Predicting Urban Tourism Flow with Tourism Digital Footprints Based on Deep Learning

Fangfang Gu[1], Keshen Jiang[1], Yu Ding[2] and Xuexiu Fan[3*]
[1] College of Economics and Management
Nanjing University of Aeronautics and Astronautics
Nanjing 211106, China
[e-mail: gff_0817@163.com, jiangksh@126.com]
[2] College of Artificial Intelligence
Nanjing Agricultural University
Nanjing 210095, China
[e-mail: yuding@njau.edu.cn]
[3] School of Foreign Language and International Business
Guilin University of Aerospace Technology
Guilin 541004, China
[e-mail: helen-f@nuaa.edu.cn]
* Corresponding author: Xuexiu Fan

Received February 26, 2023; revised March 27, 2023; accepted April 10, 2023; published April 30, 2023

Abstract

Tourism flow is not only the manifestation of tourists' special displacement change, but also an important driving mode of regional connection. It has been considered as one of significantly topics in many applications. The existing research on tourism flow prediction based on tourist number or statistical model is not in-depth enough or ignores the nonlinearity and complexity of tourism flow. In this paper, taking Nanjing as an example, we propose a prediction method of urban tourism flow based on deep learning methods using travel diaries of domestic tourists. Our proposed method can extract the spatio-temporal dependence relationship of tourism flow and further forecast the tourism flow to attractions for every day of the year or for every time period of the day. Experimental results show that our proposed method is slightly better than other benchmark models in terms of prediction accuracy, especially in predicting seasonal trends. The proposed method has practical significance in preventing tourists unnecessary crowding and saving a lot of queuing time.

Keywords: Tourism flow, deep learning methods, travel diaries, Nanjing

http://doi.org/10.3837/tiis.2023.04.007 **ISSN：1976-7277**

1. Introduction

The generation and development of big data has rapidly occupied all fields of social production and other activities, forming various forms of display types such as text, photos, audio clips, videos and so on. More and more tourists choose to tour with the help of the Internet, instead of the traditional way of tourism activities [1,2]. As an important way of life and social economic activity in modern society, tourism has been deeply "branded" by Internet big data [3].

Tourism flow refers to the one-way or two-way flow of tourist flow, material flow, energy flow, cultural flow between tourism destinations or within the destination [4-6]. As an important form of interaction and interconnection between tourist source and destination, tourism flow has always been one of the core issues of tourism research [7,8]. The tourism flow is not only an important prerequisite for obtaining the precise trajectory of tourists at the scenic scale, but also the original power and core expression formed by the urban internal scale based on spatial organization and the regional scale based on the close relationship between socio-economic and tourism activities. However, it is an important problem of obtaining tourism flow data which restricting tourism research [9]. In previous studies, as lack of long-term fine-grained spatiotemporal data [10], scholars usually choose tourist number to represent tourism flow, which directly leads to the research on tourism flow is not in-depth enough. With the continuous development of Internet technology, more and more tourists choose the Internet to express the tourism process and experience [11], forming a wide range of network information, which constitutes a digital footprint. Tourism digital footprint is a concrete manifestation of tourism behavior [12]. With the increasing improvement of big data analysis methods, tourism digital footprint has become a hot research topic for tourism scholars [13-15]. At the same time, since the flow occurs on different time and space scales, it usually produces different network spaces and complex networks composed of different nodes and networks. It is these nodes and sub-networks with different levels that play different roles prompt the flow network to show different flow characteristics and flow laws.

For the forecast of tourism, scholars have used the time series models and the econometric models [16-19]. However, these statistical models are mostly based on linear assumptions, ignored of the nonlinearity and complexity exhibited by the tourism flow, so its prediction accuracy is poor. Therefore, how to capture the spatial and temporal dependencies of tourism flow is a great challenge. With the rapid development of artificial intelligence, deep learning methods have provided new vitality for tourism prediction. Deep learning constructs complex nonlinear network structures by modeling from the perspective of time and space, so as to better represent the rich change laws inherent in transactions [20], and has been widely used in many fields. Koprinska et al. [21] investigated the application of convolutional neural networks for energy time series forecasting. Huang et al. [22] used Graph Convolutional Neural Network (GCN) and gated recurrent unit to predict the daily demand of hotels. Huang et al. [23] proposes Spatio-Temporal Graph Convolutional Neural Network based on Periodic Component (Periodic ST-GCN) to predict the passenger glide at public transportation stations. Among the existing articles, few have pulled deep learning methods into the tourism flow prediction Consequently how to combine deep learning and tourism flow forecasting depth needs to be studied urgently.

Based on this, combined with deep learning methods such as GCN and Convolutional Neural Network (CNN) models, this study uses the travel diaries of domestic individual tourists in Nanjing collected by the "Octopus Collector" to discover and reveal the process and regular characteristics of tourist flow, furthermore, it also predicts urban tourism flow. The results show that: compared with common prediction models, the model proposed in our paper effectively improves the accuracy of urban tourism flow prediction.

2. Related Work

Tourism flow has always been an important research topic in the field of tourism [3,24]. Multi-disciplinary and multi-perspective communication and interaction promote the deepening of research, involving a wide range of research content, from the initial qualitative transformation to quantitative, and then to the combination of qualitative and quantitative transformation process. Throughout the research on tourism flow, we can found that the research content mainly focuses on the theoretical system of tourism flow, tourism flow scale, flow spatial pattern, flow influencing factors and so on [25-28]. The above research provides a good basis and material for the further study of tourism flow, but objectively, it can also be seen that the research on tourism flow relies more on questionnaire collection or statistical data [29,30]. Tourist flow is a dynamic process of change, which should not only be reflected in the results of flow, but also pay attention to the process analysis and the prediction of flow, which provides a new direction and perspective for the study of tourism flow. In the era of “Internet +” tourists prefer to share their travel diaries on the social media platform, which forms the tourism digital footprints [31]. As an electronic trace with geotag or location information formed by tourists' activities, tourism digital footprint can reflect the spatial movement of tourists in tourist destinations, and enable scholars to identify the spatial structure and evolution law of tourism flow with space as the entry point.

Traffic flow prediction, which is like tourism flow structure, has achieved some research results due to its important role in people's livelihood. In recent years, spatiotemporal models combining GCN and Recurrent Neural Network (RNN) have become a research hotspot in traffic prediction and have made significant progress [32,33]. Deep learning has attracted much attention due to its advantages in capturing nonlinear and complex patterns. Many deep learning methods have been applied to traffic prediction, such as RNN based on sequence prediction to predict traffic flow. However, RNN suffer from short-term memory problems. So, researchers have turned to the Long Short-term Memory Network (LSTM) and the Gated Recursion Unit (GRU) [34-37]. Although these models can capture the temporal correlation in traffic flow, researchers gradually realize the importance of spatial correlation and improve the prediction accuracy by introducing CNN to extract spatial information and combining it with LSTM. Since CNN are designed for Euclidean Spaces such as images and grids, they have limitations in transportation networks with non-Euclidean topologies, and thus cannot essentially characterize the spatial dependence of traffic flows. Emerging GCN is dedicated to dealing with network structure, which can better simulate the spatial dependence of road segments on the traffic network [38,39].

Based on the above research, scholars have carried out a series of studies on the application of digital footprint in tourism flow. However, these studies lack the analysis of exploring the tourists flow from temporal and spatial perspective, and moreover, lack the research on the prediction of tourism flow using big data. Therefore, this paper uses the collected travel diaries samples to mine the tourism spatial information of individual tourists in Nanjing, and uses the deep learning method to analyze the spatio-temporal dynamic changes from the perspective of tourists' behavior, and predicts the future tourism flow, as to provide some theoretical reference for urban tourism.

3. The Proposed Method

Tourist flow prediction of tourism nodes requires deep learning of temporal and spatial features. In this section, a model including spatial graph convolution network and temporal convolution network is constructed to study the spatial dependence of tourism nodes and the changes of characteristics at different time points, to achieve the prediction of passenger flow volume at future time points.

