PM2.5 concentration forecasting using Long Short-Term Memory Neural Network and Multi-Level Additive Model

Lixin Tao

Capital Medical University

Feifei Tian

Capital Medical University

Lijuan Wu

Capital Medical University

Mengyang Liu

Capital Medical University

Yuan Ma

Capital Medical University

Zhiyuan Wu

Capital Medical University

Haibin Li

Capital Medical University

Xiuhua Guo (✉ statguo@ccmu.edu.cn)

Capital Medical University <https://orcid.org/0000-0001-6657-6940>

Research article

Keywords: PM2.5, Long-short term memory neural network, Time series, Pollutant forecasting, Additive model

Posted Date: October 3rd, 2019

DOI: <https://doi.org/10.21203/rs.2.15521/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background PM 2.5 concentration predication can provide an effective way to protect public health by early warning. Though there are many methods available, the comparison between multi-level additive model (AM) and long short-term memory (LSTM) neural network in predicting PM 2.5 concentration is limited. This study aimed to compare the performance of multi-level AM and LSTM in predicting hourly and daily PM 2.5 concentration.

Methods Air pollution data from Jul 1, 2016 to Dec 31, 2017 were obtained from Beijing Municipal Environmental Monitoring Center, and meteorological data were derived from the National Meteorological Science Data Sharing Service. Multi-level AM and LSTM were developed to estimate the regional hourly and daily concentration of PM 2.5 .

Results In the prediction of hourly PM 2.5 concentrations, LSTM achieved a better performance than multi-level AM (range of R^2 : 0.76-0.92 for LSTM, 0.59-0.78 for multi-level AM; range of root mean square error (RMSE): 6.20-17.58 $\mu\text{g}/\text{m}^3$ for LSTM, 19.19-30.81 $\mu\text{g}/\text{m}^3$ for multi-level AM; range of mean absolute error (MAE): 4.50-13.42 $\mu\text{g}/\text{m}^3$ for LSTM, 13.55-22.35 $\mu\text{g}/\text{m}^3$ for multi-level AM; range of mean absolute percentage error (MAPE): 0.18%-0.55% for LSTM, 0.50%-0.87% for multi-level AM). While in the prediction of daily PM 2.5 concentrations, multi-level AM showed a higher predictive accuracy than LSTM (range of R^2 : 0.43-0.93 for LSTM, 0.74-0.98 for multi-level AM; range of RMSE: 32.46-46.82 $\mu\text{g}/\text{m}^3$ for LSTM, 4.83-20.98 $\mu\text{g}/\text{m}^3$ for multi-level AM; range of MAE: 24.32-34.89 $\mu\text{g}/\text{m}^3$ for LSTM, 3.67-16.33 $\mu\text{g}/\text{m}^3$ for multi-level AM; range of MAPE: 0.92%-1.74% for LSTM, 0.11%-0.45% for multi-level AM).

Conclusion LSTM showed better performance than the multi-level AM when there is a large amount of data, while multi-level AM showed better performance than LSTM when the amount of data is relatively small.

Background

Fine particulate matter (particulate matter with an aerodynamic diameter less than or equal to 2.5 μm , $\text{PM}_{2.5}$) can increase the risk of occurrence and death of multiple respiratory [1, 2], circulatory systemic diseases [3, 4] and can increase the morbidity and mortality of tumors [5, 6]. In addition, heavy metals carried by $\text{PM}_{2.5}$ can accumulate in human body and cause chronic hazards. Microorganisms adhered to $\text{PM}_{2.5}$ can cause sensitization, which do great harm to human body [7]. The revised new standards for ambient air quality in 2012 have added $\text{PM}_{2.5}$ to the air pollution monitoring project. The monitoring data of air pollutants published by the Beijing Municipal Environmental Protection Bureau showed that the average annual concentration of $\text{PM}_{2.5}$ in Beijing from 2013 to 2017 were 89.5 $\mu\text{g}/\text{m}^3$, 85.9 $\mu\text{g}/\text{m}^3$, 80.6 $\mu\text{g}/\text{m}^3$, 73 $\mu\text{g}/\text{m}^3$, and 58 $\mu\text{g}/\text{m}^3$ respectively. Although the annual mean concentration of $\text{PM}_{2.5}$ showed a decline trend, the decline was small, and it is much higher than the ambient air quality secondary standard, which is annual mean concentration of 35 $\mu\text{g}/\text{m}^3$. In addition, the annual mean concentration limit for $\text{PM}_{2.5}$ in the European Union is 25 $\mu\text{g}/\text{m}^3$, and in the United States is 15 $\mu\text{g}/\text{m}^3$. It

can be seen that PM_{2.5} pollution in Beijing cannot be ignored. Therefore, the accurate PM_{2.5} concentration prediction is very important for controlling air pollution and preventing health hazards caused by it.

