International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n2.08
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Retrieval of PM2.5 Using MODIS Aerosol Products Based on Depth Learning

Xueke Zheng[1], Ziqian Zhang[2], Yichi Zhang[3]

[1]School of Survey and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454003, China

[2]School of Software, Henan University of Science and Technology, Luoyang 471003, China

[3]Hohai University School of Artificial Intelligence and Automation, Nanjing 211100, China

ABSTRACT

Most of the current PM2.5 concentration estimation models are insufficient, such as narrow research scope, poor representativeness and complex model. To improve the above deficiencies, this paper tries to explore a simpler and more efficient PM2.5 concentration estimation model by using deep neural network. In this paper, the aerosol optical thickness (Aerosol Optical Depth, AOD), which has been widely used in the field of PM2.5 concentration estimation model, and other known variables related to the distribution of PM2.5, wind speed, relative humidity, build a small neural network using dense connection network, and estimate the PM2.5 concentration in 2020. The results proved that deep learning has efficient regularity mastery in PM2.5 estimation. After about 20 minutes of iterative training in 8 months of data, the average absolute error was 7.8μg/m3 in the annual data. And with the deepening of the number of model layers and the number of iterations, the accuracy will be steadily improved. Through the estimation of the four quarters of 2020, the relatively expected results have been achieved.

KEYWORDS

Deep Learning; Deep Neural Network; MODIS; AOD; PM2.5

1. INTRODUCTION

PM2.5 refers to the particulate matter with aerodynamic equivalent diameter less than or equal to 10 μ m in ambient air, also known as inhalable particulate matter. According to the national standard, PM2.5 is measured by microoscillation balance method or β -ray method, and the 24-hour average concentration limit is 75 μ/m3 in residential areas, commercial traffic areas, cultural areas and residential areas, industrial areas and rural areas[1]. The current PM2.5 concentration monitoring, mainly depends on the ground environment air pollutants monitoring, the most accurate data records, but there are more these sites, such as limited number and uneven distribution (especially in the plateau), can represent area is limited and mostly concentrated in urban areas, etc., not well reflect the large scale PM2.5 spatial distribution. In the atmospheric field, studies have shown that sensors from satellite platforms can effectively estimate PM2.5 concentration on the surface. Among the products produced by satellite sensors, aerosol optical thickness (Aerosol Optical Depth, AOD) is the most widely used to study the spatial distribution of PM2.5[2]。

Before this, a large number of relevant studies at home and abroad have shown that there is a strong correlation between AOD and PM2.5 concentration, and more outstanding research results have been achieved. Many studies have used different data sources and different scales of AOD data to establish

many linear or non-linear PM2.5 concentration estimation models, mainly including regression statistical models, such as Yang et al[2]The proposed linear mixed-effects model, Pawan Gupta et al[3]The proposed multiple linear regression model, and the Ma et al[4]The proposed geographically weighted regression model; the machine learning model, such as Shen et al[5]For the proposed random forest model, Chen et al[6]The proposed gradient hoist (eXtreme Gradient Boosting, XGBoost) model. Wei Jing et al[7]Proposed extreme random tree model for space-time. These models, while establishing the relationship between AOD and PM2.5 concentration, Meteorological data (for example, wind speed, wind direction, relative humidity, boundary layer height, Precipitation, etc.), air composition data (such as carbon monoxide, sulfur dioxide, nitrogen oxide, etc.), or surface information Digital Elevation Model DEM), the Normalized Differential Vegetation Index (NDVI), surface cover type) and other auxiliary data, As a correction to the model, And the good regression effect has been achieved. However, most models generally have the problem of small research areas, that is, the research areas are mostly concentrated in one city or one province, which leads to the reduced applicability of the model and the greater chance in the model accuracy verification, and can not well represent the AOD PM2.5 relationship in a larger range or other regions.

