

## Data Driven based PM<sub>2.5</sub> Concentration Forecasting

Haiqin LI

Ningbo University  
Ningbo, Zhejiang, China  
e-mail: lihaiqin543@163.com

Xuhua SHI

Ningbo University  
Ningbo, Zhejiang, China  
e-mail: shixuhua@nbu.edu.cn

**Abstract**—A PM<sub>2.5</sub> concentration prediction approach using data-driven model is proposed in this paper, which uses support vector machine regression (SVR) and SVR combined with Particle Swarm Optimization (PSO) respectively. The forecast results have a certain advantage by comparing PSO-SVR prediction model and single SVR model. In PSO-SVR prediction model, the parameters of the SVR are optimized by particle swarm optimization algorithm. Then PM<sub>2.5</sub> concentration of regional air can be precisely forecasted by using this method. The simulation results show that the PSO-SVR model is a good forecasting method.

**Keywords**—SVR modeling; PM<sub>2.5</sub> concentration forecasting; data-driven

### I. INTRODUCTION

Today air pollution affected people's daily lives seriously and endanger people's health. According to statistics, the total number of dust-haze days have reached the highest haze of 52 years in 2013, and hindered China's sustainable economic development. The pollutants such as carbon monoxide (CO), PM<sub>2.5</sub>, PM<sub>10</sub>, SO<sub>2</sub>, NO<sub>2</sub>, and ozone (O<sub>3</sub>), are life-threatening and cause difficulties in breathing, headache, and dizziness in short time, as well as the different cancers and heart attack in long time. Therefore, the prediction of the air concentration of particulate matter (PM<sub>2.5</sub>) has a very large significance, which provides useful information for air pollution prevention and treatment policies.

Many researches concerned prediction of air pollutions have been done so far. A lot of scholars have launched research to it, with development more and more scholars apply it on the different atmospheric pollution in the field of simulation. It has now developed the third-Generation Air Quality Modeling [1][2][3]. Nie [4] and others adopted a system of air quality model to carry on the numerical simulation for nuclear leakage accident in Japan on March 11, 2011, the results show that the model system of nuclear accident has a good ability to simulate and study the extent of this impact on countries waters. Shi [5] and others also used a modeling system of multi-scale air quality model to simulate the observation of the NO<sub>2</sub> during the North America and in comparison with NO<sub>2</sub> regional chemical transport model, then assess the performance of CMAQ. Zhang [6] and others used to discharge mode and multi-scale air quality models to establish air quality modeling platform for the Pearl River Delta region, in order to simulate and verify the air quality of the Pearl River Delta region, the results show that the simulation system can perfectly

simulate the change trends of pollutants of the Pearl River Delta region. Wang [7] and others adopted multi-scale air quality model to study the dust weather and simulate the space-time change of air pollutants and the interconnectedness between them in North China. The results are consistent with the observed values. Thus, it indicating that the system can be used to study weather dust pollutants. Wang [8] and others used models-3 simulation system to carry out a numerical simulation for Shenyang air pollutants and results showed that the simulated variation of atmospheric pollutants are very similar with actual variation, which indicated models-3 simulation system is highly suitable and also further explore the regulation of changes for atmospheric pollutants in Shenyang. Ma[9] and others set up MM5 - models - 3 / CMAQ regional air quality simulation system for the concentration of pollutants in the atmosphere in China, simulated the particle concentration of SO<sub>2</sub>, NO<sub>2</sub>, the results show that the simulation system can well simulate the space-time change trend of pollutants, has a strong ability of simulation.

Although numerical models of pollutants in the air have a strong modeling capabilities and have high scientific explanatory, it requires detailed pollutant emissions and meteorological model data and it is not easy to get these data in actual life. So it is not suitable for using of numerical models to predict the concentration of air pollutants. This paper proposes a data-driven model to forecast the concentration of atmospheric pollutants and the requirement of sample data is much lower than the numerical models.

