Neural Computing and Applications (2023) 35:6457–6470
https://doi.org/10.1007/s00521-022-07876-0

ORIGINAL ARTICLE

Human migration-based graph convolutional network for PM2.5 forecasting in post-COVID-19 pandemic age

Choujun Zhan[1,2] · Wei Jiang[2] · Hu Min[2] · Ying Gao[3] · C. K. Tse[4]

Received: 17 March 2022 / Accepted: 21 September 2022 / Published online: 22 November 2022

Abstract

Due to the coronavirus disease 2019 pandemic, local authorities always implanted non-pharmaceutical interventions, such as maintaining social distance to reduce human migration. Besides, previous studies have proved that human migration highly influenced air pollution concentration in an area. Therefore, this study aims to explore whether human migration can work as a significant factor in the post-pandemic age to help PM2.5 concentration forecasting. In this work, we first analyze the variations of PM2.5 in 11 cities of Hubei from 2015 to 2020 and further compare PM2.5 trends with the migration trends of Hubei province in 2020. Experimental results indicate that the human migration indirectly affected the urban PM2.5 concentration. Then, we established a graph data structure based on the migration network describing the migration flow size between any two areas in the Hubei province and proposed a migration attentive graph convolutional network (MAGCN) for forecasting PM2.5. Combined with the migration data. The proposed model can attentively aggregate the information of neighbor nodes through migration weights. Experimental results indicate that the proposed MAGCN can forecast PM2.5 concentration accurately.

Keywords COVID-19 · Air pollution · Graph neural network · Deep learning

1 Introduction

The novel coronavirus disease 2019 (COVID-19) pandemic has become a tremendous thread globally since the end of 2019. Previous studies indicate that intercity migrations can accelerate the spread of the COVID-19 [1, 2]. Hence, local authorities always implement non-pharmaceutical interventions (NPIs) and even lockdown the whole city to contain the spread of the COVID-19 epidemic. Previous studies indicate that the concentration of air pollutants in megacities of many countries, including India, China, Japan, Italy, and the USA, reduced due to the restriction of human activities (or human migration) caused by the lockdown of the city [3–8]. From the end of 2019 to mid-2020, Wuhan suffered from the COVID-19 pandemic, while the local government took measures to lock the city down to restrict human migration. At the same time, the air quality improved in Wuhan and even in the entire Hubei province [9, 10]. Multiple factors influence the air quality in an area, such as sand storms, factory exhaust emissions, transportation exhaust emissions, agricultural incineration, and waste incineration [11, 12]. Previous studies reveal that

✉ Hu Min
minh@nfu.edu.cn

Choujun Zhan
zchoujun2@gmail.com

Wei Jiang
jwwweee0115@gmail.com

Ying Gao
gaoying@scut.edu.cn

C. K. Tse
cktse@ieee.org

[1] School of Computer, South China Normal University, Guangzhou, Guangdong, China

[2] School of Electrical and Computer Engineering, Nanfang College Guangzhou, Guangzhou, Guangdong, China

[3] School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong, China

[4] Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China

these factors can be utilized to enhance the performance of the air quality forecasting method [13, 14]. However, in the post-pandemic age, human migration is also one of the most significant factors for forecasting air pollution due to the frequent lockdown of cities. Many studies have achieved accurate predictions of air pollution based on human activities during the COVID-19 epidemic [15, 16].

PM2.5 is considered as one of the main air pollutants in environmental science [17]. With the development of machine learning (ML) and deep learning (DL), many studies utilize improved DL algorithms to predict the PM2.5 concentration or other air pollutants. Some early studies use deep neural networks (DNNs) to predict air pollution concentration [18–20]. Recurrent neural networks (RNNs) are one of the common methods adopted for predicting air pollutants, which can extract temporal features. For instance, RNNs are used to forecast PM10, PM2.5 and CO_2 concentration [21–23]. Convolution neural networks (CNNs) are also used to convolute the temporal features for predicting PM2.5 [24]. In addition, studies use wavelet decomposition (WD) and complete ensemble empirical mode decomposition (CEEMD) filters to preprocess air pollution data, and then combine these two methods with DL algorithms to predict air quality index (AQI) [25, 26]. However, none of the neural network models is effective in all air quality forecasting problems and can overwhelm all the other models. Instead, these proposed various methods all have their advantages and disadvantages.

Graph embedding has been widely used in processing network (or graph) tasks. Scarselli et al. first proposed the graph neural network (GNN) [27]. Bruna et al. first proposed the GCN according to the spectral method, using the Laplacian matrix to transform the graph information into the spectral domain through the Fourier Transform filter for convolution calculation [28]. Then Defferrard et al. replaced the Fourier Transform filter with Chebyshev polynomials and proposed ChebyNet [29]. Kipf et al. used the first-order Chebyshev polynomial approximation to convolute, which is the most common GCN, and its calculation at the node-level can be considered as the calculation in spatial domain [30]. Velivckovic et al. proposed a graph attention network (GAT) [31]. This algorithm normalizes the extraction of the weight features of the edges and further filters the neighbor nodes in the aggregation process.

Graph embedding models can combine features with a hidden layer to aggregate important information from different nodes, which is more effective in dealing with network data. Hence, it is worth exploring the potential use of GCN in air quality forecasting. Using GCNs to aggregate the information of air pollution concentration in both the temporal domain and spatial domain can obtain an accurate prediction. Qi et al. first use GCN to aggregate the air pollution and meteorological information from different observation stations in an area to predict the PM2.5 trends [32]. Using GCN to predict air pollution requires data-driven processing of air pollution data. Additionally, constructing reasonable networks from aggregated information, including air pollution, meteorological and other data, to reveal the key features is one of the most important steps for developing a GCN model predicting air pollution treads. For example, Zhou et al. used wind direction data in an area to build a wind-filed as the network for air pollution forecasting [33]. Wang et al. established a network to predict air pollution based on the regional and functional stations [34]. Wang et al. built a network based on the relationship of multiple factors such as weather, temperature, and wind direction to predict air pollution [35]. However, to our knowledge, few researchers consider the influence of human activity on constructing migration networks to develop GCNs models for air pollution concentration forecasting.

In this study, we adopt 11 cities in Hubei province as the study area. We first compare the 2020 annual PM2.5 concentration in Hubei with the average PM2.5 concentration from 2015 to 2019, and find that the PM2.5 concentrations in Hubei cities all reduced. In order to confirm that the reduction in PM2.5 concentration in 2020 is related to the lockdown of cities, we analyze the PM2.5 concentration trends in each city in 2020 and compare them with the migration pattern. Then, we construct a graph data structure based on the relationship between migration patterns and air pollution. Finally, we propose a migration attentive graph convolutional network (MAGCN) for predicting the PM2.5 concentration of each city. The model extracts and aggregates the migration information of the 11 cities to construct weighted migration networks. The prediction results show that the proposed model combining with migration data is better than the results of the baseline models, including ChebyNet, GGNN, GCN, and GAT. The main contributions of this paper are as follows:

- We deeply analyze the relationships between human migration and the concentration of PM2.5, and clearly show that the migration flow indirectly affects the air pollution variation during the COVID-19 pandemic.
- We find the characteristics of the migration network in Hubei province, and utilize the characteristics to combine the air pollution dataset and human migration dataset with time steps, reconstructing them into a new dynamic graph data structure based on the migration network, which is named migration air graph (MAG).
- We propose a migration attentive graph convolutional network (MAGCN) based on the migration attentive coefficient (MAC) with consideration of the human

migration data. The MAGCN achieves better performance by considering human migration data.

2 Data description

2.1 Air pollution data and human migration data

Air pollution data are provided by the Ministry of Ecology and Environment and downloaded from the website (https://www.aqistudy.cn). This dataset records daily air pollution concentration of 6 air pollutants, including PM2.5, PM10, CO_2, NO_2, SO_2, O_3, and daily climate data such as temperature, humidity, wind level. Both two kinds of data are recorded by the observation stations located in the 168 cities of China from January Jan 1, 2015, to Dec 31, 2020.