Present PM_{2.5} predication methods could be divided into two categories generally. One is deterministic methods [8, 9], in which Community Multiscale Air Quality (CMAQ) model [10] and the latest WRF-Chem model [11] are currently used internationally, while Nested Air Quality Prediction Modeling System (NAQPMS) [12] is commonly used in China. The mechanism of deterministic methods is clear, but complex prior knowledge is needed, and there are various application restrictions. Such as the forecasting process should take a long time, and the potential effects of the associated factors of PM_{2.5} is not fully considered [13, 14]. The other is statistical methods, which could avoid sophisticated theoretical models and apply statistical-based models simply to predict the concentration of air pollutants rapidly, because of no need to describe the physical and chemical processes of pollutants. Statistical methods can predict the concentration of air pollutants including PM_{2.5} by analyzing air quality related data and have received extensive attention from scholars. Statistical methods include traditional multiple linear regression (MLR) [15] models, autoregressive moving average (ARIMA) [16] models, land use regression (LUR) [17] models, generalized additive models (GAM) [18, 19], support vector regression (SVR) models [20], artificial neural network (ANN) models [21] in machine learning, recurrent neural network (RNN) [22, 23], convolutional neural networks (CNN) [24, 25] and long-term memory neural network (LSTM) [26, 27] in deep learning [28]. In these models, LSTM have nearly all the advantages of ANN and RNN, such as the ability of performing nonlinear mapping, the characteristic of adaptability and robustness, high performance in the field of temporal series predication. In addition, LSTM [29] neural networks have the ability of studying long temporal series, at the same time it will not be affected by the problem of gradient vanishing. These features are important in PM_{2.5} concentration predication, because PM_{2.5} concentration is related to its previous concentration. Li et al [30] extended the LSTM model to predict the PM_{2.5} hourly concentration basing on data from 12 air quality monitoring stations from January 2014 to May 2016. In PM_{2.5} concentration predication, GAM can identify nonlinear relationship and interaction of the associated factors with PM_{2.5}, and is suitable for the analysis of temporal series including PM_{2.5} concentration predication. The results of the both two studies showed that GAM has a significant improvement in prediction efficiency compared to simple linear regression models [31, 32]. However, in most studies the collinearity has not been well solved, and it needs to be improved to choose the degree of freedom of each variable in GAM. Since LSTM and AM both have several advantages, and the comparison between them is limited, this study aimed to compare performance of multi-level AM and LSTM in hourly and daily PM_{2.5} concentration predication based on data from 16 districts of Beijing from July 2016 to December 2017.

Methods

Air Pollution and meteorological data

The air pollution data including PM_{2.5}, CO, NO₂, O₃, and SO₂ were obtained from Beijing Municipal Environmental Monitoring Center (<http://www.bjmemc.com.cn/>). This study collected data on hourly and daily mean concentrations of PM_{2.5}, CO, NO₂, O₃, SO₂ recorded in 35 monitoring stations which covered 16 districts of Beijing from Jul 1, 2016 to Dec 31, 2017.

Meteorological data were obtained from the National Meteorological Science Data Sharing Service (<http://data.cma.cn/>), which include air pressure, sea level pressure, maximum pressure, minimum pressure, maximum wind speed, instant maximum wind speed, 10-minute mean wind speed, wind direction, temperature, maximum temperature, minimum temperature, relative humidity, minimum relative humidity, vapor pressure and precipitation, a total of 15 variables. This study collected meteorological hourly and daily mean concentration of 16 districts in Beijing from Jul 1, 2016 to Dec 31, 2017.

Statistical Analysis

The median and inter-quartile range was used to describe the distribution of variables which were not normally distributed. Spearman rank correlation analysis was used to assess the association between meteorological factors and PM_{2.5}.

The general form of multi-level AM can be specified as [33]:

$$Y = \sum gX + \alpha \quad [1]$$

Multi-level AM in this study was modified as follows:

$$Y = \sum \beta X + \sum sX + \alpha \quad [2]$$

Here Y refers to dependent variable, X refers to independent variables, g(X) denotes the fitting function, $\beta(X)$ denotes linear function, s(X) denotes smoothing function, and α is intercept term.

In this study, PM_{2.5} concentration with a certain time lag was selected as dependent variable, the date was used as a time series to control the time confounding factors, meanwhile, season, rainfall, holidays, weekends were used as categorical variables. And some selected meteorological factors enter into multi-level AM to solve the multi-collinearity between the associated factors and PM_{2.5}. Penalty cubic spline smoothing function was selected to perform the non-parametric fitting in gaseous pollutants and most meteorological factors, while thin plate spline smoothing function was selected for wind speed and direction because of interaction. Considering time as level 1 and districts and counties as level 2, a multi-level AM was established including weekends, holidays and season to choose variables for multi-level AM. Degrees of freedom of variables were selected based on the partial autocorrelation function.