While fully understanding and borrowing the mechanism and shortcomings of the above models, A MODIS(Moderate Resolution Imaging Spectroradiometer) in mainland China from 01 January 2020 to 31 December 2020 MAIAC (Multi-Angle Implementation of Atmospheric Correction) product, PM2.5 ground monitoring station data, ERA5 reanalysis meteorological data and surface condition data (DEM and NDVI). Previous machine learning methods (such as random forest, gradient hoist, etc.) can only transform the input data into one or two continuous representation spaces (so it is also called shallow learning), and when the data becomes complex, these shallow technologies often fail to support the ability to master complex data.

2. DATA SOURCES

2.1. MODIS MAIAC AOD Data

In this study, the AOD data is a MAIAC algorithm product based on MODIS data, with a spatial resolution of 1km and the MODIS family product named MCD19A2. The MAIAC algorithm is a new aerosol inversion algorithm, which assumes that the space is uniform and stable, and calculates AOD by dynamically decoupling aerosol and terrestrial reflection contributions using time series data while considering the influence of bidirectional surface reflectivity[8]. On May 30, 2018, the official release of aerosol products with 1km spatial resolution. The product uses an improved MAIAC algorithm to improve on the spectral regression coefficient, cloud detection, aerosol pattern determination, and the overall optimization of the global aerosol inversion process[9]。

In this study, MAIAC 550nm AOD data in Chinese mainland from January 1 to December 31,2020 were collected with 10,248 data (HDF format) with 1 to 10 bands each, each data recorded during the transit of Aqua or Terra satellites, the transit time is concentrated between 7:00 to 18:00 (23:00 to 10:00 the day before UTC). In the pre-processing of data, only extracting the quality assurance (Quality Assurance, QA) data retains the "high quality" images (QACloud Mask=Clear and QAA djacency Mask=Clear)。The average value of each band in each scene is taken as the data of the scene, and finally the daily transit average image is stitched together into a map of the whole country to facilitate further extraction and calculation in the later stage.

2.2. Ground observation data

All ground observation data of PM2.5 concentration are from the National Urban Air Quality Real-time release platform of China Environmental Monitoring Station, with a time resolution per hour. One separate file every day, each file contains many data items, including PM2.5,

PM10、SO2、NO2、O3 And the respective 8-and 24-hour mean data. However, due to the problems of monitoring site data uploading instruments, the quality of the data set is not good, that is, a site is not recorded for 24 hours, and there is no value in the study time. Therefore, only the sites with PM2.5 concentration recorded for the daily duration of the study were selected to obtain PM2.5 records for 1524 sites throughout 2020. To correspond to the MAIAC data, the data at 818 at each site were then averaged as the daily average of the day. Figure 1 shows the distribution of all the sites.

