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.

Computers and Electrical Engineering 93 (2021) 107235

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

A graph spatial-temporal model for predicting population density of key areas

Zhihao Xu [a], Jianbo Li [a,b,*], Zhiqiang Lv [a], Yue Wang [b], Liping Fu [b], Xinghao Wang [b]

[a] School of Computer Science and Technology, Qingdao University, Qingdao 266000, Shandong, China
[b] Institute of Ubiquitous Networks and Urban Computing, Qingdao University, Qingdao 266001, Shandong, China

ARTICLE INFO

Keywords:
Population density
Key areas
WE-STGCN
Feature component

ABSTRACT

Predicting the population density of key areas of the city is crucial. It helps reduce the spread risk of Covid-19 and predict individuals' travel needs. Although current researches focus on using the method of clustering to predict the population density, there is almost no discussion about using spatial-temporal models to predict the population density of key areas in a city without using actual regional images. We abstract 997 key areas and their regional connections into a graph structure and propose a model called Word Embedded Spatial-temporal Graph Convolutional Network (WE-STGCN). WE-STGCN is mainly composed of the Spatial Convolution Layer, the Temporal Convolution Layer, and the Feature Component. Based on the data set provided by the DataFountain platform, we evaluate the model and compare it with some typical models. Experimental results show that WE-STGCN has 53.97% improved to baselines on average and can commendably predicting the population density of key areas.

1. Introduction

At present, a part of big cities in China has been completed and some are in the second round of population gathering and expansion of the built-up area. This means that the population will accelerate into the big cities, making the built-up areas of big cities continue to expand. In the background of continuous population growth in Beijing, monitoring and forecasting the population density of certain key areas is an important task, which can prevent the spread of Covid-19 and other infectious diseases to some extent, avoid the occurrence of the stampede and alert unexpected abnormal gathering of people. Besides, the vehicle dispatch system of the taxi company can dispatch taxies [1] according to people's travel needs reflected by the population density of different areas.

1.1. Background

Population density plays a crucial role in researches related to cities [2,3], it is an important indicator of the level of cities' development without a reliable data set describing the daily activities of residents. Generally speaking, the data that we easily obtain is provincial-level census data, but the provincial-level census data is often of limited use in analyzing the rationality of urban internal road planning, evaluating the integrity of infrastructure construction, and implementing regional emergency plans. However, it is of great significance to study the population density of these areas after dividing the regions within the city. A survey from the Office of

This paper is for regular issues of CAEE. Reviews processed and approved for publication by the co-Editor-in-Chief Huimin Lu.

* Corresponding author.
E-mail address: lijianbo@qdu.edu.cn (J. Li).

https://doi.org/10.1016/j.compeleceng.2021.107235
Received 15 April 2021; Received in revised form 12 May 2021; Accepted 25 May 2021
Available online 3 June 2021

Management and Budget (OMB) shows that the area with the highest population density in a city is usually the area with the best traffic condition, the highest level of infrastructure construction, and the best economy, that is, the central business district (CBD).

As the process of urbanization continues to accelerate, many researchers in developing countries have begun to pay attention to the role of urban functional areas. Da Silva et al. [4] used road coverage and population density variables to define functional areas of a city in Brazil and found that the development of the urban transport system is closely related to people. Balakrishnan [5] proposed a method for urban population density prediction at 30 m resolution, using data such as the height of the building and street conditions of a city in India to analyze the internal conditions of the city. A detailed understanding of population density is of great significance for improving excessive population crowding and improving urban services reasonably. Many researchers used the population density function [6] to describe the internal regional structure of metropolises and non-metropolises to explore the scale of urban development. A survey towards Shenzhen shows that regional population density can intuitively reflect people's travel needs. Generally speaking, people's demand for taxies is positively correlated with regional population density.

1.2. Methods of studying population density

The rapid development of neural networks [7] has provided researchers with new ideas for studying regional population density. Shesteperov et al. [8] used artificial neural networks (ANN) to predict the density of the biological population in certain areas; the structure of ANN [9] can simulate the interactive response of the biological nervous system due to the stimulation of real-world objects, and it has shown its practicality in many disciplines such as biology and engineering [10], but the problems of ANN have always existed, such as excessive parameters, slow training speed, and difficulty in tuning; the long-short term memory (LSTM) [11] makes the recurrent neural network (RNN) not only remember the past information but also selectively forget some unimportant information through the gating mechanism for modeling long-term semantic information; the gated recurrent unit (GRU) [12] is based on the principle of LSTM to reduce the phenomenon of gradient disappearance while retaining long-term sequence information. Therefore, for time-series data, GRU and LSTM can effectively capture the semantic association among long sequences; the spatial features of the data can be extracted by the convolutional neural network (CNN) [13], and one convolution kernel is used to extract a feature of the data. When all the convolution kernels can effectively complete the task of feature extraction, the complex system (CNN) composed of many convolution layers can effectively extract and process multifarious spatial features; the connections among areas are complex and interlaced. We can build a graph structure with areas as vertices and connections among areas as edges. However, CNN cannot handle data with the non-Euclidean structure such as the graph structure. To solve this problem, Wu et al. [14] used the graph convolutional network (GCN) to extend traditional CNN to the graph structure and merge the node features with the graph topology.

1.3. Contributions

Based on the above researches, we clean the data set about the population density of key areas in Beijing provided by the DataFountain platform and screened out the data with obvious features. Then we transformed the spatial relationship of these areas into the graph structure, used the population density and association strength of these areas as the vertex matrix and the adjacency matrix respectively, and introduced a graph spatial-temporal model called WE-STGCN to reflect the temporal dependence and spatial correlation of the data. Experiments show that our work can well complete the task of predicting the population density of key areas with different urban functions [15]. The main contributions of our work are as follows:

●We predict the population density of certain key areas within the city from the perspective of data mining. Compared with current studies, we use the graph spatial-temporal network to model the regional population density data without considering actual images of key areas. This is the first work to use the graph spatial-temporal model to predict the population density of key areas within the city without the help of actual images of areas. It is of great significance to the construction of safe cities, the rational deployment of urban resources, the control of the spread of the epidemic, the intelligent dispatch of taxies, and the construction of intelligent cities. According to the phenomenon of the crowd gathering in different areas at different time points, city managers will make diverse decisions. For example, if city managers observe excessive population density of a certain commercial area during the epidemic prevention and control period, they will deploy police in advance to restrict the flow of people in certain commercial areas with relatively high population density. The taxi dispatch system can also make decisions based on actual road conditions and regional population density, and guide taxi drivers to certain nearest and densely crowded areas.

●We have introduced a graph spatial-temporal model called WE-STGCN. It is mainly composed of three parts, which are the temporal convolutional network (TCN) for expressing time-series information, GCN for extracting information about nodes and edges in the graph structure, and the Feature Component including some vital attributes (such as weather conditions, what day of the week, migration conditions and regional natures) that affect the population density of the area. TCN and GCN can extract spatial-temporal features of data, and the Feature Component can enhance the importance of other attributes.

●We used the data set about the population density of key areas in Beijing to evaluate the performance of WE-STGCN and used six typical models such as GRU, LSTM, and GCN as baselines for conducting comparative experiments. Experiments show that WE-STGCN has advantages in the accuracy of regional population density prediction.

The rest of the paper is organized as follows. In Section 2, we introduce the basis of the graph structure and the graph spatial-temporal models. In Section 3, we introduce the process of data preprocessing and the application of WE-STGCN in predicting the population density of key areas. In Section 4, we carry out the experiment and compare our model with other typical models. In Section 5, we summarize the work. In Section 6, we discuss future work.

2. Related work

2.1. Contributing factors of regional population density

People's travel behavior determines the population density. Therefore, the regional population density is constantly changing under the influence of the periodicity of people's travel behavior [16]. Generally, the population density of the area is relatively high on weekdays and decreases rapidly on weekends. At the same time, the population density also shows obvious periodicity in the hourly level changes. The distribution of the population density at different time points of the day is different. In the morning, especially around nine o'clock, the population density of some areas is relatively high, while the population density of certain areas will decrease at night. In different periods of each day, the growth trend of regional population density varies greatly. During the day, especially at noon, the growth rate is fast, while the growth rate is extremely slow in the early morning. Besides, the change of population density is also restricted by the nature of the area. For example, the population density of scenic spots and parks is high while the population density of working areas is low on Sundays. On the contrary, the population density of working areas is higher than that in parks on weekdays. Some researchers [17] explored the influence of weather on the density of biological population, and then Jones et al. [18] found that there is a clear positive correlation between the population of a kind of pest and the maximum temperature, which is related to the average value of relative humidity. The results can help researchers design a reasonable pest control model. The Baidu Migration provides a good platform for us to understand the immigration and emigration of the urban population. Based on the data provided by the platform, we can understand the urban travel intensity of a certain city every day, as shown in Fig. 1. The horizontal axis represents the date, and the vertical axis represents the urban travel intensity of Beijing. The urban travel intensity here does not refer to the specific number of people traveling outside, but a value that can indicate the approximate proportion of the number of people traveling outside. Generally speaking, there is a positive correlation between the travel intensity of a city and the population density of key areas in the city.