LSTM model is composed of an input layer, an output layer, and a series of cyclically connected hidden layers, namely memory blocks, each block consists of some self-recurrent memory cells and three multiplicative units (input gates, output gates, and forget gates) which provide functions to read, write, and reset continuously for the cells. Self-recurrent memory cells can block any external disturbance, so the state can remain unchanged from one step to the next, enabling LSTM to solve the problem of gradient vanishing. Forget gates can reset the memory block when the state is out of date, meanwhile, prevent the gradient from exploding. In addition, the input gate allows the incoming signal to modify the state of the cells, and the output gate allows or prevents the state of cells from affecting other cells. The basic structure of LSTM is shown in S1 Fig.

Hourly or daily PM_{2.5} concentration was used as dependent variable, and the independent variables, namely, input characteristics, included 4 gaseous pollutants, 15 meteorological factors, and 5 time variables. The number of LSTM layers, fully connected layers, cells in each layer, batch size and epochs were adjusted according to training and testing losses. The neural networks were disconnected with a probability of 0.01 to avoid over-fitting. The loss function was mae, the optimizer was adam, and the time step was set to 1, which indicated that PM_{2.5} concentration at the next time point (next hour or day) was predicted based on historical data.

Multi-level AM and LSTM were established based on the data of each district and county in Beijing to predict hourly and daily PM_{2.5} concentration respectively. The first two-thirds of the data was used as training set, while the latter one-third was used as testing set. The efficiency of the two types of models was evaluated with determination coefficient (R^2), root mean square error (RMSE), mean absolute error (MAE) and the mean absolute percentage error (MAPE).

The descriptive statistical analysis of the pollutant and meteorological data was performed using Arcgis10.2 and R3.4.3 software. The fitting of multi-level AM was performed using R3.4.3, and the Python 3.6 software was used to fit LSTM.

Results

The distribution of air pollutants and meteorological factors were described in S1 Table. PM_{2.5} was 45.0 (71.0) $\mu\text{g}/\text{m}^3$ with a range of 2.0 and 1004.0 $\mu\text{g}/\text{m}^3$, CO was 0.8 (0.8) mg/m^3 with a range of 0.1 and 16.7 mg/m^3 , NO₂ was 41.0 (47.0) $\mu\text{g}/\text{m}^3$ with a range of 1.0 and 300.0 $\mu\text{g}/\text{m}^3$, O₃ was 40.0 (69.0) $\mu\text{g}/\text{m}^3$ with a range of 1.0 and 504.0 $\mu\text{g}/\text{m}^3$, and SO₂ was 4.0 (8.0) $\mu\text{g}/\text{m}^3$, the range was from 1.0 to 307.0 $\mu\text{g}/\text{m}^3$.

The spatial distribution of PM_{2.5} was shown in S2 Fig, it can be observed that PM_{2.5} concentration varied in districts and counties, meanwhile gradually decreased from south to north. So we predicted PM_{2.5} concentration based on the data in every district and county (16 in total) of Beijing.

The time series distribution of daily PM_{2.5} concentration in the districts and counties with the highest (Fangshan) and lowest (Miyun) during the study period are shown in S3 Fig. Of the total 529 days, daily PM_{2.5} concentration levels exceed the second standard in China on 191 and 115 days. Meanwhile, daily PM_{2.5} concentration levels seem to fluctuating randomly, whereas, there was a decline trend during the study period.

The association between PM_{2.5} and meteorological factors were shown in S2 Table. All the correlation coefficients were statistically significant. If spearman rank correlation coefficients (r_s) among several correlated meteorological factors were above 0.60, the one which has the highest spearman rank correlation coefficient with PM_{2.5} should be selected. Finally, Minimum relative humidity ($r_s = 0.29$, $P < 0.001$), sea level pressure ($r_s = -0.09$, $P < 0.001$), maximum wind speed ($r_s = -0.16$, $P < 0.001$), temperature ($r_s = 0.01$, $P < 0.001$) were under consideration to be involved in the construction of multi-level AM.

The result of multi-level AM was shown in Table 1, it demonstrated that every variable is statistically correlated with PM_{2.5} ($\beta = -15.05-33.73$, $P < 0.05$). Finally, minimum relative humidity, sea level pressure, maximum wind speed and wind direction, temperature, rainfall, CO, NO₂, SO₂, and O₃ were selected to be involved in the construction of multi-level AM.

Table 1. Associated factors with PM_{2.5}

variables	β	SE	t	P	variables	β	SE	t	P
warm season	-12.22	3.84	-3.19	0.001	CO	33.73	0.69	48.80	<0.001
weekend	-7.72	1.28	-6.03	<0.001	NO ₂	1.36	0.03	50.18	<0.001
holiday	16.75	2.47	6.79	<0.001	SO ₂	2.43	0.08	28.72	<0.001
rainfall	-12.69	1.35	-9.37	<0.001	O ₂	-0.07	0.03	-2.57	0.010
wind direction (north as control)					min relative humidity	1.43	0.04	37.72	<0.001
east	28.10	1.75	16.10	<0.001	sea level pressure	-0.79	0.12	-6.78	<0.001
south	31.07	1.56	19.96	<0.001	temperature	0.37	0.17	2.22	0.026
west	8.08	1.78	4.53	<0.001	wind speed	-15.05	0.36	-42.33	<0.001

Note: All estimates are from multi-level additive model.

