Air pollution modeling over Shanghai and Guangzhou

TY Zhang¹, Z Chen², Y Zhu³, RH Zong⁴

¹ School of Statistics, Jiangxi University of Finance and Economics, Qingshanhu, Nanchang, China
² School of Life Science and Technology, University of Electronic Science and Technology of China, Chen Du, Sichuan, China
³ School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China
⁴ School of Computer Science, Sichuan University, Chengdu, Sichuan, China
Email: ztycreative@outlook.com

Abstract. As one of the most prominent pollutants that threaten human health over big cities, fine particle matter (PM2.5) has largely attracted public and researches’ attention. Critical challenges are unresolved regarding how to effectively predict atmospheric PM2.5 concentrations. Here, our team aimed to capture the PM2.5 high-frequency dynamics over Shanghai and Guangzhou using advanced machine learning technique. Our results showed that PM2.5 concentration could be forecasted with historical PM2.5 record and meteorology forcings including temperature, humidity, precipitation, pressure, wind speed, and due point temperature. Our sensitivity analyses further revealed that the prediction was robust against critical model parameters across cities. The underlying sink processes could be different despite of their similar temporal features.

1. Introduction:
China has been quickly growing during the recent decades in terms of economic and social development. Accompanied with the rapid industrialization, emissions of air pollutants (e.g., fine particle matter and N₂O) have been largely increased, most of which sink remotely and generate negative impacts on human health and environmental conditions [1]. Furthermore, land surface properties (e.g., natural forest cover versus urban area) are greatly modified due to the rapid urbanization. Land cover changes modify land surface energy balance and could induce regional climate changes that increase the residence time of air pollutant [2]. Mega cities, such as Beijing, Shanghai, and Guangzhou, have been suffering from periodic sever air pollution events, in which fine particle matter (PM) pollution, especially PM2.5 is one of the most prominent ones [3]. PM2.5 is defined as particles less than 2.5 micrometers in diameter. It could strongly impact human health especially respiratory system and lead to higher mortality. For example, in toxicological studies, PM2.5 is cytotoxic to cell death which may cause respiratory inflammation [4]. Multiple factors convolute together and contribute a high ambient PM2.5 condition. They include anthropogenic factors, such as, industrial emission with the relatively long-term influence (monthly or yearly) and vehicle emission with a clear diurnal pattern (e.g., high during rush hours). In addition, environmental factors could also contribute to severe PM2.5 pollution. Meteorological driver such as high pressure could stabilize the boundary condition, and prevent PM2.5 from being removed out of atmosphere. High wind speed may have a double-edged effects. It could either decrease the concentration of the PM2.5 by quickly propagating the pollutant to adjacent regions or increase the
PM2.5 emission by blowing up fine dust from the land surface. Humidity and precipitation negatively affect the density of particulate matters by inducing wet deposition [5,6]. To tackle the critical challenge of understanding and predicting PM2.5 over large cities, previous efforts have been focused on three major avenues: 1) physical modeling; 2) simple statistical inference; 3) advanced machine learning techniques. Physical models were developed based on inner physical processes of particulate matter formation, emission, transport, chemical reaction, and removal. Statistical models could also be applied to infer PM2.5 dynamics, mainly based on regressed relationship between target pollutant (PM2.5) concentration and drivers (e.g., wind speed, precipitation, and air temperature). However, the underlying processes behind PM2.5 dynamics is extremely complex, thus restricts the accuracy of the simple statistical inference prediction. Recently, many researchers used advanced machine learning algorithms to reproduce this complex processes from a data-driven perspective. For example, existing studies used neural network techniques to either directly predict the air pollutant concentration or classify the air pollutant level [7,8]. In this work, our team apply advanced neural network architecture, called Recurrent Neural Network (RNN) to forecast the PM2.5 concentration over Shanghai and Guangzhou. Unlike other feedforward neural networks models, RNN mechanismically considers the historical information or internal memory of the system itself, which is more appropriate to time series tasks like PM2.5 concentrations.