2.2. Graph structure

The connections among areas are complex and diverse, and different areas have different features. The graph is a data structure [19, 20] that can represent certain areas and their connections. We can consider the areas themselves as nodes, and the spatial correlation among nodes as edges. Topological relations can be converted into matrices for the calculation of the graph network. Fig. 2 shows the graph structure, the adjacency matrix (W matrix or edge matrix), and the vertex matrix (V matrix) constructed according to areas and their connections. Each row of the V matrix represents the population density of ten areas (0–9) at a certain moment; the W matrix represents the association strength among areas. For two areas without the direct relationship, we consider the association strength to be 0. For convenient computation, we think that the association strength is also 0 between a certain area and oneself.

2.3. Graph spatial-temporal network

Traditional models in machine learning can achieve high prediction accuracy by modeling complex data, but the huge amount of population density data forces us to choose deep learning methods for data processing and data analysis. The complex interaction relationship among areas can be represented by the graph structure, but CNN cannot handle the non-Euclidean structure of graph structure. With the development of the neural network, the graph convolutional network (GCN) proposed by Wu et al. [14] extends CNN to data with graph structure and integrates node features, edge features, and graph topology. Many kinds of data do not have a regular spatial structure, such as knowledge graphs and social networks. For these irregular data objects, the effect of ordinary convolutional networks is poor. Therefore, for the data with the graph structure, we have to consider not only the node information but also the structure information. GCN can not only automatically learn the features of nodes, but also learn the association information among nodes, and it is efficient to process data with graph structure.

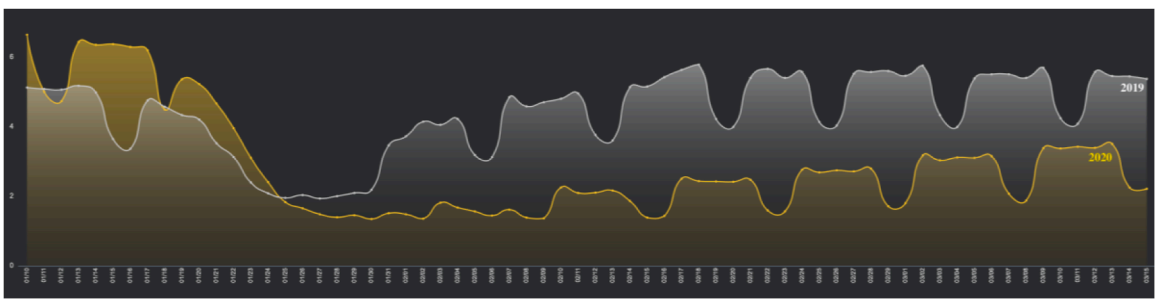

Fig. 1. Changes in urban travel intensity in Beijing during the Spring Festival in 2019 and 2020.

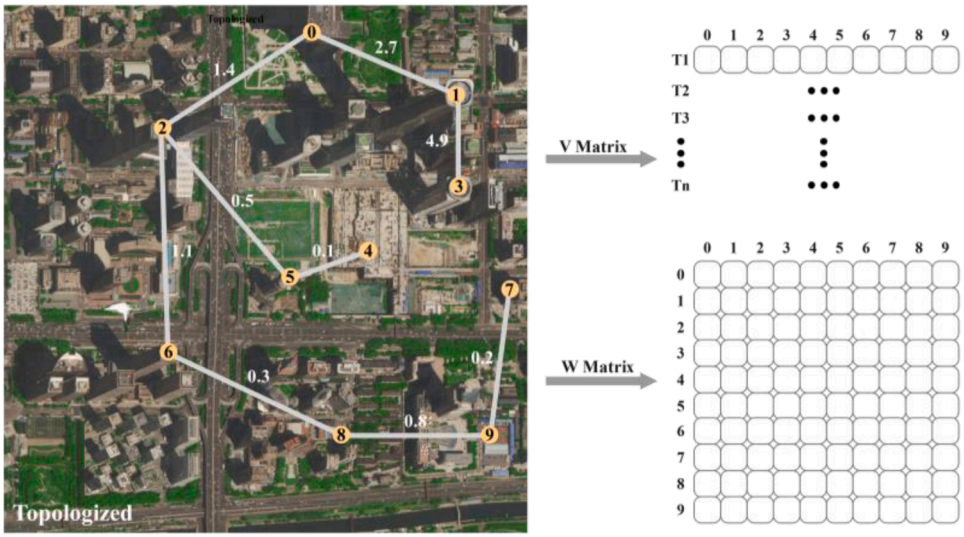

Fig. 2. Examples of graph structure (topologized structure), V matrix, and W matrix.

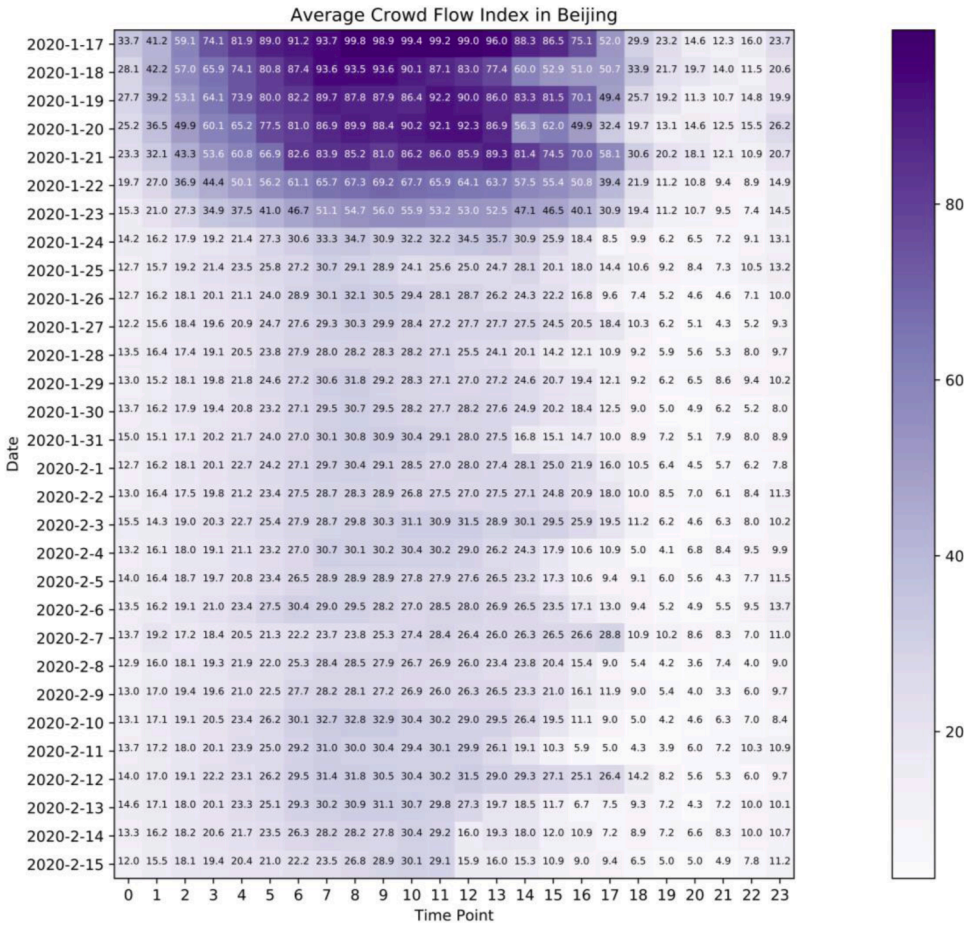

Fig. 3. Average crowd flow index in Beijing from 0:00 on January 17, 2020 to 23:00 on February 15, 2020.