3.1. Tourist flow forecasting problem definition

To analyze the spatio-temporal characteristics of Nanjing's tourism flow and predict the tourist flow, this paper defines the network structure of Nanjing tourist attractions as graph $G = (V, E)$. $V = (v_1, v_2, \cdots, v_N)$ is a vertex set that represents N travel nodes in the sample travel diary. E is the set of edges, $e_{ij} = (v_i, v_j) \in E$ means that v_i and v_j are neighbors in the graph. The adjacency matrix $A \in R^{N \times N}$ represents the connection relationship between the tourism nodes. As shown in **Fig. 1**, the corresponding position value of the adjacency matrix is 0 to indicate that there is no tourist flow between the two tourism nodes corresponding to the sample data, and the corresponding position value of the adjacency matrix is 1 to demonstrate that there are tourist traveling between the two tourism nodes. Hence, it can be expressed by the formula:

$$\begin{cases} A_{ij} = 1\,, if\ e_{ij} \in E \\ A_{ij} = 0, if\ e_{ij} \notin E \end{cases} \tag{1}$$

The tourist flow volume data of each tourism node is represented as the graph signal on the graph. Let x_t denotes the features of Nanjing tourism nodes for the time interval. The characteristics of all nodes within the time interval t are $x_t \in R^{N \times \widehat{D}}$, where N represents the number of tourist nodes and $\widehat{D}$ denotes the characteristic dimension of each node. Tourist flow includes the number of inflow and outflow, and we then add the four features of historical popularity, landscape area, distance to city centre and month popularity to the tourist nodes, so $\widehat{D} = 6$. The method of tourist flow prediction is to learn a function $f(\cdot)$, which can predict the graph signal of the next time step T:

$$[X^{t-S:t}, G] \overset{f(\cdot)}{\rightarrow} [X^{t+1:t+T}] \tag{2}$$

where $X^{(t-S):t}$ and $X^{t+1:t+T} \in R^{N \times \widehat{D} \times T}$, S refers to the historical time step and T indicates the forecast time step.

	The Sun Yat-sen Mausoleum	Ming Xiaoling Mausoleum	Meiling Palace	Linggu Temple	Confucius Temple	Zhanyuan Garden	Lion Bridge	Old East Gate	Nanjing Museum	Nanjing Presidential Palace	1912 Block	Jiming Temple	The Nanjing Circumvallation	Xuanwu Lake	⋮	Underwater World	Chaotian Palace	The Nanjing Yangtze River Bridge	Nanjing Eye Footbridge	Ginkgo Lake Park	Meiyuan Xincun Memorial Hall	Nanjing Hongshan Forest Zoo	Zhongshan Botanical Garden	Nanjing Museum of Paleontology	Nanjing Olympic Sports Center	Jiuhua Mountain	Stone City Park	Tianfei Palace
The Sun Yat-sen Mausoleum	0	1	0	0	1	0	0	1	0	0	0	1	0	1		0	0	0	0	0	0	1	0	1	0	0	0	1
Ming Xiaoling Mausoleum	1	0	0	0	1	0	1	0	0	1	0	1	0	1		0	1	0	1	0	0	0	0	0	0	1	0	0
Meiling Palace	0	0	0	0	1	0	0	0	0	0	1	1	0	1		0	0	0	0	0	0	0	0	0	0	0	0	0
Linggu Temple	0	0	0	0	0	0	1	0	0	0	0	0	0	1		1	0	0	0	0	1	0	0	0	0	0	0	0
Confucius Temple	1	1	1	0	0	0	1	0	0	1	1	1	0	1		0	1	1	0	1	0	0	1	0	1	1	0	1
Zhanyuan Garden	0	0	0	0	0	0	0	0	0	0	0	1	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0
Lion Bridge	0	1	0	1	1	0	0	0	0	1	0	0	0	1		0	0	0	0	0	0	1	0	0	0	0	0	0
Old East Gate	1	0	0	0	0	0	0	0	0	0	0	1	0	0		0	0	0	0	0	0	0	0	0	1	0	0	0
Nanjing Museum	0	0	0	0	0	0	0	0	0	0	0	0	1	1		0	0	0	1	0	0	1	0	0	0	0	1	0
Nanjing Presidential Palace	0	1	0	0	1	0	1	0	0	0	0	1	0	1		0	0	0	0	0	0	0	0	0	0	0	0	0
1912 Block	0	0	1	0	1	0	0	0	0	0	0	1	0	0		0	0	0	0	0	0	0	1	0	0	0	0	0
Jiming Temple	1	1	1	0	1	1	0	1	0	1	1	0	0	1		0	0	1	0	0	0	0	0	1	0	0	0	0
The Nanjing Circumvallation	0	0	0	0	0	0	0	0	1	0	0	0	0	1		1	0	0	0	0	1	0	0	0	1	0	1	0
Xuanwu Lake	1	1	1	1	1	0	1	0	1	1	0	1	1	0		1	1	0	0	1	0	1	0	1	0	1	0	0
⋮															0													
Underwater World	0	0	0	1	0	0	0	0	0	0	0	0	1	1		0	1	0	0	0	1	0	0	0	0	0	0	0
Chaotian Palace	0	1	0	0	1	0	0	0	0	0	0	0	0	1		1	0	0	0	0	1	0	1	0	0	0	1	0
The Nanjing Yangtze River Bridge	0	0	0	0	1	0	0	0	0	0	0	1	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0
Nanjing Eye Footbridge	0	1	0	0	0	0	0	0	1	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0
Ginkgo Lake Park	0	0	0	0	1	0	0	0	0	0	0	0	0	1		0	0	0	0	0	0	1	0	0	0	1	0	0
Meiyuan Xincun Memorial Hall	0	0	0	1	0	0	0	0	0	0	0	0	1	0		1	1	0	0	0	0	0	0	0	0	0	0	0
Nanjing Hongshan Forest Zoo	1	0	0	0	0	0	1	0	1	0	1	0	0	1		0	1	0	1	0	0	0	1	0	0	1	0	0
Zhongshan Botanical Garden	0	0	0	0	1	0	0	0	0	0	1	0	0	0		0	1	0	1	0	0	0	0	0	0	0	0	0
Nanjing Museum of Paleontology	1	0	0	0	0	0	0	0	0	0	0	1	0	1		0	0	0	0	0	0	0	0	0	0	0	0	0
Nanjing Olympic Sports Center	0	0	0	0	1	0	0	1	0	0	0	0	1	0		0	0	0	0	0	0	0	0	0	0	0	0	0
Jiuhua Mountain	0	1	0	0	1	0	0	0	0	0	0	0	0	1		0	0	0	0	1	0	1	0	0	0	0	0	0
Stone City Park	0	0	0	0	0	0	0	0	1	0	0	0	1	0		0	1	0	0	0	0	0	0	0	0	0	0	0
Tianfei Palace	1	0	0	0	1	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 1. Graph adjacency matrix A.

3.2 Framework of tourist flow prediction model

The tourist flow of tourism nodes not only has the attribute of time dimension, but also has a strong dependence on the distribution of tourism nodes in the city, that is, the spatial structure. In order to obtain temporal and spatial characteristics, as shown in **Fig. 2**, we construct a tourist flow prediction model that includes temporal convolution and spatial convolution, which consists of stacked multiple spatio-temporal convolutional layers and two fully connected layers. The spatio-temporal convolutional layer includes the Temporal Convolutional Network (TCN) module and GCN module. The selected travel diary samples with a time span of one year are divided into 12 moments by month. For example, x_1 represents all the characteristics of tourism nodes from July to August, and x_{12} represents all the characteristics of nodes from June to July of the next year. Firstly, the time series feature $X = [x_1, x_2, \cdots, x_{12}]$is input into the temporal convolution network, and the processed time series feature $H(X)$ is obtained through extended causal convolution, which is used as the input $H^{(l)}$of the graph convolution network in Equation (6). In summary, the spatio-temporal feature sequence is extracted through the spatio-temporal convolutional layer, and then input it into the fully connected network to predict the passenger flow at time T in the future.