Human migration data are provided by AutoNavi Big Data (https://trp.autonavi.com/home.html). The migration dataset consists of the migration routes from one city to another city in a province, while the migration flow size from one city to another city is denoted by the AutoNavi Migration Index (AMI). According to the migration flow size from the AutoNavi data, a migration network can be established, which is shown in Fig. 1.

2.2 Adopted cities

In this study, we focus on the air quality and migration in cities of the Hubei province. The details of the air pollution dataset and AutoNavi migration dataset are shown in Table 1. In this study, the air pollution dataset of 11 cities, recorded from January 1st, 2015, to December 31st, 2020, is adopted for analysis. AutoNavi migration dataset contains the migration flow size of 16 cities. The period of this dataset is from December 1st, 2019, to November 30th, 2020. The sampling interval of these two datasets is 24 h. Note that the air pollution dataset covers 11 cities, while the migration dataset covers 16 cities. Hence, we just adopted 11 cities, both in the migration and air pollution datasets, for investigation. These 11 cities are Suizhou, Ezhou, Xianning, Jingzhou, Jingmen, Xiaogan, Wuhan, Yichang, Xiangyang, Huangshi, and Huanggang. Additionally, the analysis period of these two datasets is from December 1st, 2019, to November 30th, 2020.

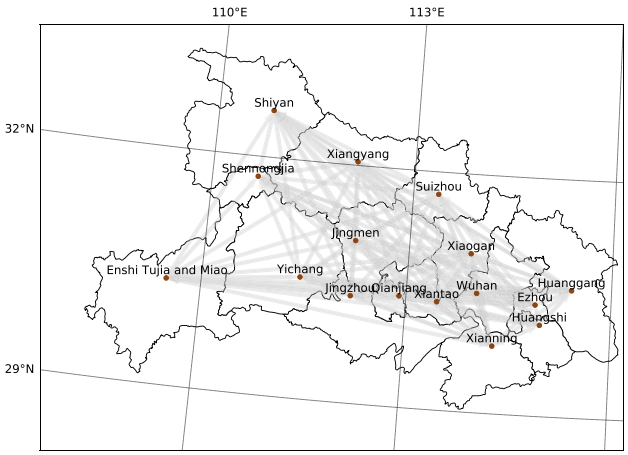


Fig. 1 AutoNavi migration network in Hubei province, red points stand for cities in Hubei province, and the gray lines with arrows are the directed routes of migrations. Thus, the migration network is a fully connected direct graph (color figure online)

3 Data analysis

3.1 Annual PM2.5 variation in Hubei cities

We first analyze the annual variations of PM2.5 concentration. The daily PM2.5 concentration $p(t)$ of each city is utilized to derive the annual average PM2.5 concentration $\bar{p}$ in a certain year according to the Eq. (1):

$$\bar{p} = \frac{1}{T}\sum_{t=1}^{T} p(t), \tag{1}$$

where T is the number of days in a year, namely, $T = 365$ (or $T = 366$ in 2016, 2020). The annual average PM2.5 concentration for a total of 6 years from 2015 to 2020 are represented by $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_6$, respectively.

Then, we utilize the annual average PM2.5 concentrations $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_6$ to derive the annual variation of PM2.5 in 11 cities of Hubei province. There are two types of variation:

- **Type 1**: Variation between annual PM2.5 concentration in 2020 and the concentration in 2019:

$$r_{2019} = \frac{\bar{p}_5 - \bar{p}_6}{\bar{p}_5} \times 100\%, \tag{2}$$

  where r_{2019} is the reduction rate of PM2.5 concentration between 2019 and 2020.
- **Type 2**: Variation between the annual concentration in 2020 and the average concentration of past five years (2015–2019):

$$\begin{aligned} r_f &= \frac{\frac{1}{M}\sum_i^M \bar{p}_i - \bar{p}_j}{\frac{1}{M}\sum_i^M \bar{p}_i} \times 100\% \\ &= 1 - \frac{M\bar{p}_j}{\sum_i^M \bar{p}_i} \times 100\%, \end{aligned} \tag{3}$$

  where r_f is the reduction rate ofthe PM2.5 concentration between 2020 and past 5 years (2015–2019), $M = 5$, $i = 2, 3, \ldots, 6$ and $j = 1, 2, \ldots, 5$.

Table 1 Dataset information

Dataset	Cities	Period	Sampling interval
Air pollution	Suizhou, Ezhou, Xianning, Jingzhou, Jingmen, Xiaogan, Wuhan, Yichang, Xiangyang, Huangshi, Huanggang	2015/01/01–2020/12/31	24 h
AutoNavi migration	Suizhou, Ezhou, Xianning, Jingzhou, Jingmen, Xiaogan, Wuhan, Yichang, Xiangyang, Huangshi, Huanggang, Xiantao, Qianjiang, Shiyan, Shennongjia, Enshi Tujia and Miao	2019/12/01–2020/11/30	24 h

Results indicate that the PM2.5 concentration of 11 cities in 2020 all decreased (shown in Fig. 2), especially for Jingzhou city, the reduction rate is high as 33.89%, compared with the past 5 years. Additionally, the PM2.5 concentration in Yichang, Jingzhou, and Jingmen cities, respectively, reduced by 21.94%, 20.39%, and 19.38% in 2020. Approximately 20% reduced the PM2.5 concentration in these cities compared with the concentrations in 2019. The annual variations of PM2.5 concentration in the 11 cities are presented in Table 2.

In order to further prove the annual PM2.5 reduction of each city, we plot the annual average PM2.5 concentration in each city for comparison based on original daily data. The comparison results are shown in Fig. 3, and the maximum, minimum, mean and median are shown in Table 3.

3.2 PM2.5 concentration trends from 2015 to 2020

The Moving Average (MA) of the daily PM2.5 concentration from 1st January to 31st December in the past 5 years and the MA of PM2.5 trend from 1st January to 31st December 2020 are derived, respectively. The PM2.5 concentration in a city at time t is represented by $x(t)$, then, the PM2.5 MA can be calculated using the Eq. (4):

$$\bar{x}(t) = \frac{1}{W}\sum_{k=0}^{W-1} x(t-k), \tag{4}$$

where W is the window size of MA. Here, we set the window size to $W = 7$, and k is the step of the window. The original daily PM2.5 concentration and the MA PM2.5 concentration in 2020 and the past 5 years are shown in Fig. 4.

The PM2.5 concentration in 2020 rose before the lockdown of Wuhan from 1st January to 25th January. However, after 25th January, the Wuhan lockdown restricted large-scale human migration. Therefore, the trend of PM2.5 concentration in Hubei cities in 2020 reduced sharply. This phenomenon continued until the end of the lockdown in Wuhan. Obviously, the PM2.5 concentrations from 25th January to 6th April in 2020 were smaller than the same period in the past five years because of the COVID-19 lockdown and the restriction of human migration. Due to the COVID-19 epidemic, the PM2.5 concentration in each city remained lower than the average concentration in the past five years.

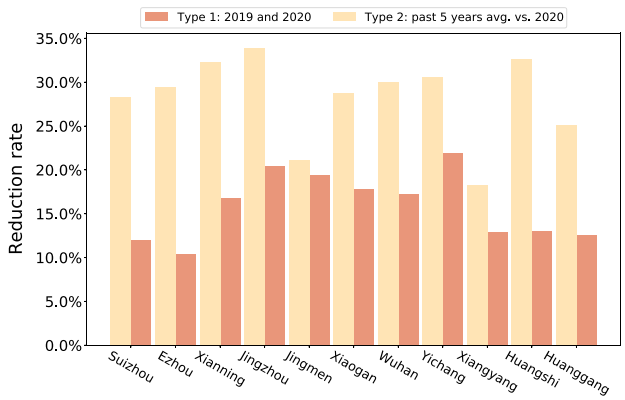


Fig. 2 PM2.5 reduction rate in each city. The yellow bar represents the reduction rate of the city's annual PM2.5 between the past 5 years and 2020. The orange bar represents the reduction rate of PM2.5 concentration between 2020 and 2019 (color figure online)

3.3 Comparative analysis between PM2.5 and migration trends

We combine the monthly averages of PM2.5 concentration in 11 cities in 2020 with the monthly average trends of AMI data for comparative analysis, is shown in Fig. 5. Due to the COVID-19 epidemic, Wuhan began to lockdown the city from 23rd January 2020 until 8th April 2020. In addition, Wuhan is the capital city of Hubei province, which inevitably affects migration flow size in other surrounding cities. This leads to a sharp decrease in the migration population in all most cities of Hubei province from January to April. Until April, the AMI slowly recovered to 0.3. Thus, the decline of the monthly PM2.5 trend from January to Jun 2020 is related to the lockdown

Table 2 Variations of PM2.5 concentration in each city

City	Type 1: 2019 and 2020 (%)	Type 2: past 5 years avg. versus 2020 (%)
Suizhou	−11.94	−28.28
Ezhou	−10.39	−29.32
Xianning	−16.82	−32.31
Jingzhou	−20.39	−33.89
Jingmen	−19.38	−21.12
Xiaogan	−17.80	−28.79
Wuhan	−17.21	−29.95
Yichang	−21.94	−30.50
Xiangyang	−12.87	−18.82
Huangshi	−13.00	−32.62
Huanggang	−12.58	−25.08

of the Wuhan. After human migration recovered in April, the overall PM2.5 concentration in Hubei province gradually rose in August.

