2583 |
|                 | tourism      | 5066.05 | 2476.41 | 1464.5          | 1382   | 15,246 |      |
|                 | scenic spot  | 352.46 | 128.27 | 321.5           | 78     | 795 |      |
| Siguniang Mountain | Weather in Siguniang Mountain | 1419.19 | 494.72 | 1311.5          | 717    | 4640 |      |
|                 | Travel guide of Siguniang Mountain | 485.76 | 331.28 | 375            | 66     | 2583 |
|                 | Mountain altitude | 131.58 | 49.70 | 138            | 0      | 343 |      |
|                 | Mountain travel | 210.73 | 92.69 | 193.5          | 59     | 900 |      |
|                 | Tickets for Siguniang Mountain | 49.74 | 56.41 | 59           | 0      | 221 |      |
|                 | Mountain location | 102.45 | 56.01 | 126            | 0      | 296 |      |
|                 | Mountain scenic area | 332.28 | 182.89 | 296        | 65     | 947 |      |
|                 | mountain accommodation | 6.91 | 19.56 | 0            | 0      | 120 |      |
|                 | Mountain scenic area | 45.03 | 45.63 | 59           | 0      | 205 |      |

Table A3
Descriptive statistics of weather data.

| Attraction name | weather data name | Count | Mean | Standard deviation | Median | Min | Max |
|-----------------|-------------------|-------|------|--------------------|--------|-----|-----|
| Jiuzhaigou      | Max               | 599   | 7.92 | 7.92              | 20     | 0   | 36  |
|                 | Min               | 599   | 7.42 | 7.42              | 10     | -13 | 22  |
|                 | Good              | 599   | 0.40 | 0.40              | 1      | 0   | 1   |
|                 | Moderate          | 599   | 0.36 | 0.36              | 0      | 0   | 1   |
|                 | Bad               | 599   | 0.20 | 0.20              | 0      | 0   | 1   |
| Sichuan Province | Max               | 599   | 7.23 | 7.23              | 13     | -6  | 36  |
|                 | Min               | 599   | 8.17 | 8.17              | 1      | -20 | 24  |
|                 | Good              | 599   | 0.41 | 0.41              | 1      | 0   | 1   |
|                 | Moderate          | 599   | 0.37 | 0.37              | 0      | 0   | 1   |
|                 | Bad               | 599   | 0.23 | 0.23              | 0      | 0   | 1   |

Table A4
Descriptive statistics of confirmed cases in COVID-19.

| Statistical scope | Count | Mean | Standard deviation | Median | Min | Max |
|-------------------|-------|------|--------------------|--------|-----|-----|
| China             | 599   | 28.30 | 30.39             | 19     | 0   | 353 |
| Sichuan Province  | 599   | 1.19  | 1.66              | 1      | 0   | 12  |