In multi-level AM, the degrees of freedom of the independent variables are determined one by one according to the principle of minimizing the partial autocorrelation function (pacf). One example was

given in S4 Fig, when the k was 20, the pacf was minimum, and then the degree of freedom of this variable in this model could be determined as 20.

In LSTM, when the training and testing loss tend to be stable, it indicated that the model trained well (S5 Fig), the epoch, batch size, number of cells, number of network layers were adjusted to optimize models, and the final LSTM models performed well with 1 LSTM layer with 20 cells, a fully connected layer with 1 cell, the number of epoch is 40, and batch size is 20.

Hourly PM_{2.5} concentration for the next hour was predicted based on the current hourly data of meteorological factors and air pollutants, and the prediction results are shown in Table 2. The R² of LSTM is in the range of 0.70~0.92, which is generally higher than multi-level AM (0.59~0.80), RMSE is among 6.20 and 17.58µg/m³, which is lower than multi-level AM (19.19~30.81µg/m³), MAE varies from 4.50 to 13.42µg/m³, which is lower than multi-level AM (13.55~22.35µg/m³), MAPE is in the range of 0.18%~0.55%, which is lower than multi-level AM (0.50%~0.87%). The results suggested that LSTM performs better in hourly PM_{2.5} concentration predication than multi-level AM. The comparison of observed and predicted PM_{2.5} hourly concentration based on multi-level AM and LSTM in Daxing (best predication result in LSTM) and Fangshan (worst predication result in LSTM) is shown in Fig 1. It can be seen that the predicated hourly PM_{2.5} concentration using LSTM are more consistent with the observed than multi-level AM, which could also suggest that LSTM performs better than multi-level AM in hourly PM_{2.5} concentration predication.

Table 2. Comparison of AM and LSTM on hourly PM_{2.5} concentration predication.

district and county	AM				LSTM			
	R ²	RMSE	MAE	MAPE	R ²	RMSE	MAE	MAPE
Dongcheng	0.78	20.88	15.56	0.55	0.86	7.36	5.86	0.24
Xicheng	0.75	22.41	16.21	0.54	0.88	6.20	5.10	0.27
Chaoyang	0.67	26.37	19.19	0.65	0.78	9.99	7.40	0.21
Haidian	0.70	22.18	15.99	0.62	0.77	9.62	6.97	0.20
Fengtai	0.59	27.92	20.26	0.66	0.85	6.71	5.08	0.18
Shijingshan	0.74	22.65	15.66	0.60	0.78	11.62	9.58	0.48
Fangshan	0.61	29.34	21.33	0.64	0.70	17.58	13.42	0.31
Daxing	0.76	21.39	15.40	0.57	0.89	5.72	4.50	0.18
Tongzhou	0.59	30.81	22.35	0.73	0.83	7.47	5.47	0.18
Shunyi	0.80	20.44	14.15	0.67	0.86	8.04	6.22	0.39
Changping	0.71	21.68	15.46	0.72	0.89	7.54	6.77	0.55
Mentougou	0.65	26.31	17.62	0.80	0.81	9.65	7.61	0.47
Pinggu	0.63	29.12	17.87	0.70	0.85	7.39	5.87	0.34
Huairou	0.65	25.66	16.98	0.87	0.81	10.61	7.46	0.39
Miyun	0.71	20.93	14.45	0.70	0.92	9.15	7.04	0.39
Yanqing	0.74	19.19	13.55	0.50	0.76	10.86	8.94	0.40

Note: Predicating efficiency of AM and LSTM on hourly PM_{2.5} concentration.

Daily PM_{2.5} concentration for the next day was predicted based on the current data of meteorological factors and air pollutants, and the prediction results are shown in Table 3. The R² of LSTM is in the range of 0.43~0.93, which is generally lower than multi-level AM (0.67~0.98), RMSE is among 32.46 and 46.82µg/m³, which is higher than multi-level AM (4.83~20.98µg/m³), MAE varies from 24.32 to 34.89µg/m³, which is higher than multi-level AM (3.67~16.33µg/m³), MAPE is in the range of 0.92%~1.74%, which is higher than multi-level AM (0.11%~0.45%). The results suggested that multi-level AM performs better in predicting PM_{2.5} daily concentration than LSTM. The comparison of observed and predicted PM_{2.5} daily concentration based on multi-level AM and LSTM in Fengtai (best predication result in multi-level AM) and Fangshan (worst predication result in multi-level AM) is shown in Fig 2. It can be seen that the predicated PM_{2.5} hourly concentration plots of multi-level AM are more consistent with the

observed than LSTM, which could also suggest that multi-level AM performs better than LSTM in predication of daily concentration of PM_{2.5}.