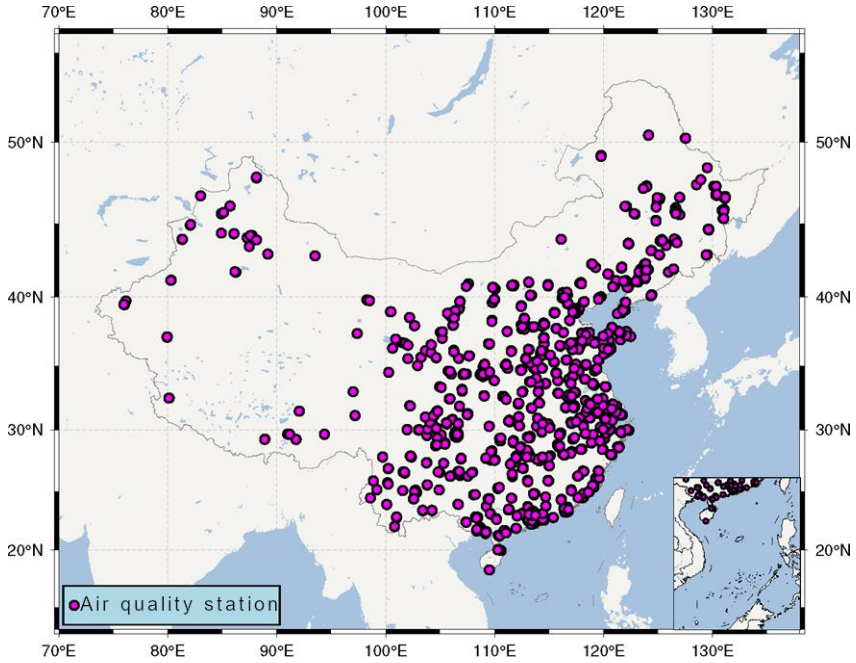

Figure 1 Distribution of the 1524 air quality sites used in this study

2.3. Auxiliary data

2.3.1. The ERA 5 data

ERA 5 Reanalysis Meteorological Data is the latest generation of meteorological reanalysis data created by the Copernican Climate Change Service and operated by the European Centre for Medium-term Weather Forecast (ECWMF). ECWMF Can now provide ERA 5 data from 1979 to today, with quality assurance data and a three-month lag update. Preliminary data can be updated by 5 days[15]. In terms of data acquisition, all ERA 5 element data can be downloaded for free. You can select the elements, year, month, date, hour, data format and region online, and submit orders online for online download. Data can also be downloaded through the official Python API. In this study, hourly data were downloaded from 2020 to 18:00 (0:0 to 10:00 UTC), including six elements: north component wind speed (V), east component wind speed (U), 2m dew-point temperature (Td), 2m temperature (T), boundary layer height (BLH), and surface pressure (SP). Where Td and T (both in degrees Celsius) can be used to calculate the relative humidity (RH), see formula (1).

$$RH = 100 \cdot \exp\left(\frac{17.67 \cdot Td}{243.5+Td} - \frac{17.67 \cdot T}{243.5+T}\right) \tag{1}$$

The two component wind speed can be used to calculate the wind speed (WS) and wind direction (WD) at this point, see formula (2)

$$WS = \sqrt{U^2 + V^2}$$

$$WD = \begin{cases} \arctan\left(\frac{U}{V}\right) \cdot \frac{180}{\pi} & V < 0 \\ \arctan\left(\frac{U}{V}\right) & V > 0 \end{cases} \tag{2}$$

A single file in the format of GRIB 2 was downloaded, with 366×11×7 = 28182 layers. In this study, the data from each day (8:00 to 18:00,11 hours) were averaged daily, finally extracting the daily meteorological elements of each site.

2.3.2. Digital elevation model

In August 2015, NASA opened the free download channel for the SRTMv3 1 arc-second resolution DEM. In addition, the International Agricultural Research Advisory Group (CGIAR-CSI) of the Spatial Information Alliance used different interpolation algorithms and auxiliary DEM to fill the void data to release the SRTMv4.1 version on the basis of the 3 arc second resolution data product[10, 11]. The DEM data used in this study is version SRTMv4.1, with a spatial resolution of 90 m. From the CGIAR-CSI web site (https://srtm.csi.cgiar.org) downloaded 62 DEM data with a range of 5°×5°, which to cover China (except the South Sea Islands), and extracted the elevation values of all PM2.5 sites in Figure 1.

2.3.3. Normalized Differential Vegetation Index

Considering the inhibitory effect of vegetation on PM2.5 diffusion, NDVI was introduced in the model in this study. NDVI is a kind of ratio data obtained based on the near-infrared band and infrared band operation. Its value is between-1 and 1. This study was conducted from the NASA Goddard Space Center (https://ladsweb.modaps.eosdis.nasa.gov/) downloaded the MOD13A3 data of 226 scenes in 2020.

