Joint prediction of CO, NOx, NMHC pollutant concentrations in urban area

JL Zhao¹, MX Sun², Q Li³

¹ Hong-Kong Baptist University, Kowloon Tong, Hong Kong
² University of Essex, Colchester, UK
³ Duke University, NC, USA
E-mail: 1921938704@qq.com

Abstract. Primary air pollutants could directly and indirectly (through generating secondary air pollutants) threaten natural or human systems. In particular, urban air pollution issue becomes more and more significant in the recent decades, as a result of rapid urbanization. However, the estimation of multiple-pollutant concentrations is limited by high spatial and temporal variations, which hinders the accuracy of mechanistic modeling of air pollution. In this study, we employed an Artificial Neural Network (ANN) model to jointly predict multiple primary pollutants, including CO, NOx, and nan-methane hydrocarbons (NMHC), over the urban area of Italy. The results showed that performances of the ANN model (MSE and Pearson correlation) in joint prediction cross multiple pollutants were much better than in prediction of any single pollutant individually, indicating the joint measurements of multiple pollutants could favor the machine-learning model by providing useful information from one pollutant to predict another.

1. Introduction
Air pollution occurs when the air is contaminated by harmful chemical, physical or biological agents from natural or human sources. In the aspect of natural resource, dust from dry land, gas emissions from forest fires, ash from volcano eruptions and so on could significantly contribute to air pollution. Anthropogenic emissions through burning fossil fuel and biomass could also negatively affect the air quality. Air pollutants can be categorized into primary pollutants and secondary pollutants. Primary pollutants are the harmful substances directly entered into the air, including e.g., Carbon monoxide (CO), Nitrogen Oxides (NOₓ), Nanmethane hydrocarbons (NMHC), and others. However, secondary pollutants are formed when primary pollutants react with atmospheric chemicals. CO is colorless and odorless pollutant, which is produced by incomplete burning of fuels [1]. The primary source of CO is from motor vehicle exhausts. Thereby, CO usually accumulates in areas with heavy traffic. CO diffuses into human’s blood and impair the ability of hemoglobin to carry oxygen in red blood cells, which can result vital organs failure, such as brain. NOₓ is a common mixture pollutants over the urban area pollutants, includes nitrogen monoxide (NO) and nitrogen dioxide(NO₂).Vehicle exhausts, industrial activities, and electricity producing contribute to the high concentration of NOₓ. For example, NOₓ are often generated when fuel is burned at a high temperature. NO could be extremely harmful when it dissolved into water in the atmosphere and from acid rain, which damages buildings, threatens the existence of plants and animals including human [2] [3]. Non-methane hydrocarbons (NMHCs) are long-lived chemical species in the atmosphere. They play an important role in regulating the air quality via reacting with hydroxyl radical (OH) and producing many harmful oxygenic compounds, and result in the formation secondary air...
pollutants in the atmosphere [4]. Vehicular emissions, biofuel combustion, biomass burning and industrial emissions were the major contributors to NMHCs. Air those three major pollutants of interest could induce adversely short and long term effect on human healthy depending on the degree of concentration, length of exposure, type of pollutants and individual characteristics. In order to better understanding the complex process of dispersion of air pollutants and to make the connection between concentrations and the sources, models have been widely used to project the pollutant concentrations at present day and also in the future [5] including regional photochemical grid models and local-scale dispersion models[6]. For example, regional photochemical grid models are able to directly simulate the transport and formation of pollutants with atmospheric chemical reactions (i.e., Community Multiscale Air Quality(CMAQ) model). These models are computationally expensive and have been used at regional scale, but are not easily applicable to local-scale dynamics [6]. The data-driven models such Machine Learning (ML) models provide a promising way to learn purely from existing data and based on which generate computationally efficient prediction in the future. In this study, our objective is to accurately predict concentrations of multiple air pollutants including CO, NOx, and NMHCs.

2. Methodology
We used observations of multiple pollutants from Italy, with hourly temporal resolution, from March 2004 to April 2005. Measurements were collected in an urban area, where was characterized by heavy car traffic. Multiple sensors measured concentrations of CO ($\mu$g/m$^3$), non-metanic hydrocarbons (NMHC) ($\mu$g/m$^3$), NOx(ppb), and the temperature and relative humidity. At the measure site, conventional analyzers were used to provide ground truth of the target pollutants. Then, multisensor device is used to jointly measure all pollutants. In summary, observed CO concentration is much higher than NOx and NMHC (Figure 1 left panel); and the joint distribution of density between CO concentration and NMHC/NOx indicated strong correlations (Figure 1 right panel).

We employed an Artificial Neural Network (ANN) to model the concentration dynamics of CO, NOx, and NMHC. ANN is a fully connected neural network (Figure 2), in which each neuron calculate a linear combination of previous layer neurons’ output and embed with a non-linear activation function:
\[ node = \sigma(\sum_{i} x_i \cdot w_i) \]  

(1)

where \( \sigma \) is sigmoid activation function \( \phi(x) = \frac{1}{1 + e^{-x}} \), \( x \) and \( w \) represent values from previous layers’ node and corresponding weighting parameters.