3. Materials and methods

3.1. Data preparation, data cleaning, and data processing

In the background of the raging Covid-19 epidemic and the low level of safety management in cities, to further grasp the movement and gathering of people and deal with emergencies, we take public health and urban safety as research purposes and explore a new method for predicting the population density of key areas in a city. We choose to use the data set of population density of key areas in Beijing provided by the DataFountain platform, which includes the daily in and out of Beijing, the association strength among areas, and the population density of key areas. The time range of the data set is from 0:00 on January 17, 2020 to 23:00 on February 15, 2020.

The time is an integer from 0 to 23. Considering the privacy of data, the data set uses the migration index to represent the actual migration rate in Beijing. The migration index is positively correlated with the number of people moving in or out. For the sake of convenience, we designate all migration indexes out of Beijing as negative numbers. The data set uses the human flow index to indicate the amount of regional human flow, and the human flow index is positively correlated with the regional human flow; the association strength is determined by the strength of the interaction between two areas. Generally, the greater the association strength, the stronger the interaction between two areas. The data set provided by the DataFountain platform is based on the population density data of 997 key areas provided by 11 units including the Beijing Municipal Development, the Beijing Reform Commission, and the Beijing Municipal Bureau of Economics and Information Technology. The data set is obtained after fitting, cross-validation and weighting. It can scientifically and objectively reflect the interaction among different types of areas at a certain time point and the population density of each area and then portray people's travel rules and travel needs from the side, and at the same time provide early warning of the abnormal gathering behavior.

Fig. 3 shows changes in the average crowd flow index in Beijing from 0:00 on January 17, 2020 to 23:00 on February 15, 2020 (Considering the privacy of data, the true values are replaced by indexes.). Combining the analysis of Fig. 1, it is not difficult to find that since the first confirmed case of the Covid-19 epidemic in Beijing on January 20, 2020, Beijing's urban travel intensity and human flow index have both declined significantly. Besides, due to the influence of the Spring Festival, the flow of people and the urban travel intensity before the Spring Festival (January 25, 2020) were extremely high at certain moments. To avoid the influence of festivals on the population density as much as possible, we choose to train the population density index (PDI, the specific calculation method is shown in Formula (1)) of 997 areas from 0:00 on January 26, 2020 to 23:00 on February 6, 2020, and to evaluate the population density index of 997 areas from 0:00 on February 7, 2020 to 23:00 on February 15, 2020. Fig. 4 shows the interaction among all grids in area B and A1 in area A. A1Bj is the association strength between A1 and Bj.

$$PDI = \frac{CFI}{S \times 10^{-6}} \tag{1}$$

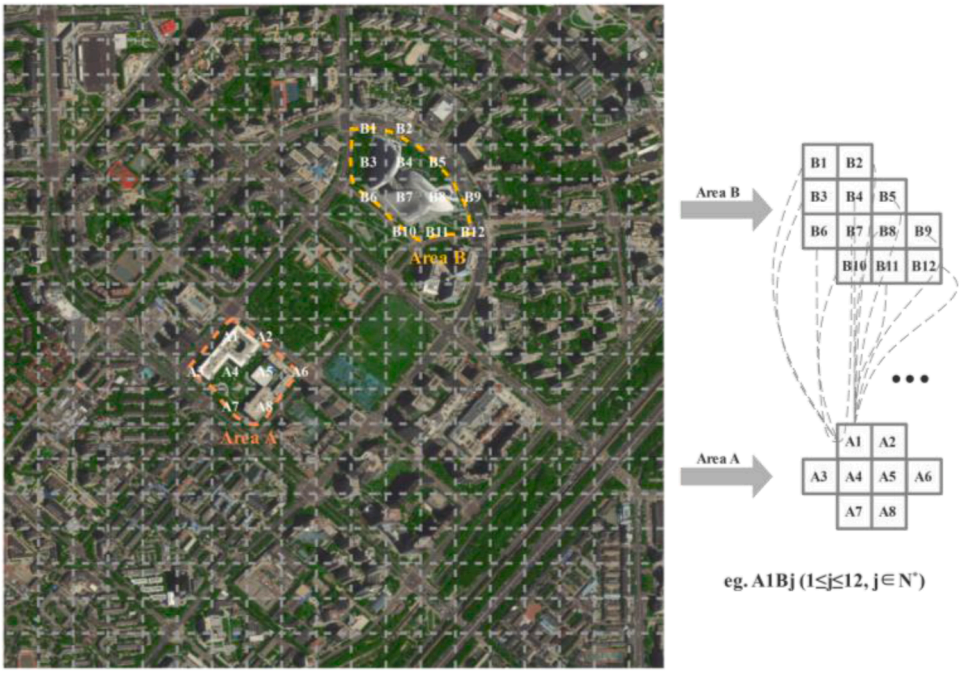

Fig. 4. The association strength among 8 grids in area A and 12 grids in area B.

$$AS_{AB} = \frac{\sum A_i B_j}{i_{\max} \cdot j_{\max}} (i \in [1,8], j \in [1,12], i \in N^*, j \in N^*) \quad (2)$$

Formula (1) is the expression of the population density index in a certain area. The *PDI* represents the population density index; the *CFI* represents the index of human flow; the *S* represents acreage and the unit of measurement is square meter.

After finishing the data cleaning, we process the data set. Firstly, we divide 997 areas into 5 categories, which are the commercial land, the land for transportation facilities, the science and education park, the public service land, and the medical land. Secondly, the acreage range of 997 key areas is [10,000.47, 21,658,303.75], we convert it to [0.01, 21.66], and the unit of measurement is square kilometer. Then, for the human flow index, we combine the acreage of the area to convert it to the regional population density index and calculate that the range of the regional population density index is [0, 214.08]; for the migration index, we set the migration index of going to Beijing as a positive value and count its range as [0.00003, 0.62208], then we set the migration index of moving out of Beijing as a negative value and count its range as [−0.29134, −0.00003]. Therefore, the range of the migration index is [−0.29, 0.62]. Finally, because the data set divides the entire Beijing into many 200(meters)*200(meters) grids, and gives the data of the association strength among the grids, we need to convert the association strength data among the grids to the association strength data among areas. The specific method is shown in Fig. 4 and Formula (2).

Formula (2) is an expression for the association strength between area A and area B. The AS_{AB} represents the association strength; the *A* represents area A; the *B* represents area B; the A_iB_j represents the association strength between the grid A_i and the grid B_j; the i_{max} represents the maximum value of i; the j_{max} represents the maximum value of j.

3.2. Baselines

We set up three categories of baselines to verify the performance of the model, including traditional recurrent neural networks (such as GRU and LSTM), spatial-temporal networks, and graph spatial-temporal networks. All models are trained and evaluated on the same data set, and the experimental results are the average of these training and evaluation results. The six baselines used in the experiment are as follows:

- ●LSTM: It controls the transmission state of data through a gating mechanism. Compared with ordinary RNN which only mechanically superimposes one kind of memory, the principle of LSTM is to retain the long-term memory and forget the unimportant information.
- ●GRU: Its input and output structures are similar to ordinary RNN, and its processing logic unit is similar to that in the structure of LSTM. Compared with LSTM, GRU has two gates and fewer parameters, but it can achieve similar functions and accuracy to LSTM. Considering the computing power and time cost of hardware, GRU is the first choice of many researchers.
- ●Spatial-temporal Dynamic Network (STDN) [21]: It uses local CNN and LSTM to process spatial-temporal information, and uses the attention mechanism to model long-term information.
- ●GCN: The essential purpose of GCN is to extract the spatial features of topological graphs. The core idea is to use edge information to aggregate node information to generate new node information.
- ●Spatial-temporal Graph Convolutional Network (STGCN) [22]: STGCN is a general framework for processing structured time series. The spatial-temporal convolution block combines graph convolution and temporal convolution, which can extract the most useful spatial features and capture the most basic temporal features coherently. The STGCN is completely composed of the convolution structure, which is parallelized at the input, with fewer parameters and fast training speed. More importantly, this structure allows the model to process large-scale data with higher efficiency.
- ●Attention Based Spatial-temporal Graph Convolutional Network (ASTGCN) [23]: ASTGCN uses the attention mechanism to extract spatial-temporal features. Then it uses graph convolution to extract spatial features and uses standard convolution to extract temporal features. Finally, it stacks the attention layer and convolution layers to complete the construction of the spatial-temporal block

3.3. Problem definition

The topological relationship among areas can be described as a graph structure: *G*=(V, W). The *V* represents the population density index of each area node per time step, and the *W* represents the adjacency matrix. The population density prediction problem can be defined as the prediction of the data in the next time from historical data ([X_{t-n}, …, X_{t-2}, X_{t-1}], the *n* represents the number of historical data, we choose $n = 7$ in the experiment). The predicted data X_t is shown in Formula (3).

$$X_t = M_f(X_{t-n}, \cdots, X_{t-2}, X_{t-1}) \quad (3)$$

The *M* represents the modeling method, the *t* represents the time point; the *f* represents external factors including weather conditions, migration index, natures of areas, and what day of the week. The data input to the model has four dimensions, namely [Batch Size, Node Number, Time Step, Channel]. The *Batch Size* represents the number of samples selected in a round of training, and we choose Batch Size=16 in the experiment; the *Node Number* represents the number of areas; the *Time Step* represents the length of the time step. In the experiment, we choose each step as 1 hour. The *Channel* represents the number of features of the data.

