Prediction of local particle pollution level based on artificial neural network

Jie Xiong[1,2,*], *Runming* Yao[1,3], and *Baizhan* Li[1,2]

[1]Joint International Research Laboratory of Green Buildings and Built Environments (Ministry of Education), Chongqing University, Chongqing 400045, China
[2]National Centre for International Research of Low-carbon and Green Buildings (Ministry of Science and Technology), Chongqing University, Chongqing 400045, China
[3]School of the Built Environment, University of Reading, Reading RG6 6DF, UK

Abstract. Citizens eager to know the local pollution level to prevent from air pollution. The real-time measurement for everywhere is a very expensive way, a statistical model based on artificial neural network is applied in this research. This model can estimate particle pollution level with some influencing factors, including background pollution level, weather conditions, urban morphology and local pollution sources. The monitoring from regulatory monitoring sites is considered as the background level. The field measurements of 20 locations are conducted to feed the output layer of ANN model. The average relative error of prediction compared with measurement is 9.24% for PM10 and 18.90% for PM2.5.

1 Introduction

In recent years, the air pollution issue has drawn widespread public attention. Citizens eager to know the local pollution level, i.e. the concentrations of some main air pollutants, which can advise them to make some protection for their outdoor activities [1] or to decide whether or not to apply the natural ventilation for energy-efficiently comfortable indoor environment [2]. It is also very important data for the risk assessment of some environmental-related diseases [3].

There are two main approaches to acquire the particle concentrations: measurement and prediction. Measurement is the method with the most accuracy, it directly reflects the true value of the sampling point when ignoring the system errors, but the cost, including equipment, maintenance and labour costs, is much higher. Other than this, quantities of models, further divided into the numerical model and statistical model, were developed to estimate the dispersion and concentration of particulate matters. For the statistical model, the multiple linear regression (MLR) and the artificial neural network (ANN) are two mainstream approaches.

This research will develop a fast-speed prediction model of particle concentrations at any location of urban area. Some variables regarding the outdoor weather status, the local pollution sources and the urban features will be input into this model.

2 Method

2.1 Artificial neural networks (ANN)

Artificial neural networks (ANN), in an entirely different way from the conventional algorithms, are computing systems vaguely inspired by the biological neural networks that constitute human brains [4]. The structure of a fully connected feedforward neural network is consisting of the input layer, the hidden layers and the output layer (Fig. 1). The activation of a_j^l (the jth neuron in the lth layer) is related to the neurons in the (l-1)th layer by the equation:

$$a_j^l = f(\sum_{k=1}^{n_{l-1}} w_{jk}^l a_k^{l-1} + b_j^l)$$

where a_k^{l-1} is the kth neuron in the (l-1)th layer, n_{l-1} is the total number of neuron in the (l-1)th layer, w_{jk}^l is the weight for the connection from the kth neuron in the (l-1)th layer to the jth neuron in the lth layer, b_j^l is the bias of the jth neuron in the lth layer, and $f(*)$ is the activation function, which determines its nonlinear properties.

* Corresponding author: j.xiong@cqu.edu.cn

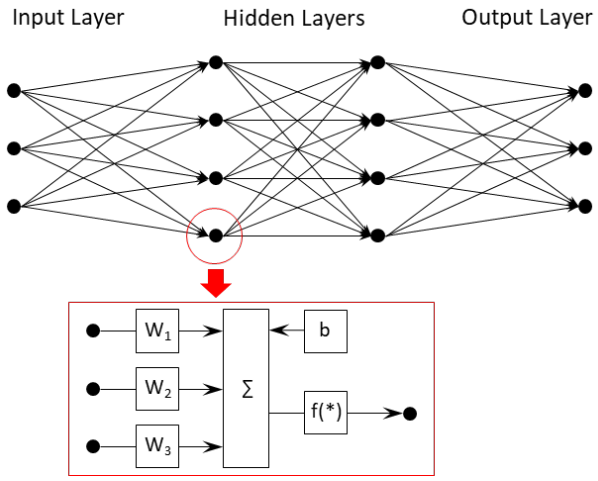


Fig. 1. The comparison between predicted values and measured values of a) PM10 and b) PM2.5.