Table 3. Comparison of AM and LSTM on daily PM_{2.5} mean concentration predication.

district and county	AM				LSTM			
	R ²	RMSE	MAE	MAPE	R ²	RMSE	MAE	MAPE
Dongcheng	0.97	6.60	5.21	0.17	0.70	44.66	33.69	1.34
Xicheng	0.98	5.78	4.72	0.15	0.83	46.03	34.89	1.45
Chaoyang	0.94	8.95	6.87	0.20	0.86	45.23	33.82	1.33
Haidian	0.86	12.34	9.66	0.32	0.64	37.31	28.71	1.19
Fengtai	0.98	4.83	3.84	0.11	0.93	46.82	33.56	1.26
Shijingshan	0.98	5.54	4.36	0.17	0.73	40.12	32.12	1.35
Fangshan	0.67	20.98	16.33	0.38	0.93	44.09	33.25	0.98
Daxing	0.97	6.45	5.10	0.15	0.68	45.61	32.97	1.44
Tongzhou	0.95	8.38	6.46	0.18	0.63	42.96	31.45	0.92
Shunyi	0.97	6.86	5.20	0.20	0.65	42.86	33.08	1.67
Changping	0.97	5.84	4.61	0.20	0.71	38.46	29.65	1.65
Mentougou	0.92	9.53	7.44	0.28	0.79	41.41	32.07	1.74
Pinggu	0.98	4.83	3.67	0.13	0.50	41.72	31.89	1.21
Huairou	0.79	15.16	11.34	0.44	0.59	36.31	27.55	1.46
Miyun	0.95	6.83	5.53	0.22	0.43	32.46	24.32	1.14
Yanqing	0.74	15.79	12.80	0.45	0.62	33.20	26.79	1.16

Note: Predicating efficiency of AM and LSTM on daily PM_{2.5} concentration

Discussion

LSTM and multi-level AM in this study were used to predict hourly and daily PM_{2.5} concentration. The results showed that LSTM performed better in PM_{2.5} hourly concentration predication, while multi-level AM performed better in PM_{2.5} daily mean concentration predication. It was indicated that LSTM is suitable for the predication of long time series, while multi-level AM is more suitable for shorter time series.

The PM_{2.5} concentration prediction model based on statistical methods has become a research trend due to its simplicity and speed. Linear regression can only identify linear relationships, while GAM can identify nonlinear connections and reduce bias, meanwhile GAM based PM_{2.5} concentration prediction model can generally consider more factors than linear regression such as time. Multi-level AM in this study had better performance (R^2 , 0.67–0.98; RMSE, 4.83–20.98 $\mu\text{g}/\text{m}^3$) for daily PM_{2.5} concentration predication than linear regression (R^2 , 0.88; RMSE, 23.42 $\mu\text{g}/\text{m}^3$) in the study of Zhao et al [34]. In recent years, with the emphasis on environmental issues in China, the increase in environmental monitoring inputs, the monitoring data of atmospheric pollutants and meteorological factors have been accumulated for a long time. In the context of environmental big data, deep learning can be utilized in data with large quantities and wide sources to learn the correlation between PM_{2.5} and influencing factors, finally improve the accuracy of PM_{2.5} concentration prediction. Moreover, the deep learning model, such as the cyclic neural network, has strong scalability, and other methods can be integrated into the deep learning model to avoid the defects of the single model. LSTM have better prediction accuracy and prediction efficiency in existing models. Li et al [30] extended LSTM to predict PM_{2.5} concentration with meteorological factors such as temperature, humidity, wind speed and visibility, R^2 reached 98%, and RMSE, MAE, MAPE were significantly lower than SVR, ARMA, and TDNN, indicating that LSTM has a higher predictive power. Zhao et al [35] indicated that the MAE and RMSE of LSTM model for 1–6 hours PM_{2.5} concentration predication were 25.05 $\mu\text{g}/\text{m}^3$ and 39.23 $\mu\text{g}/\text{m}^3$, for 25–48 hours were 51.80 $\mu\text{g}/\text{m}^3$ and 73.69 $\mu\text{g}/\text{m}^3$, generally higher than MAE and RMSE of LSTM model in this study (4.50–13.42 $\mu\text{g}/\text{m}^3$ and 6.20–17.58 $\mu\text{g}/\text{m}^3$ for hourly PM_{2.5} concentration predication; 24.32–34.89 $\mu\text{g}/\text{m}^3$ and 32.46–46.82 $\mu\text{g}/\text{m}^3$ for daily PM_{2.5} concentration predication).