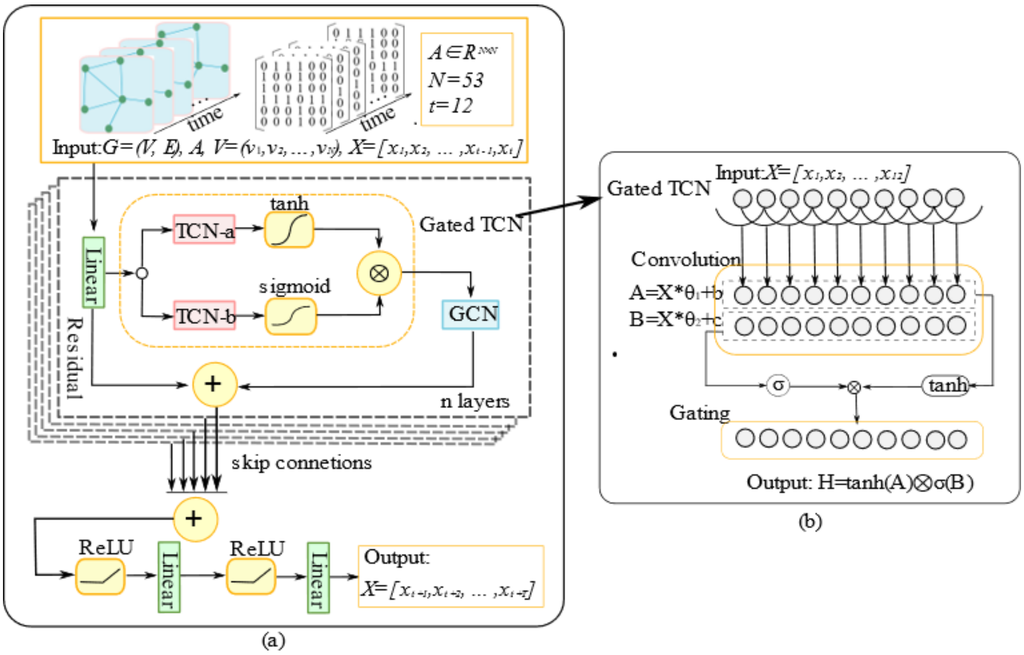

Fig. 2. (a): Architecture of the tourist flow prediction model. (b): Gated TCN schematic diagram. Where N is the number of tourism nodes in the sample, and n is the number of spatio-temporal convolution layers.

3.3 Temporal feature extraction

The TCN module in this paper is based on the Gated Tanh Unit, which is divided into two parallel temporal convolution layers, TCN-a and TCN-b, as shown in the **Fig. 2**. TCN-a is a linear convolution with tanh activation function, TCN-b is the gating part, which controls what information can be passed to the next layer through sigmoid function. Given the characteristics $X = [x_1, x_2, \cdots, x_t]$ of tourism nodes at each time t as input, the convolution of the hidden layer can be expressed as:

$$H(X) = tanh(X * \theta_1 + b) \otimes \sigma(X * \theta_2 + c) \tag{3}$$

where θ_1, θ_2, b and c are learnable parameters of the model, $g(\cdot)$ is the activation function of the output layer, σ is the sigmoid function that maps any number between 0 and 1, $\otimes$ is the element-wise product between matrices.

Recurrent neural network (RNN) is generally used to process time series data, which can learn the dependence of time dimension, but with the increase of time, RNN will generate the problem of gradient disappearance or gradient explosion, and our TCN based on CNN model adopts dilated causal convolution to help alleviate these problems. Moreover, dilated convolution can exponentially expand the receptive field, making CNN capture long-term dependence with fewer layers. As shown in the **Fig. 3**, the causal convolution at time t of each layer is obtained by the corresponding value at time t of the previous layer and the value before it. Dilated convolution is to inject cavities on the basis of standard convolution map to increase reception fields. Suppose that is a filter $F = (f_1, f_2, \cdots, f_K)$, sequence $X =$

$(x_1, x_2, \cdots, x_t)$, the causal convolution and dilated convolution with dilatation rate of d at x_t are respectively as follows:

$$(F * X)_{x_t} = \sum_{k=1}^{K} f_k x_{t-K+k} \tag{4}$$

$$(F *_d X)_{x_t} = \sum_{k=1}^{K} f_k x_{t-(K-k)d} \tag{5}$$

where d is a hyperparameter dilation rate, which refers to the number of kernel intervals, and dilatation rate is equal to 1 in standard CNN. K is the size of the convolution kernel that determines the number of inputs selected.

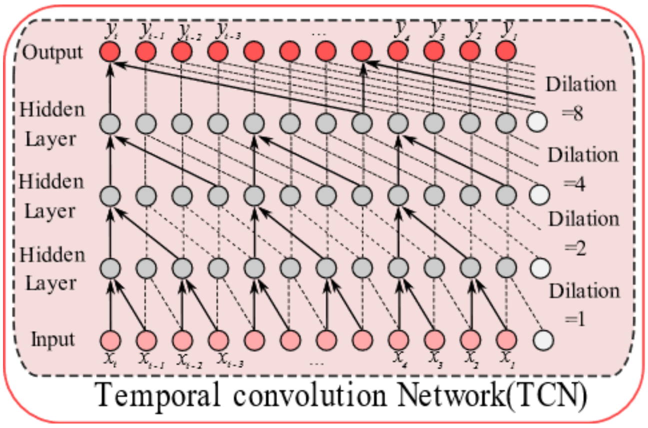

Fig. 3. A general framewor of TCN.

3.4 Spatial feature extraction

Because of the limitation of traditional CNN for supplied data, only Euclidean data with spatial rules can be received and further processed. Graph and manifolds, for instance, input of non-Euclidean data, are challenged to be operated by convolution because of their non-translational invariance and inconsistent number of neighbor nodes. To this end, Kipf et al. proposed the GCN, which applies convolution operations to the graph structure to better extract the implicit spatial relationships in the graph. **Fig. 4** is a hidden structure of the graph convolutional network used herein. In our model, the sequence feature $H \in R^{N \times \hat{D} \times \hat{L}}$ and the adjacency matrix $A \in R^{N \times N}$ on the time series of the tourism node graph are used as inputs to GCN, where N is the number of tourism nodes, $\hat{D}$ is the dimension of the feature, which is 6, and $\hat{L}$ represents the sequence length.

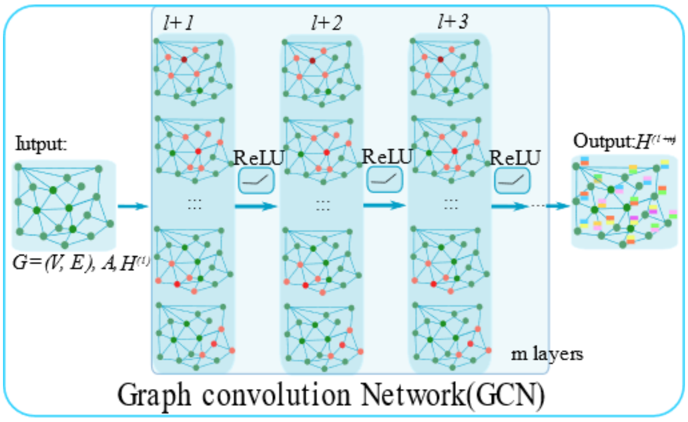

Fig. 4. A general framework of GCN.