3. DEEP LEARNING PM2.5 ESTIMATION MODEL BUILDING

In this study, samples can be regarded as simple vector data of AOD, DEM, NDVI and meteorology, and PM2.5 as$y$. These vector data are stored in 2D tensors and are usually processed with a dense connected layer [densely connected layer, also called a fully connected layer (fully connected layer) or a dense layer (dense layer)]. This study used dense connected networks. To facilitate network building, the model for this study was implemented using the Python-based Keras. Keras is a popular framework in the industry to build neural network models using Python language, which can easily define and train almost all types of deep learning models. In Keras' framework the dense connection layers correspond to the Dense object. Building a dense link network using Keras is very simple, just adding a Dense layer to the model all the time. This network is usually used to process vector data. This network assumes that the input features have no specific structure and that each cell within the layer is connected to all other cells. Ddensely connected networks are most commonly used as classification data (such as in this model, the input is a series of PM2.5 influence factors and representational factors). In this study, a densely connected network with ten hidden layers was built. The input of the network is 8 variables: AOD, WSRH, TEM, BLH, SP, NDVI, and DEM (normalized), and the output is PM2.5 concentration, as shown in Figure 2.

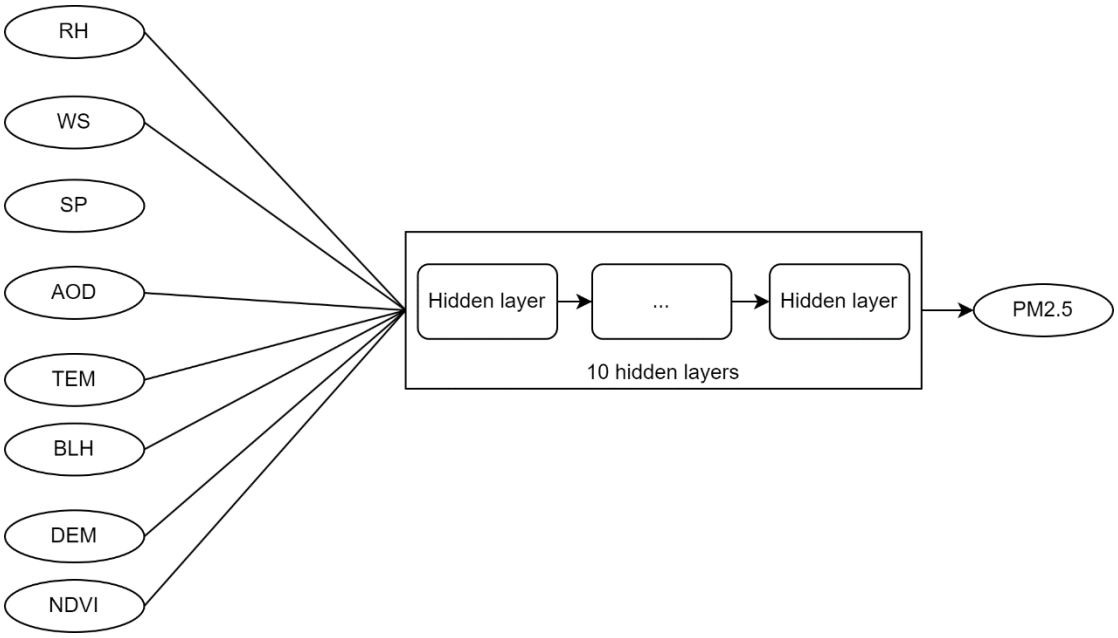

Figure 2 Schematic diagram of the model

Data aspects. After deleting the invalid and unreliable values of satellite data, a total of 77564 sets of data were obtained. Two months of data were selected for training in each quarter, so 49571 samples of 8 months participated in the training. Due to the large amount of data available, multi-fold cross-validation was chosen instead of setting aside the validation set. Split out 70% for training, 30% for validation and output validation accuracy. In addition, four quarters of 2020 were tested.

The various parameter settings of the model established in this study are shown in Table 1.