LSTM is suitable for predication basing on long temporal series, previous PM_{2.5} concentration predictions based on neural network mostly used PM_{2.5} hourly concentration. While AM is generally used to study the influence factors of PM_{2.5} [36], and the research data is the daily concentration, that is, the data with shorter time series. In this study, in the hourly PM_{2.5} concentration prediction, the time steps were long, LSTM performed better than multi-level AM, consistent with previous study [37] which indicted that deep learning methods have less error (MAE) than the non-deep learning methods in 1st hour predication. While in the PM_{2.5} daily concentration prediction, the time steps were short, multi-level AM performed better than LSTM. LSTM is a type of deep learning and require a large number of training set to obtain relatively stable models. And in this study, there were a relatively large amount of data in the PM_{2.5} hourly concentration prediction, while the PM_{2.5} daily concentration prediction has a relatively small amount of data, whereas, the results in this study need to be confirmed by further research. In addition, it needs less time to construct LSTM than multi-level AM, because there is only several parameters need to be determined in LSTM, while the amount of parameters in multi-level AM is consistent with the amount of variables. So it can be concluded that LSTM is suitable for longer time series prediction, and multi-level AM is better to predict shorter time series data. However, the sample size of 529 days in daily PM_{2.5} concentration in this study is not large and the time series is not long enough. Therefore, a longer and

more complete time series should be used to predict the $PM_{2.5}$ daily concentration based on multi-level AM and LSTM, and verify the conclusions of this study. In addition, multi-level AM needs a complex variables selection, while LSTM can study the input information automatically.

In this study, it can be indicated that the concentration of CO , NO_2 , SO_2 is positively correlated with $PM_{2.5}$, while O_3 is negatively correlated with $PM_{2.5}$, there might be the reason that CO , NO_2 , SO_2 could form the ingredients of $PM_{2.5}$, and the generation process of O_3 could be weakened by $PM_{2.5}$. The minimum relative humidity and temperature were positively correlative with $PM_{2.5}$, while sea level pressure, wind speed and rainfall were negatively correlated with $PM_{2.5}$.

The limitation is that, in the predication of daily $PM_{2.5}$ concentration, the sample size was small and time series was not long enough. However, this study conducted LSTM and multi-level AM for hourly and daily $PM_{2.5}$ concentration predication at the same time. In view of the difference of hourly and daily $PM_{2.5}$ concentration, daily $PM_{2.5}$ concentration should be predicted in longer time series basing on LSTM and multi-level AM to verify the conclusions of this study.

Conclusion

It can be concluded that LSTM performed better than the multi-level AM when there is a large amount of data, while multi-level AM performed better than LSTM when the amount of data is relatively small.

List Of Abbreviations

$PM_{2.5}$: Fine Particulate Matter

LSTM: Long Short-Term Memory

AM: Additive Model

MLR: Multiple Linear Regression

ARIMA: Autoregressive Moving Average

LUR: Land Use Regression

GAM: Generalized Additive Models

SVR: Support Vector Regression

ANN: Artificial Neural Network

RNN: Recurrent Neural Network

CNN: Convolutional Neural Networks

R²: Determination Coefficient

RMSE: Root Mean Square Error

MAE: Mean Absolute Error

MAPE: Mean Absolute Percentage Error

R_s: Spearman Rank Correlation Coefficients

Declarations

Ethics approval and consent to participate (Not applicable)

Consent for publication (Not applicable)

Availability of data and material

The datasets generated and/or analysed during the current study are available in the Beijing Municipal Environmental Monitoring Center [<http://www.bjmemc.com.cn/>] and the National Meteorological Science Data Sharing Service [<http://data.cma.cn/>].

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by National Natural Science Foundation of China (81872708, 81502886, and 81530087) and Natural Science Foundation of Beijing Municipality (Z160002).

There is no role of the funding body in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript should be declared.

Authors' contributions

LT, FT and LW analyzed the data and were major contributors in writing the manuscript. ML, YM, ZW and HL performed the data collection and organization. XG performed the modification of this manuscript. All authors read and approved the final manuscript.

Acknowledgements (Not applicable)

Authors' information (Not applicable)