The calculation formula of graph convolution is given in GCN, as shown in Equation (6), where $H^{(l)}$ represents the feature vector of the node at the l layer, which is the feature sequence convolved by the TCN module. $H^{(l+1)}$ constitutes the feature vector of the node at the $l+1$ layer after convolution, $W^{(l)}$ represents the convolution parameter at the l layer, and δ represents the activation function. $D \in R^{N\times N}$ is the degree matrix of the graph, and all other elements except the diagonal are 0. The diagonal values are calculated as Equation (7). D_{ii} indicates the number of nodes connected to node i. The part composed of matrices A and D is a kind of Laplace matrix. I_N is the identity matrix, that is, the matrix whose diagonal is 1 and others are 0.

$$H^{(l+1)} = \delta(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{6}$$

$$D_{ii} = \sum_j A_{ij},\ \tilde{A} = A + I_N,\ \widetilde{D}_{ii} = \sum_j \tilde{A}_{ij} \tag{7}$$

GCN can be divided into two main algorithms, spatial graph convolution network and spectral graph convolution network. We will focus on spectral convolution. For the signal $x \in R^n$, with the symmetric normalized Laplacian $L = I_N - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, to be Fourier transformed, orthogonal matrix $U = (\vec{u}_1, \vec{u}_2, \cdots, \vec{u}_n)$, which is the matrix of eigenvectors of L, is used to combine x linearly. In Laplace matrix L's decomposition: H^{l+1}. The Fourier transform is a process which can split a function e.g. $f(t)$ satisfying certain conditions into the sum of countless sine waves which possess various frequencies or linear combination of integral forms, which has been used widely in signal processing.

$$F(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)e^{\ (-i\varpi t)}\ dt \tag{8}$$

where $F(\omega)$ denotes the coefficient of a wave with angular frequency ω, which can be seen as the coordinate on the corresponding base of ω, when function $f(t)$ is projected to basis function $e^{i\omega t}$. For several forms of Fourier transform, Discrete-time Fourier series (DFS) is that the Fourier transform takes a discrete form in both the time and frequency domains in which the coefficient series is periodic. For a sequence y(n) with a period N, its discrete Fourier series expansion is,

$$y(n) = \sum_{k=0}^{N-1} cke^{j2\pi\frac{k}{N}n} \tag{9}$$

Since the graph structure is not a standard discrete Fourier transform, it is derived by analogy with this definition. The Fourier transform is a projection based on the eigenvectors of the Laplace operator, then the graph Fourier transform is a projection based on the eigenvectors of the Laplace matrix. Signal $x \in R^n$ can be expressed as a linear combination of Laplace's eigenvectors:

$$x = \hat{x}(\lambda_1)\vec{u}_1 + \hat{x}(\lambda_2)\vec{u}_2 + \cdots + \hat{x}(\lambda_n)\vec{u}_n \tag{10}$$

where, $\hat{x}(\lambda_n)$ is Fourier coefficient. $\hat{x}$ represents x's signal in spectral domain. Thus, we get the definition of the inverse Fourier transform and Fourier transform on graph:

$$x = U\hat{x},\ \hat{x} = U^T x \tag{11}$$

For the convolution kernel g_θ, we also convert it to the spectral domain. Thus, the convolution in graph can be defined by analogy with the traditional convolution definition:

$$g_\theta * x = U((U^T g) \odot (U^T x)) = U g_\theta(\Lambda) U^T x \tag{12}$$

where $g_\theta(\Lambda)$ is the convolution kernel, θ is the convolution kernel parameter that the model needs to learn. The whole convolution is the original signal is Fourier transformed to the spectral domain, in which convoluted with the convolution kernel, and then reduced to the spectral domain by doing the inverse Fourier transform. However, this approach is computationally intensive. A truncated expansion in terms of Chebyshev polynomials $T_k(x)$ up to K^{th} order can accurately approximate the value of $g_\theta(\Lambda)$ by Hammond et al. (2011). By this method, the convolution definition can be updated:

$$g_\theta * x = \sum_{k=0}^{K} \theta_k T_k(\tilde{L})x \tag{13}$$

where K is the parameter of convolution kernel. θ_k is now a vector of Chebyshev coefficients. $\tilde{L} = \frac{2}{\lambda_{max N}}$ is the normalized Laplacian matrix. λ_{max} is the maximum eigenvalue of L. Specially, $K = 1$, λ_{max}, after re-normalization trick Eq. (11) can be transformed into:

$$g_\theta * x = \theta(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}})x \tag{14}$$

The final fast graph convolution definition is obtained after adding the activation function:

$$H^{(l+1)} = f(H^l, A) = \delta(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{15}$$

where $H^{(l)} \in R^{N\times\widehat{D}\times\hat{L}}$, N is the number of tourism nodes, $\widehat{D}$ represents the feature dimension of nodes, and $\hat{L}$ represents the sequence length.

4. Case Study and Result

4.1 Case study region in Nanjing, China

Nanjing is the capital of Jiangsu Province, located in eastern China and the lower reaches of the Yangtze River (**Fig. 5**). Affected by subtropical monsoon climate, Nanjing has four seasons with distinct characteristics. As a crucial node city in the Yangtze River Delta, Nanjing is an important gateway connecting the central and western regions, with developed regional transportation and high level of economic development. At the same time, as the ancient capital of the Six Dynasties, Nanjing has a long history and has many historical and cultural landscape resources and natural landscape resources. It is a city suitable for tourism and attracts tourists at home and abroad. Therefore, this paper takes Nanjing as a case site of tourism study, investigating into the network structure of tourism flow, hoping to provide some basis for the future construction and development of Nanjing tourism.

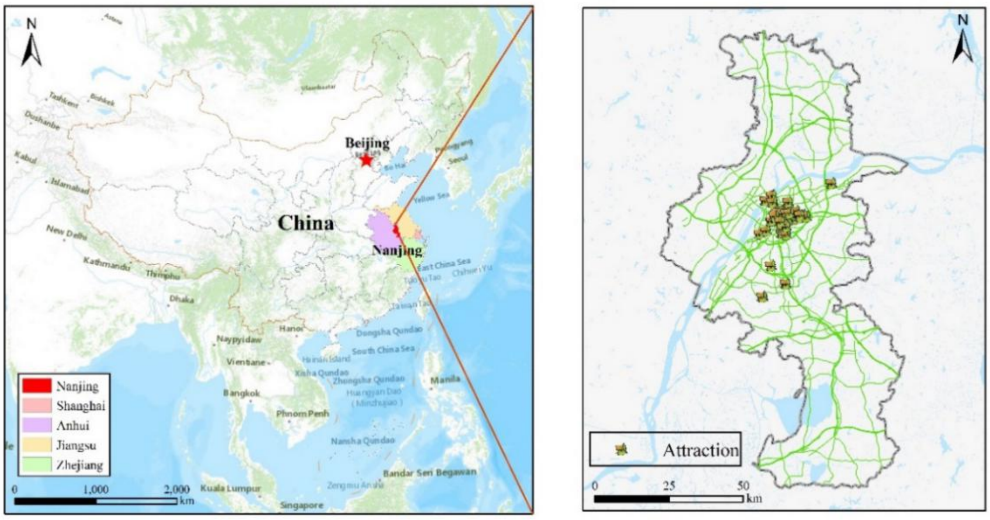

Fig. 5. Distribution of attractions in Nanjing, China.

4.2 Data acquisition

This paper uses the web scraping software "Octoparse" to collect the travel diaries of Nanjing released by tourists who tour around Nanjing on the Mafengwo website. Mafengwo.cn (http://www.mafengwo.cn/) is a leading travel search engine and entertainment community in China used more by the young generation. It is one of the largest Chinese social media sharing platforms for online travel diaries. The data from Mafengwo.cn has been applied to the study of post replying behavior [40], tourists' behavior [41], and co-visitation network [42].

In the actual collection, enter "Nanjing" on the travel diaries page of the Mafengwo website, and arrange the searched travel diaries of Nanjing according to the latest update time. This paper studies the tourism digital footprint of Nanjing in recent one year. Considering the time difference between the release time of tourists and the actual travel time, the published travel diaries from July 1, 2017 to August 1, 2018 are selected for data collection. 1994 travel diaries are collected and screened in Excel, and 670 travel diaries meets the research standards are finally counted. Screening criteria of travel diaries are: (1) the authors of travel diaries must be non-local independent tourists visiting Nanjing; (2) the travel diaries should be able to reflect the complete footprint journey; (3) photos and description are needed in the travel diaries to verify the authenticity of the footprints; (4) the travel diaries should reflect the specific travel time and duration of stay, and the travel time should be between July 1, 2017 and July 1, 2018.

