Day-ahead Forecast of PM2.5 in Chengdu with Combination Model

–Using ARMA and SVR

Zhong Hongwei

June 22, 2020

Abstract

Air-quality forecasting is difficult and complex when using mechanism models. However, a combination forecasting model with both linear and nonlinear models may be appropriate to represent the complex behavior of a heterogeneous time series data set. In this paper, we try to propose some methodology to integrate the individual forecasting models of regular econometrics and that of machine learning and to forecast PM2.5 of Chengdu based on times series PM2.5 and panel data of other variables. In the model, day-ahead data are used to predict the next-day PM2.5, and the result shows that SVR helps ARMA to capture the daily fluctuation, improving the performance of single model.

Keywords: PM2.5 forecast; Combination model; ARMA; SVR

Contents

1 Introduction

It has been known for over sixty years that particulate matter air pollution harms human health, and the fine particles penetrate deeply into the lungs (Cobourn, 2010). As China's economic development enters a new stage, air quality has become one of the most concerned livelihood issues. Chengdu's air quality has generally improved since 2014, but we can still feel the impact of smog on people's work and life every winter. According to the Chengdu Municipal Bureau of Statistics, the city had 115 polluted days in 2016, including 44 days with moderate and heavy pollution. Around November 20 of the same year, heavy smog last for over a dozen days. During this period, Chengdu's air pollution surpassed that of Beijing and other heavily polluted cities, ranking third in the world. This alerts us that it is urgent to pay attention to the air quality regulation. As the primary pollutants affecting air quality in China, accurate prediction of PM2.5 concentration is of great significance for predicting heavy pollution weather, formulating emergency plan and initiating mechanism, and optimizing enterprises' plans to limit production.

The United States, Australia and other countries as early as the 20th century has carried out large-scale urban air quality forecast research (Tao and Zhang, 2015). In our country, since 2001 began to publish air quality forecast, forecast models are divided as mechanism models and non-mechanism models (Sun et al., 2017). The mechanism models input real-time interpolation and meteorologic pollution monitoring data, build the initial field and join the ground meteorological and pollution emission data. Through a complex network and deductive process, it releases the next 24-48h of the pollutants concentration prediction value (Xue and Fu, 2014). The model covers the complex physical and chemical processes from the generation to diffusion of pollutants, and its field construction is very difficult to initiate and the prediction process is complex. For the ordinary researchers who have not used the corresponding system and are lack of atmospheric physics knowledge, it is almost impossible to operate the system.

However, non-mechanism modeling in pollutant forecast helps scholars and other analyst break the barrier of complex theories and systems, facilitating the analysis in other fields. Non-mechanism model, using the historical data to determine the law of pollutant concentration change, mainly includes the statistical model and machine

learning model (Zhang et al., 2013). In these models, we select pollutants concentration, temperature, humidity, air pressure, wind speed, vapor pressure to predict PM2.5 concentrations (Yin et al., 2018). Non-mechanism model does not split the formation process and diffusion process of pollutants under the influence of meteorological conditions ,and does not separate the controllable production behavior and uncontrolled weather Phenomenon.

In addition, with pollutants concentration as the independent variable, we reckon PM2.5 concentration and SO_2, NO_x and other pollutants have direct correlation, ignoring the exact source of PM2.5 like industrial production, fuel combustion, transportation, dust. We also did not pay attention to the nature of the cause of air pollution of human behavior. Forecasted values of Ohter pollutants are put into the model as variables, which plays an important role in the prediction (Xue and Fu, 2014).

In this paper, we try to use non-mechanism models combining linear and non-linear models to forecast the PM2.5 of Chengdu with climate data and other pollutants' data. Based on the present study, we take special weather characteristics of Chengdu into consideration to train the model and compare the forecast performance with other single models.