3.4. Model design

To perform scale compression and feature transformation on the regional population density index data, we input the data into the first Temporal Convolution Layer in a graph format instead of conduct the operations of linearization and concatenation. The association strength among areas is mapped to an adjacency matrix. The adjacency matrix is combined with the output of the first Temporal Convolution Layer to calculate the spatial correlation of the population density index of each area. The output of the Feature Component and the output of the GCN are combined and input into the three-layer GRU structure to calculate the temporal dependence of fused data. Since WE-STGCN is almost entirely composed of convolutional networks, BatchNorm2D is used to prevent the disappearance and explosion of the gradient in convolutional networks. Finally, the Fully-connected Layer maps features of the data to the sample space. Fig. 5 shows the structure of WE-STGCN. This model is mainly composed of three parts [24], which are the Feature Component, the Temporal Convolution Layer, and the Spatial Convolution Layer.

Temporal Convolution Layer

For the time-series data with the graph structure, we choose to use Temporal Convolution Layer to extract temporal features. The Temporal Convolution Layer is designed based on TCN [25]. The design of TCN is very clever, unlike the convolutional long-short term memory (ConvLSTM). ConvLSTM introduces a convolution operation so that LSTM can process image information. Its convolution only operates on input images at a time, while TCN directly extracts features across time steps.

TCN retains all historical information and uses causal convolution to obtain long-term historical information. The formula of causal convolution is shown in Formula (4). $\{X_1, X_2, \ldots, X_t\}$ is the input sequence; $\{Y_1, Y_2, \ldots, Y_t\}$ is the output sequence; $\{f_1, f_2, \ldots, f_k\}$ represents the filter; the k is a constant. However, there is a problem with causal convolution, that is, it requires many layers or a large

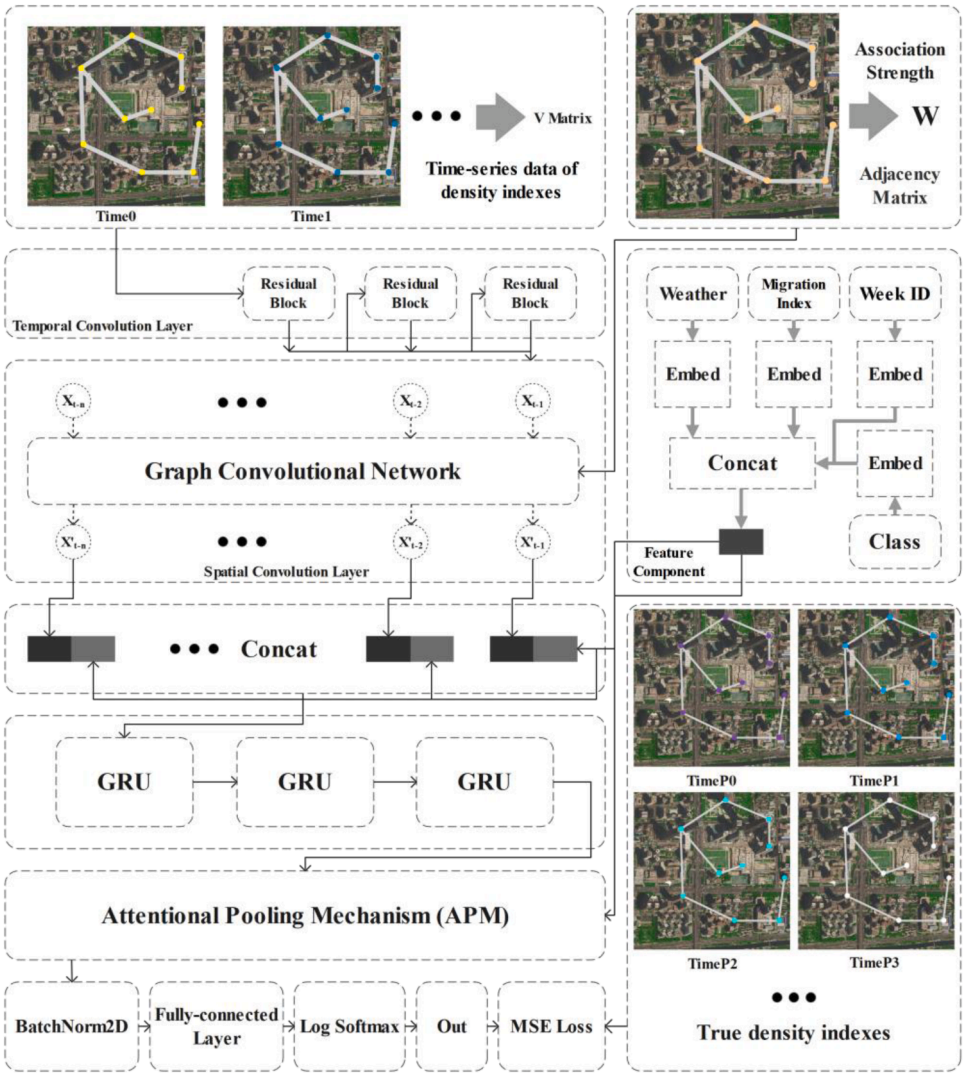

Fig. 5. The structure of WE-STGCN.

filter to expand the receptive field.

Therefore, dilated convolution is used to expand the receptive field. The expression of the dilated convolution is shown in Formula (5). X is the input; Y is the output; $\{f_0, f_1, \ldots, f_{k-1}\}$ represents the filter; the k is a constant. The d represents the dilated divisor, which changes according to the way as shown in Fig. 6 according to the depth of the network. Increasing d or KS can increase the range of the receptive field. Dilated convolution allows the filter to be applied to an area larger than the length of the filter itself by skipping part of inputs so that the problems caused by causal convolution can be solved. In Fig. 6, the convolutional receptive field has been enlarged 1, 2, and 4 times in turn. Dilated convolution can make the model have a very large receptive field when the number of layers is not large. Besides, the dilated divisor changes according to the convex function, so that the model will not lose local information.

$$Y_t = \sum_{i=1}^{k} f_i \cdot X_{t-k+i} \tag{4}$$

$$Y_t = \sum_{i=0}^{k-1} f_i \cdot X_{t-i \cdot d} \tag{5}$$

Fig. 6 shows the structure of the Temporal Convolution Layer. On the left is the process of dilated causal convolution. The *KS* is the size of the convolution kernel; the d is the dilated divisor which represents the size of the expansion; the *Dilations* represents the sequence of changes in the dilated divisor. On the right is the residual structure between every two layers. The *S(i-1)* represents the (i-1)-th state, and the *S(i)* represents the i-th state.