4.3 Temporal characteristic analysis

4.3.1 Analysis on travel time

After processing the statistics, it is found that there are obvious seasonal characteristics as for the travel time of non-local independent tourists in Nanjing, shown in the changing curve in **Fig. 6**. The sample data shows there are two obvious tourism peaks in October 2017 and April 2018, and more independent tourists pay visits to Nanjing in springs and autumns. A possible explanation is that Nanjing has a changeable climate that spring and autumn is shorter compared to winter and summer. It is not a very comfortable holiday city. In April, the weather turns warmer and the climate is more comfortable. At the same time, it coincides with International Labour Day and Tomb Sweeping national holiday, during which time tourists have leisure time to travel. In October, the climate is slowly cool, and promoted by the National Day Golden Week, the number of tourists gradually increases. After the Golden Week, the number of tourists decreases slowly as the temperature gradually becomes cold. The overall number of tourists is greatly limited by leisure time and the climate change.

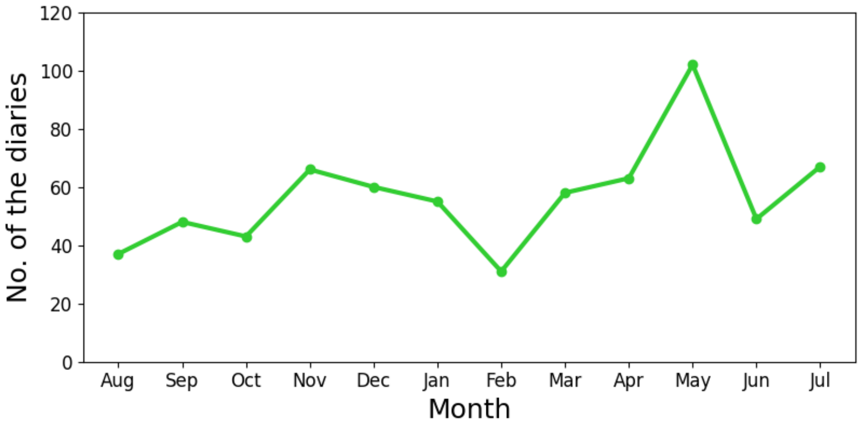

Fig. 6. Travel time of domestic independent tourists in Nanjing.

4.3.2 Analysis on duration of stay

The analysis (**Fig. 7**) shows that the stay time of non-local independent tourists in Nanjing ranges from 1 day to 10 days, and the average length of stay time is 2.88 days. Among them, the proportion of samples staying for 1~3 days, 4~5 days, 6~7 days and more than seven days in the total samples are 77.16%, 20.15%, 2.24% and 0.45% respectively. The number of tourists staying for three days is the largest, accounting for 37.61% of the total, which is consistent with the average length of stay in Nanjing. Most independent tourists stay in Nanjing for 1~5 days, with a short average stay time.

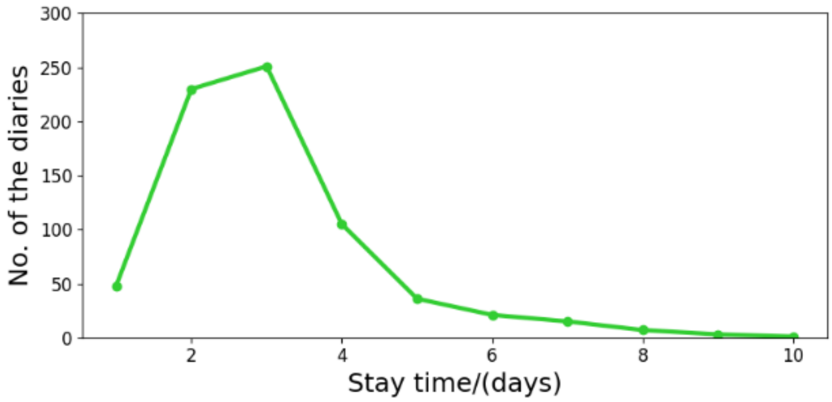

Fig. 7. Length of stay time of non-local independent tourists in Nanjing.

4.4 Spatial Characteristic Analysis

In the process of extracting actual footprints, it is found that tourists have gone to many unconventional tourist attractions, such as IBRAIRIE AVANT-GARDE, Lion Bridge, Xinjiekou, etc. Meanwhile, some colleges and universities in Nanjing have also tun into tourist attractions. In order to ensure the integrity of the tourism flow footprint, these unconventional scenic spots are also classified together with traditional tourist attractions as "Tourism node". For Example, the author finds that tourists who visit the Confucius Temple will also normally go to the Wuyi Lane, Imperial Examination Museum of China and other places, so they are merged into Confucius Temple node. While tourists going to Pukou Railway Station will generally pass through Zhongshan Wharf and Pukou Wharf, so they are merged into the Pukou Railway Station node.

According to statistics (**Table 1**), there are altogether 53 tourism nodes in the travel diaries sample. The top 20 nodes ranked in the order of frequency of occurrence are as follows. There are 13 nodes with frequency greater than 20%, 7 nodes with frequency greater than 10% but less than 20%, and the frequencies of the other 33 tourism nodes are less than 10%, and most of them are less than 3%, showing obvious long-tail characteristics. The preliminary results show that 13 tourism nodes with a frequency of more than 20% can be regarded as the core tourism nodes in Nanjing, among which the Sun Yat-sen Mausoleum, Ming Xiaoling Mausoleum and Meiling Palace constitute Zhongshan tourism area, the Xuanwu Lake, Jiming Temple and Nanjing Presidential Palace constitute central urban tourism area, and the Confucius Temple and Old East Gate constitute Confucius Temple-Qinhuai tourism area. This is basically consistent with Jin Cheng's finding that Nanjing scenic spots are mainly divided into three systems: the Sun Yat-sen Mausoleum, Confucius Temple and central urban scenic spot system. In addition to these landmark scenic spots in Nanjing, the tourism flow of IBRAIRIE AVANT-GARDE (Wutaishan head office) - Nanjing University - Southeast University attracts more young people who pursue a sense of cultural experience, while as an old CBD area, Xinjiekou, like Confucius Temple, is a tourism node where tourists would rather visit at night.

Table 1. Ranking of tourism node frequency and probability statistics.

Serial number	Tourism node	Frequency	Probability (%)	Serial number	Tourism node	Frequency	Probability (%)
1	Confucius Temple	486	72.54%	11	Xinjiekou	164	24.48%
2	The Sun Yat-sen Mausoleum	312	46.57%	12	Meiling Palace	152	22.69%
3	Nanjing Presidential Palace	308	45.97%	13	The Nanjing Circumvallation	136	20.30%
4	Nanjing Museum	260	38.81%	14	Nanjing University	109	16.27%
5	Old East Gate	240	35.82%	15	1912 Block	97	14.48%
6	Jiming Temple	233	34.78%	16	Line Friends	77	11.49%
7	Xuanwu Lake	231	34.48%	17	Zhanyuan Garden	76	11.34%
8	Ming Xiaoling Mausoleum	227	33.88%	18	Lion Bridge	76	11.34%
9	The Memorial Hall of the Victims in Nanjing Massacre by Japanese Invaders	207	30.90%	19	Linggu Temple	70	10.45%
10	IBRAIRIE AVANT-GARDE	188	28.06%	20	Southeast University	68	10.15%

4.5 Model evaluation metrics

In this paper, mean absolute error (MAE) and root mean square error (RMSE) are adopted as the evaluation indexes of the prediction model, and the specific calculation formula is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{16}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{17}$$

where n is the number of samples, y_i is the true value and $\hat{y}_i$ is the predicted value. MAE and RMSE are used to reflect the deviation between the predicted value and the real value. The smaller the MAE and RMSE values are, the better the prediction performance of the model is. We compare the model used in this paper with other models. The performance comparison results with other models are shown in **Table 2**. In the data set extracted from travel diaries, our model is slightly better than other benchmark models in accuracy. We have minimal MAE and RMSE. **Fig. 8** shows the predicted results of our dataset using different models. The fitting degree of our model and STGCN model is better, and the error between predicted value and real value is the smallest. The fitting effect of LSTM and DCRNN is relatively poor.

Table 2. RMSE and MAPE values for different models.