2 Literature Review

With non-linear models, many studies indicate that an integration of different models can improve predictive performance, especially when the models in the ensemble are quite different (Babu and Reddy, 2014; Yan et al., 2016; Gairaa et al., 2016). The combination model of ARIMA (Autoregressive Integrated Moving Average) and radial basis function network (RBFN) performed better than ARIMA or RBFN alone, which was tested for the prediction of monthly groundwater level (Yan et al., 2016). Babu and Reddy (2014) explored the nature of volatility using a moving-average filter, and then an ARIMA and an ANN (Artificial Neural Network) model were suitably applied. Zhang et al. (2013) employed a hybrid ARIMA and ANN to contrast a model for predicting three different data sets. Empirical results suggest that the hybrid model outperforms greatly each component model used in isolation. Wang et al. (2012) presented a novel text mining approach by combining ARIMA and SVR (Support Vector

Regression) to forecast quarterly ROEs (Return of Equity) of six security companies. Zhu and Wei (2013) constructed a novel hybrid methodology that exploits the unique strength of the ARIMA.

The combination models are very suitable for dealing with PM2.5 concentrations time series data consisting of one-dimension series data and multi-dimension panel data. Traditionally, the ARMA method is used for estimating the one-dimension series data. The ANN model or SVM(Support Vector Machine) model is selected to process the panel data. It is reasonable that the forecast values of the different model should be combined in one way or another.

3 Data

3.1 Area description

Chengdu, located in *the Sichuan Basin* (shown in Figure 1), has a subtropical monsoon climate with obvious seasonal and regional characteristics of human production and living activities. Therefore, haze pollution varies significantly with time. In this study, the air quality data from Jan 1, 2016 to April 13, 2020 and the meteorological data from Jan 1, 2016 to April 13, 2020 were obtained by selecting the data of nearly five years and selecting the observation time reasonably. The reasion we choose the timespan is that Chengdu experience the most polluted period in 2016. We tend to include this period to make our model more general and sensitive to extreme values. Meanwhile, the missing data were filled to analyze the temporal distribution characteristics of pollutant concentration in Chengdu.

3.2 Data overview

The air quality data and the meteorological data used in this paper come from Blog of Wang X. L. [1]. In air quality data, the indexes of $PM2.5$, $PM10$, CO, NO_2, O_3 and SO_2 are mainly included in the urban area of Chengdu and Wenjiang stations. The selected period is from 1/1/2016 to 4/17/2020. Among them, all air quality index data are daily average data, calculated according to the arithmetic average of hourly data[2]. Meteoro-

[1] https://quotsoft.net/air/

logical data include air temperature, dew point temperature, wind speed and wind direction (AIR_TEMPERATURE, DEW_TEMPERATURE, WIND_SPEED and WIND_DIRECTION). The selected period is January 1, 2016 to April 17, 2020.

Table 1 and Table 2 describe the statistics characteristics of processed data. Notations in Table 2 are shown in Table 3. In the model, we use the averge data of two stations for we focus on the temporal distribution characteristics. From 1/1/2016 to 4/17/2020, there are 1569 entries with at most 22 missing values in some features. Observing the smooth fluctuations for all features, we simply fill the missing values with that of one day before the missing one. For example, the change of $PM2.5$ is regular and smooth as is shown in Figure 2.

According to the analysis of air quality index from 2016 to 2019 in Figure 3, it can be concluded that winter is the season with frequent haze weather, while summer haze weather is less likely to occur, and spring is more prone to haze pollution than autumn. There are many reasons for this seasonal difference, but it is mainly due to geographical factors and human social activities. Chengdu is located in *The Sichuan Basin*. In winter, the weather is wet and cold, and the air humidity is relatively high. Therefore, the combination of water molecules and small particles in the air is easy to form haze pollution. But in the summer, long sunshine duration, high surface temperature and low air humidity make it difficult to form haze weather (Yu, 2019). On the other hand, a sharp increase in burning in autumn and winter, coupled with an increase in the number of fireworks and bamboo sets off during the Spring Festival, has exacerbated smog pollution.

4 Models and Methodology

4.1 Model

4.1.1 ARMA model

As a time series model, ARMA model has been widely concerned and applied, especially in finance econometrics and electric power. ARMA model is composed of

[2]Note that climate daily data of two stations are calculated separately for they have diffrent frequency and there exist missing values at diffrent periods.

autoregression model and moving average model, and the general expression is

$$\begin{aligned} u_t = c + \phi_1 u_{t-1} + \phi_2 u_{t-2} + ... + \phi_p u_{t-p} \\ +\varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + ...\phi_q \varepsilon_{t-q}, \\ (t = 1, 2, ..., T) \end{aligned} \tag{1}$$

where c is constant and $\phi_1, \phi_2, ..., \phi_p$ are coefficients of autoregression model, p is the order. ε_t is a white noise series with σ^2 variance, $\phi_1, \phi_2, ..., \phi_q$ are coefficients of moving average model. This model (also named as ARMA(p, q)) requires stationary series as the dependent variable. Therefore, stationarity test and difference are introduced before using ARMA.