To reduce the complexity of the training process, the residual structure is used to replace the gated structure in the traditional recurrent neural network. The residual structure can overcome some shortcomings, which are the traditional recurrent neural network does not support parallel computing, the training speed is slow, and the gradient disappearance. The residual structure mainly includes a two-layer convolutional network and a nonlinear mapping process. Weight Norm can speed up the model convergence, and the Dropout layer can ignore a certain number of neurons to reduce over-fitting. After the convolution process, we choose Tanh as the activation function. The range of the regional population density index is [0, 214.08], the minimum value is −1.69 after the Z-Score standardization process, and the maximum value is 1.32. The expression of Z-Score is Formula (6), the X is the input data, the *Mean* is the average, and the *Std* is the standard deviation. The average value is nearly equal to 0 after the Z-Score standardization process, the number larger than the average number is positive after the Z-Score standardization process, and the number smaller than the average number is negative after the Z-Score standardization process. Since the output from Tanh is in the interval [−1, 1], the average value of the output is 0, which is close to the average value of the standardized regional population density index, the convergence speed is faster than sigmoid, and the training efficiency is high, we choose Tanh rather than other activation functions. Besides, since the number of channels between *S(i-1)* and the output (*R_Output*) of the two-layer convolutional network may be different, we design a convolution layer (KS=1) so that the transformed *S(i-1)* and *R_Output* can be directly added. The calculation method of this step is shown in Formula (7), the *Act* represents a certain activation function; the *S(i)* represents the result.

$$Z = \frac{X - Mean}{Std} \tag{6}$$

$$S(i) = Act(S(i-1) + R\_Output) \tag{7}$$

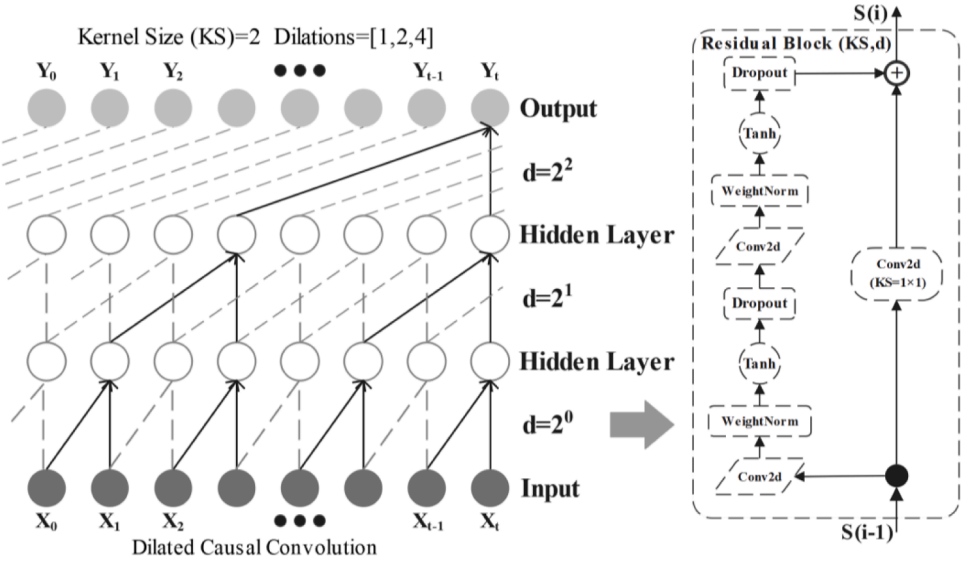

Fig. 6. The structure of the Temporal Convolution Layer.

Spatial Convolution Layer

We use GCN as the Spatial Convolution Layer, the main function of this layer is to extract the spatial correlation among various areas. The spatial correlation is established based on the association strength and is represented by the graph structure. GCN exquisitely designs a method to extract features from graph data, so that we can use these features to perform node classification, graph classification, boundary value prediction on graph data, and also get the embedding representation of the graph by the way. The core of GCN is to use the edge information to aggregate the node information to generate a new node representation. Each vertex in the graph will always be affected by its neighboring nodes or other nodes and constantly change its state until it reaches a balanced state. The adjacency matrix only defines which nodes are connected to the node, and there is no information about the node itself, which is reflected in the matrix that the diagonals of the adjacency matrix are all 0. We want to express the information of the node and the adjacency information of the node in a matrix at the same time, then we introduce the Laplacian matrix. The Laplacian matrix is the degree matrix (the degree matrix is the diagonal matrix formed by the degree of each node) minus the adjacency matrix, and the calculation method is shown in Formula (8). The L represents the Laplacian matrix; The D is the degree matrix; the A is the adjacency matrix.

However, to prevent performance degradation due to different scales during training, we need to normalize the eigenvalues of the Laplacian matrix, so that we will get the Symmetric Normalized Laplacian matrix used by us in the experiment, as shown in the formula (9). The B represents the Symmetric Normalized Laplacian matrix; the D is the degree matrix; the A is the adjacency matrix.

Besides, the Fourier Transform in the graph convolution process makes the calculation time complicated. To avoid the existence of the Fourier Transform, many researchers have introduced the Chebyshev Polynomial. We also choose to use the Chebyshev Polynomial to fit the convolution kernel to reduce complexity. The convolution kernel $g_\theta(B)$ can be approximated by a truncated shifted k-order Chebyshev Polynomial (as shown in Formula (10)). The β_k is the coefficient of the Chebyshev Polynomial; the K is a constant; the B represents the Symmetric Normalized Laplacian matrix; the $T_k(\widetilde{B})$ can be expressed by Formula (11). The λ_{max} is the maximum eigenvalue of the Laplacian matrix; the I is the identity matrix. The reason for this shift transformation is that the input of the Chebyshev Polynomial should be in the range of $[-1, 1]$. According to the properties of Chebyshev Polynomials, recursive formula (Formula (11)) can be obtained, which is the recursive definition of Chebyshev Polynomial, also known as K-localized convolution algorithm, which ensures that the current node only considers the influence of nodes within a specified range.

$$L = D - A \tag{8}$$

$$B = D^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}} \tag{9}$$

$$g_\theta(B) = \sum_{k=1}^{K} \beta_k T_k\left(\widetilde{B}\right) \tag{10}$$

$$T_k\left(\widetilde{B}\right) = 2\widetilde{B} \cdot T_{k-1}\left(\widetilde{B}\right) - T_{k-2}\left(\widetilde{B}\right), T_0\left(\widetilde{B}\right) = 1, T_1\left(\widetilde{B}\right) = \widetilde{B}, \widetilde{B} = \frac{2}{\lambda_{\max}} B - I \tag{11}$$

Feature Component

We have introduced the Feature Component that integrates various attributes that affect the population density of the area, including weather conditions (Weather, such as "sunny"), the nature of the area (Class, such as Class 0, we transfer it into "Zero"), what day of the week (Week ID, such as "Monday") and the migration index (Migration Index, such as 0.5, we transfer it into "zero point five"). Each attribute in the Feature Component is processed by word embedding and concatenated. The result is input into some places of the model to enhance the importance of these attributes. The population density of an area is affected by many complex attributes, such as weather conditions, regional nature, and urban travel intensity. Therefore, considering these attributes is very important to predict the population density.

The first step of word embedding is to encode certain words or sentences through indexes, that is, to assign an index to each different word or sentence. Next, the embedding matrix will be created. We have to decide how many latent attributes to assign to each index. Compared with one-hot encoding, the word embedding method has two main advantages. Firstly, since the vocabulary size of the classification value may be very large, the word embedding method effectively reduces the input dimension, so the calculation efficiency is high. Secondly, some studies found that classification values with similar semantic meanings are usually embedded in similar locations, so the embedding method helps to find and share similar patterns among different regions. As far as the experiment itself is concerned, we can give different weights to each attribute in the Feature Component. For example, we can give larger dimensions to the factors that affect the results to a large extent.

GRU

GRU is a variant of traditional RNN. Like LSTM, it can effectively capture the semantic association among long sequences and alleviate the phenomenon of gradient disappearance or gradient explosion caused by many layers of convolutional networks (mainly TCN and GCN). At the same time, its structure and calculation process are simpler than LSTM, and its calculation efficiency is higher than that of LSTM. Besides, the number of parameters of GRU is not very large, which reduces the risk of over-fitting. Compared with LSTM, GRU has a faster convergence speed. Based on the above theory, we choose to use a three-layer GRU structure (Experiments show that the effect of the three-layer GRU is the best.) to store and filter the information obtained after the Temporal Convolution layer, Spatial Convolution Layer, and the Feature Component. GRU does not clear the previous information as time progresses, on the contrary, it retains the relevant information and passes it to the next unit.

Attentional pooling mechanism

Fig. 7 is the structure of the attentional pooling mechanism (APM). APM consists of two inputs and one output. The A is the output of the Feature Component; the B is the output of the three-layer GRU structure. The A is processed by the linear layer and Tanh activation function to change the size and shape. As a result, the AP undergoes a dimensional upgrade and conducts batch matrix multiplication (bmm) operation with B. All the data will be amplified after going through the bmm layer, and then we can make the value distribute in the range of [0, 1], and the final output is Con.

$$Bmm^{1,2}_{[a,b,d]} = B^{1}_{[a,b,c]} \circ B^{2}_{[a,c,d]} \tag{12}$$

Formula (12) is the formula of bmm operation. The B^1, B^2, and $B^{1,2}$ are three matrices; the *Bmm* represents bmm matrix multiplication; [*x, y, z*] represents the shape of the matrix. The parameters of the bmm operation are two three-dimensional matrices. The first dimension of the two matrices must be the same, and the latter two dimensions must satisfy the dimensional transformation condition of matrix multiplication in terms of dimension. APM has two obvious advantages. For one thing, the attentional pooling mechanism compresses the number of features of the output generated by the upper layers, reduces the parameters, and simplifies the calculation of the model in the post-processing stage. For another, the attentional pooling mechanism enables the output via the upper layers network to continuously pay attention [26] to the influence of certain necessary features. The attentional pooling mechanism assigns different weight parameters to each element of the input to highlight useful information.