To prove that the lockdown of Wuhan affects the migration flow size in surrounding cities, we plot the migration network of Hubei. We use Eq. (5) for averaging the AMI weight of each edge of the migration network for an entire year:

$$\bar{a}_{ij} = \frac{1}{T}\sum_{t}^{T} a_{ij}(t) \tag{5}$$

where T is the days of a year, and $\bar{a}_{ij}$ is the annual average AMI of each edge. The annual average AMI of all edges of the migration network is shown in Fig. 6, which indicates the variations in the migration population of 11 cities in Hubei province in 2020.

Obviously, most of the travelers in Hubei migrate from Wuhan to Xiaogan, Wuhan to Huanggang, Wuhan to Ezhou, and Wuhan to Xianning. It can be found that most of the routes with a relatively large AMI are all related to Wuhan. This proves that if the scale of human migration between Wuhan and surrounding cities is still very large, once Wuhan is in lockdown, the migration flow size of surrounding cities will inevitably be affected. The results of migration flow in the network provide an important reference for PM2.5 prediction. Moreover, the migration flow also is an additional feature of the prediction model, which reveals the information of nodes that the proposed model should aggregate. Hence, the analysis of migration flow is significant and helpful for the improvement of the proposed model.

4 PM2.5 concentration prediction based on migration attentive graph convolutional network

4.1 Migration-air graph representation

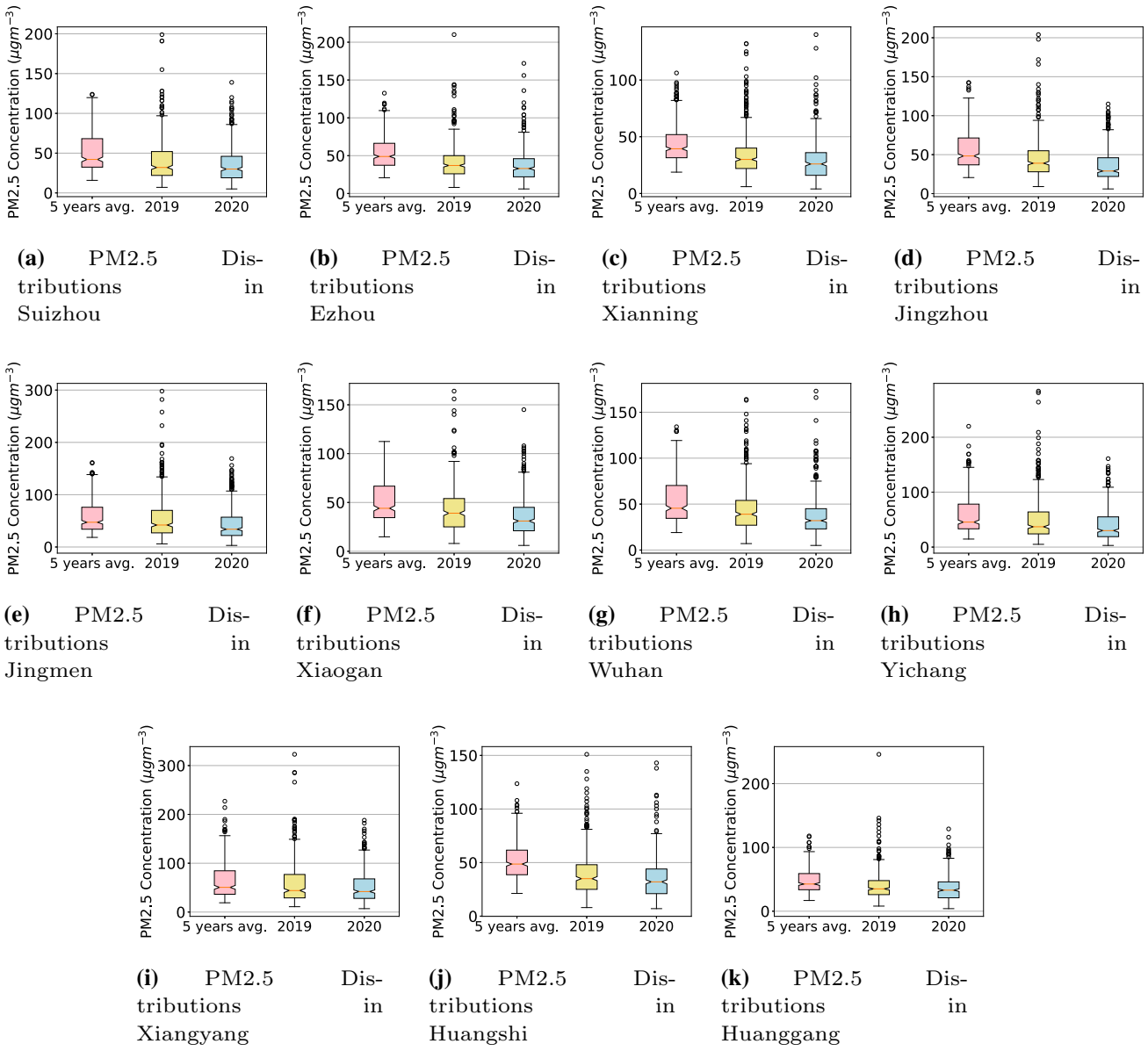
Here, we first develop weighted migration networks $G = (\mathcal{V}, \mathcal{E})$ for representing the migration flow size from one city to another city. In this network, node-j stands for the migration flow size from city-i to city-j at time t. Note that the migration network of cities in Hubei is a fully connected network. In order to facilitate prediction, we combine migration flow size a_{ij} and a_{ji} to derive the total migration flow size between city-i and city-j as follows:

$$m_{ij} = a_{ij} + a_{ji} \tag{6}$$

Then, we can derive a simplified migration network, with $m_{ij} = m_{ji}$, which is shown in Fig. 7.

Cities can be represented as a set $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$, and the number of nodes is $|\mathcal{V}| = N$. The migration network is transformed from a directed graph to an undirected graph, which is represented as $G = (\mathcal{V}, \mathcal{E}, \mathcal{M})$. $\mathcal{M}$ is the edge weight set formed by the AMI data, $m_{ij} \in \mathcal{M}$, which represents the weight of the edge connecting node-i and node-j. Additionally, we adopt the air pollution concentrations and climate data of each city in Hubei province as features h_i of each node in the migration network graph. Then, we can define migration-air graphs (MAGs), which are shown in Fig. 8.

Obviously, a MAG consists of multiple city nodes $\mathcal{V}$, and each city node has a feature set of air pollution and climate data $h_{v_i}(t) \in \mathbb{R}^{1\times F}$ at the time step t, where F is the number of features. Then the feature set combination matrix of all nodes can be expressed as $H(t) = \{h_{v_1}(t), h_{v_2}(t), \ldots, h_{v_N}(t)\}, H(t) \in \mathbb{R}^{N\times F}$, each MAG is represented by $G(t)$ at time step t.



