AQI multi-point spatiotemporal prediction based on K-mean clustering and RNN-LSTM model

Jia Zhu1,a*, Baofeng Li1,b, Hong Chen1,c
1School of Architecture and Urban Planning of Huazhong University of Science and Technology, Wuhan, China
a*email: d201780844@hust.edu.cn, bemail: libaofeng_1956@hust.edu.cn, cemail: chhwh@hust.edu.cn

Abstract: The short term air quality index can usually be predicted by statistical and numerical methods, but for the multi-point prediction of AQI, the traditional methods are often inaccurate. In this paper, a new hybrid multi-point prediction method was proposed by combining K-means clustering with the circulating neural network long and short time memory (RNN-LSTM) model. Based on this prediction method, the air quality index in Dezhou was predicted 1-5 days in advance by using 28 multi-point pollution monitoring sensor data from January 1, 2018 solstice to August 31. The prediction results show that the model not only improves the accuracy and effectiveness of the prediction, but also reveals the relationship between land use patterns and air quality index (AQI), which provides important information for land use planning, air pollution mitigation and urban intelligent governance.

1. Introduction
The forecasting of air quality in different locations is important for urban residents to make plans for outdoor activities. This research is to find a more accurate and effective forecasting method of air quality few days in advance.

The study of air quality prediction was mostly conducted by numerical methods. Different complexity Gaussian models as AERMOD (Gibson et al., 2013), PLUME (Vernon et al., 2018) are widely used by industries. Lagrangian models study a trajectory of an air parcel (Ma et al., 2006). Eulerian models use a gridded system to monitor atmospheric properties in specific points of interest (Super et al., 2017). These models combine atmospheric science and multi-processor computing techniques, highly relying on resources like real-time meteorological data (Heo et al., 2008). However, geophysical characteristics might complicate the implementation of these models (Yang et al., 2015). The problem is that high resolution patterns, high computational costs are the biggest disadvantage of them. The new method machine learning algorithm in the fields of statistics and computer science has boomed grown. Some researchers have used it to fix geographic models to predict PM$_{2.5}$ (Zhou et al., 2014). Artificial neural network (ANN) was used to train their independent models. But general ANN (Ravi et al., 2007)! and fuzzy neural network FNN (Ugalde et al., 2015) are not completely suitable for learning time series.

At present, most of these studies fail to predict the different regions of a city, or at the same time provide a more accurate prediction of a sub-region 3-5 days or even a week in advance. Therefore, we choose to use a clustering analysis algorithm combined with a deep learning model, like RNN-LSTM, instead of general ANN.
2. Experiment area and data

Dezhou is a city located in the northwest of China, which was selected as the experiment area. In this study, the intelligent sensor include 8 pollutant indexes in Table 1. The experimental data consist of 28 main monitoring sites in Dezhou, including the traffic junctions of main streets. For each monitoring point, the frequency of pollutant monitoring numerical measurement is interval 24 hours, from 0 a.m. to 24 p.m. every day. In addition, weather forecast data are collected from China Meteorological Administration and used as detection indicators.

| Monitoring site | PM$_{2.5}$ $\mu$g/m$^3$ | PM$_{10}$ $\mu$g/m$^3$ | SO$_2$ $\mu$g/m$^3$ | NO$_2$ $\mu$g/m$^3$ | CO mg/m$^3$ | O$_3$ $\mu$g/m$^3$ | O$_3$-8H $\mu$g/m$^3$ | AQI |
|----------------|----------------|----------------|----------------|----------------|---------|----------------|----------------|-----|
| 1              | 109            | 182            | 35             | 87             | 2.101   | 51             | 45             | 143 |
| 2              | 99             | 179            | 37             | 79             | 2.29    | 50             | 45             | 130 |
| 3              | 102            | 177            | 41             | 78             | 2.339   | 56             | 45             | 134 |
| 4              | 96             | 173            | 37             | 68             | 2.359   | 61             | 52             | 127 |
| 5              | 106            | 181            | 35             | 83             | 2.164   | 53             | 46             | 139 |
| 6              | 87             | 170            | 32             | 71             | 2.126   | 43             | 40             | 115 |

3. Methodology

3.1 Model building and data analysis

In the first step, we do data cleaning by replacing the data loss points with interval averages. Then we take the correlation analysis of each index data conducted by SPSS. The result shows the correlation between pollutants and AQI and correlation between pollutants in Fig. 1.