2.2 Measurement of real-time pollutant concentrations

In order to get real-time pollutant concentration data, the field measurement was carried out on the central urban area with the dense built environment in Chongqing from July 2015 to January 2016 (covering summer, autumn and winter). There are totally 20 dwellings i.e. locations selected in the area covering five central districts, namely Yuzhong District, Jiangbei District, Shapingba District, Yubei District and Jiulongpo District. Continuous 4~5 days monitoring data were collected for each location successively (totally 84 days).

The measuring variables include temperature, relative humidity, concentrations of particulate matter (including PM10 and PM2.5). All the monitoring equipment are set-up to log data in 1-min interval, and the collected data could be processed for specific purposes readily.

2.3 The predicting variables

2.3.1 Background pollution level

Local emission, dispersion and deposition status contribute to the overall air pollution level in a macro scale, in return, the local air pollution level can be considered using overall air pollution level adding the features influencing the production and movement of pollutants. The background pollution level is related to factors such as socio-economic development, and not the focus of this study. So, the particle pollution monitoring data from regulatory sites are used to reflect the background pollution level. The hourly PM10 and PM2.5 data are obtained from the National Air Quality Real-time Release Platform (http://106.37.208.233:20035/) [5] by the China National Environmental Monitoring Centre. There are 6 sites selected from the case-study. Not statistics (like average, maximum, minimum) of those sites, instead data from all the selected sites were directly entered the predicting dataset, which would help the model learn its spatial associations, making this part of variables play its role in spatial interpolation, and the other variables to capture the local features.

2.3.2 Meteorological conditions

Daily and hourly observations from the China Meteorological Administration (http://data.cma.cn/) [6] are obtained. The observation site chosen is called Shapingba (57516), where located in the urban area, it is the closest to all field measurement points. The daily temperature, relative humidity, wind speed, sunshine hours and precipitation are analysed to capture some characteristics of the case-study area. Additionally, the hourly temperature and relative humidity were measured in the measurement campaign as mentioned above. The measurement data and official data are compared, and the hourly temperature and relative humidity from field measurement enter the predicting dataset, and without measurement of wind speed and precipitation on-site, instead the weather station data are applied.

2.3.3 Urban morphology

Building coverage ratio (BCR) is the percentage of total area covered by building in a target land, indicating the compactness of infrastructures horizontally, which is the most commonly used indices for quantifying the building density at land lot scale [7]:

$$BCR = \frac{\sum_{i=1}^{n} A_i}{S}$$

where S is total area of target land, A_i is the coverage area of the building i, and n is the total number of building in the target land.

The building coverage ratio at different heights is calculated to express the urban form with the density of the buildings, and reflect its changes in the vertical direction, using a set of values to depict more details of the three-dimensional morphological characteristics of the urban.

2.3.4 Local pollution sources

Roads are one of the sources of pollutant emission in an urban area. The statistics of transportation facilities and information from the real-time release platform of road condition are utilised to symbolize the pollutant emission level of the local area and its surroundings. The transportation facilities are recognised using the satellite image provided by the software Google Earth Pro (vision 7.3.2) during the field measurement period (21th Oct. 2015). The length of each road on 500 m*500 m buffer area centred on the sampling point can be measured, and the number of lanes for each road can be counted.

A large amount of dust generated from the construction site can carry for a wide range of area over a long period of time. The construction site within 500 m based on the sampling point is also recognised with the

satellite image provided by the software Google Earth Pro. The area of construction sites and the distance from the sampling point are input into the model as the estimators for local traffic emissions. If there is no construction site appearing in the surrounding area, the area of construction sites is set as 0 m2, and the distance is set as 10 km.

2.4 Model evaluation

The effectiveness of the prediction can be evaluated by statistics measuring how well the observed outcomes are replicated by the model. The root mean square error (RMSE) and the mean absolute error (MAE) is the most common indicators used with prediction models. RMSE use the square root of the second sample moment of the differences between predicted values and measured values to represent the overall accuracy.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - M_i)^2}{n}}$$

where P_i is the ith predicted value, M_i is the ith measured value, and n is the volume of the datasets to compare.

The Pearson correlation coefficient (r), a value between -1 and +1, is a measure of the linear correlation between predicted values and measured values.

$$r = \frac{\sum_{i=1}^{n}(P_i - \bar{P})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2}\sqrt{\sum_{i=1}^{n}(M_i - \bar{M})^2}}$$

where $\bar{M}$ is the average of measured values, and $\bar{P}$ is the average of predicted values.