Post-processing

Unlike traditional machine learning [27] algorithms, the post-processing process in deep learning models is extremely important. Before the nonlinear transformation is performed, since the upper layer network has a deeper depth and has undergone multiple nonlinear transformations, the distribution of values output by the upper layer network will shift or change. The reason why the convergence speed of the training process is slow is that the overall distribution gradually approaches the upper and lower limits of the value range of the nonlinear function. The process of batch normalization (BatchNorm2D) forces the value distribution back to the standard normal distribution with a mean value of 0 and a variance of 1 through a certain normalization method. In this way, a small change in the value will cause a large change in the loss function, avoiding the problem of vanishing gradient. Besides, the larger the gradient means the faster convergence speed, which can greatly speed up the training speed.

The expression of BN is shown in Formula (13), the *Mean* and *S* are the mean and variance of the input data respectively; the *ele* is the stability divisor added to increase the stability of the calculation, and its value is equal to 10–5; the *Q* and *T* are learnable coefficient matrices; the *X* is the input data; the *Y* is the output. Finally, the Fully-connected Layer and log softmax activation function map the distributed features calculated by the network to the sample target space. The difference between the predicted value of the model and the true value is measured by the mean square error loss (MSE Loss). The formula of MSE Loss is shown in Formula (14). The *Y* is the predicted value; the *Ytr* is the true value; the *n* represents the number of samples.

$$Y = \frac{X - Mean}{\sqrt{S} + ele} \times Q + T \tag{13}$$

$$MSELoss(Y, Ytr) = \frac{1}{n}\sum (Y - Ytr)^2 \tag{14}$$

4. Results

4.1. Evaluation indexes

We use three commonly used evaluation indicators to evaluate the performance of the model. They are root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The respective expressions are Formula (15),

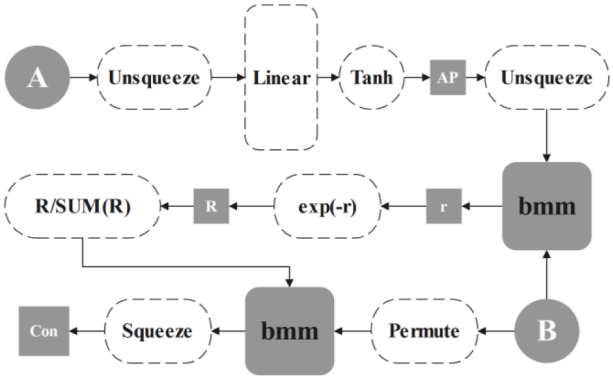

Fig. 7. The structure of the attentional pooling mechanism (APM).

Formula (16), and Formula (17) (The *Ypr* represents prediction value; the *Ytr* represents true value; the *n* represents the number of samples.).

We use Nvidia Tesla V100 GPU to train different models for 2000 rounds, and the evaluation results are shown in Table 1. GRU and LSTM are traditional recurrent neural networks, they have a high level of processing time-series data, but they cannot be used to process data with the graph structure, so they perform the worst. STDN adds a convolutional network based on the traditional recurrent neural network to extract spatial features, and the accuracy is slightly improved compared to GRU and LSTM. However, traditional convolutional networks cannot handle data with the non-Euclidean structure, so STDN does not affect this experiment. GCN can capture the spatial relationship among data with graph structure and extract spatial features, but it cannot obtain the temporal dependence among these data, so its performance is average. Both STGCN and ASTGCN belong to graph spatial-temporal models, and their essence is to use GCN to extract spatial features of data with graph structure and to use recursive neural networks to extract temporal features. Due to the consideration of both temporal and spatial features, STGCN and ASTGCN perform relatively well. Compared with the above-mentioned baselines, WE-STGCN performs best. This is because WE-STGCN not only considers the spatial correlation and temporal features of the population density of each area, but also the Feature Component is introduced to enhance the influence of some important external attributes on the regional population density. Besides, the evaluation results of WE-STGCN without the Feature Component (WE-STGCN(N)) are worse than the evaluation results of graph spatial-temporal networks (STGCN and ASTGCN).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Ypr_i - Ytr_i)^2} \tag{15}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Ypr_i - Ytr_i| \tag{16}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Ypr_i - Ytr_i}{Ytr_i}\right| \tag{17}$$

4.2. Under-fitting phenomenon

Firstly, we want to explain the under-fitting problem of marginal values, which refers to the difference between true values and the model's prediction of small values and large values in the regional population density index. This problem is reflected in Fig. 8. When the regional population density index is in the range of 20 to 160, the predicted value is consistent with the true value; while in the range of 15 to 20 and in the range of 160 to 172, the predicted results are extremely poor. This problem occurs because the regional population density index calculated according to the data set has more intermediate values, so the features of this part of the value are more obvious, and the data volume of the regional population density index with large and small values is small, which leads to unobvious features. Fig. 9 is the prediction results of GCN, STGCN, and ASTGCN. Because GCN can model the data with the graph structure, the predicted value has a certain fitting ability in the entire numerical range, so it effectively relieves the under-fitting problem of marginal values. It is worth mentioning that some models that are dependent on historical data will also have a certain degree of the under-fitting problem of marginal values because the weight of marginal values in historical data is too small.

Secondly, we want to talk about the under-fitting problem of local peaks, which refers to the under-fitting phenomenon of the predicted value near the extreme value when the real data has a local extreme value, as shown in Fig. 10. GCN can extract the spatial correlation of data with the graph structure, but GCN lacks consideration of the temporal relationship in data with the graph structure. The traditional recurrent neural network has significant advantages in extracting temporal features. The graph spatial-temporal network represented by STGCN and ASTGCN shows the advantages of predicting margin values and local peaks. They use GCN to solve the problem of under-fitting of margin values and use recurrent neural networks to solve the under-fitting phenomenon of local peaks.

The prediction effects of WE-STGCN and WE-STGCN(N) are shown in Fig. 11. Fig. 11 shows the superiority of WE-STGCN in capturing temporal features in a continuous period (17:00 on February 8, 2020 to 20:00 on February 12, 2020). The accuracy of WE-STGCN is better than traditional recurrent neural networks because the Spatial Convolution Layer can establish the spatial correlation

Table 1
Different evaluation indexes.

Model	RMSE	MAE	MAPE
LSTM	6.51	4.02	4.35%
GRU	6.55	3.82	4.32%
STDN	6.50	3.71	4.29%
GCN	6.44	3.66	4.01%
WE-STGCN(N)	3.35	3.01	2.12%
STGCN	3.01	2.49	2.86%
ASTGCN	2.99	2.50	1.86%
WE-STGCN	**1.32**	**1.95**	**1.59%**