Figure 2. ANN Model structure. A 4-layer neural network with two inputs, three hidden layers of 10 neurons each and one output layer.

The ANN structure is designed as input-h1-h2-h3-output. Temperature and relative humidity were used in input variable. h represents hidden layer. And the output variable is pollutant. We constructed four independent ANN models: ANN model 1 predicts CO concentration; ANN model 2 predicts NOx; ANN model 3 simulates NMHC dynamics; and the fourth ANN jointly predicts all pollutants together.

The experimental data were divided into training, cross-validation, and testing purposes. By gradient decent search method, the training process progressively find the minimum of mean square error (MSE). Our hypothesis is that pollutant concentration prediction could benefit from joint prediction across multiple targets.

3. Results and discussion

3.1. Model prediction of single pollutant

All ANN models were set up with the same configuration, using 3 hidden layer, 10 neurons each layer, number of epoch was 150, and batch size was 1000. To avoid overfitting, the data set was randomly split into two sets, 80% for model training (to compute the gradient and updating of the network parameters, such as weights and biases—the training set) and 20% for model testing. The type of hidden (and output) unit activation function is relu function, and the type of optimization algorithms is Adam Optimization Algorithm [7].

Individual modeling of CO, NOx, and NHMC with ANN is not satisfactory in this case. CO, NOx, and NHMC only models had relatively large MSE of 0.027, 0.025, and 0.030, respectively. Moreover, the correlation between observed and predicted CO, NOx, and NHMC were smaller than 0.5 in general, indicating potential model deficiency (Figure 3).
3.2. Joint prediction of multiple pollutants
Using the same model configuration as well as the same model complexity, we also jointly trained and predicted CO, NOx, and NMHC dynamics altogether. We found that the joint prediction could lower the MSE down to 0.00018 and enhance the model-data correlation to 0.97 (Figure 3). It supported our hypothesis that CO, NOx, NMHC in urban area were generated from the same source (e.g., vehicle emissions) could be jointly simulated. And more importantly, although three different chemical species, their time series encode similar information thus could benefit ANN modeling when the model enable the prediction of any species through learning from dynamics from other species.

3.3. Limitation and future work
In this study, we limited the pollutants of interest to be CO, NOx and NMHC, and used them to train ANN models. However, there are other pollutants could also be potentially useful to help the joint prediction of CO, NOx, and NMHC dynamics. Furthermore, ANN model performance is highly dependent on parameters tuning. We did not go through all possibilities of parameters, such as the total possible value of nodes in hidden layers to determine the structure of ANN model. Appropriate network structure is achieved through experience and many times of trial and error, which require a large diversity of training for operation. In addition, we also acknowledged that we used a dataset from only one signal site, which is one of the main street in the centre of an Italian city. The spatial and temporal characteristics are limited based on just one site for one year, which may not present overview of the variations in the pollution situation.

In the future, we will focus on more pollutants at different sites to improve the accuracy of prediction and generalize our ANN modeling of multiple pollutant dynamics. In addition, to achieving appropriate network structure is to list as more possibilities as possible and to use trial and error way to find the good parameters.

4. Conclusions
Monitoring and modeling the temporal evolution of multiple primary pollutants over urban area is critically important to better understand their source-sink dynamics. However, most of the available observations were made for each individual pollutant at different locations, in which case the quality of data is highly sensitive to instrumental malfunction. In this study, we used advanced machine learning technique to demonstrate that joint measurements of
multiple primary pollutants not only benefit the spatial consistency of different pollutant measurements, but also favored the modeling effort if joint modeling of multiple pollutants was employed. We showed that by jointly modeling CO, NOx and NMHC, machine-learning model could learn information of one pollutant from the time series dynamics of another pollutant and improve the overall modeling accuracy.

5. References

[1] Levy, R. J. (2015). Carbon monoxide pollution and neurodevelopment: a public health concern. *Neurotoxicology and teratology*, 49, 31-40.

[2] Capşa, D., Barsan, N., Felegeanu, D., Stanila, M., Joita, I., Rotaru, M., & Ureche, C. (2016). Influence of climatic factors on the pollution with nitrogen oxides (NOx) in city, Romania. *Environmental Engineering & Management Journal (EEMJ)*, 15(3).

[3] De Vito, S., Piga, M., Martinotto, L., and Di Francia, G. (2009). CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic Bayesian regularization. *Sensors and Actuators B: Chemical*, 143(1), 182-191.

[4] Rappenglück, B., & Fabian, P. (1999). Nonmethane hydrocarbons (NMHC) in the greater Munich area/Germany. *Atmospheric Environment*, 33(23), 3843-3857.

[5] Shaban, K. B., Kadri, A., & Rezk, E. (2016). Urban air pollution monitoring system with forecasting models. *IEEE Sensors Journal*, 16(8), 2598-2606.

[6] Arunachalam, S., Valencia, A., Akita, Y., Serre, M. L., Omary, M., Garcia, V., & Isakov, V. (2014). A method for estimating urban background concentrations in support of hybrid air pollution modeling for environmental health studies. *International journal of environmental research and public health*, 11(10), 10518-10536.

[7] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.