Totally 40 predicting variables, including time periodicity, background pollution level, weather conditions, urban morphology and local pollution sources are considered in this model, it is a spatial interpolation model considering comprehensively the local divergence, including metrological conditions, urban morphologies and emission sources (SC0). Other three input variable schemes are put forward to discuss the impact of inputs on the prediction performance. SC1 only omits background pollution level, it can be used when there is no knowledge of real-time pollutant concentration on certain locations surrounded. SC2, considering the meteorological conditions and local pollution sources, is the most common input variable scheme in the previous studies. SC3 only used official-released data measured from 6 regulatory sites, which is the application of ANN to the spatial interpolation.

3 Results

The prediction results of ANN model with background pollution level, weather conditions, urban morphology and local pollution sources well presents the measuring data (Fig. 2). The mean square error for PM10 is 13.12 μg m^{-3}, and 13.44 μg m^{-3} for PM2.5. It shows linear relationship between predicted values and measured values, with the Pearson coefficient of 0.937 for PM10, and 0.948 for PM2.5. The bias is very small for PM10, but it shows certain negative bias for PM2.5 (Table 1 and Table 2), however, the positive errors appear in the higher concentration, and the negative bias mainly caused in lower concentration.

The prediction performances are also compared to other 3 input variables scheme (Table 1, Table 2 and Fig. 3). SC1 omits background pollution level, it also shows very good performance of prediction with the Pearson coefficient of 0.948 for PM10 and 0.927 for PM2.5. This input scheme can be used to predict the pollution level when there is no knowledge of real-time pollutant concentration in certain locations surrounded. SC2 considers the meteorological conditions and local pollution sources, the prediction performance has become lower without the input of urban morphologies information. But the worst performance is appeared with SC3, only using official-released data measured from 6 regulatory sites. The mean square error reaches around 20 μg m^{-3}, and the Pearson coefficient is only 0.851 for PM10 and 0.708 for PM2.5. This result indicates the application of ANN to the spatial interpolation is relatively limited, it must consider some local information relevant to the generation and dispersion of air pollutants. The distributing of relative error of PM10 and PM2.5 respectively using predicted value compared with the measured value. The relative error is most concentrated around 0 for SC0, and most scattered for SC3.

Table 1. The statistics for the prediction performance of models with different predicting variable schemes for PM10.

Predicting variable scheme	*RMSE* (μg m^{-3})	*r*	*Bias* (μg m^{-3})	*Average relative error*
SC0	13.12	0.937	0.83	9.24%
SC1	11.89	0.948	0.10	19.59%
SC2	14.92	0.916	0.46	24.77%
SC3	19.99	0.851	0.45	31.06%

Table 2. The statistics for the prediction performance of models with different predicting variable schemes for PM2.5.

Predicting variable scheme	*RMSE* (μg m^{-3})	*r*	*Bias* (μg m^{-3})	*Average relative error*
SC0	13.44	0.948	-6.30	18.90%
SC1	15.33	0.927	-6.37	21.50%
SC2	14.65	0.937	-6.81	21.94%
SC3	20.06	0.708	-7.50	29.89%