References

1. Polezer G, Tadano YS, Siqueira HV, Godoi A, Yamamoto CI, de Andre PA. Assessing the impact of PM_{2.5} on respiratory disease using artificial neural networks. *Environmental Pollution*. 2018; 235: 394–403.
2. Chien LC, Chen YA, Yu HL. Lagged influence of fine particulate matter and geographic disparities on clinic visits for children's asthma in Taiwan. *International journal of environmental research and public health*. 2018; 15, PMID: 29690596, <https://doi.org/10.3390/ijerph15040829>.
3. Zeng X, Liu J, Du X, Zhang J, Pan K, Shan W, et al. The protective effects of selenium supplementation on ambient PM_{2.5}-induced cardiovascular injury in rats. *Environ Sci Pollut Res Int*. 2018; 25: 22153–22162.
4. Wang H, Shen X, Tian G, Shi X, Huang W, Wu Y, et al. Ampk α 2 deficiency exacerbates long-term pm_{2.5} exposure-induced lung injury and cardiac dysfunction. *Free Radical Biology and Medicine*. 2018; 121:202–214, <https://doi.org/10.1016/j.freeradbiomed.2018.05.008>.
5. Li R, Zhou R, Zhang J. Function of pm_{2.5} in the pathogenesis of lung cancer and chronic airway inflammatory diseases. *Oncology letters*. 2018; 15:7506–7514, PMID: 29725457, <https://doi.org/10.3892/ol.2018.8355>.
6. Gao Y, Yang L, Chen J, Li Y, Jiang P, Zhang J, et al. Nitro and oxy-pahs bounded in pm_{2.5} and pm_{1.0} under different weather conditions at mount tai in eastern china: Sources, long-distance transport, and cancer risk assessment. *Science of The Total Environment*. 2018; 622–623:1400–1407, <https://doi.org/10.1016/j.scitotenv.2017.11.200>.
7. Ouyang Y, Xu Z, Fan E, Li Y, Miyake K, Xu X, et al. Changes in gene expression in chronic allergy mouse model exposed to natural environmental pm_{2.5}-rich ambient air pollution. *Scientific reports*. 2018; 8:6326, PMID: 29679058, <https://doi.org/10.1038/s41598-018-24831-z>.
8. Kim Y, Fu JS, Miller TL. Improving ozone modeling in complex terrain at a fine grid resolution: Part i—examination of analysis nudging and all pbl schemes associated with Isms in meteorological model. *Atmospheric Environment*. 2010; 44:523–532, <https://doi.org/10.1016/j.atmosenv.2009.10.045>.
9. Jeong JI, Park RJ, Woo J-H, Han Y-J, Yi S-M. Source contributions to carbonaceous aerosol concentrations in korea. *Atmospheric Environment*. 2011; 45:1116–1125, <https://doi.org/10.1016/j.atmosenv.2010.11.031>.
10. Chen J, Lu J, Avise JC, DaMassa JA, Kleeman MJ, Kaduwela AP. Seasonal modeling of pm_{2.5} in california's san joaquin valley. *Atmospheric Environment*. 2014; 92:182–190, <https://doi.org/10.1016/j.atmosenv.2014.04.030>.
11. Saide PE, Carmichael GR, Spak SN, Gallardo L, Osses AE, Mena-Carrasco MA, et al. Forecasting urban pm₁₀ and pm_{2.5} pollution episodes in very stable nocturnal conditions and complex terrain using wrf–chem co tracer model. *Atmospheric Environment*. 2011; 45:2769–2780, <https://doi.org/10.1016/j.atmosenv.2011.02.001>.
12. Wang Q, Zeng Q, Tao J, Sun L, Zhang L, Gu T, et al. Estimating PM_{2.5} Concentrations Based on MODIS AOD and NAQPMS Data over Beijing(-)Tianjin(-)Hebei. *Sensors (Basel)*. 2019; 19(5).

13. Vautard R, Builtjes PHJ, Thunis P, Cuvelier C, Bedogni M, Bessagnet B, et al. Evaluation and intercomparison of ozone and pm10 simulations by several chemistry transport models over four european cities within the citydelta project. *Atmospheric Environment*. 2007; 41:173–188, <https://doi.org/10.1016/j.atmosenv.2006.07.039>.
14. Stern R, Builtjes P, Schaap M, Timmermans R, Vautard R, Hodzic A, et al. A model inter-comparison study focusing on episodes with elevated pm10 concentrations. *Atmospheric Environment*. 2008; 42:4567–4588, <https://doi.org/10.1016/j.atmosenv.2008.01.068>.
15. Li C, Hsu NC, Tsay S-C. A study on the potential applications of satellite data in air quality monitoring and forecasting. *Atmospheric Environment*. 2011; 45:3663–3675, <https://doi.org/10.1016/j.atmosenv.2011.04.032>.
16. Zhang X, Zhang T, Young AA, et al. Applications and comparisons of four time series models in epidemiological surveillance data. *PLOS ONE*. 2014; 9(2): e88075.
17. Liu C, Henderson BH, Wang D, Yang X, Peng ZR. A land use regression application into assessing spatial variation of intra-urban fine particulate matter (pm2.5) and nitrogen dioxide (no2) concentrations in city of shanghai, china. *The Science of the total environment*. 2016; 565:607–615, PMID: 27203521, <https://doi.org/10.1016/j.scitotenv.2016.03.189>.
18. Sorek-Hamer M, Strawa AW, Chatfield RB, Esswein R, Cohen A, Broday DM. Improved retrieval of PM2.5 from satellite data products using non-linear methods. *Environ Pollut*. 2013; 182: 417–23, PMID: 23995022, <https://doi.org/10.1016/j.envpol.2013.08.002>.
19. Song YZ, Yang HL, Peng JH, Song YR, Sun Q, Li Y. Estimating pm2.5 concentrations in xi'an city using a generalized additive model with multi-source monitoring data. *PloS ONE*. 2015; 10:e0142149, PMID: 26540446, <https://doi.org/10.1371/journal.pone.0142149>.
20. García Nieto PJ, Combarro EF, del Coz Díaz JJ, Montañés E. A svm-based regression model to study the air quality at local scale in oviedo urban area (northern spain): A case study. *Applied Mathematics and Computation*. 2013; 219:8923–8937, <https://doi.org/10.1016/j.amc.2013.03.018>.
21. Yan Chan K, Jian L. Identification of significant factors for air pollution levels using a neural network based knowledge discovery system. *Neurocomputing*. 2013; 99:564–569, <https://doi.org/10.1016/j.neucom.2012.06.003>.
22. Feng Y, Zhang W, Sun D, Zhang L. Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification. *Atmospheric Environment*. 2011; 45:1979–1985, <https://doi.org/10.1016/j.atmosenv.2011.01.022>.
23. Keshavarzi M, Goehring T, Zakis J, Turner RE, Moore BCJ. Use of a deep recurrent neural network to reduce wind noise: Effects on judged speech intelligibility and sound quality. *Trends in hearing*. 2018; 22:2331216518770964, PMID: 29708061, <https://doi.org/10.1177/2331216518770964>.
24. Rivas A, Chamoso P, Gonzalez-Briones A, Corchado JM. Detection of cattle using drones and convolutional neural networks. *Sensors (Basel, Switzerland)*. 2018; 18, PMID: 29954080, <https://doi.org/10.3390/s18072048>.