(a) PM2.5 Distributions in Suizhou **(b)** PM2.5 Distributions in Ezhou **(c)** PM2.5 Distributions in Xianning **(d)** PM2.5 Distributions in Jingzhou

(e) PM2.5 Distributions in Jingmen **(f)** PM2.5 Distributions in Xiaogan **(g)** PM2.5 Distributions in Wuhan **(h)** PM2.5 Distributions in Yichang

(i) PM2.5 Distributions in Xiangyang **(j)** PM2.5 Distributions in Huangshi **(k)** PM2.5 Distributions in Huanggang

Fig. 3 PM2.5 concentration in 11 cities. The red box represents the distribution of average daily PM2.5 concentration from 2015 to 2019; the yellow box represents the distribution of daily PM2.5 concentration in 2019; the blue box represents the distribution of daily PM2.5 concentration in 2020. The upper and lower edges of each box represent the upper and lower quartiles, and the red line in the middle represents the median (color figure online)

Furthermore, we use a time window k with size d, and combine the MAG time series and feature set with the same window size for adopting the input data. It means that if the window size is d, then $\mathcal{G}_k = \{G(t+d-1), G(t+d-2), \ldots, G(t)\}, t = 1, 2, \ldots, T$ and feature set $H_k = \{H(t+d-1), H(t+d-2), \ldots, H(t)\}$ are adopted as input data for training forecasting model. MAG can not only use the air pollution concentration and climate characteristics of each node, but also predict the PM2.5 concentration based on the AMI data (edge weight) and the characteristics of each neighbor.

4.2 Migration attentive graph convolutional network

In the previous sections, we analyze that human migration has an indirect impact on the trend of PM2.5 concentration (especially during the post-pandemic age), and the air quality in each city is related to its neighbors. Therefore, we propose a model that fits the situation of this research by improving the attention mechanism. In this study, based on the established MAG graph data structure, we use graph GCNs to predict PM2.5 concentration in each city of Hubei province simultaneously. We are inspired by the GCN

Table 3 PM2.5 concentration distribution of the 11 cities

City	Past 5 years avg.				2019				2020			
	Max	Min	Mean	Median	Max	Min	Mean	Median	Max	Min	Mean	Median
Suizhou	124.00	15.80	51.23	42.10	199.00	7.00	41.73	32.00	139.00	5.00	26.74	30.00
Ezhou	132.60	20.80	53.67	48.90	210	8.00	42.27	37.0	172.00	6.00	37.87	33.00
Xianning	106.00	18.80	43.89	39.40	132.00	6.00	35.72	30.00	140.00	4.00	29.71	26.00
Jingzhou	142.60	20.60	55.52	48.30	204.00	9.00	46.10	39.00	115.00	6.00	36.70	29.00
Jingmen	161.40	18.40	57.35	47.30	298.00	6.00	56.11	42.00	169.00	3.00	45.24	34.00
Xiaogan	112.40	14.80	50.20	44.00	164.00	8.00	43.49	39.00	145.00	6.00	35.75	31.00
Wuhan	134.20	19.00	53.57	45.50	164.00	7.00	45.33	39.00	173.00	5.00	37.52	32.00
Yichang	220.00	14.60	58.60	45.50	284.00	5.00	52.17	37.00	161.00	3.00	40.72	30.00
Xiangyang	227.00	19.00	64.06	50.60	323.00	11.00	60.13	44.00	188.00	7.00	52.38	42.00
Huangshi	123.60	21.20	52.05	48.60	151.00	8.00	40.31	35.00	143.00	7.00	35.07	32.00
Huanggang	118.00	16.80	47.56	42.60	246.00	8.00	40.76	35.00	129.00	4.00	35.63	33.00

proposed and summarized by Kipf et al. [30] and the GAT proposed by Velivckovic et al. [31]. In this work, we propose a migration attentive graph convolutional network (MAGCN), which is a developed GCN for the node-task regression prediction of air quality with migration networks.

The GCN aggregation layer we utilize is a 1st order approximation to ChebyNet which was proposed by Defferrard [29], which can be defined as:

$$H^{l+1} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{l}W^{l}) \tag{7}$$

where σ is the ReLU function. H^l is the hidden feature matrix in layer l. W^l is learnable parameter matrix in layer l. $\widetilde{A}$ is the self-loop adjacent matrix $\widetilde{A} = A + I$. $\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$ is the normalization in the spectral graph processing, where D is the degree matrix.

Here, we define a degree matrix D as follows:

$$D_{ij} = \begin{cases} d(v_i) & i = j, \\ 0 & \text{otherwise}, \end{cases} \tag{8}$$

where $d(v)$ represents the degree of the node v and we have $\sum_{v\in\mathcal{V}} d(v) = 2|\mathcal{E}|$.

Each node v_i has hidden layer feature set h_{v_i}. We can directly derive the aggregation of the GCN layer from the node-level, then the GCN aggregation layer can be defined as:

$$h_{v_i}^{l+1} = \sigma\left(\sum_{v_k\in\mathcal{N}(v_i)\cup v_i} \frac{1}{c_{ik}} h_{v_k}^{l} W^{l}\right), \tag{9}$$

where $c_{ik} = \frac{1}{A_{ik}}$ is the normalization constant, and $\mathcal{N}(v_i) \cup v_i$ represents the node v_i and its neighbor set and itself (self-loop).

Here, we utilize the migration index m_{ij} of each edge in MAG as the Migration Attentive Coefficient (MAC) of the model, and calculate Softmax normalization processing on the MAC m_{ij}:

$$\beta_{ij} = \text{softmax}_j(m_{ij}) = \frac{\exp(m_{ij})}{\sum_{v_k\in\mathcal{N}(v_i)} \exp(m_{ik})}. \tag{10}$$

Then, we define the MAGCN layer with the normalized MAC β_{ij} and the conventional GCN aggregation:

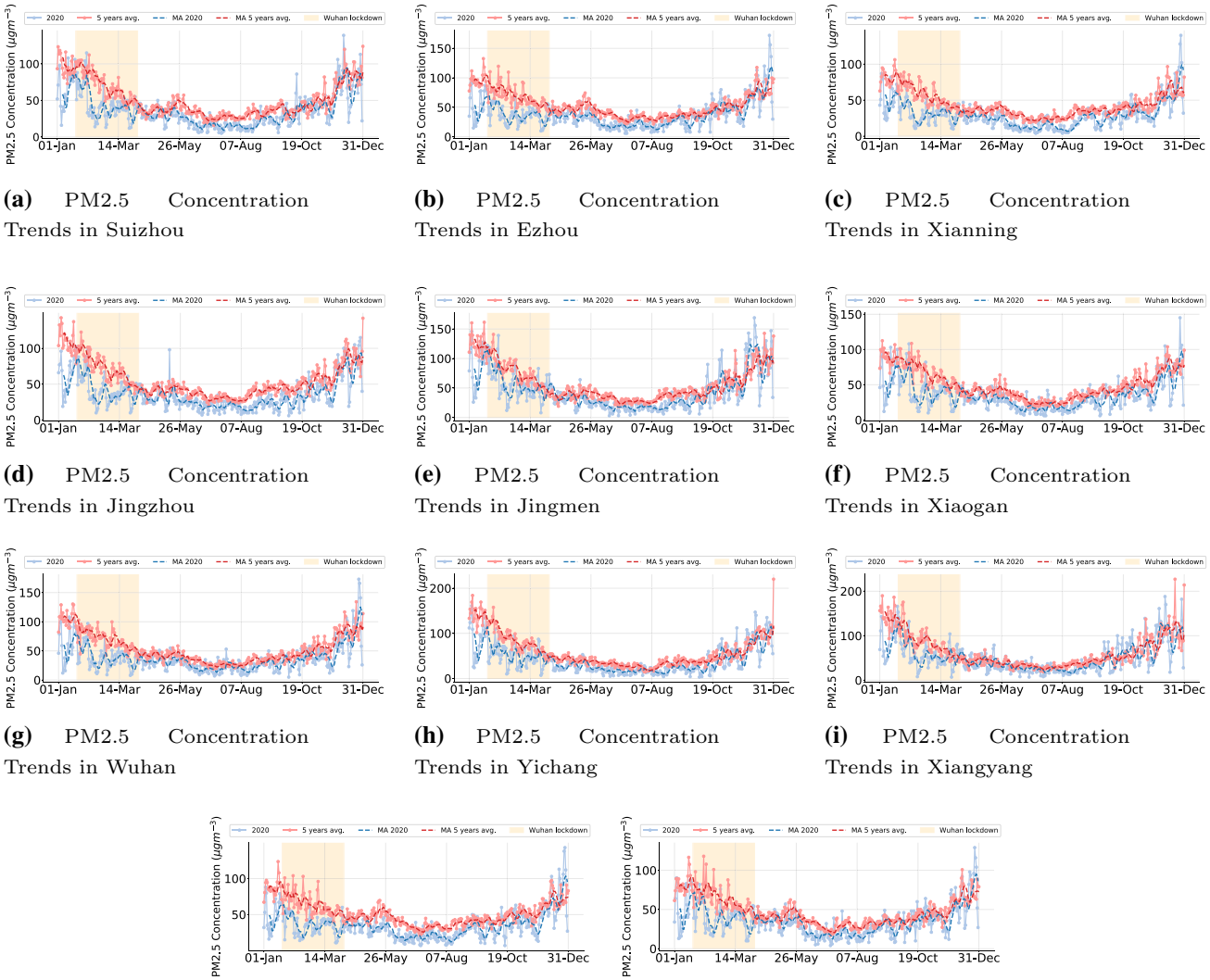
$$h_{v_i}^{l+1} = \sigma\left(\sum_{v_j\in\mathcal{N}(v_i)} \frac{1}{c_j}\beta_{ij} h_{v_j}^{l} W_{v_j}^{l} + \frac{1}{c_i} h_{v_i}^{l} W_{v_i}^{l}\right). \tag{11}$$