4.1.2 Supportive vector machines

The support vector machines (SVMs) were proposed by Vapnik (1995). Based on the structured risk minimization (SRM) principle, SVMs seek to minimize an upper bound of the generalization error instead of the empirical error as in other neural networks. Additionally, the SVMs models generate the regress function by applying a set of high dimensional linear functions (Pai and Lin, 2005). The SVM regression function is formulated as follows

$$y = w\phi(x) + b \tag{2}$$

where $\phi(x)$ is called the feature, which is nonlinear mapped from the input space x.The coefficients w and b are estimated by minimizing

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|w\|^2, \tag{3}$$

$$L_\varepsilon(d, y) = \begin{cases} \mid d - y \mid -\varepsilon & \mid d - y \mid \geq \varepsilon \\ 0 & others, \end{cases} \tag{4}$$

where both C and ε are prescribed parameters. The first term $L_\varepsilon(d, y)$ is called the ε-intensive loss function. The d_i is the actual PM2.5 in the ith period. This function indicates that errors below ε are not penalized. The term $C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i)$ is the empirical error. The second term, $\frac{1}{2}\|w\|^2$, measures the flatness of the function. C evaluates the trade-off between the empirical risk and the flatness of the model. Introducing the positive slack variables ζ and ζ^*, which represent the distance from the actual values to the corresponding boundary values of ε-tube. Eq.(3) is transformed to the following

constrained formation:

Minimize:

$$R(w,\zeta,\zeta^*) = \frac{1}{2}ww^T + C^*\left(\sum_{i=1}^{N}(\zeta_i + \zeta_i^*)\right) \tag{5}$$

Subject to:

$$w\phi(x_i) + b_i - d_i \leq \varepsilon + \zeta_i^*, \tag{6}$$

$$d_i - w\phi(x_i) - b_i \leq \varepsilon + \zeta_i, \tag{7}$$

$$\zeta_i, \zeta_i^* \geq 0, \tag{8}$$

$i = 1, 2, ..., N$.

Finally, introducing Lagrangian multipliers and maximizing the dual function of Eq.(5) changes Eq.(5) to the following form:

$$\begin{aligned} R(\alpha_i - \alpha_i^*) = \sum_{i=1}^{N} d_i(\alpha_i - \alpha_i^*) - \varepsilon\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) \\ -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*) \\ *(\alpha_j - \alpha_j^*)K(x_i, x_j) \end{aligned} \tag{9}$$

with the constraints

$$\sum_{i=1}^{N}(\alpha_i - \alpha_i^* = 0), \tag{10}$$

$$0 \leq \alpha_i \leq C, \tag{11}$$

$$0 \leq \alpha_i^* \leq C, \tag{12}$$

$i = 1, 2, ...N$.

In Eq.(9), α_i and α_i^* are called Lagrangian multipliers. They satisfy the equalities,

$$\begin{aligned} \alpha_i * \alpha_i^* = 0, \\ f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(x, x_i) + b. \end{aligned} \tag{13}$$

Here, $K(x, x_i)$ is called the kernel function. The valueof the kernel is equal to the inner product of two vectors x_i and x_j in the feature space $\phi(x_j)$, such that $K(x_i, x_j) = \phi(x_i) * \phi(x_j)$. Any function that satisfying Mercer's condition (Vapnik, 1995) can be used as the Kernel function. The Gaussian kernel function

$$K(x_i, x_j) = exp(-\|x_i - x_j\|^2/(2\sigma^2)) \tag{14}$$

is specified in this study. The SVMs were employed to estimate the nonlinear behavior of the forecasting data set because Gaussian kernels tend to give good performance under general smoothness assumptions.