References
Álvarez-Díaz, M., & Rossellid-Nadal, J. (2010). Forecasting British tourist arrivals in the Balearic islands using meteorological variables. *Tourism Economics, 16*(1), 153–168. https://doi.org/10.5367/000101908782079
Ainsa, A. G., Li, G., Song, H., & Tsianos, M. G. (2019). Modeling and forecasting regional tourism demand using the Bayesian global vector autoregressive (BGVAR) model. *Journal of Travel Research, 58*(3), 383–397.
Bi, J.-W., Liu, Y., & Li, H. (2020). Daily tourism volume forecasting for tourist attractions. *Annals of Tourism Research, 82*. https://doi.org/10.1016/j.jatour.2020.102923
Bi, J.-W., Li, C., Xu, H., & Li, H. (2021). Forecasting Daily Tourism Demand for Tourist Attractions with Big Data: An Ensemble Deep Learning Method. *Journal of Travel Research, 60*(2), 531–544. https://doi.org/10.1109/JTR.2013.2288075
Guo, W., Liu, Q., Luo, Z., & Tse, Y. (2022). Forecasts for international financial series with VMD algorithms. *Journal of Asian Economics, 80*, Article 101458. https://doi.org/10.1016/j.jaecon.2022.101458
Huang, B., & Hao, H. (2021). A novel two-step procedure for tourism demand forecasting. *International Journal of Forecasting, 37*(1), 801–813. https://doi.org/10.1016/j.ijforecast.2020.07.008
Wei, Georges, Ji, Lei, Wu, Chi Wai Don, & Tso, Kwok Fai Geoffrey. (2021). Using SARIMA–CNN–LSTM approach to forecast daily tourism demand. *Journal of Hospitality and Tourism Management, 49*, 25–33. 10.1016/j.jhtm.2021.08.080.
Hein, S. E., & Spadeck, R. E. (1988). Forecasting the daily federal funds rate. *International Journal of Forecasting, 4*(4), 581–591. https://doi.org/10.1016/0169-2070(88)90153-5
Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
Höppen, W., Eberle, T., Fuchs, M., & Lexhagen, M. (2021). Improving Tourist Arrival Prediction: A Big Data and Artificial Neural Network Approach. *Journal of Travel Research, 60*(5), 998–1017. https://doi.org/10.1177/0047287520921244
Huang, Y., & Deng, Y. (2021). A new crude oil price forecasting model based on variational mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 476*(1), 903-995. 10.1098/rspa.1998.0193.
Huang, Y., & Deng, Y. (2021). A new crude oil price forecasting model based on variational mode decomposition. *Knowledge-Based Systems, 213*, Article 106669. https://doi.org/10.1016/j.knosys.2020.106669
Huang, Norden E., Shen, Zheng, Long, Steven R., Wu, Manli C., Shih, Hsing H., Zheng, Quanan, Yue, Nai-Chyaun, Tung, Chi Chao, & Liu, Henry H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 454*(1971), 903-995. 10.1098/rspa.1998.0193.
Huang, Y., & Deng, Y. (2021). A new crude oil price forecasting model based on variational mode decomposition. *Knowledge-Based Systems, 213*, Article 106669. https://doi.org/10.1016/j.knosys.2020.106669
Johann du Preez, & Stephen, F. Witt. (2003). Univariate versus multivariate time series forecasting: An application to international tourism demand. *International Journal of Forecasting, 19*(3), 435–451.
Kilshrestha, A., Krishnaswamy, V., & Sharma, M. (2020). Bayesian BILSTM approach for tourism demand forecasting. *Annals of Tourism Research, 83*. https://doi.org/10.1016/j.annals.2020.10.025

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734.

Laaroussi, Houria, Guerouate, Fatima, & Shibhi, Mohamed. (2020). *Deep Learning Framework for Forecasting Tourism Demand*. Law, R., Li, G., Fong, D. K., Chio, Han, & Xin. (2019). Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research, 75*, 410–423. https://doi.org/10.1016/j.jaltour.2019.01.014

Li, H., Hu, M., & Li, G. (2020). Forecasting tourism demand with multivariate big data. *Annals of Tourism Research, 82*. https://doi.org/10.1016/j.jaltour.2020.10.2912

Li, S., Chen, T., Wang, L., & Ming, C. (2018). Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management, 68*, 116–126. https://doi.org/10.1016/j.tourman.2018.03.006

Li, X., Gong, J., Gao, B., & Yuan, P. (2021). Impacts of COVID-19 on tourists’ destination preferences: Evidence from China. *Annals of Tourism Research, 90*. https://doi.org/10.1016/j.jaltour.2021.102958

Li, Y., Fong, G., Tui, K.-L., & Sun, S. (2021). Forecasting influenza epidemics in Hong Kong using Google search queries data: A new integrated approach. *Expert Systems with Applications, 185*, Article 115604. https://doi.org/10.1016/j.eswa.2021.115604