A layer of aggregation process of MAGCN can be represented in Fig. 9. The green area represents the neighbor aggregated nodes in the layer of MAGCN. The red node is the target node v_i, and the color depth of the brown edge connected to the red node and its neighbor nodes represents the value of MAC β_{ij}.

The MACs β of different edges can be combined as MAC matrix B. Finally, the convolutional aggregation process of a layer of Migration Attentive aggregation in matrix form can be expressed as:

$$H^{l+1} = \sigma(\widetilde{D}^{-\frac{1}{2}}B\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{l}W^{l}) \tag{12}$$

where $M \in \mathbb{R}^{N\times N}$. The difference between the proposed MAGCN and ordinary GAT is that ordinary GAT directly uses the hidden information h to calculate the attention coefficient, while MAGCN directly uses the migration index to calculate the MAC, which means that MAGCN uses the migration index as the edge weight to extract the information of neighbor city nodes.



(a) PM2.5 Concentration Trends in Suizhou

(b) PM2.5 Concentration Trends in Ezhou

(c) PM2.5 Concentration Trends in Xianning

(d) PM2.5 Concentration Trends in Jingzhou

(e) PM2.5 Concentration Trends in Jingmen

(f) PM2.5 Concentration Trends in Xiaogan

(g) PM2.5 Concentration Trends in Wuhan

(h) PM2.5 Concentration Trends in Yichang

(i) PM2.5 Concentration Trends in Xiangyang

(j) PM2.5 Concentration Trends in Huangshi

(k) PM2.5 Concentration Trends in Huanggang

Fig. 4 PM2.5 concentration in past 5 years and 2020 in 11 cities. The light blue and light red solid lines are the original PM2.5 concentration in 2020 and the original PM2.5 concentration in the past 5 years; the dark blue dotted line and the dark red dotted line are the MA concentration in 2020 and the past 5 years, respectively. The orange region is the lockdown period of Wuhan city from January 23rd, 2020, to April 8th, 2020 (color figure online)

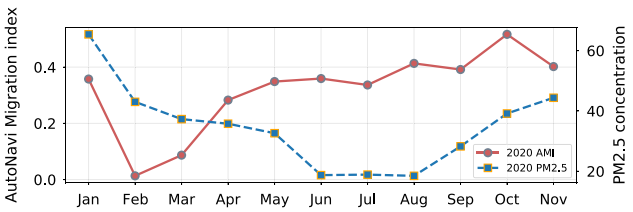


Fig. 5 Monthly PM2.5 trends and AutoNavi Migration Index. The red line is the AutoNavi Migration Index (AMI), and the blue line is the PM2.5 concentration trend. The trend range of the line segment in the figure is from December 2019 to November 2020. The y axis on the left is the specific value of the migration index, and the y axis on the right is the PM2.5 concentration (color figure online)

5 Experimental results

5.1 Algorithm settings

In this study, four Graph Neural Networks are used for comparison as the baseline models, including ChebyNet [29], GGNN [36] , GCN [30] and GAT [31]. In the optimization framework, L2 regularization is added to the Mean Square Error (MSE) loss function. The loss function of L2 regularization is as follows:

$$l(w) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{k=1}^{K}(f_{i,k} - y_{i,k})^2 + \lambda\sum_{j=1}^{M}w_j^2, \tag{13}$$

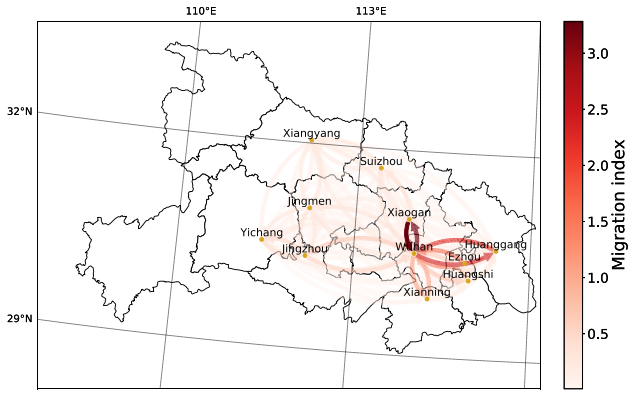


Fig. 6 2020 Average AutoNavi Migration Index on Hubei Migration Network. The darker the color of an edge in the figure, the larger the average AMI of the edge

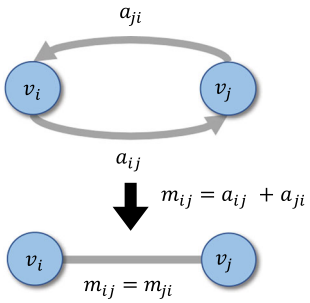


Fig. 7 Sum of direction migration indexes

where $f_{i,k}$ and $y_{i,k}$ is the predicted and observed PM2.5 concentration over a k days time windows, respectively; $l(w)$ is the error depending on the model parameter w; $\lambda \sum_{j=1}^{M} w_j^2$ is the L2 parameter of λ L2 regularization term.

We use 70% of the entire dataset as the training set and the remaining 30% dataset as the test set. Here, we adopt a grid search method to search the optimal hyper-parameters

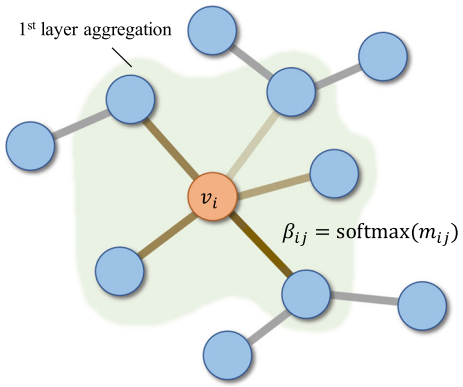


Fig. 9 The process of 1 layer MAGCN aggregation

of each model. Here, we consider three testing scenarios with different time window sizes, namely, $d = 1$, 4, and 7 days, respectively. In each scenario, the grid search method is adopted to find the optimal hyper-parameters of the four baseline models and the MAGCN model, while the detailed information of the hyper-parameters of these models is presented in Tables 4, 5 and 6, respectively, where "Order" is the order of Chebyshev polynomials.

We adopt 4 common regression task evaluation criteria for evaluating the performance of these model, including MSE, RMSE, MAE and R^2. These 4 evaluation criteria can be calculated from Eqs. (14) to (17), where f_i and y_i stand for the predicted and observed PM2.5 concentration of node v_i.