4.2 Methodology

Both the ARMA and the SVMs models have different capabilities to capture data characteristics in linear or nonlinear domains, so the combination model proposed in this study is composed of the ARMA component and the SVMs component. Thus, the model can model linear and nonlinear patterns with improved overall forecasting performance. The general idea is shown in Figure 4.

Specifically, the final model (Z_t) can then be represented as follows

$$Z_t = Y_t + N_t, \tag{15}$$

where Y_t is the linear part and N_t is the nonlinear part of the model. Both Y_t and N_t are estimated from the data set. $\tilde{Y_t}$ is the forecast value of the ARMA model at time t. Let ε_t represent the residual at time t as obtained from the ARMA model; then

$$\begin{aligned} \varepsilon_t = f(&\varepsilon_{t-1}, \varepsilon_{t-2}, ..., \varepsilon_{t-n}, \\ &d_{1,t-1}, ..., d_{1,t-n}, d_{2,t-1}, ..., d_{2,t-n}, ..., d_{k,t-n}) + \Delta_t, \end{aligned} \tag{16}$$

where f is a nonlinear function modeled by the SVMs, Δ_t is the random error and $d_{i,j}$ is the ith variable at time j. Therefore, the combined forecast is

$$\tilde{Z_t} = \tilde{Y_t} + \tilde{N_t}. \tag{17}$$

where $\tilde{N_t}$ is the forecast value of Eq.(16).

5 Result

5.1 Result of ARMA model

We use Eviews 10.0 to build the ARMA model and to capture the linear part of data from 1/1/2016 to 12/31/2019. And then data from 1/1/2020 to 4/17/2020 are im-

ported as the out-of-sample data. Firstly, we test the stationarity and decide to use the first-level difference of logarithm of $PM2.5$ with $p = 4$ and $q = 4$ in the model based on AIC value. The result is shown in Figure 5. In ARMA, we use static prediction that every value we predict is calculated with true values of the last four days. Coefficients and other parameters are presented in Table 4. It is obvious that ARMA can accomplish the forecast rather well. The results of ARMA model show that when the model is applied to forecast the $PM2.5$ concentrations, and the forecasted values enjoy a good agreement with the actual values. A conclusion can be drawn that ARMA model has successfully learned the linear part of the data. However, we can also notice that prediction of ARMA model is relarively more *s*oomth than the actual values, which suggests that larger disturbance cannot be explained by ARMA model. There still exists room for improvement to capture the non-linear part.

5.2 Result of SVR model

As a multi-dimension and non-linear learning machine, SVR formulates the regression relationship of the input data and the forecasting output. The performance of the SVM model is decided to a great extent by the selected input variables of the model. In this sense, it is reasonable to employ the factors highly relevant to the forecasting target. The SO_2, NO_2, C_O, O_3 concentrations, AQI, temperature, dew point temperature, wind speed and wind direction are selected as the input variables of the forecasting model related to forecasting target, affecting the PM2.5 concentrations more or less. We standardized the climate and pollution data and randomly separate the data as training set and test set. Note that we all use day-ahead data as input. For example, training data of 1/1/2016 corresponds to the $PM2.5$ of 1/2/2016. Key parameters and forecast performance are shown in Table 6 and Figure 5, respectively.

5.3 Result of Combination model

Different from Section 5.2, here we use day-ahead climate and pollutant data as input to forecast the error from ARMA forecast. But our training data exclude lag term of error term as shown in Eq. (16) because we found that the characteristics of residual has been caputured by ARMA model. And the rest of the residuals are nearly white

noise, the correlation of residual at different time is too little to be used as features. To be specific, in practice, data from 1/1/2016 to 12/31/2019 are used to train the model and adjust the hyper-parameters through cross validation. For example, input data of 1/1/2016 corresponds to the error of 1/2/2016 where the true error comes from the ARMA model. This ensures that the out-of-sample prediction would not include the information in the training data, thus enhancing the credibility of the forecast perfomance.

Parameters are presented in Table 6 and result is in Figure 5. We introduce three indicators MAE, MSE and R^2 to quantify the performance of three models, which is shown in Table 5. Apparently, combination model fits the out-of-sample data better than the other two models for its MSE decreases largely, compared with the other two models. In addition, we can find that SVR helps ARMA model to improve the prediciton especially in *wave crest and trough*, which is called *non-linear shock* as shown in Figure 5.