### II. DATA DRIVEN MODEL FOR AIR PM<sub>2.5</sub> CONCENTRATION

In recent years, the data driven model has been widely concerned in the prediction of atmospheric pollutant concentration. It is the goal of establishing the optimal mathematical relationship between the input and output data, such as multiple linear regression model [10] artificial neural network and support vector machine regression. Early scholars using multiple linear regression model to predict the concentration of pollutants, which the basic idea is using correlation analysis or meteorological knowledge extracted factors closely related to the concentration of pollutants as input, pollutant concentration as output, finally, according to input and output to build a regression equation [11]. Because the multiple linear regression model is difficult to excavate the useful information of the pollutant concentration data, the precision of prediction is not very high. Along with the

application of artificial neural network to the prediction of atmospheric pollutant concentration. Sang [12] and others used BP neural network to establish a forecasting model to predict the ozone concentration in the atmosphere of Seoul. According to the historical data, Prybutock and Yi.J train a BP neural network and forecast the concentration of North American Industrial Zone [13]. Although SVR can solve the problem of over fitting, it cannot make full use of the relationship between the input data to establish the corresponding model. Because the parameters of  $\bar{C}$  and  $\sigma$  in support vector machine have a great influence on the prediction effect of the model. Therefore, this paper proposes cluster and support vector machine regression (SVR) to predict the concentration of PM<sub>2.5</sub>, and using particle swarm optimization algorithm (PSO) to optimize the parameters of  $\bar{C}$  and  $\sigma$  for different SVR models.

### III. PM<sub>2.5</sub> CONCENTRATION PREDICTION

SVR (Support Vector Machine for regression) is developed by Vapnik and others on the basis of support vector machine classification. The basic idea is different from support vector machine, which is used to find an optimal hyperplane that all of the training samples near to the optimal hyperplane.

S is training samples  $\{(x_1, y_1), (x_2, y_2), \dots, (x_s, y_s)\}$ , one of  $(x_i, y_i)$  is the first I training samples.  $x_i, y_i \in R, i = 1, 2, \dots, s$ . Set of regression function, as in Eq. (1).

$$f(x) = w \cdot \Phi(x) + \bar{b} \quad (1)$$

Where,  $\Phi(x)$  is Nonlinear mapping function. The weight coefficient vector in formula (1) is obtained by convex quadratic programming problem with two times, and the introduction of relaxation factor  $\xi_i \geq 0$  and  $\xi_i^* \geq 0$ . as in Eq. (2).

$$\min \frac{1}{2} \|w\|^2 + \bar{C} \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (2)$$

Constrained condition defined as in Eq. (3).

$$\begin{aligned} y_i - w \cdot \Phi(x) - \bar{b} &\leq \varepsilon + \xi_i, i = 1, 2, \dots, n \\ -y_i + w \cdot \Phi(x) + \bar{b} &\leq \varepsilon + \xi_i^* \\ \xi_i &\geq 0, \xi_i^* \geq 0 \end{aligned} \quad (3)$$

One of  $\bar{C}$  is penalty parameter of the error term.

The parameter selection in support vector machine algorithm has great influence on the performance of the algorithm, so it is very important to select the appropriate parameters. In this paper, the (Particle Swarm Optimization) PSO algorithm is used to optimize the parameters of the SVR model, and then the optimal parameter values are obtained. The flow chart of the algorithm is shown in Fig 1.

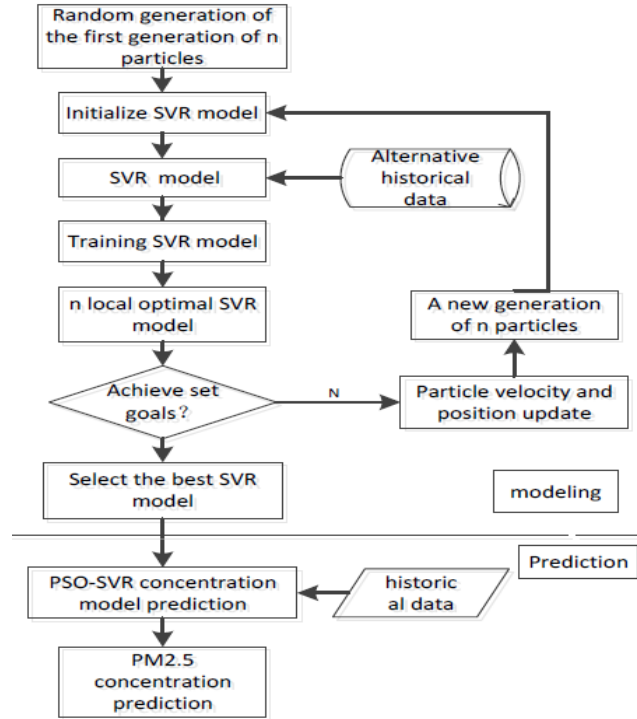


Figure 1. The algorithm flow chart of PSO-SVR model.