Fig. 1. Pearson correlation coefficient between independent and dependent variable of indicators

In the second step, we use unsupervised learning data mining algorithm k-mean clustering to classify the data of 28 sites to find the similarity degree of data between different sites. As is shown in Table 2, all the sites are classified into three categories.

| Clustering categories | Monitoring site |
|-----------------------|-----------------|
| 0                     | 7 11 12 18 19 22 25 27 |
| 1                     | 1 3 5 6 8 17 20 21 23 24 28 |
| 2                     | 2 4 9 10 13 14 15 16 26 |

Then, we take the clustering result as a matrix weight in the LSTM-RNN model. The problematic issues of vanishing gradients are solved through LSTM because it keeps the gradients steep enough and the training is relatively short and the accuracy is high. Geographical location as clustering result can be used as a feature weight to enhance attention index. We choose a training set and a validation
set by cross-validation method. We chose 4 groups of data per month, each group included 5 consecutive days to conduct random cross tests for each month. The samples in the training set, validation set, test set are 178, 20, 20 respectively for each site. In order to select relatively better parameters, we set the training epoch as 2000. Training time for each model is approximately 5 hours. Test time less than 3 seconds.

3.2 Distribution of pollutants in Dezhou city

We use unsupervised clustering classifying data sets with different characteristics. From the clustering result, we can find the pollutant numerical ranges and mean values of three categories as shown in Fig. 2.

![Box plots of important pollutants index in three categories site of the model](image_url)

Fig. 2. Box plots of important pollutants index in three categories site of the model

We mark three categories of sites according to the clustering results in the urban development land map of Dezhou, as shown in Fig. 3.

![The spatial distribution of three categories site on land use planning map in domain. Category 0 sites are marked with blue, category 1 sites are marked with green, category 2 sites are marked with red.](image_url)

Fig. 3. The spatial distribution of three categories site on land use planning map in domain. Category 0 sites are marked with blue, category 1 sites are marked with green, category 2 sites are marked with red.

It can be seen that category 0 sites in blue are located in mainly suburbs and reserve lands as grey block, with less human activity. Category 1 sites in green are mostly urban industrial use land as brown block, cultivate land as light yellow block and road use land beside red line. Category 2 sites in red are mostly areas with abundant green space as public green land and water area as sky blue block. This is highly consistent with environmental empirical results.
3.3 Prediction of AQI in multipoint of Dezhou

According to the prediction process of Dezhou air quality index and repeated iterative experiments, since this paper is based on multi-feature time series model, LSTM will be fitted by multi-variable input feature. The results of the experiments measured by MAE, RMSE and MAPE for 28 stations are presented in table 4. The best results are marked with the forecast of the first day yields lower error than the second day. This can be explained by the accumulative error, since the forecasting error for one day in advance is brought into the next day's prediction. From the aspect of absolute error, the AQI prediction of 1-5 day is as shown in Fig. 4.

![Scatterplots between forecasts and observations of AQI](image)

**Fig. 4. Scatterplots between forecasts and observations of AQI**

The hidden layer number of the model is set to 6, and the input layer and output layer are 8 and 1 respectively. The number of iterations of RNN-LSTM models is 2000, and the mean absolute error is almost 10 of five-day model, and the MAE of one-day model is nearly 3.4. Compared with non-clustering RNN algorithm, the accuracy of this prediction result is also improved as shown in Table 3.

| Forecast measure | MAE   | RMSE  | R2    | MAPE   |
|------------------|-------|-------|-------|--------|
|                  | K-mean Cluster RNN-LSTM model | K-mean RNN model | Plain RNN-LSTM model | Plain RNN model | K-mean RNN-LSTM model | Plain RNN model | K-mean RNN-LSTM model | Plain RNN model |
| +1 day           | 3.45  | 3.74  | 4.215 | 4.572  | 0.98  | 0.97  | 4.98% | 5.37% |
| +2 day           | 4.86  | 5.37  | 5.991 | 6.599  | 0.95  | 0.94  | 6.95% | 7.79% |
| +3 day           | 7.60  | 8.03  | 9.303 | 9.765  | 0.89  | 0.88  | 11.08%| 11.91%|
| +4 day           | 8.48  | 8.67  | 10.347| 10.567 | 0.86  | 0.86  | 12.49%| 12.79%|
| +5 day           | 10.16 | 10.48 | 12.455| 12.965 | 0.80  | 0.78  | 14.52%| 15.01%|