Model	MAE (person-time)	RMSE (person-time)
Our Model	5.43	6.63
STGCN	5.56	6.78
LSTM	5.83	7.12
DCRNN	5.74	6.95

Note: STGCN: Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. LSTM: Long short-term memory. DCRNN: Diffusion convolution recurrent neural network.

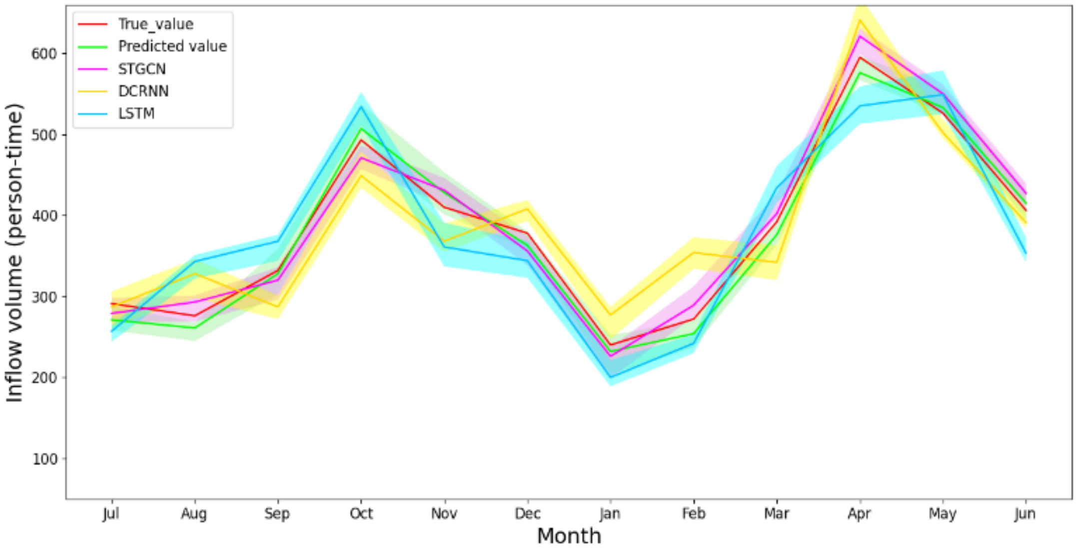

Fig. 8. The true values of tourist flow recorded in the tourist diary from August to July of the next year in Nanjing was compared with the predicted values of each model.

4.6. Analysis of tourist flow prediction results

Fig. 9 shows the prediction results of tourist flows of Confucius Temple, Sun Yat-sen Mausoleum and Yuhuatai Scenic Spot in Nanjing according to the tourism diary data, and compares them with the actual values. The popularity of these three scenic spots is divided into three levels, which is convenient to analyze the model prediction results more clearly. The model is more accurate in predicting the trend of passenger flow with seasonal changes.

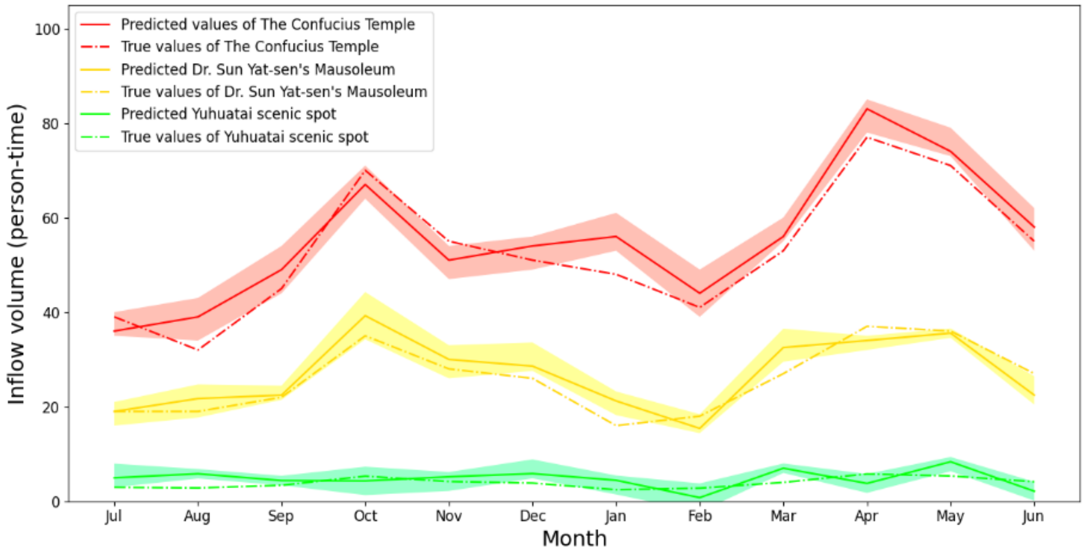

Fig. 9. Comparison of predicted and real tourist flow of Confucius Temple, Sun Yat-sen Mausoleum and Yuhuatai scenic spots.

Fig. 10 provides a visual representation of the predicted flow of visitors to attractions in Nanjing. Confucius Temple is a very popular attraction and the destination of tourists going out from Confucius Temple is scattered in several other attractions. Sun Yat-sen Mausoleum is a famous attraction with the characteristics of the Republic of China, visitors who come here to visit have a certain interest history of the Republic of China, so visitors can be seen flowing from Sun Yat-sen Mausoleum to such Republic of China buildings as 1912 Block and Meiling Palace.

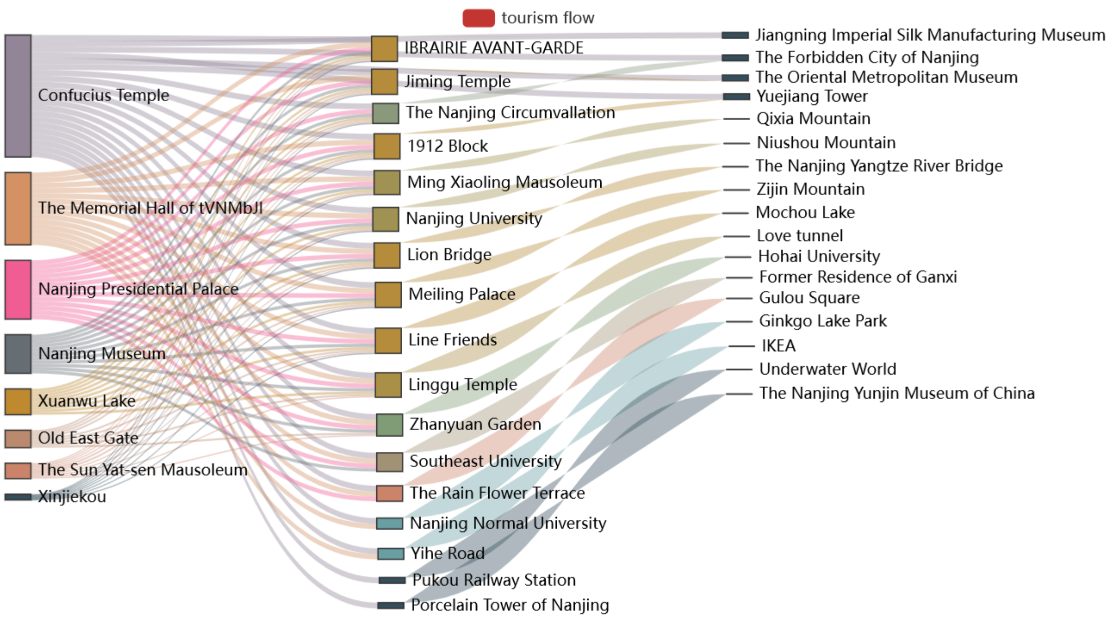

Fig. 10. Tourist flow map of scenic spot.

In a stable external environment, the other four characteristics of tourism nodes: historical popularity, landscape area, distance to city center and month popularity, have very weak temporal and spatial dependence, and the results after model processing do not change from before. **Fig. 11** shows the hierarchical heatmap of some scenic spots in Nanjing. The historical popularity, landscape area, and the distance to city center and month popularity is divided into 1 to 10 levels.