### IV. PREDICTION RESULTS

The basic data of this paper is based on the monitoring data from January 1, 2013 to June 6th, 2013 in Ningbo city. The basic data include atmospheric pollutant concentration data and meteorological data. The data of atmospheric pollutant concentration is SO<sub>2</sub>, NO<sub>2</sub>, PM<sub>10</sub>, CO, O<sub>3</sub> and PM<sub>2.5</sub>. And meteorological data contains wind speed, wind direction, temperature, pressure, humidity, and visibility. In the total 3500 groups of data, the first 3250 groups are used as the training sample, the other 250 groups are used as the test sample.

In this paper, the PM<sub>2.5</sub> prediction results of single SVR model and PSO-SVR model are compared with the results of single SVR model shown in Fig 2 and Fig 3.

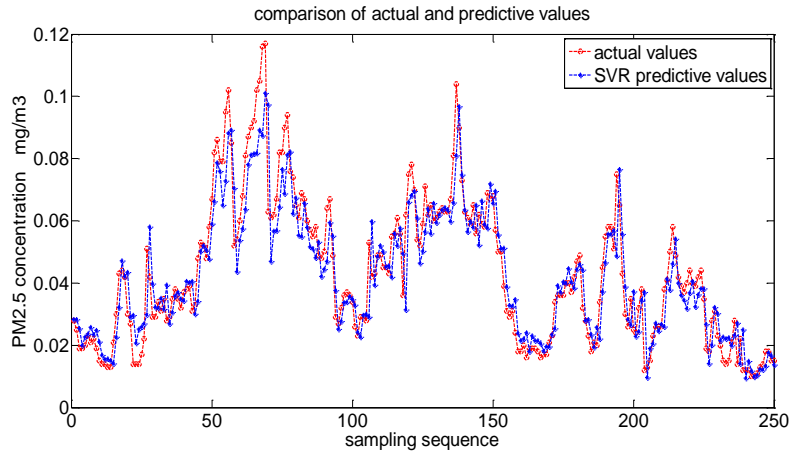


Figure 2. The results of SVR forecast.

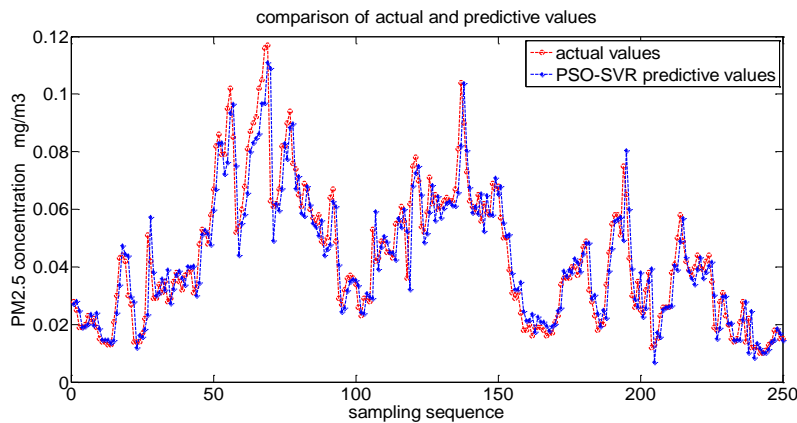


Figure 3. The results of PSO-SVR forecast.

In the figure, the horizontal coordinate is the test data sequence for 250 groups. The vertical coordinate is the PM<sub>2.5</sub> concentration, the curve is the simulation result of the actual value and the predicted value. In order to reflect the actual value and predictive value of the simulation results better, the definition of the sum of residual absolute value, as in Eq.(4).

$$absErrSum = \sum_{i=1}^{n_2} |y_i - y_{outi}| \quad (4)$$

The formula  $n_2$  indicates that the number of forecast value, the  $y_i$  is the first  $i$  actual value, and the  $y_{outi}$  is the first  $i$  forecast value.