2. Methodology

2.1. Air pollutant and relevant meteorology data

The data is continuous high-frequency measurements of atmospheric PM2.5 concentration collected over the cities of Shanghai and Guangzhou [9]. The temporal period covers from 2011 to 2015. High-frequency measurements of meteorology drivers were also used, including temperature, pressure, relative humidity, precipitation, wind speed, and dew point (temperature that induces the saturated water). At the data pre-processing stage, the missing values and extreme values (outliers) were removed, so that the data can be safely used for model training. The reason to choose Guangzhou and Shanghai to simultaneously model PM2.5 is that their annual variation as well as a seasonal cycle of PM2.5 are relatively consistent (Figure 1).

![Figure 1. Time series of PM2.5 concentration over Shanghai and Guangzhou from 2011 to 2015. The left panels show annual variation and the right panels show seasonal dynamics](image)

2.2. Predictive modeling

In this study recurrent neural network was employed to model atmospheric PM2.5 concentration, and particularly with long-short-term-memory (LSTM) technique first proposed by Hochreiter et al. [10]. LSTM was originally designed to help conserve historical information within a neural network, which cannot be handled by feedforward neural networks, such as convolutional neural network or fully connected neural network. Similarly, here our aim is to conserve a certain fraction of historical
information about PM2.5 concentrations as well as relevant meteorology factors to help understand and predict future fate of PM2.5. The first step in the LSTM is to decide which information will be discarded from the cell state. This decision is accomplished by forget gate layer. This gate will read previous time step output \( h_{t-1} \) and current time step input \( x_t \) to output a value between zero and one, and later on will be applied to previous time step cell state \( C_{t-1} \). Here, one means "complete reservation" and zero means "complete abandonment".

\[
f_t = \sigma(W_f \cdot [h_{t-1}, x_t])
\]

The next step determines which newly input information should be stored in the cell state, which is calculated in two steps. First, the sigmoid neural network layer (input gate) decides how much new information in terms of percentage will be used. Then, a tanh neural network layer creates a candidate value vector of in internal state. Finally, two components are combined to generate a new update of internal state.

\[
i_t = \sigma(W_o \cdot [h_{t-1}, x_t])
\]

\[
\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t])
\]

The new internal state \( C_t \) is updated using previous internal state \( C_{t-1} \) and the newly generated candidate state \( \tilde{C}_t \).

\[
C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t
\]

Finally, the model output is based on the new internal cell state. First, a sigmoid neural network layer determines which part of the cell state will be outputted. Then, running the cell state through tanh activation function (get a value between -1 and 1) and multiplying it with the output from previous sigmoid layer will get the output at current time step \( h_t \):

\[
o_t = \sigma(W_c \cdot [h_{t-1}, x_t])
\]

\[
h_t = o_t \cdot \tanh(C_t)
\]

The optimization method is Adam (Adaptive Moment Estimation), a momemtent based stochastic gradient descent method, to train LSTM, and (Mean Squared Error) MSE to evaluate the optimization and prediction.

\[
RMSE = \sqrt{\frac{1}{N} \sum_{t} (\hat{z}_t - z_t)^2}
\]

where \( \hat{z}_t \) is the prediction value, \( z_t \) is the real value

### 3. Results and discussion

#### 3.1. Model performance at Shanghai and Guangzhou

Two independent Recurrent Neural Network (RNN) models at Shanghai and Guangzhou was trained, using first 80% of the observed time series. Input features are historical PM2.5 and meteorology forcings; and target variable is the present time step PM2.5 concentration. Figure 2 showed the evaluation of the two models using the rest 20% data. In general, the results demonstrated that LSTM was an effective modeling tool to capture the temporal evaluation of atmospheric PM2.5 concentration over Shanghai (Figure 2 left panel) and Guangzhou (Figure 2 right panel).

In addition, sensitivity test of the model performance against the number of LSTM layers answers the question of how deep the LSTM network is necessarily needed, in order to best capture the observed PM2.5 concentrations. Figure 2 indicated that three LSTM layers was sufficient enough to predict PM2.5 dynamics, while increase the number of LSTM layers did not significantly benefit the overall model performance and sometime even decreased the model performance due to introducing more parameters (Figure 1 left panel). Then the number of LSTM layers was fixed at three, while another sensitivity test of the model performance against the number of neuron at each LSTM layer was
conducted for how much internal LSTM complexity is required. Figure 3 showed that increasing the number of neurons from ten to fifty did not improve model performance at all.

![Figure 2](image2.png)

**Figure 2.** Model evaluation at Shanghai and Guangzhou. The overall performance has been tested against the number of layers in LSTM.