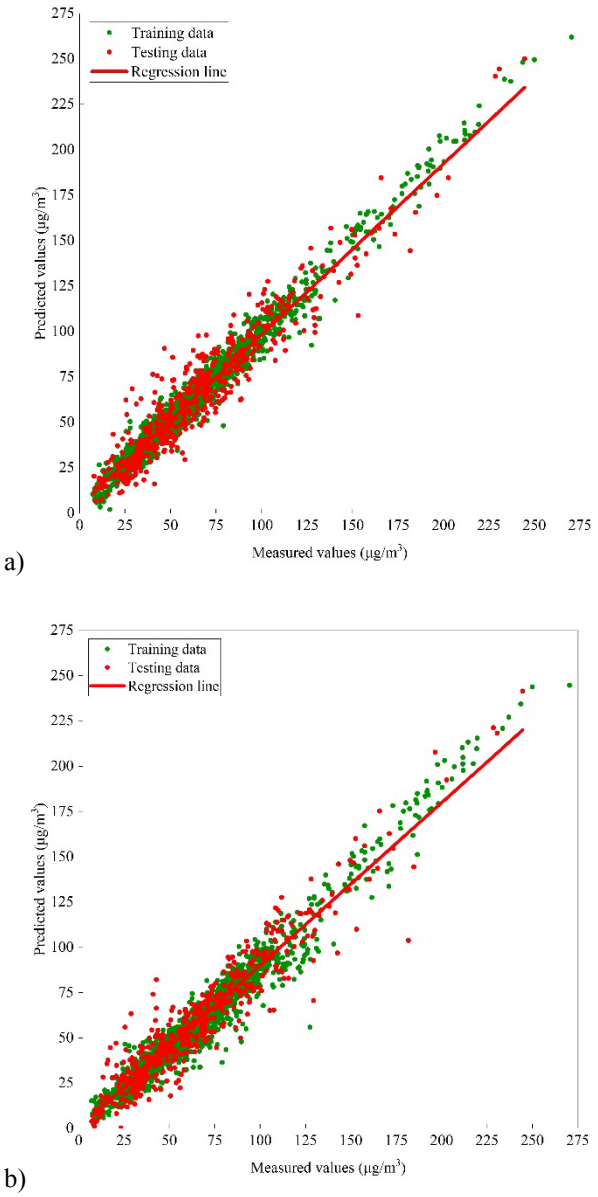
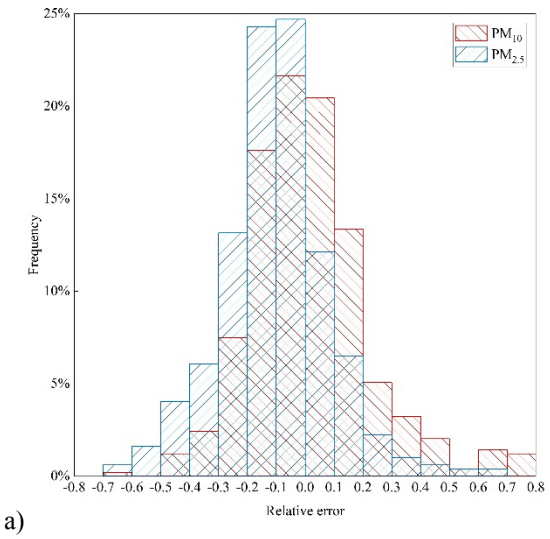


Fig. 2. The comparison between predicted values and measured values of a) PM10 and b) PM2.5.