- Mean square error (MSE):

$$\text{MSE} = \frac{1}{K}\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N}(f_{i,k} - y_{i,k})^2. \tag{14}$$

- Root mean square error (RMSE):

Fig. 8 Migration-air pollution Graph Representation

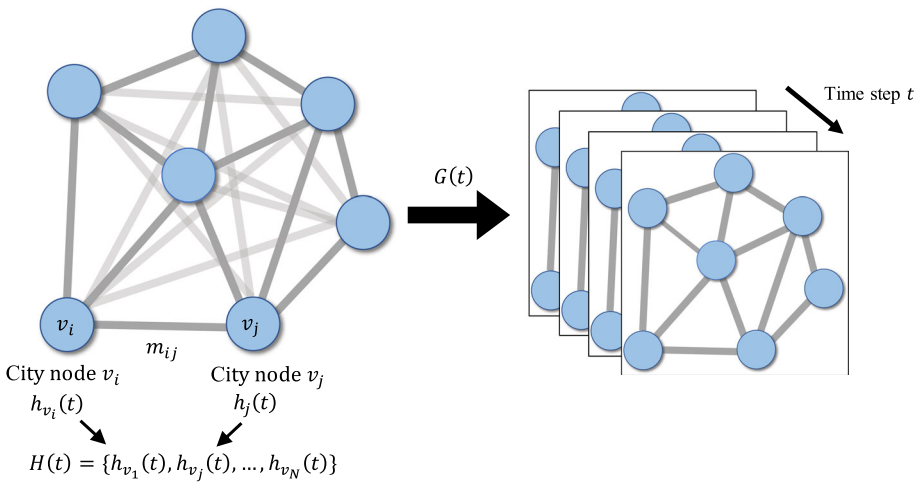


Table 4 Hyper-parameters in scenario-I with window size $d = 1$ day

Model	Batch size	Neuron	Layer num	Order	Learning rate	L2
ChebyNet	8	64	4	2	0.1	0.0001
GGNN	8	8	1	NA	0.1	0.0001
GCN	16	64	3	NA	0.1	0
GAT	16	64	3	NA	0.1	0
MAGCN	32	8	1	NA	0.1	0

Table 5 Hyper-parameters in scenario-II with window size $d = 4$ day

Model	Batch size	Neuron	Layer num	Order	Learning rate	L2
ChebyNet	8	8	4	1	0.1	0.01
GGNN	32	32	3	NA	0.1	0.1
GCN	16	64	1	NA	0.1	0.0001
GAT	8	8	1	NA	0.1	0.1
MAGCN	8	128	1	NA	0.1	0.0001

Table 6 Hyper-parameters in scenario-III with window size $d = 7$ day

Model	Batch size	Neuron	Layer num	Order	Learning rate	L2
ChebyNet	8	8	1	4	0.1	0.01
GGNN	32	56	2	NA	0.1	0
GCN	8	64	1	NA	0.01	0.0001
GAT	8	64	1	NA	0.01	0.0001
MAGCN	8	64	1	NA	0.1	0.001

$$\text{RMSE} = \frac{1}{K}\frac{1}{N}\sum_{k=1}^{K}\sqrt{\sum_{i=1}^{N}(f_{i,k} - y_{i,k})^2}. \tag{15}$$

- Mean absolute error (MAE):

$$\text{MAE} = \frac{1}{K}\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N}|f_{i,k} - y_{i,k}|. \tag{16}$$

- R square (R^2):

$$R^2 = \frac{1}{K}\sum_{k=1}^{K}1 - \frac{\sum_{i=1}^{N}(f_{i,k} - y_{i,k})^2}{\sum_{i=1}^{N}(y_{i,k} - \bar{y_k})^2}. \tag{17}$$

  where $\bar{y}$ is the average of observed PM2.5 concentrations.

For MSE, MAE, and RMSE, they are smaller for the better model. The range of R^2 is $(-\infty, 1]$, then the closer R^2 to 1, the better the prediction result.

5.2 Forecasting results

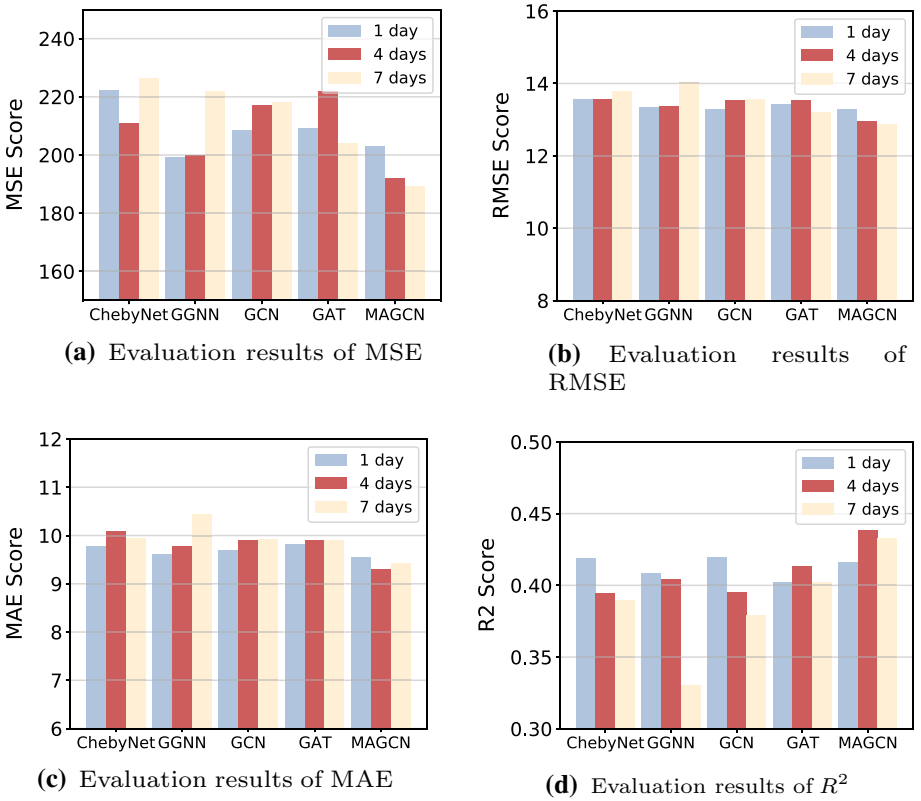
We set 3 different window sizes ($d = 1$, 4, and 7 days) as 3 scenarios for training and use MSE, RMSE, MAE and R^2 as criteria to evaluate these models, including ChebyNet, GGNN, GCN, GAT and the proposed MAGCN model. The experimental results are shown in Table 7 and Fig. 10. the bold texts in the table stand for the best scores. Forecasting results after tests denote that the proposed MAGCN model can provide adequate performance when $d = 1$ day with MAE as 9.5444. If the window sizes are $d = 4$ and $d = 7$ days, the proposed MAGCN model outperforms other models. When the window size is $d = 4$ days, the best forecasting models are MAGCN with MSE as 191.9483, RMSE as 12.9515, MAE as 9.3074, and R2 as 0.4383. Similarly, in the scenario with $d = 7$ days, the MSE, RMSE, MAE and R2 of the MAGCN prediction PM2.5 concentration are 189.2669, 12.8878, 9.4314, 0.4328, which is also the best results.

In detail, we compare the evaluation criteria of MAGCN model in different scenarios $d = 1, d = 4, d = 7$, as shown in Fig. 11. The evaluation results show that, in scenario $d = 1$, all criteria are far worse than in scenario $d = 4$ and $d = 7$. In the scenario with $d = 4$, the MAE of MAGCN is better than the scenario with $d = 7$, which is 9.3074. Besides, in scenario $d = 7$, the R^2 is higher than the scenario $d = 4$, which reaches 0.4383. On the contrary, both MSE and RMSE score in scenario $d = 7$ are better than the results in scenario $d = 7$. These results show that the proposed MAGCN forecasts the PM2.5 concentration can perform better results in scenarios $d = 4$ and $d = 7$.