![Figure 3](image3.png)

**Figure 3.** Model evaluation at Shanghai and Guangzhou. The overall performance has been tested against the number of nodes used in each LSTM layer.

3.2. Model performance across cities

The observed temporal patterns of PM2.5 concentration are generally similar at Shanghai and Guangzhou (Figure 1). However, the underlying mechanisms that govern the emergent patterns could be different. Fortunately, it was able to test this hypothesis by applying the RNN model trained at one city (Shanghai or Guangzhou) to another (Guangzhou or Shanghai). Figure 3 summarized the cross-evaluation results of RNN models. Firstly, applying RNN trained at Guangzhou to Shanghai generally underestimated PM2.5 concentration (Figure 4 left two panels). Further such underestimation is sensitive to the number of LSTM layers but not number of neurons per LSTM layer. More LSTM layers will prevent the model being effectively applied to other cities. Secondly, applying RNN trained at Shanghai to Guangzhou could either overestimate or underestimate PM2.5 concentrations (Figure 4 right two panels) according to the different model setup (number of LSTM, number of neurons per LSTM). For example, the RNN with five LSTM layers and 30 neurons per LSTM layer (blue dots) tended to overestimate PM2.5 at Guangzhou, while an underestimation was observed for the RNN with seven LSTM layers (purple dots, upper right panel). In summary, it
indicated that although observed PM2.5 dynamics were quite similar across different cities, having a general model that is successful at different cities could be challenging because the underlying predictive relationship and PM2.5 sink and source processes could be largely different.

Figure 4. The comparison of the trend in cross-prediction by changes in layers and nodes

4. Conclusion and Limitation

4.1. Conclusion
Predictive modeling of important air pollutant is challenging due to the complex nature of the pollutant source/sink dynamics. In this paper, a recurrent neural network architecture called long-short-term-memory (LSTM) was employed to model the atmospheric PM2.5 concentrations over Shanghai and Guangzhou. The model successfully forecasted the PM2.5 dynamics over the two cities, furthermore, the performance was quite robust and stable with different neural network settings. More importantly, cross-city application suggested that although the observed PM2.5 temporal features are similar, their underlying source/sink processes could be different.

4.2. Limitation and future work
This study is limited by computational resources. In the RNN modeling, the number of the epochs and batch size was constant, although the training of RNN could be sensitive to both of them. Future work will be carried out for a full set of values for epochs and batch size in order to find the generally optimal combination of these hyperparameters. Secondly, this study only tested our RNN framework at two cities, while the methodology should also be applicable to other big cities. It is possible for us to expand our modeling efforts to other cities of interest, such as New York or Los Angeles.
5. Reference

[1] Wang, J, et al. (2017). Large-scale transport of PM 2.5 in the lower troposphere during winter cold surges in China. Scientific reports, 7(1), 13238.

[2] Kalnay, Eugenia et al., (2003). Impact of urbanization and land-use change on climate. Nature 423.6939: 528.

[3] Lu, F, et al. (2015). Systematic review and meta-analysis of the adverse health effects of ambient PM2.5 and PM10 pollution in the Chinese population. Environmental research, 136, 196-204.

[4] Xing Y F, et al. (2016) The impact of PM2.5 on the human respiratory system[J]. J Thorac Dis,, 8(1):E69.

[5] Yang L et al., (2011). Seasonal and diurnal variations in N(2)O concentrations and fluxes from three eutrophic rivers in Southeast China[J]. JOURNAL OF GEOGRAPHICAL SCIENCES 20(5):820-832.

[6] Gibson M et al., (198). Ground effects on pressure fluctuations in the atmospheric boundary layer. Journal of Fluid Mechanics, 86.3: 491-511.

[7] Fu, M et al. (2015). Prediction of particular matter concentrations by developed feed-forward neural network with rolling mechanism and gray model. Neural Comput. Appl, 26, 1789–97.

[8] Kleine D, et al. Modeling (2017). PM2.5 urban pollution using machine learning and selected meteorological parameters. Journal of Electrical and Computer Engineering

[9] Liang X, et al. (2016). PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities[J]. Journal of Geophysical Research Atmospheres, 121.

[10] Hochreiter S, et al., (1997). Long Short Term Memory, Neural computation,, 9(8): 1735-80.