Fig.2 shows the only SVR model for predicting, the sum of residual absolute value is 1.7576, and the relative error is 17.772%. Fig3 is the PSO-SVR model the sum of residual absolute value is 1.5348, and the relative error is 16.503%. From the above result, we can see that the prediction accuracy of PSO-SVR approach is higher than that of only -SVR approach.

## V. CONCLUSIONS

In this paper, the support vector machine regression method is used to predict the concentration of PM<sub>2.5</sub>. The algorithm considers all kinds of influence factors of PM<sub>2.5</sub>, and selects the cluster samples to cluster the training samples by correlation calculation. Each type of data after clustering is added to the corresponding meteorological factors as input variables are trained by SVR model. The parameters of different SVR models are optimized by using PSO optimization algorithm. The simulation results show that the PSO-SVR model is a good forecasting method.

## ACKNOWLEDGMENT

This work is supported in part by the Natural Science Foundation of China (61503204), Natural Science Foundation of Zhejiang (LY14F030004), Science and Technology Planning Project of Zhejiang (2015C31017), the Natural Science Foundation of NingBo (2016A610092).

## REFERENCES

- [1] Dennis R L, Byun D W, Novak J H, et al. The next generation of integrated air quality modeling: EPA's Models-3[J]. Atmospheric Environment, 1996, 30(12): 1925-1938.

- [2] Jang C.J. and N. B. Chang. Development and Applications of U.S. EPA's Regulatory Air Quality Modeling Systems, Submitted to J. Of Chinese Environment Engineering. 1999.
- [3] Science algorithms of the EPA Models-3 community multiscale air quality (CMAQ) modeling system [M]. Washington, DC, USA: US Environmental Protection Agency, Office of Research and Development, 1999.
- [4] Xinwang Nie, Yibai Wang, Shouxun Sun, et al. Application of Models-3/CMAQ to the numerical study of nuclear leakage in Japan's Fukushima nuclear disaster in March 2011 [J]. Meteorology, 2012, 38 (10): 1182-1188, In Chinese.
- [5] Shi C, Zhang B. Tropospheric NO<sub>2</sub> columns over Northeastern North America: Comparison of CMAQ model simulations with GOME satellite measurements [J]. Advances in Atmospheric Sciences, 2008, 25: 59-71.
- [6] Lijun Zhang. Study on Simulation and calibration of air quality in the Pearl River Delta region based on [D]. Models-3/CMAQ South China University of Technology, 2010, In Chinese.
- [7] Yibai Wang, Jianfang Fei, Xiaogang Huang. Preliminary study on the application of Models-3/CMAQ model to a strong sand dust weather in North China [J]. Meteorology, 2009, 35 (6): 46-53, In Chinese.
- [8] Yangfeng Wang, Hongchao Zuo, Ma Yanjun, et al. Numerical simulation study on air quality of Shenyang city by using Models-3 model system [J]. Journal of environmental science, 2007, 27 (3): 487-493, In Chinese.
- [9] Fang Ma, Litao Wang, Xuemei Pan. Simulation of atmospheric pollution in China based on [J]. MM5-Models-3/CMAQ Journal of Hebei University of Engineering: Natural Science Edition, 2010, 27 (4): 46-51, In Chinese.
- [10] Milionis A E, Davies T D. Regression and stochastic models for air pollution—I. Review, comments and suggestions [J]. Atmospheric Environment, 1994, 28(17): 2801-2810.
- [11] Abdul-Wahab S A, Bakheit C S, Al-Alawi S M. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations [J]. Environmental Modelling & Software, 2005, 20(10): 1263-1271.
- [12] Sohn S H, Oh S C, Jo B W, et al. Prediction of ozone formation based on neural network [J]. Journal of environmental engineering, 2000, 126(8): 688-696.
- [13] Yi J and Prybutok V R. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. Environmental Pollution. 1996, 92(3): 349-357.