In order to further compare the PM2.5 concentration prediction performances of five models in the scenario $d = 4$ and $d = 7$, Fig. 12 are shown to represent predicted and

Table 7 Forecasting precision indexes with different window size

Model	Window size $d = 1$				Window size $d = 4$				Window size $d = 7$			
	MSE	RMSE	MAE	R^2	MSE	RMSE	MAE	R^2	MSE	RMSE	MAE	R^2
ChebyNet	222.3544	13.5801	9.7783	0.4187	210.9322	13.575	10.0912	0.3941	226.613	13.7837	9.9426	0.3897
GGNN	**199.1878**	13.3448	9.6168	0.4087	199.8944	13.3703	9.778	0.4043	222.02	14.0395	10.4437	0.3303
GCN	208.6679	**13.3026**	9.692	**0.4197**	217.1349	13.5393	9.9001	0.3952	218.2019	13.5783	9.9258	0.379
GAT	209.3639	13.4263	9.8171	0.4019	221.929	13.5432	9.8969	0.4134	204.1847	13.2148	9.9102	0.402
MAGCN	203.1844	13.305	**9.5444**	0.4164	**191.9483**	**12.9515**	**9.3074**	**0.4383**	**189.2669**	**12.8878**	**9.4314**	**0.4328**



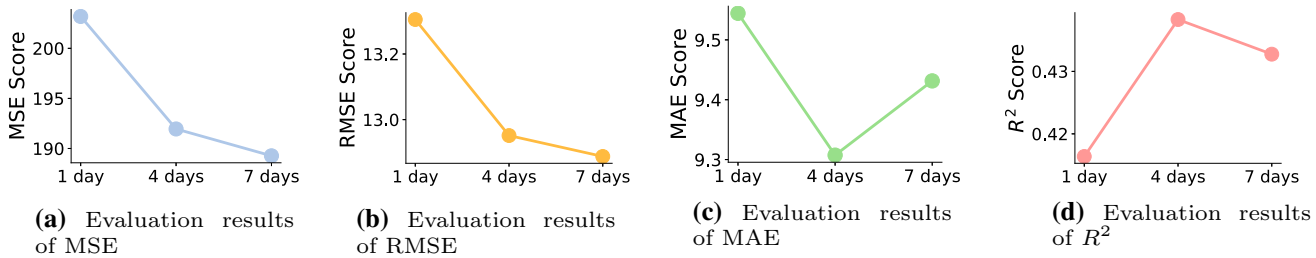
(a) Evaluation results of MSE

(b) Evaluation results of RMSE

(c) Evaluation results of MAE

(d) Evaluation results of R^2

Fig. 10 Evaluation criteria. The y axis represents the score of each evaluation criteria, the blue bar represents the result tested with the window size $d = 1$, and the red bar represents the test result of the dataset using the window size $d = 4$ and the yellow bar represents the test result of the dataset with the window size $d = 7$ (color figure online)



(a) Evaluation results of MSE

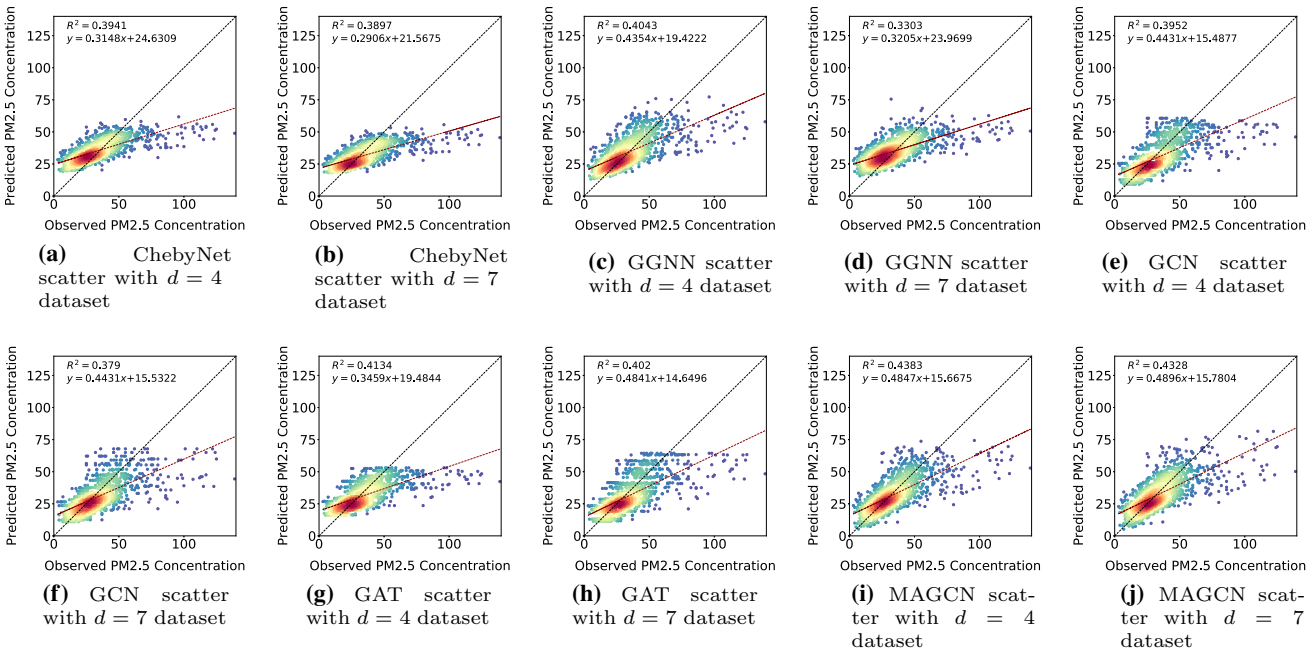
(b) Evaluation results of RMSE

(c) Evaluation results of MAE

(d) Evaluation results of R^2

Fig. 11 MAGCN test with different window sizes of datasets

observed values by scatters and lines. The scatters distribute closely around the diagonal. The slopes of the scatter trend line of all these models are all less than 1.0 and the intercepts are positive. Note that the scatter trend slope is more close to 1.0, which stands for the prediction performance better. The ChebyNet shows the worst forecasting performance (Fig. 12a, b), the slopes of scatter trend in both scenario $d = 4$ and $d = 7$ are smallest (0.3148



(a) ChebyNet scatter with $d = 4$ dataset **(b)** ChebyNet scatter with $d = 7$ dataset **(c)** GGNN scatter with $d = 4$ dataset **(d)** GGNN scatter with $d = 7$ dataset **(e)** GCN scatter with $d = 4$ dataset

(f) GCN scatter with $d = 7$ dataset **(g)** GAT scatter with $d = 4$ dataset **(h)** GAT scatter with $d = 7$ dataset **(i)** MAGCN scatter with $d = 4$ dataset **(j)** MAGCN scatter with $d = 7$ dataset

Fig. 12 Scatter plots with the comparison models. The x axis is the true PM2.5 concentration (Observed Value), and the y axis is the predicted PM2.5 concentration value (Predicted Value). The color of the scatters represents the density of the scatters. The red dotted line is the fitting trend of scatters, and gray dotted line is the diagonal (color figure online)

and 0.2906). For GCN and GAT in scenario $d = 7$ (Fig. 12f, h), although the slopes of the scatter trend are 0.4431 and 0.4841, the distribution of scatters between the predicted and observed values deviate from the diagonal. Comparing with all the models, Fig. 12i, j show that the proposed MAGCN in this paper performs the best fit between the predicted and observed values, the slopes of scatter trend in both scenario $d = 4$ and $d = 7$ are highest (0.4847 and 0.4896), while the R^2 of the proposed model are also highest at 0.4383 and 0.4328. The experimental results indicate that considering the human migration data in graph embedding models like GCN and can achieve a better prediction performance.

6 Conclusion

In this study, we propose an MAGCN model for PM2.5 concentration forecasting considering human migration data. In order to prove the human migration is influenced by the COVID-19 outbreak and affects the variation of PM2.5 concentration, we conduct a series of data analysis. Above all, we analyze the annual PM2.5 concentration variation in 11 Hubei cities from 2015 to 2020. We find that the PM2.5 concentration in 2020 reduces in all cities, compared with the past 5 years. Besides comparing the human migration flow size with PM2.5 concentration and finding that in January 2020, both of two variables simultaneously reduced until April 2020. Based on the results of data analysis, we establish a human migration network migration-air graph (MAG) by utilizing human migration flow size from the AutoNavi and air pollution datasets. Then, we adopt four graph embedding models ChebyNet, GGNN, GCN, GAT as baseline models, and we test all of the models, including our proposed MAGCN model, in three scenarios with $d = 1$, 4, 7. Experimental results indicate that the proposed MAGCN model can predict more accurately than other baselines, especially in scenarios with $d = 4$, 7.