Nicholas, A. (2021). Forecasting US overseas travelling with univariate and multivariate models. *Journal of Forecasting, 40*(6), 963–976. https://doi.org/10.1002/for.2760

Palmer, A., Montaino, J. J., & Sesé, A. (2006). Designing an artificial neural network for forecasting tourism time series. *Tourism Management, 27*(5), 781–790. https://doi.org/10.1016/j.tourman.2005.05.006

Pan, B., Wu, D. C., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology, 3*(3), 196–210. https://doi.org/10.1108/17579881211264886

Prilli, Suci Karunia, Permanasari, Adhaidya Erna, & Fausiati, Silmi. (2021). The Effect of the COVID-19 Pandemic and Google Trends on the Forecasting of International Tourist Arrivals in Indonesia.

Sun, S., Wei, Y., Tui, K.-L., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management, 70*, 1–10. https://doi.org/10.1016/j.tourman.2018.07.010

Szekely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics, 35*(6), 2769–2794. https://doi.org/10.1214/009055607000000050

Tang, Ling, Zhang, Chengyu, Li, Tingfei, & Li, Ling. (2021). A novel BEMD-based method for forecasting tourist volume with search engine data. *27*(5), 1015-1038. 10.1177/1354816620929595

Volchek, Katerina, Liu, Anya, Song, Haiyan, & Buhalis, Dimitrios. (2019). Forecasting tourist arrivals at attractions: Search engine empowered methodologies. *25*(3), 425-447. 10.1177/1354816618811558.

Xie, G., Qian, Y., & Wang, S. (2020). A decomposition-ensemble approach for tourism forecasting. *Annals of Tourism Research, 81*, Article 102891. https://doi.org/10.1016/j.jaltour.2020.10.2891

Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management, 46*, 386–397. https://doi.org/10.1016/j.tourman.2014.07.019

Zhang, B., Li, N., Shi, F., & Law, R. (2020). A deep learning approach for daily tourist flow forecasting with consumer search data. *Asia Pacific Journal of Tourism Research, 25*(3), 323–339. https://doi.org/10.1080/10941665.2019.1709876

Zhang, Chong, Liu, Hanfei, Chen, Zhiyang, & Tang, Haiyan. (2021). *Tourism Forecast Based on Web Search Data and Sentiment Analysis of Social Network*. Paper presented at The 2nd International Conference on Computing and Data Science, Stanford, CA, USA. Doi: 10.1145/3448734.3450807.

Zhang, C., Tian, Y.-X., Fan, Z.-P., Liu, Y., & Fan, L.-W. (2020). Product sales forecasting using macroeconomic indicators and online reviews: A method combining prospect theory and sentiment analysis. *Soft Computing, 24*(9), 6213–6226. https://doi.org/10.1007/s00500-018-03742-1

Zhang, C., Tian, Y.-X., & Fan, Z.-P. (2022). Forecasting sales using online review and search engine data: A method based on PCA–DSFOA–BPNN. *International Journal of Forecasting, 38*(3), 1005–1024. https://doi.org/10.1016/j.ijforecast.2021.07.010

Zhang, Q., Wu, J., Ma, Y., Li, G., Ma, J., & Wang, C. (2022). Short-term load forecasting method with variational mode decomposition and stacking model fusion. *Sustainable Energy, Grids and Networks, 30*, Article 100622. https://doi.org/10.1016/j.sregen.2022.100622

Zhang, X., Lei, K. K., & Wang, S.-Y. (2008). A new approach for crude oil price analysis supported by PCA and improved BPNN using Baidu index. *Expert Systems, 25*(3), 279–286. 10.1016/j.eswa.2007.02.012

Zhao, E., Du, P., & Sun, S. (2022). Historical pattern recognition with trajectory similarity for daily tourist arrivals forecasting. *Expert Systems with Applications, 203*, Article 117427. https://doi.org/10.1016/j.eswa.2022.117427

C. Zhang and Y.-X. Tian