25. Min X, Zeng W, Chen S, Chen N, Chen T, Jiang R. Predicting enhancers with deep convolutional neural networks. *BMC bioinformatics*. 2017; 18:478, PMID: 29219068, [https://doi.org/ 10.1186/s12859-017-1878-3](https://doi.org/10.1186/s12859-017-1878-3).
26. Vlachas PR, Byeon W, Wan ZY, Sapsis TP, Koumoutsakos P. Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proc Math Phys Eng Sci*. 2018; 474: 20170844, PMID: 29887750, <https://doi.org/10.1098/rspa.2017.0844>.
27. Tsiouris K, Pezoulas VC, Zervakis M, Konitsiotis S, Koutsouris DD, Fotiadis DI. A long short-term memory deep learning network for the prediction of epileptic seizures using eeg signals. *Computers in biology and medicine*. 2018; 99:24–37, PMID: 29807250, <https://doi.org/10.1016/j.compbimed.2018.05.019>.
28. Ong BT, Sugiura K, Zettsu K. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting pm2.5. *Neural computing & applications*. 2016; 27:1553–1566, PMID: 27418719, <https://doi.org/10.1007/s00521-015-1955-3>.
29. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural computation* 9:1735–1780, PMID: 9377276.
30. Li X, Peng L, Yao X, Cui S, Hu Y, You C, et al. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environ Pollut*. 2017; 231: 997–1004, PMID: 28898956, <https://doi.org/10.1016/j.envpol.2017.08.114>.
31. Zhou X, Cai J, Zhao Y, Chen R, Wang C, Zhao A, et al. Estimation of residential fine particulate matter infiltration in Shanghai, China. *Environ Pollut*. 2018; 233: 494–500, PMID: 29102879, <https://doi.org/10.1016/j.envpol.2017.10.054>.
32. Johnson T, Long T, Ollison W. Prediction of hourly microenvironmental concentrations of fine particles based on measurements obtained from the baltimore scripted activity study. *Journal of exposure analysis and environmental epidemiology*. 2000; 10:403–411, PMID: 11051530.
33. Liu Y, He K, Li S, Wang Z, Christiani DC, Koutrakis P. A statistical model to evaluate the effectiveness of pm2.5 emissions control during the beijing 2008 olympic games. *Environment International*. 2012; 44:100–105, <https://doi.org/10.1016/j.envint.2012.02.003>.
34. Zhao R, Gu X, Xue B, Zhang J, Ren W. Short period PM2.5 prediction based on multivariate linear regression model. *PLoS ONE*. 2018; 13: e0201011, <https://doi.org/10.1371/journal.pone.0201011>.
35. Zhao J, Deng F, Cai Y, Chen J. Long short-term memory - Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction. *Chemosphere*. 2019; 220: 486–492.
36. Wang X, Qian Z, Wang X, Hong H, Yang Y, Xu Y, et al. Estimating the acute effects of fine and coarse particle pollution on stroke mortality of in six Chinese subtropical cities. *Environ Pollut*. 2018; 239: 812–817, PMID: 29751339, [https://doi.org/ 10.1016/j.envpol.2018.04.102](https://doi.org/10.1016/j.envpol.2018.04.102).
37. Wen C, Liu S, Yao X, Peng L, Li X, Hu Y, et al. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *Sci Total Environ*. 2019; 654: 1091–1099.

Figures

Figure 1

AM and LSTM plots of observed and predicted PM_{2.5} hourly concentration in Daxing and Fangshan.

Figure 2

AM and LSTM fitting plots of observed and predicted PM_{2.5} daily concentration in Fengtai and Fangshan.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [S5Fig.tif](#)
- [S4Fig.tif](#)
- [S3Fig.tif](#)
- [S1Fig.tif](#)
- [S1Text.docx](#)
- [S2Fig.tif](#)