Our MAGCN model shows a better result for forecasting PM2.5 concentration while considering human migration. Human migration is not the only factor for PM2.5 concentration forecasting in the post-COVID-19 pandemic age. We will explore and combine more other potential factors that could improve the accuracy of forecasting PM2.5 concentration while using graph embedding models in future work.

Acknowledgements This work is partially supported by the Natural Science Foundation of Guangdong Province, China (2020A1515010761), and Key R & D Program in Key Areas of Guangdong Province under Grants, China (2020B010166001).

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declaration

Conflict of interest The authors declared that they have no conflicts of interest to this work.

References

1. Zhan C, Tse C, Fu Y, Lai Z, Zhang H (2020) Modelling and prediction of the 2019 coronavirus disease spreading in china incorporating human migration data. PLoS ONE 15:e0241171
2. Zhan C, Chi KT, Lai Z, Chen X, Mo M. General model for COVID-19 spreading with consideration of intercity migration, insufficient testing and active intervention: application to study of pandemic progression in Japan and USA. medRxiv
3. Gayen A, Haque SM, Mishra SV (2021) COVID-19 induced lockdown and decreasing particulate matter (PM10): an empirical investigation of an Asian megacity. Urban Clim 36:100786
4. Gao C, Li S, Liu M, Zhang F, Achal V, Tu Y, Zhang S, Cai C (2021) Impact of the COVID-19 pandemic on air pollution in Chinese megacities from the perspective of traffic volume and meteorological factors. Sci Total Environ 773:145545
5. Jiaxin C, Hui H, Feifei W, Mi Z, Ting Z, Shicheng Y, Ruoqiao B, Nan C, Ke X, Hao H (2021) Air quality characteristics in Wuhan (China) during the 2020 COVID-19 pandemic. Environ Res 195:110879
6. Zhang H, Zhang L, Yang L, Zhou Q, Zhang X, Xing W, Hayakawa K, Toriba A, Tang N (2021) Impact of COVID-19 outbreak on the long-range transport of common air pollutants in Kuwams. Chem Pharm Bull 69(3):237–245
7. Toro R, Catalán F, Urdanivia FR, Rojas JP, Manzano CA, Seguel R, Gallardo L, Osses M, Pantoja N, Leiva-Guzman MA (2021) Air pollution and COVID-19 lockdown in a large South American city: Santiago metropolitan area, Chile. Urban Climate 36:100803
8. Baldasano JM (2020) COVID-19 lockdown effects on air quality by NO2 in the cities of Barcelona and Madrid (Spain). Sci Total Environ 741:140353
9. Huang C, Wang T, Niu T, Li M, Liu H, Ma C (2021) Study on the variation of air pollutant concentration and its formation mechanism during the COVID-19 period in Wuhan. Atmos Environ 251:118276
10. Dong L, Chen B, Huang Y, Song Z, Yang T (2021) Analysis on the characteristics of air pollution in china during the COVID-19 outbreak. Atmosphere 12(2):205
11. Cole MA, Neumayer E (2004) Examining the impact of demographic factors on air pollution. Popul Environ 26(1):5–21
12. Zhang T, Wooster MJ, Green DC, Main B (2015) New field-based agricultural biomass burning trace gas, PM2.5, and black carbon emission ratios and factors measured in situ at crop residue fires in Eastern China. Atmos Environ 121:22–34
13. Huang L, Zhang C, Bi J (2017) Development of land use regression models for PM2.5, SO2, NO2 and O3 in Nanjing, China. Environ Res 158:542–552
14. Jiang P, Yang J, Huang C, Liu H (2018) The contribution of socioeconomic factors to PM2.5 pollution in urban China. Environ Pollut 233:977–985
15. Lovrić M, Pavlović K, Vuković M, Grange SK, Haberl M, Kern R (2021) Understanding the true effects of the COVID-19 lockdown on air pollution by means of machine learning. Environ Pollut 274:115900
16. Tiwari A, Gupta R, Chandra R (2021) Delhi air quality prediction using LSTM deep learning models with a focus on COVID-19 lockdown. arXiv preprint. arXiv:2102.10551
17. Mustafić H, Jabre P, Caussin C, Murad MH, Escolano S, Tafflet M, Périer M-C, Marijon E, Vernerey D, Empana J-P et al (2012) Main air pollutants and myocardial infarction: a systematic review and meta-analysis. JAMA 307(7):713–721
18. Zhu H, Lu X (2016) The prediction of PM2.5 value based on ARMA and improved BP neural network model. In: International conference on intelligent networking and collaborative systems (INCoS), IEEE, pp 515–517
19. Voukantsis D, Karatzas K, Kukkonen J, Räsänen T, Karppinen A, Kolehmainen M (2011) Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki. Sci Total Environ 409(7):1266–1276
20. Mahajan S, Chen L-J, Tsai T-C (2017) An empirical study of PM2.5 forecasting using neural network. In: IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, internet of people and smart city innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE, pp 1–7
21. Chen Z, Ye X, Huang P (2018) Estimating carbon dioxide (CO_2) emissions from reservoirs using artificial neural networks. Water 10(1):26
22. Biancofiore F, Busilacchio M, Verdecchia M, Tomassetti B, Aruffo E, Bianco S, Di Tommaso S, Colangeli C, Rosatelli G, Di Carlo P (2017) Recursive neural network model for analysis and forecast of PM10 and PM2.5. Atmos Pollut Res 8(4):652–659
23. Tsai Y-T, Zeng Y-R, Chang Y-S (2018) Air pollution forecasting using RNN with LSTM. In: IEEE 16th international conference on dependable, autonomic and secure computing. 16th International conference on pervasive intelligence and computing. 4th International conference on big data intelligence and computing and cyber science and technology congress (DASC/PiCom/DataCom/CyberSciTech). IEEE, pp 1074–1079
24. Tao Q, Liu F, Li Y, Sidorov D (2019) Air pollution forecasting using a deep learning model based on 1d convnets and bidirectional GRU. IEEE Access 7:76690–76698
25. Cheng Y, Zhang H, Liu Z, Chen L, Wang P (2019) Hybrid algorithm for short-term forecasting of PM2.5 in China. Atmos Environ 200:264–279
26. Niu M, Wang Y, Sun S, Li Y (2016) A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM2.5 concentration forecasting. Atmos Environ 134:168–180
27. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G (2008) The graph neural network model. IEEE Trans Neural Netw 20(1):61–80
28. Bruna J, Zaremba W, Szlam A, LeCun Y (2013) Spectral networks and locally connected networks on graphs. arXiv preprint. arXiv:1312.6203
29. Defferrard M, Bresson X, Vandergheynst P (2016) Convolutional neural networks on graphs with fast localized spectral filtering. arXiv preprint. arXiv:1606.09375
30. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint. arXiv:1609.02907
31. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. arXiv preprint. arXiv:1710.10903
32. Qi Y, Li Q, Karimian H, Liu D (2019) A hybrid model for spatiotemporal forecasting of PM2.5 based on graph convolutional

neural network and long short-term memory. Sci Total Environ 664:1–10
33. Zhou H, Zhang F, Du Z, Liu R (2021) Forecasting PM2.5 using hybrid graph convolution-based model considering dynamic wind-field to offer the benefit of spatial interpretability. Environ Pollut 273:116473
34. Wang C, Zhu Y, Zang T, Liu H, Yu J (2021) Modeling inter-station relationships with attentive temporal graph convolutional network for air quality prediction. In: Proceedings of the 14th ACM international conference on web search and data mining, pp 616–634
35. Wang S, Li Y, Zhang J, Meng Q, Meng L, Gao F (2020) PM2.5-GNN: a domain knowledge enhanced graph neural network for PM2.5 forecasting. In: Proceedings of the 28th international conference on advances in geographic information systems, pp 163–166
36. Li Y, Tarlow D, Brockschmidt M, Zemel R (2015) Gated graph sequence neural networks. arXiv preprint. arXiv:1511.05493