The distribution pattern can be clearly seen from the spatial error map of multiple sites as shown in Fig. 5.
4. Discussions and Conclusions

In this paper, a new hybrid model is proposed that is capable of predicting the daily AQI of multiple monitoring sites 1-5 days in advance. It is built by applying the K-mean clustering into RNN-LSTM model. Combined with multipoint spatial-temporal coefficient, the hybrid model is considered to be an effective tool to improve the forecasting accuracy of AQI.

One significant novelty of this approach is that using this hybrid model can forecast AQI of multipoint area in a big city at the same time. The study results show that it can get a more accurate prediction within MAE as much as 3.45 one-day in advance, 4.86% two-day in advance, and 10.16 five-day in advance, even if there is discontinuous or missing data.

Another novelty of this approach is that the relationship between air pollution quality and land-use can be reflected from this model. However, there are still some limitations to be improved. For example, there are many other impacts on AQI including other climate parameters such as wind speed, rainfall, temperature, traffic conditions, industry distribution and other factors. Therefore, more influence factors could be taken into account for forecasting, and more experiments should be performed in different cities with different climates.

Acknowledgments

This work was supported by the National natural science foundation of China (No. 51538004) National natural science foundation of China (No.51778251), and the China Scholarship Council.

References

[1] Mark D.Gibson, Soumita Kundu, Mysore Satish, 2013, Dispersion model evaluation of PM2.5, NO2 and SO2 from point and major line sources in Nova Scotia, Canada using AERMOD Gaussian plume air dispersion model, Atmospheric Pollution Research, 4 (2): 157-167.

[2] Vernon, C. J., Bolt, R., Canty, T., et al., 2018, The impact of MISR-derived injection height initialization on wildfire and volcanic plume dispersion in the HYSPLIT model, Atmospheric Measurement Techniques, 11(11): 6289-6307.

[3] Ma, Y.J., Wang, Y.F., Liu, N.H., 2006, Numerical simulation of spatial and temporal distribution characteristics of major atmospheric pollutants in urban agglomerations in central Liaoning, journal of meteorology and environment, 22(2): 6-10.

[4] Super, I., Van der Gon, H.A.D., et al., 2017, A multi-model approach to monitor emissions of CO2 and CO from an urban–industrial complex. Atmospheric Chemistry and Physics, 17 (21): 13297-13316.

[5] Chen, C., Zhu, Z.J., Liu, D. et al., 2013, Pollution characteristics and source analysis of atmospheric PM2.5 in Zhengzhou [J]. China environmental monitoring, 29 (5): 47-52.
[6] Heo J B, Hopke P K, Yi S M, 2008, Source apportionment of PM2.5 in Seoul, Korea [J]. Atmos, Chen. Phys. 8 (6): 4957-4971

[7] Yang, D.Y., Liu, B.X., Zhang, D.W., et al., 2015, Spatial and temporal distribution law and correlation analysis of water-soluble ions in PM2.5 in Beijing during 2012-2013 [J]. Environmental science, 36 (3): 768 773.

[8] Zhou, Q.P., Jiang, H.Y., Wang, J.Z., Zhou, J.L., 2014, A hybrid model for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network, Science of The Total Environment, 10, 264-274.

[9] Ravi Kumar P, Ravi V, 2007, Bankruptcy prediction in banks and firms via statistical and intelligent techniques a review. Eur J Oper Res. 180 (1): 1–28.

[10] Ugalde HMR, Carmona J-C, Reyes-Reyes J, Alvarado VM, Corbier C, 2015, Balanced simplicity–accuracy neural network model families for system identification. Neural Comput & Applic, 26 (1): 171–186.

[11] Wang, X., Wu, J., Liu, C., et al. 2018, Exploring LSTM based recurrent neural network for failure time series prediction. Journal of Beijing University of Aeronautics and Astronautics, 44 (4): 772-784.