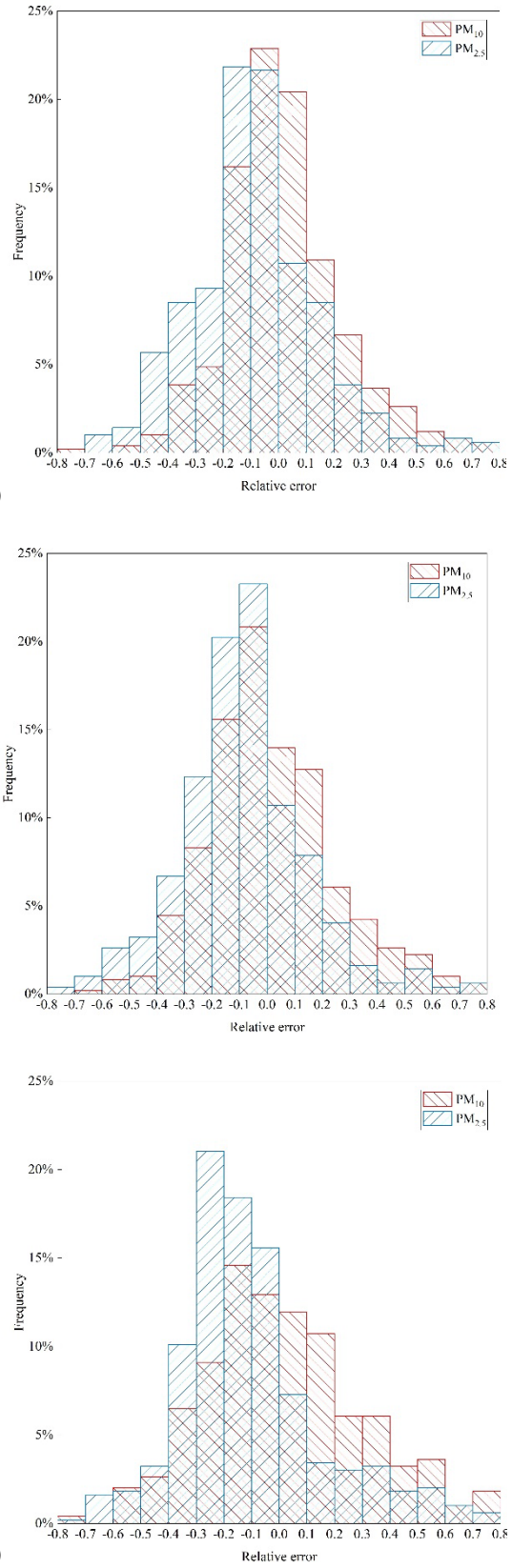


Fig. 3. The histogram of the relative errors (predicted value compared with measured value) from models with different input variable schemes: a) SC0 - time periodicity, background pollution level, meteorological conditions, urban morphology and pollution sources; b) SC1 - time periodicity, meteorological conditions, urban morphology, pollution sources; c) SC2 - time periodicity, meteorological conditions and pollution sources and d)SC3 - background pollution level.

4 Conclusion

A fast-speed estimation model of particle pollution considering some influencing factors, including background pollution level, weather conditions, urban morphology and local pollution sources, is presented. This model is a statistical model based on artificial neural network.

The background pollution level from regulatory monitoring sites replaces most macro-scale development factors that influence the overall pollution level. Local pollution sources directly show how many particle pollutant generates locally. The weather information may influence the whole process of particle contamination, generation, dispersion, transformation and deposition. Urban morphology is displayed using generic indexes to show the urban texture, which will impact the dispersion and deposition of particulate matters. All these factors are input into the ANN model, and the estimation performance is validated with testing dataset. This model can be used for spatial interpolation of particle concentrations. And it can be further used as an operational tool for air quality forecasting with suitable adaptations in any other dense urban areas.

The research work is based on the UK-China collaborative research project 'Low carbon climate-responsive Heating and Cooling of Cities (LoHCool)' supported by the National Natural Science Foundation of China [NSFC Grant No. 51561135002] and the UK Engineering and Physical Sciences Research Council [EPSRC Grant No. EP/N009797/1]. The research is associated with the China National Key R&D Programme 'Solutions to Heating and Cooling of Buildings in the Yangtze River Region' [Grant No. 2016YFC0700300] and the Global Innovation Initiative (GII) project "The impact of ambient air pollution on indoor air quality in China: Evaluation of a practical intervention" supported by the Institute of International Education [Grant No. EGA/A.S/S-13-05]. The authors would like to thank Prof. Howard Kipen and Dr Qingyu Meng from EOHSI, Rutgers University for providing the technical guidance on the field measurement, and also appreciate Dr Wei Yu, Dr Han Wang, Ms Tujingwa Zhang, Mr Zhu Chen and Mr Sheng Zhang's participating in the field measurement campaign.

References

[1] Pacitto A, Amato F, Salmatonidis A, Moreno T, Alastuey A, Reche C, et al. Effectiveness of commercial face masks to reduce personal PM exposure. Sci Total Environ 2019;650:1582–90. doi:10.1016/J.SCITOTENV.2018.09.109.

[2] Yao R, Costanzo V, Li X, Zhang Q, Li B. The effect of passive measures on thermal comfort and energy conservation. A case study of the hot summer and cold winter climate in the Yangtze River region. J Build Eng 2018;15:298–310. doi:10.1016/j.jobe.2017.11.012.

[3] Künzli N, Kaiser R, Medina S, Studnicka M, Chanel O, Filliger P, et al. Public-health impact of outdoor and traffic-related air pollution: a European assessment. Lancet 2000;356:795–801. doi:10.1016/S0140-6736(00)02653-2.

[4] Haykin SO. Neural Networks and Learning Machines: A Comprehensive Foundation. 3rd Editio. Pearson Education; 2009.

[5] China National Environmental Monitoring Centre. National Air Quality Real-time Release Platform n.d. http://106.37.208.233:20035/ (accessed September 15, 2018).

[6] China Meteorological Administration. Dataset of Daily Surface Observation Data in China. China Meteorol Data Serv Cent n.d. http://data.cma.cn/data/cdcdetail/dataCode/SURF_CLI_CHN_MUL_DAY_V3.0.html (accessed July 1, 2017).

[7] Yu B, Liu H, Wu J, Hu Y, Zhang L. Automated derivation of urban building density information using airborne LiDAR data and object-based method. Landsc Urban Plan 2010;98:210–9. doi:10.1016/j.landurbplan.2010.08.004.