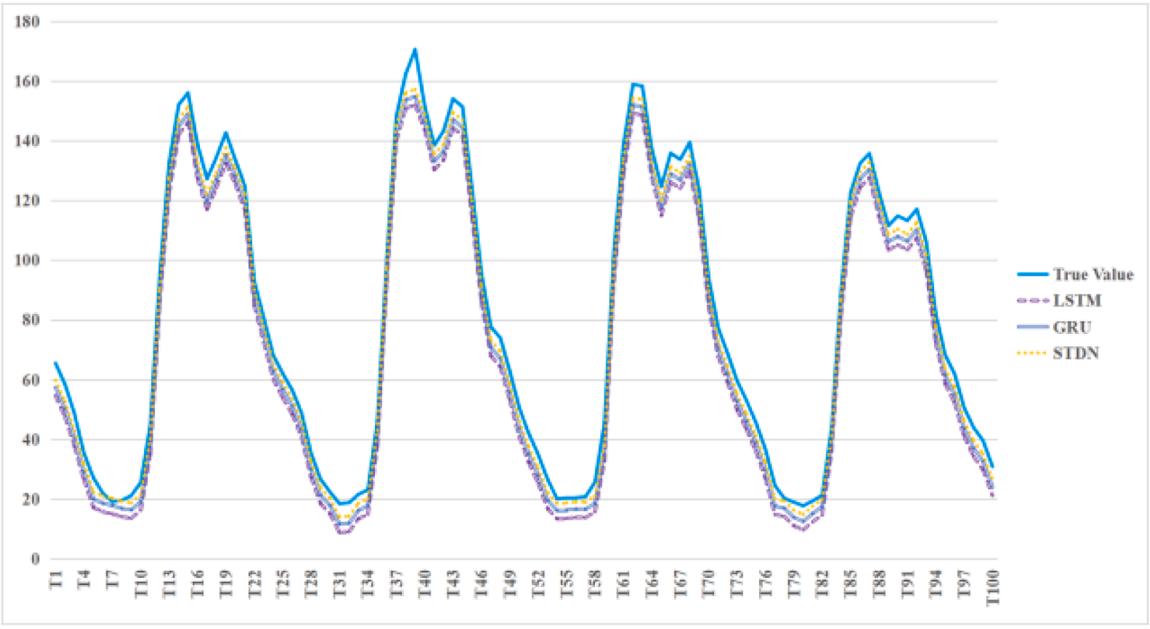

Fig. 8. GRU, LSTM, and STDN forecast the population density index of area A in 100 consecutive time points.

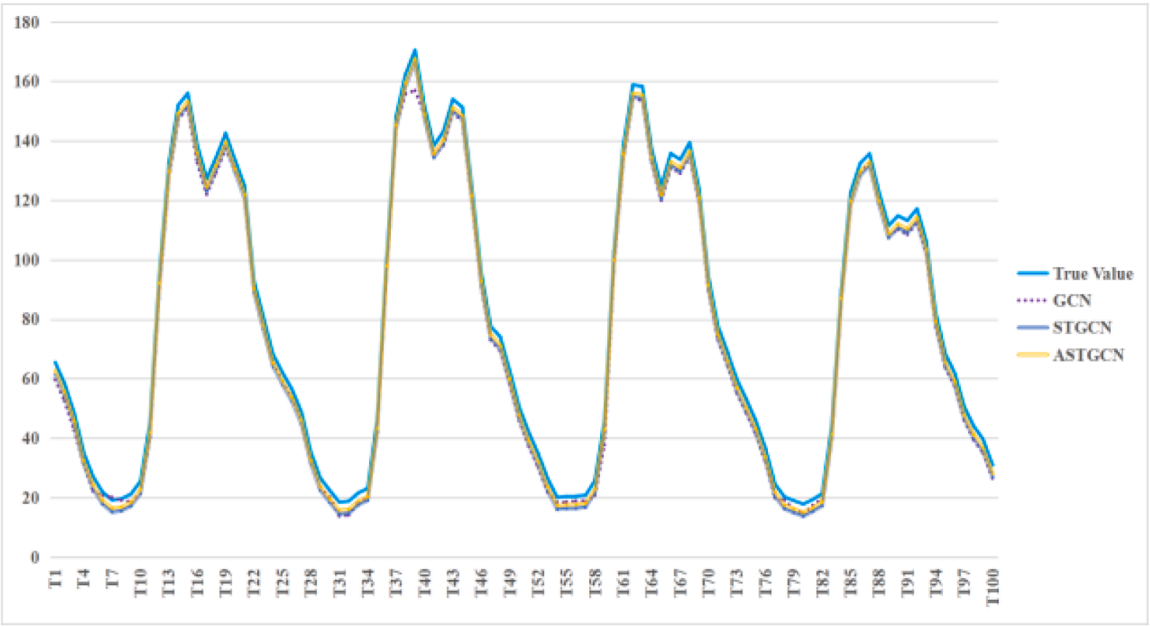

Fig. 9. GCN, STGCN, and ASTGCN forecast the population density index of area A in 100 consecutive time points.

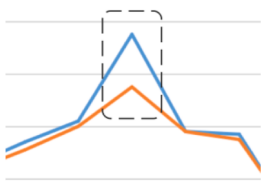

Fig. 10. Under-fitting of the local peak.

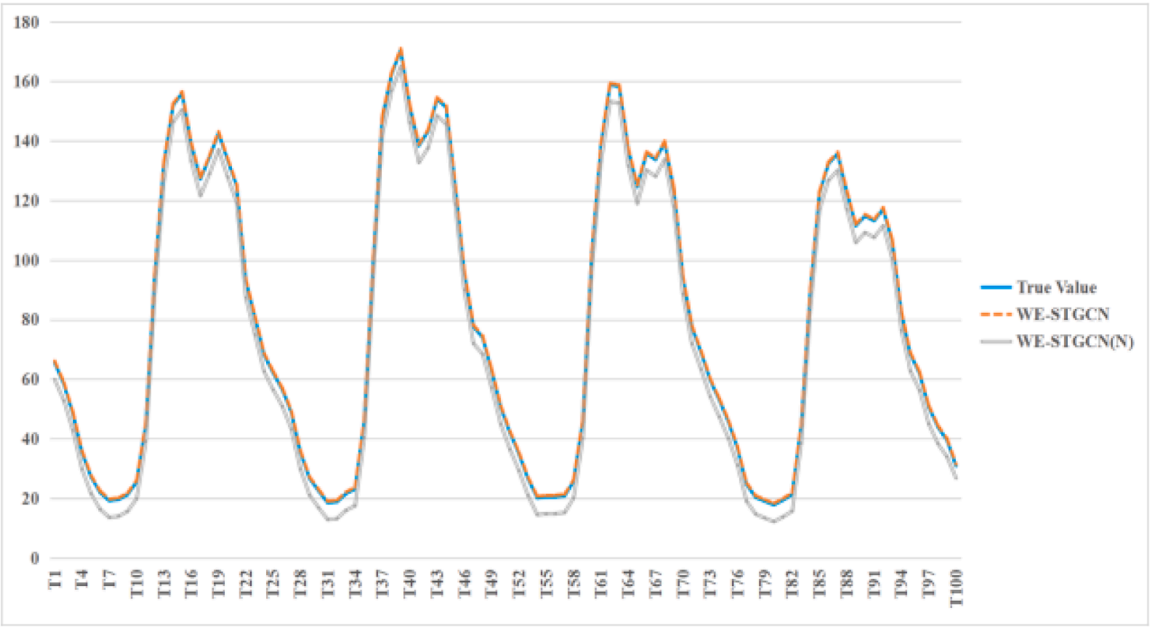

Fig. 11. WE-STGCN and WE-STGCN(N) forecast the population density index of area A in 100 consecutive time points.

among areas; the accuracy of WE-STGCN is better than GCN because the Temporal Convolution Layer has established a wide range of receptive fields, and relies on the multi-layer residual structure to overcome the excessive temporal dependence of the gating mechanism on historical data; the accuracy of WE-STGCN is better than that of the graph spatial-temporal network because the Feature Component integrate some attributes with the regional population density index, making the external attributes that affect the regional population density more important.

5. Conclusions

We proposed a new method for predicting regional population density based on the graph spatial-temporal network (WE-STGCN) without the help of actual images of areas. Considering the complex relationship among areas, we abstracted the 997 key areas and the regional connections among them into a graph structure, taking the population density of the areas as the vertex matrix, and the association strength among the areas as the adjacency matrix. Besides, to highlight attributes that affect the density of regional populations such as weather, migration condition, and regional nature, we designed the Feature Component for highlighting these external attributes in the form of word embedding. Based on the topological features of the data with the graph structure, we used slightly modified GCN and TCN to extract the spatial correlation and temporal dependence among areas respectively. WE-STGCN not only has high accuracy but also has stable performance. In general, the WE-STGCN model has the best performance compared with baselines. When baselines have the under-fitting problem, WE-STGCN can still predict steadily. Therefore, using WE-STGCN can complete the work of predicting the population density of key areas. Our work is of great significance in launching regional emergency plans, reducing the spread risk of infectious diseases such as Covid-19, and predicting people's travel needs.

6. Future work

There are still some issues that need further study. First of all, the regional population density data we selected is the data during the Covid-19 epidemic. We just used the model to predict the population density of each key area during the epidemic. However, we did not consider changes in the population density during the non-epidemic period. The data during the epidemic period should be compared with the data during the non-epidemic period to analyze the impact of the Covid-19 epidemic on the population density of areas with different urban functions. The impact can further reflect the impact of the Covid-19 epidemic on various economic forms. Then we should consider using the model for predicting population density of the large region, such as the prediction of population density of certain city-level areas in the Guangdong-Hong Kong-Macao Greater Bay Area of China.