6 Conclusion and further discussion

6.1 Conclusion

In this paper, pollutant and climate daily data were collected from Chengdu on the southwest of China. The combined model is applied to forecast the next-day $PM2.5$. The results demonstrate that besides meterological factors, $PM2.5$ concentrations show a significant correlation with air polluttants factors and the results clearly show that the proposed method is more accurate than the other forecast methods.

6.2 Further discussion

Though the result has suggested the success of the methodology, much work can still be accomplished, compared with other papers.

Using other non-linear models to compare the performance. As is shown in Section 2, other algorithms are available to be combnined and trained. Due to lack of time, we view SVR as a relatively better model for several papers only present the results of SVR as well (Zhu and Wei, 2013; Pai and Lin, 2005; Wang et al., 2012).

Collecting data from more stations to analyze the distinctions of each other. As Tan et al. (2010) and Song et al. (2018) suggest, data of different stations and different periods can be estimated separately because of diversed characteristics of data in different time and space. In this paper we only considered the time-variant characteristics and take the average number of two stations due to lack of data from more stations.

Testing the heteroscedasticity of the combination. Wang et al. (2017) raised the heteroscedasticity problem in the combination. Because in the pure addition of residual and linear forecast, we take the equal weights of two models and the significance of the coefficients as granted[3]. However, series may appear heteroscedasticity because the series can be influenced by exogeneous factors, which requires the GLS (Generalized Least Square) to adjust the weights and the test for significance of the coefficients. This indicates that there exists space for futher improvement and modification for the combination model.

[3]In Eq. (17), we view two components as equal weight in the simulation and did not test the significance of coefficients.

References

Babu, C.N., Reddy, B.E., 2014. A moving-average filter based hybrid arimaann model for forecasting time series data. Applied Soft Computing 23, 27 – 38. URL: http://www.sciencedirect.com/science/article/pii/S1568494614002555, doi:https://doi.org/10.1016/j.asoc.2014.05.028.

Cobourn, W.G., 2010. An enhanced pm2. 5 air quality forecast model based on nonlinear regression and back-trajectory concentrations. Atmospheric Environment 44, 3015–3023.

Gairaa, K., Khellaf, A., Messlem, Y., Chellali, F., 2016. Estimation of the daily global solar radiation based on BoxJenkins and ANN models: A combined approach. Renewable and Sustainable Energy Reviews 57, 238–249. URL: http://www.sciencedirect.com/science/article/pii/S136403211501494X, doi:https://doi.org/10.1016/j.rser.2015.12.111.

Pai, P.F., Lin, C.S., 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega 33, 497–505. doi:10.1016/j.omega.2004.07.024.

Song, G., Guo, X., Yang, X., Liu, S., 2018. ARIMA-SVM combination prediction of PM2.5 concentration in Shenyang. China Environment Science 38, 4031–4039. doi:10.19674/j.cnki.issn1000-6923.2018.0445.

Sun, B.L., Sun, H., N., Z.C., 2017. Forecast of air pollutant concentrations by bp neural network. Acta Scientiae Circumstantiae 37, 1864–1871.

Tan, Z., Zhang, J., Wang, J., Xu, J., 2010. Day-ahead electricity price forecasting using wavelet transform combined with ARIMA and GARCH models. Applied Energy 87, 3606–3610. URL: http://dx.doi.org/10.1016/j.apenergy.2010.05.012, doi:10.1016/j.apenergy.2010.05.012.

Tao, J., Zhang, L., 2015. Control of PM2.5 in Guangzhou during the 16th Asian Games period: Implication for hazy weather prevention. Science of The Total Environment 508, 57–66. doi:https://doi.org/10.1016/j.scitotenv.2014.11.074.

Vapnik, V.N., 1995. The nature of statistical learning theory .

Wang, B., Huang, H., Wang, X., 2012. A novel text mining approach to financial time series forecasting. Neurocomputing 83, 136–145. URL: http://www.

sciencedirect.com/science/article/pii/S0925231211007302, doi:https://doi.org/10.1016/j.neucom.2011.12.013.