Table 1 Model parameter settings

epoch	activation function	hidden layer	loss function	optimizer	regularization	Sample batch
4000	ReLU	10	MSE	Adam	Dropout	128

4. PM2.5 ESTIMATION AND ACCURACY ANALYSIS

4.1. Model verification

This study used the MAE and evaluated and monitored the performance of the model during training, using R2 To evaluate the performance of the model on the PM2.5 concentration estimates. The validation results of the model on the training data are shown in Figure 3. As can be seen, on the validation dataset, the validation accuracy started to flatten after 3000 rounds, and the MAE finally reached 9.1, meaning that the mean absolute error of the model is 9.1 μ g / m3, Which shows that the model has a relatively accurate accuracy in the validation data. In this study, the PM2.5 concentration of PM2.5 in each quarter was obtained by entering the site location in the four quarters, and the PM2.5 observation obtained the scatter plot of PM 2.5 concentration in the four quarters (Figure 4) and its statistical indicators (as shown in Table 2. As can be seen from the scatter chart, the PM2.5 level in the whole year of 2020 is relatively low, with the third quarter being the lowest, which is mostly concentrated at 100 μ g / m3 below, PM2.5 concentration did not differ much except in other quarters.

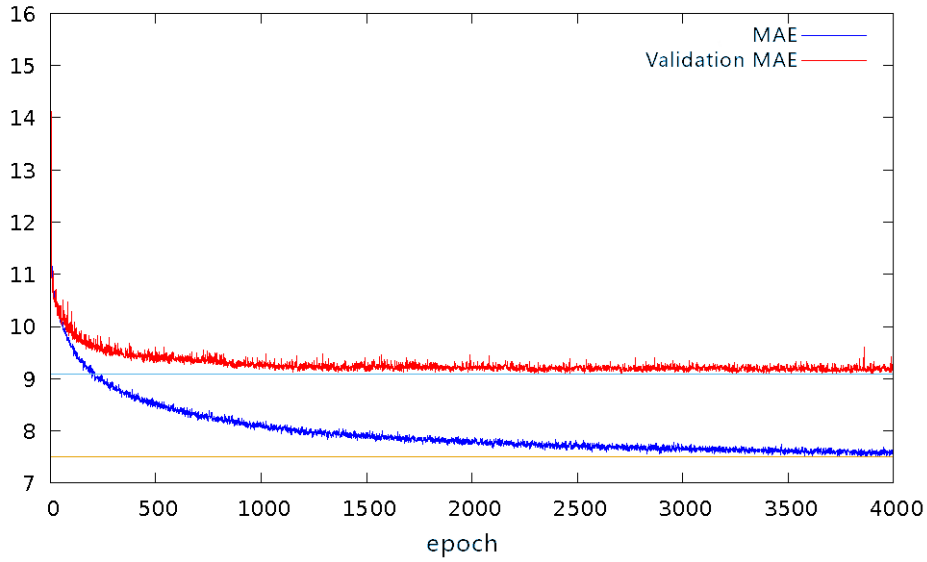

Figure 3 Validation metrics for each round of the model during training

Table 2 Model Verification Table for the four quarters of 2020

season	data size	R^2	MAE(μg/m3)
first quarter	21368	0.67	11.3
The second quarter	20369	0.72	7.94
The third quarter	11633	0.56	6.11
The fourth quarter	24195	0.72	9.90