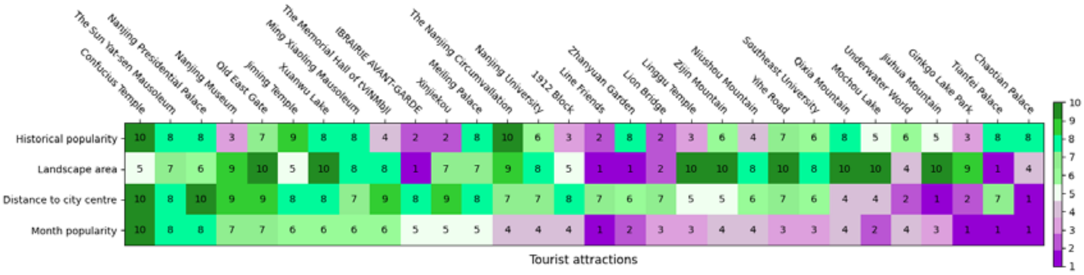

Fig. 11. Hierarchical heatmap.

Our current predictions and verification are based on the travel diaries of Nanjing released by tourists who tour around Nanjing on the Mafengwo website. But its application in real life can provide convenience to tourists and help plan the development of tourism. For example, as shown in **Fig. 12**, our model can be applied to predict the tourist flow of tourist attractions on every day of a year or in each time period of a day, so as to save tourists unnecessary crowding and a lot of queuing time. At the same time, it can also play an important role in traffic planning.

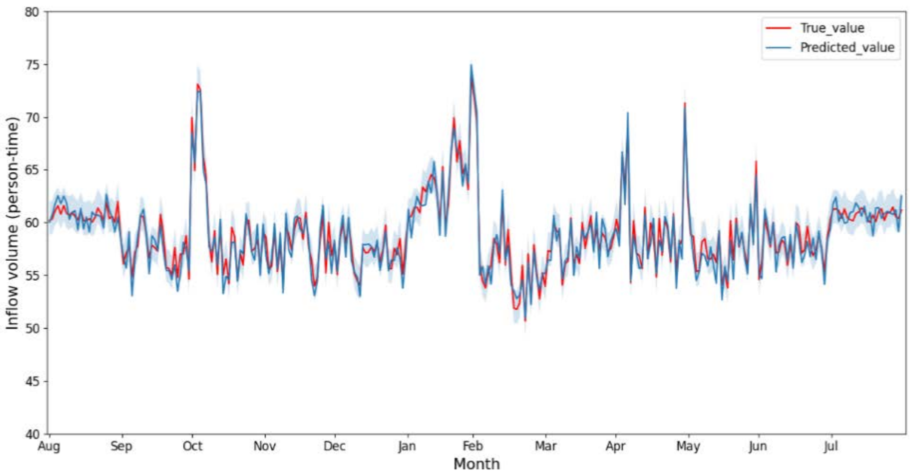

Fig. 12. Daily tourist flow forecast.

5. Conclusion and Discussion

In this paper, a spatio-temporal convolutional network model is constructed to obtain the spatio-temporal dependence relationship to predict the tourist flow of tourism nodes. The results show that compared with STGCN neural network used for traffic prediction, the model used in this paper is more suitable for tourism node tourist flow prediction, and has better performance in terms of prediction accuracy. Hence, in the era of big data, tourism is also developing in the direction of popularization and networking. Through the collection and statistics of tourism digital footprints, it can yet be regarded as a renewed research method to obtain the most direct data from large number of samples and big data, and then study the changing characteristics of tourism flow or tourists' temporal and spatial behaviors. Since 2020, the tourism industry has stagnated due to the normalized impact of the global COVID-19. How to scientifically predict the destination tourism flow is more important for the development of the local tourism industry. There are still some deficiencies in this study. For example, in terms of data sets, our paper only collects the travel diaries of domestic tourists in Nanjing, but does not collect the foreign tourists’. In the next step, it is necessary to expand the collection scope for experimental verification, so as to improve the prediction model.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. (42001155). The authors are grateful to the anonymous reviewers whose comments helped to improve this paper.

References

[1] F. Gu, K. Jiang, X. Hu, and J. Yang, “Deep learning-based image geolocation for travel recommendation via multi-task learning,” *Journal of Circuits, Systems and Computers*, vol. 31, no. 7, 2250127, Jan. 2022. Article (CrossRef Link)
[2] K. Zhang, Y. Chen, and C. Li, “Discovering the tourists' behaviors and perceptions in a tourism destination by analyzing photos' visual content with a computer deep learning model: The case of Beijing,” *Tourism Management*, vol. 75, no. 6, pp. 595-608, Dec. 2019. Article (CrossRef Link)
[3] K. Pavlovich, “The evolution and transformation of a tourism destination network: The Waitomo Caves, New Zealand,” *Tourism Management*, vol. 24, no. 2, pp. 203-216, Apr. 2003. Article (CrossRef Link)
[4] B. Kadar and M. Gede, “Tourism flows in large-scale destination systems,” *Annals of Tourism Research*, vol. 87, no. 2, 103113, Mar. 2021. Article (CrossRef Link)
[5] N. Mou, Y. Zheng, T. Makkonen, and T. Yang, “Tourists’ digital footprint: The spatial patterns of tourist flows in Qingdao, China,” *Tourism Management*, vol. 81, no. 6, 104151, Dec. 2020. Article (CrossRef Link)
[6] J. Zhang and C. Jensen, “Comparative advantage: Explaining Tourism Flows,” *Annals of Tourism Research*, vol. 34, no. 1, pp. 223-243, Jan. 2007. Article (CrossRef Link)
[7] L. Okafor, C. Tan, and U. Khalid, “Does panda diplomacy promote Chinese outbound tourism flows?,” *Journal of Hospitality and Tourism Management*, vol. 49, no. 4, pp. 54-64, Dec. 2021. Article (CrossRef Link)
[8] B. Prideaux, “Factors affecting bilateral tourism flows,” *Annals of Tourism Research*, vol. 32, no. 3, pp. 780-801, Jul. 2005. Article (CrossRef Link)
[9] R. Ahas, A. Aasa, A. Roose, U. Mark, and S. Silm, “Evaluating passive mobile positioning data for tourism surveys: An Estonian case study,” *Tourism Management*, vol. 29, no. 3, pp. 469-486, Jun. 2008 Article (CrossRef Link)

[10] X. Xu, Z. Huang, G. Hou, and C. Zhang, "The spatial spillover effects of haze pollution on inbound tourism: evidence from mid-eastern China," *Tourism Geographies*, vol. 22, no. 1, pp. 83-104, Apr. 2019. Article (CrossRef Link)

[11] Z. Liu, A. Wang, K. Weber, E. H. Chan, and W. Shi, "Categorization of cultural tourism attractions by tourist preference using location-based social network data: The case of central, Hong Kong," *Tourism Management*, vol. 90, no. 3, 104488, Jun. 2022. Article (CrossRef Link)

[12] F. Girardin, F. Calabrese, F. D. Fiore, C. Ratti, and J. Blat, "Digital footprinting: Uncovering tourists with user-generated content," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 36-43, Oct. 2008. Article (CrossRef Link)

[13] N. Mou, T. Yang, H. Zhang, J. Tang, and T. Makkonen, "Exploring spatio-temporal changes of city inbound tourism flow: The case of Shanghai, China," *Tourism Management*, vol. 76, no. 1, 103955, Feb. 2020. Article (CrossRef Link)

[14] M. Salas-Olmedo, B. Moya-Gómez, J. García-Palomares, and J. Gutierrez, "Tourists' digital footprint in cities: comparing big data sources," *Tourism Management*, vol. 66, no. 3, pp. 13-25, Jun. 2018. Article (CrossRef Link)

[15] Y. Yang, Y. Fan, L. Jiang, and X. Liu, "Search query and tourism forecasting during the pandemic: When and where can digital footprints be helpful as predictors?," *Annals of Tourism Research*, vol. 93, no. 2, 103365, Mar. 2022. Article (CrossRef Link)

[16] H. Song and S. Lin, "Impacts of the financial and economic crisis on tourism in Asia," *Journal of Travel Research*, vol. 49, no. 1, pp. 16-30, Jan. 2010. Article (CrossRef Link)