Wang, P., Zhang, H., Qin, Z., Zhang, G., 2017. A novel hybrid-Garch model based on ARIMA and SVM for PM2.5 concentrations forecasting. Atmospheric Pollution Research 8, 850–860. URL: http://dx.doi.org/10.1016/j.apr.2017.01.003, doi:10.1016/j.apr.2017.01.003.

Xue, W., Fu, F., 2014. Numerical study on the characteristics of regional transport of pm2.5 in china. China Environmental Science 34, 1361–1368.

Yan, Q., Tang, S., Gabriele, S., Wu, J., 2016. Media coverage and hospital notifications: Correlation analysis and optimal media impact duration to manage a pandemic. Journal of Theoretical Biology 390, 1–13. URL: http://www.sciencedirect.com/science/article/pii/S0022519315005366, doi:https://doi.org/10.1016/j.jtbi.2015.11.002.

Yin, J.G., Peng, F., Xie, L.K., 2018. The study on the prediction of the pm2.5 concentration based on model of the least squares support vector regression under wavelet decomposition and adaptive multiple layer residuals correction. Acta Scientiae Circumstantiae 38, 2090–2098.

Yu, H.S., 2019. The space-time distribution, prediction and diffusion analysis of PM2.5 in Chengdu.

Zhang, Y., Guanhong, M., Jianshi, Y., Jingli, W., 2013. Multivariate statistical analysis and prediction of factors influencing pm2.5. Resource conservation and environmental protection 11, 135–136.

Zhu, B., Wei, Y., 2013. Carbon price forecasting with a novel hybrid ARIMA and least squares support vector machines methodology. Omega 41, 517–524. URL: http://www.sciencedirect.com/science/article/pii/S0305048312001004, doi:https://doi.org/10.1016/j.omega.2012.06.005.

Table 1: Descriptive statistics of processed air quality data

	AQI	**CO**(mg/m^3)	**NO_2**($\mu g/m^3$)	**O_3**($\mu g/m^3$)	**PM2.5**($\mu g/m^3$)	**PM10** ($\mu g/m^3$)	**SO_2** ($\mu g/m^3$)
count	1569	1569	1569	1569	1569	1569	1569
mean	77.77	0.94	47.89	50.11	53.19	85.55	10.21
std	44.11	0.31	17.05	28.69	36.80	55.34	4.87
min	14.29	0.37	7.79	5.08	4.96	9.38	3.67
25%	48.05	0.73	35.79	27.46	26.83	46.00	6.83
50%	66.38	0.87	45.88	46.38	43.83	71.95	8.96
75%	95.50	1.07	58.71	68.83	68.58	109.17	12.33
max	397.79	2.85	120.42	147.42	329.62	486.12	32.92

Table 2: Descriptive statistics of processed climate data

	wj_air($^\circ C$)	**wj_dew**($^\circ C$)	**wj_wind_dir**	**wj_wind_spd**(m/s)	**cd_air**($^\circ C$)	**cd_dew**($^\circ C$)	**cd_wind_dir**	**cd_wind_spd**(m/s)
count	1569	1569	1569	1569	1569	1569	1569	1569
mean	16.38	12.6	196.98	1.47	17.86	13.17	117.63	1.53
std	7.35	7.5	57.82	0.42	7.33	7.33	54.45	0.75
min	-0.69	-7.45	22.5	0.5	1.38	-6.96	0	0.08
25%	9.58	6.24	161.25	1.2	11	6.96	75.62	1.04
50%	16.56	13.03	203.75	1.38	18.12	13.71	114.71	1.38
75%	22.88	19.25	240	1.62	24.25	19.62	154.67	1.83
max	30.15	26.51	326.25	4.25	32.5	25.75	351.25	5.83

Table 3: Climate data notation

Notation	**Specification**
wj_air	Air temperature in Wenjiang station
wj_dew	Dew point temperature in Wenjiang station
wj_wind_dir	Wind direction in Wenjiang station[4]
wj_wind_spd	Wind peed in Wenjiang station
cd_air	Air temperature in Chengdu station
cd_dew	Dew point temperature in Chengdu station
cd_wind_dir	Wind direction in Chengdu station
cd_wind_spd	Wind speed of Chengdu station

[4]We denote north as 0 and number increases as the direction changes clockwise.