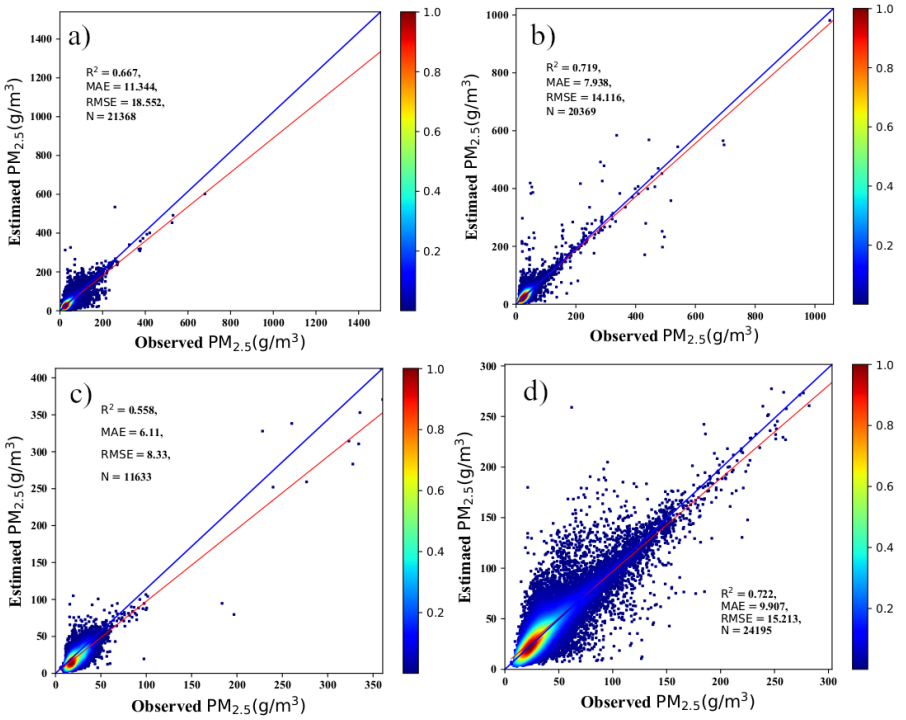

Figure 4 Site PM2.5 scatter density map in 2020 a) first quarter, b) second quarter, c) third quarter and d) fourth quarter

4.2. Model application

In order to reflect the significance of this study, the spatial distribution map of PM2.5 mean concentration in four quarters in most parts of the country was obtained by the meteorological data obtained by satellite and estimated by the model. As shown in Figure 6. Of the Qinghai-tibet plateau due to AOD thin, AOD data in the region quality is not reliable, and in the second chapter data processing by quality control module, and according to figure 1, the Qinghai-tibet plateau air quality monitoring station, even with AOD data will be excluded because of no site data, so the Qinghai-tibet plateau area data is not reliable. As can be seen from the figure, in 2020, the overall air quality preference of other regions except for the overall average value of the Taklimakan Desert region in Xinjiang is relatively high. Especially in the third quarter, the national PM2.5 concentration was at relatively low levels. Since the fourth quarter, the national PM2.5 situation began to rise, but most of them were within a reasonable range. It can be seen that under the influence of the epidemic, in the second and third quarter, PM2.5 pollution in most parts of China eased in most parts of China compared with before and after the epidemic.

In order to reflect the above nationwide estimation accuracy, the PM2.5 observed value of each site as the true value and the estimated variable as the input value obtained the PM2.5 concentration of the site location. After calculation R^2, the estimation accuracy of the site location of the four quarters was obtained, as shown in Figure 5. It can be seen that in most areas, the estimation accuracy has a high level, especially in the first quarter of the central region R^2 on average around 0.7. In the third quarter, due to the overall low PM2.5 situation, there had some impact on the accuracy of the model.