[17] H. Song, A. Liu, G. Li, and X. Li, "Bayesian bootstrap aggregation for tourism demand forecasting," *International Journal of Tourism Research*, vol. 23, no. 5, pp. 914-927, Apr. 2021. Article (CrossRef Link)

[18] W. Hong, Y. Dong, L. Chen, and S. Wei, "SVR with hybrid chaotic genetic algorithms for tourism demand forecasting," *Applied Soft Computing*, vol. 11, no. 2, pp. 1881-1890, Mar. 2011. Article (CrossRef Link)

[19] D. Gounopoulos, D. Petmezas, and D. Santamaria, "Forecasting tourist arrivals in Greece and the impact of macroeconomic shocks from the countries of tourists' origin," *Annals of Tourism Research*, vol. 39, no. 2, pp. 641-666, Apr. 2012. Article (CrossRef Link)

[20] Z. Gu, C. Chen, J. Zheng, and D. Sun, "Traffic flow prediction based on STG-CRNN," *Control and Decision*, vol. 2022, no. 3, pp. 645-653, Mar. 2022. Article (CrossRef Link)

[21] I. Koprinska, D. Wu, and Z. Wang, "Convolutional neural networks for energy time series forecasting," in *Proc. of International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, Brazil, pp. 1-5, Jul. 2018. Article (CrossRef Link)

[22] L. Huang, C. Li, and W. Zheng, "Daily hotel demand forecasting with spatiotemporal features," *International Journal of Contemporary Hospitality Management*, vol. 35, no. 1, pp. 26-45, Jan. 2023. Article (CrossRef Link)

[23] W. Huang, B. Cao, X. Li, and M. Zhou, "Passenger flow prediction for public transportation stations based on spatio-temporal graph convolutional network with periodic components," *Journal of Circuits, Systems and Computers*, vol. 31, no. 7, 2250134, Jul. 2022. Article (CrossRef Link)

[24] D. Dredge, "Networks, conflict and collaborative communities," *Journal of Sustainable Tourism*, vol. 14, no. 6, pp. 562-581, Mar. 2006. Article (CrossRef Link)

[25] A. Kimbu and M. Ngoasong, "Centralized decentralization of tourism development: A network perspective," *Annals of Tourism Research*, vol. 40, no. 1, pp. 235-259, Jan. 2013. Article (CrossRef Link)

[26] S.-H. Lee, J.-Y. Choi, S.-H. Yoo, and Y.-G. Oh, "Evaluating spatial centrality for integrated tourism management in rural areas using GIS and network analysis," *Tourism Management*, vol. 34, no. 1, pp. 14-24, Feb. 2013. Article (CrossRef Link)

[27] S.-G. María, J. Francisco, and J. Ledesma-Rodríguez, "International trade and tourism flows: An extension of the gravity model," *Economic Modelling*, vol. 52, no. 2, pp. 1026-1033, Jan. 2016. Article (CrossRef Link)

[28] N. Scott, C. Cooper, and R. Baggio, "Destination networks: four Australian cases," *Annals of Tourism Research*, vol. 35, no. 1, pp. 169-188, Jan. 2008. Article (CrossRef Link)
[29] B. Liu, S. Huang, and H. Fu, "An application of network analysis on tourist attractions: The case of Xinjiang, China," *Tourism Management*, vol. 58, no. 1, pp. 132-141, Feb. 2017. Article (CrossRef Link)
[30] X. Yang and Q. Wang, "Exploratory space-time analysis of inbound tourism flows to China cities," *International Journal of Tourism Research*, vol. 16, no. 3, pp. 303-312, Mar. 2014. Article (CrossRef Link)
[31] F. Girardin, F. D. Fiore, C. Ratti, and J. Blat, "Leveraging explicitly disclosed location information to understand tourist dynamics: A case study," *Journal of Location Based Services*, vol. 2, no. 1, pp. 41-56, Sep. 2008. Article (CrossRef Link)
[32] L. Xing, W. Zi-Li, R. Yi, Y. De-Zhen, F. Qiang, S. Bo, and L. Du, "Traffic network resilience analysis based on the GCN-RNN prediction model," in *Proc. of International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE)*, Zhangjiajie, China, pp. 96-103, Aug. 2019. Article (CrossRef Link)
[33] H. Zhu, Y. Xie, W. He, C. Sun, K. Zhu, G. Zhou, and N. Ma, "A novel traffic flow forecasting method based on RNN-GCN and BRB," *Journal of Advanced Transportation*, vol. 2020, no. 1, 7586154, Oct. 2020. Article (CrossRef Link)
[34] R. Fu, Z. Zhang, and L. Li, "Using LSTM and GRU neural network methods for traffic flow prediction," in *Proc. of 31st Youth Academic Annual Conference of Chinese Association of Automatic*, Wuhan, China, pp. 324-328, Nov. 2016. Article (CrossRef Link)
[35] E. Tuna and A. Soysal, "LSTM and GRU based traffic prediction using live network data," in *Proc. of IEEE Conference on Signal Processing and Communications Applications (SIU)*, Istanbul, Turkey, pp 1-4, Jun. 2021. Article (CrossRef Link)
[36] S. Wang, J. Zhao, C. Shao, C. Dong, and C. Yin, "Truck traffic flow prediction based on LSTM and GRU methods with sampled GPS data," *IEEE Access*, vol. 8, no. 1, pp. 208158-208169, Jan. 2020. Article (CrossRef Link)
[37] C. Zeng, C. Ma, K. Wang, and Z. Cui, "Parking occupancy prediction method based on multi-factors and stacked GRU-LSTM," *IEEE Access*, vol. 10, pp. 47361-47370, Apr. 2022. Article (CrossRef Link)
[38] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, "T-GCN: A temporal graph convolutional network for traffic prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 9, pp. 3848-3858, Sep. 2020. Article (CrossRef Link)
[39] J. Zhu, Q. Wang, C. Tao, H. Deng, L. Zhao, and H. Li, "AST-GCN: Attribute-Augmented spatiotemporal graph convolutional network for traffic forecasting," *IEEE Access*, vol. 9, no. 1, pp. 35973-35983, Nov. 2020. Article (CrossRef Link)
[40] J. Fang, L. Chen, X. Wang, and B. George, "Not all posts are treated equal: An empirical investigation of post replying behavior in an online travel community," *Information & Management*, vol. 55, no. 7, pp. 890-900. Nov. 2018. Article (CrossRef Link)
[41] X. Lv, R. Zhang, and Q. Li, "Value co-destruction: The influence of failed interactions on members' behaviors in online travel communities," *Computers in Human Behavior*, vol. 122, no. 9, 106829, Sep. 2021. Article (CrossRef Link)
[42] Y. Sun, Y. Shao, and E. Chan, "Co-visitation network in tourism-driven peri-urban area based on social media analytics: A case study in Shenzhen, China," *Landscape and Urban Planning*, vol. 204, no. 12, 103934, Dec. 2020. Article (CrossRef Link)

Fangfang Gu is currently pursuing the Ph.D. degree in Management Science and Engineering at Nanjing University of Aeronautics and Astronautics, Nanjing, China. She received the M.E. degree from Nanjing Normal University in 2012. Her research interest includes tourism management, tourism economics and tourism big data.

Keshen Jiang is a professor of College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His research interest includes regional economics and tourism economics.

Yu Ding received the B.S. degree from Nanjing University of Aeronautics and Astronautics in 2006, M.E degree from University of Science and Technology of China in 2013, and Ph.D. degree in computer science from Army Engineering University of PLA in 2021. From 2022, he has working in the College of Artificial Intelligence, Nanjing Agricultural University. His main research interests include artificial intelligence and data mining.

Xuexiu Fan received the B.S. degree and M.E. degree from Guilin University of Technology in 2007 and 2012, she is a Ph.D. degree student in College of Economics and Management at Nanjing University of Aeronautics and Astronautics. She is a senior lecturer in school of foreign language and international business at Guilin University of Aerospace Technology. Her research interests include industrial economics and tourism economics.