Author statement

Zhihao Xu: Writing, Conceptualization, Software. Jianbo Li: Supervision. Zhiqiang Lv: Visualization, Resources. Yue Wang: Investigation. Liping Fu: Data Curation. Xinghao Wang: Software.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgements

This research is supported in part by National Key Research and Development Plan Key Special Projects [grant number 2018YFB2100303]; Shandong Province colleges and universities youth innovation technology plan innovation team project [grant number 2020KJN011]; Shandong Provincial Natural Science Foundation [grant number ZR2020MF060]; Program for Innovative Postdoctoral Talents in Shandong Province [grant number 40618030001]; National Natural Science Foundation of China [grant number 61802216]; and Postdoctoral Science Foundation of China [grant number 2018M642613].

References

[1] Lu H, Zhang Y, Li Y, Jiang C, Abbas H. User-Oriented Virtual Mobile Network Resource Management for Vehicle Communications. IEEE Trans Intell Transport Syst 2020;1(1):1–12.
[2] Boyko CT, Cooper R. Clarifying and re-concep tualising density. Prog Plann 2011;76(1):1–61.
[3] Kadi N, Khelfaoui M. Population density, a factor in the spread of COVID-19 in Algeria statistic study. Bull Natn Res Cent 2020;44(1):1–7.
[4] Silva Da, R. AN, Manzato GG, Pereira HTS. Defining functional urban regions in Bahia, Brazil, using roadway coverage and population density variables. J. Transp Geogr 2014;36(2):79–88.
[5] Balakrishnan K. A method for urban population density prediction at 30m resolution. Cartogr Geogr Inf Sci 2020;47(3):193–213.
[6] Bazyari A, Mousavi N. Estimation of Population Density Function in Line Transect Sampling Method with Detection Functions. J Stat Sci 2019;12(2):365–83.
[7] Li Y, Jiang Y, Tian D, Hu L, Lu H, Yuan Z. Ai-enabled emotion communication. IEEE Netw 2019;33(6):15–21.
[8] Shesteperov AA, Lukyanova EA, Bondarev AA. Application the Hassell model for prediction the population density of golden nematode of potato after growing Globodera resistant variety of potato. Rossiĭskiĭ Parazitologicheskiĭ Zhurnal 2019;1(1):90–6.
[9] Chaudhary A, Raheja JL. Bent fingers' angle calculation using supervised ANN to control electro-mechanical robotic hand. Comput Electr Eng 2013;39(2):560–70.
[10] Maier HR, Dandy GC. Application of artificial neural networks to forecasting of surface water quality variables: issues applications and challenges. Artif Neural Netw Hydrol 2020;36(2):287–309.
[11] Asif N, Huang Z, Wang S. Convolutional LSTM based transportation mode learning from raw GPS trajectories. IET Intell Transp Syst 2020;14(6):570–7.
[12] Zhang D, Kabuka M. Combining weather condition data to predict traffic flow: a GRU-based deep learning approach. IET Intell Transp Syst 2018;12(7):578–85.
[13] Lu H, Li B, Zhu J, Li Y, Li Y, Xu X, He L, Li X, Li J, Serikawa S. Wound intensity correction and segmentation with convolutional neural networks. Concurr Comp-Pract E 2017;29(6):3927–36.
[14] Wu F, Souza A, Zhang T, Fifty C, Yu T, Weinberger K. Simplifying graph convolutional networks. In: Proc. International conference on machine learning; 2019. p. 6861–71.
[15] Yao L, Wei WEI, Yu Y, Xiao J, Chen L. Rainfall-runoff risk characteristics of urban function zones in Beijing using the SCS-CN model. J Geogr Sci 2018;28(5):656–68.
[16] Zhang X, Cheng L, Li B. Too far to see? Not really! —Pedestrian detection with scale-aware localization policy. IEEE Trans Image Process 2018;27(1):3703–15.
[17] Rödel HG, Dekker JJA. Influence of weather factors on population dynamics of two lagomorph species based on hunting bag records. Eur J Wildlife Res 2012;58(6):923–32.
[18] Jones LM, Koehler AK, Trnka M, Balek J, Challinor AJ, Atkinson HJ, Urwin PE. Climate change is predicted to alter the current pest status of Globodera pallida and G. rostochiensis in the United Kingdom. Glob Chang Biol 2017;23(11):4497–507.
[19] Kaur R, Singh S. A comparative analysis of structural graph metrics to identify anomalies in online social networks. Comput Electr Eng 2017;57(2):294–310.
[20] Mathew AB. Data allocation optimization for query processing in graph databases using Lucene. Comput Electr Eng 2018;70(1):1019–33.
[21] Yao H, Tang X, Wei H, Zheng G, Li Z. Revisiting spatial-temporal similarity: a deep learning framework for traffic prediction. In: Proc. AAAI conference on artificial intelligence; 2019. p. 5668–75.
[22] Hammond DK, Vandergheynst P, Gribonval R. Wavelets on graphs via spectral graph theory. Appl Comput Harmon A. 2011;30(2):129–50.
[23] Guo S, Lin Y, Feng N, Song C, Wan H. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In: Proc. AAAI Conference on Artificial Intelligence; 2019. p. 922–9.
[24] Chen Z, Lu H, Tian S, Qiu J, Kamiya T, Serikawa S, Xu L. Construction of a Hierarchical Feature Enhancement Network and Its Application in Fault Recognition. IEEE Trans Ind Informat 2020;17(7):4827–36.
[25] Hewage P, Behera A, Trovati M. Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. Soft Comput 2020;24(21):16453–82.
[26] Lu H, Yang R, Deng Z, Zhang Y, Gao G, Lan R. Chinese image captioning via fuzzy attention-based DenseNet-BiLSTM. ACM T Multim Comput 2021;17(1s):1–18.
[27] Zheng Q, Zhu J, Tang H, Liu X, Li Z, Lu H. Generalized Label Enhancement with Sample Correlations. IEEE T Knowl Data En 2021;12(8):1–14.

Zhihao Xu was born in Binzhou city, Shandong province, China, in 1999. He received a bachelor's degree in digital media technology from Qingdao University, China, in 2020. From 2017 to 2019, he did data cleaning and paper translation work in the Qingdao Institute of Bioenergy and Bioprocess Technology, the Chinese Academy of Sciences. He is currently studying for a master's degree at Qingdao University, China, majoring in computer technology. His main research directions are deep learning, urban computing, intelligent transportation, crowd density research, and crowd interest prediction.

Jianbo Li was born in Weifang city, Shandong province, China, in 1980. He received a doctor's degree in computer science and technology department from the University of Science and Technology of China in 2009. From 2013 to 2014, he was a visiting scholar at Fordham University. He is currently a professor at the college of the computer science and technology at Qingdao University. He is the chairman of ACM Qingdao Branch, a senior member of China Computer Federation, and member of the Internet of Things Professional Committee of China Computer Federation. His research interests include urban computing, mobile social networks, and data offloading.

Zhiqiang Lv was born in Weifang city, Shandong province, China, in 1995. From 2018 to 2019, he studied in the software parallel group of the State Key Laboratory of Computer Architecture, Institute of Computer Technology Chinese Academy of Sciences. Now he works at the Institute of Ubiquitous Networks and Urban Computing, Qingdao. He has 1 invention patent and many university student scholarships. His main research directions are traffic demand research based on traffic trajectory and travel time, traffic data forecast research based on traffic flow, traffic congestion, high-performance parallel computing research based on deep learning and reinforcement learning.

Yue Wang was born in Zibo city, Shandong province, China in 1997. She received a bachelor's degree in computer science and technology from Harbin Normal University in June 2020. Now, she is studying for a master's degree at Qingdao University, China, majoring in computer technology. She has won many university student scholarships. Her research focuses on using deep learning to predict traffic flow among regions.

Liping Fu was born in Weifang city, Shandong province, China, in 1994. She graduated from Qingdao University in 2017 with a master's degree in computer science and technology and now works at the Institute of Ubiquitous Networks and Urban Computing, Qingdao. Her main research directions are urban computing, intelligent transportation, and traffic flow prediction.

Xinghao Wang was born in Weifang city, Shandong Province in 1994. He graduated from Qingdao University in 2017 with a master's degree in computer science and technology and now works at the Institute of Ubiquitous Networks and Urban Computing, Qingdao. His main research direction is urban computing.