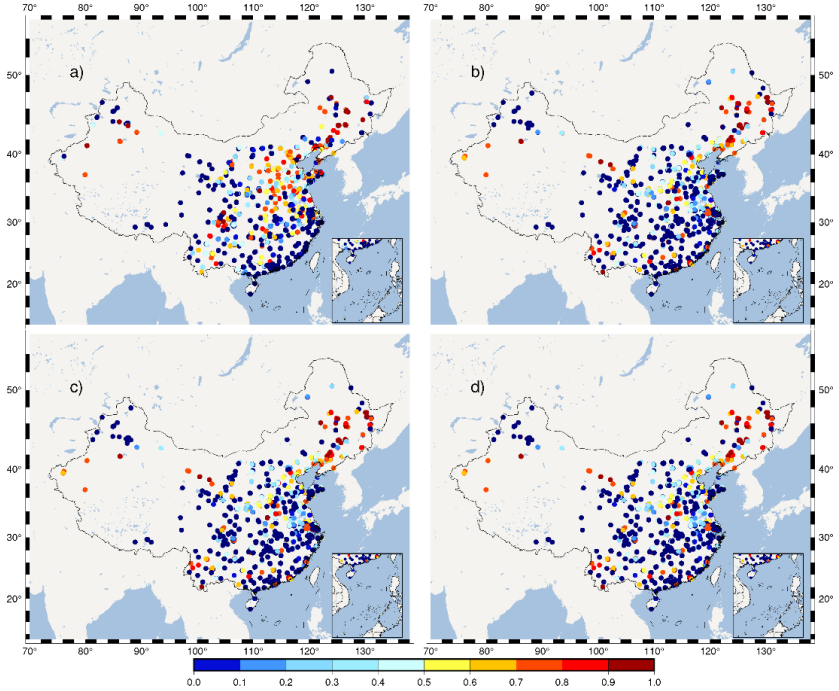

Figure 5 PM2.5 concentration at some sites in 2020 a) Q 1, b) Q 2, c) Q 3 and d) Q 4 Estimate performance R^2

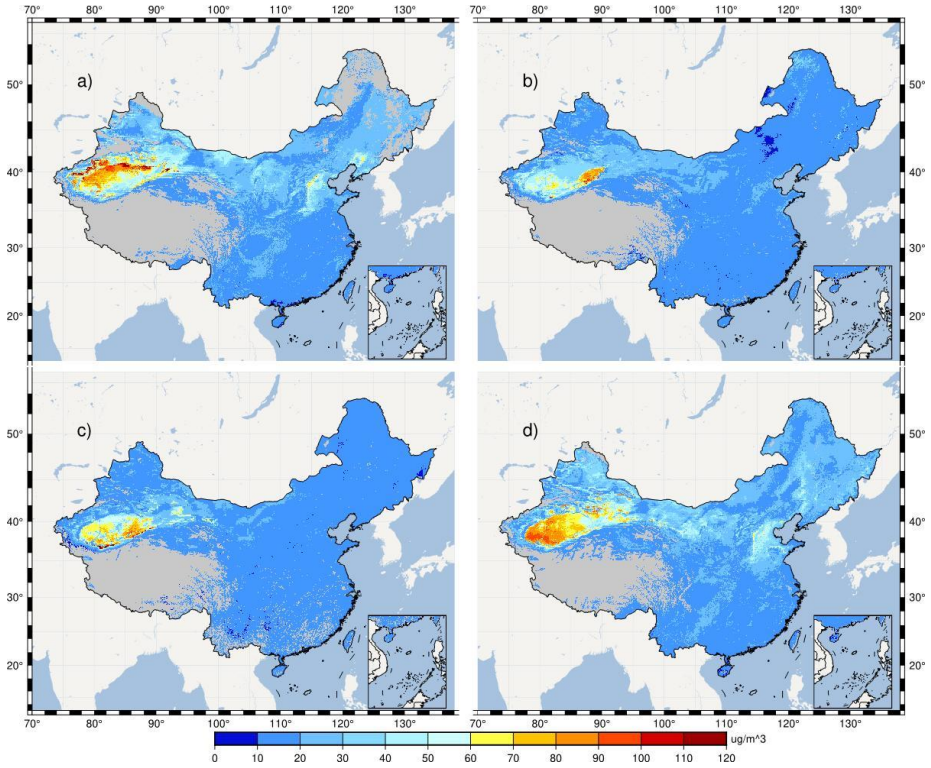

Figure 6 2020), first quarter a), second quarter b), third quarter c) and fourth quarter d) PM2.5 mean concentration distribution

5. CONCLUSIONS

In recent years, with the deep learning in the field of computer vision and data mining, represented by deep learning of a new batch of modeling means applied to science central stage, and applied to including big data, wisdom city, human-computer interaction, many fields, and whether it is also difficult to grasp the law of weather and environment still have strong advantages, is the current environmental science of major career.

The present study is a small exploration of them. Through research, proved that the deep neural network in PM2.5 and AOD and other influence factors, have a powerful ability, use eight months of a year data, through 20 minutes of iterative training, can achieve single digit error on the training data, and will increase with the number of iterations and become more robust. After calculation, the model has an average coefficient of determination of 0.6 and an average absolute error of 7.8 on the data throughout the year. Combined with the impact of the 2020 outbreak, the model of this study can also better reflect the air pollution comparison before and after the outbreak. On different quarterly scales, the best performance in the second and fourth quarters, R^2 both above 0.7. Moreover, it can be seen that the response of the model on the large value is especially sensitive and can accurately grasp the large value.

However, this model only makes a very preliminary attempt to estimate PM2.5 by using the dense connection network, and outputs the exploration stage of deep learning, so the current estimation level of PM2.5 concentration is not perfect. In the field of deep learning, there are some such as one-dimensional convolutional neural network, recurrent neural network (recurrent neural network, RNN) principle more complex network architecture, they are also commonly used in time series data modeling (such as temperature prediction, stock prices, house prices, etc.), these models have more and more complex parameters, to further improve the concentration of PM2.5 estimation accuracy has more positive significance. In the future, we will focus on these two networks to further improve

the precision of PM2.5 concentration estimation, so as to provide guidance for the goal of "carbon peak" and "carbon neutral".

REFERENCE

[1] Ministry of Environmental Protection, State Administration. GB 3095- -2012 Ambient Air Quality Standard [S]. China Environmental Science Press, 2012.

[2] Yang Lijuan, Xu Hanqiu, Jin Zhifan. MODIS Satellite remote sensing estimates the PM_ (2.5) concentration in Fuzhou [J]. Journal of Remote Sensing, 2018,22 (01): 64-75.

[3] GUPTA P, CHRISTOPHER S A.Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach[J].Journal of Geophysical Research: Atmospheres, 2009,114(D14).

[4] MA Z, HU X, HUANG L, et al.Estimating ground-level PM2. 5 in China using satellite remote sensing[J].Environmental science \& technology, 2014,48(13): 7436-7444.

[5] Shen Yuan, Chen Chaoliang, Qian Jing, etc. High-resolution PM_ (2.5) remote sensing inversion based on random forest —— Take Guangdong Province as an example [J]. Integration Technology, 2018,7 (03): 31-41.

[6] CHEN Z, ZHANG T, ZHANG R, et al.Extreme gradient boosting model to estimate PM2. 5 concentrations with missing-filled satellite data in China[J].Atmospheric Environment, 2019,202: 180-189.

[7] WEI J, LI Z, CRIBB M, et al. Improved 1 km resolution PM 2.5 estimates across China using enhanced space--time extremely randomized trees[J]. Atmospheric Chemistry and Physics, 2020,20(6): 3273-3289.

[8] LYAPUSTIN A, WANG Y, LASZLO I, et al.Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm[J].Journal of Geophysical Research: Atmospheres, 2011,116(D3).

[9] LYAPUSTIN A, WANG Y, KORKIN S, et al.MODIS collection 6 MAIAC algorithm[J].Atmospheric Measurement Techniques, 2018,11(10): 5741-5765.

[10] JARVIS A, REUTER H I, NELSON A, et al.Hole-filled SRTM for the globe Version 4[J].available from the CGIAR-CSI SRTM 90m Database (http://srtm.csi.cgiar.org), 2008,15: 25-54.

[11] Tang Xinming, Li Shijin, Li Tao, etc. Global digital elevation product overview [J]. Journal of Remote Sensing, 2021,25 (01): 167-181.