Table 4: Coefficients and other statistics of ARMA(4,4) on DLOG(PM2.5)

Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**Prob.**
C	-0.000158	0.002303	-0.068449	0.9454
AR(1)	0.498239	0.229189	2.173924	0.0299
AR(2)	0.409240	0.273912	1.494054	0.1354
AR(3)	0.306011	0.159717	1.915955	0.0556
AR(4)	-0.326527	0.084998	-3.841586	0.0001
MA(1)	-0.778120	0.228250	-3.409069	0.0007
MA(2)	-0.619993	0.331178	-1.872083	0.0614
MA(3)	-0.122037	0.182377	-0.669145	0.5035
MA(4)	0.540284	0.130915	4.126990	0.0000
SIGMASQ	0.173268	0.006247	27.73734	0.0000

Table 5: Performance comparison with data from 1/1/2020 to 4/17/2020

Model	**MAE**	**MSE**	R^2
ARMA	14.48	327.05	0.45
SVR	12.16	310.67	0.60
ARMA-SVR	11.01	220.39	0.55

Table 6: Key parameters of SVR models

	C	**gamma**
SVR	13	0.002
Combine_SVR	4	0.007

Figure 1: Location of Chengdu

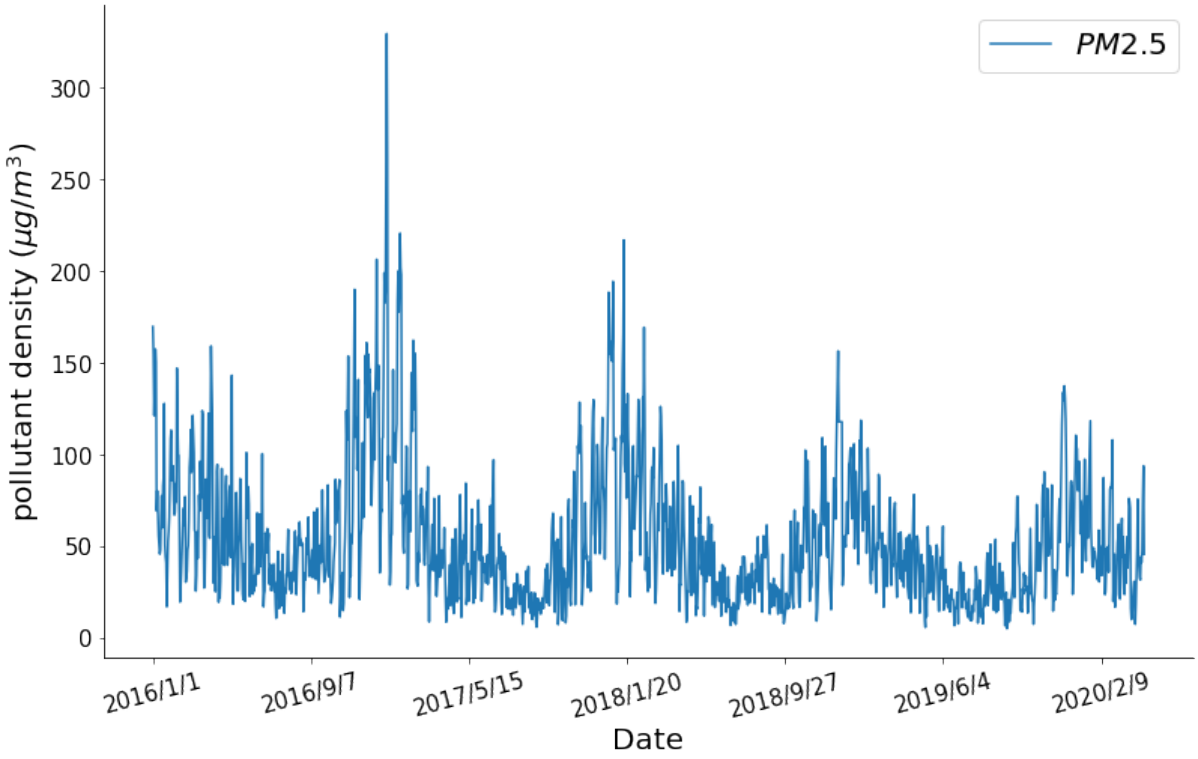

Figure 2: Daily series of $PM2.5$

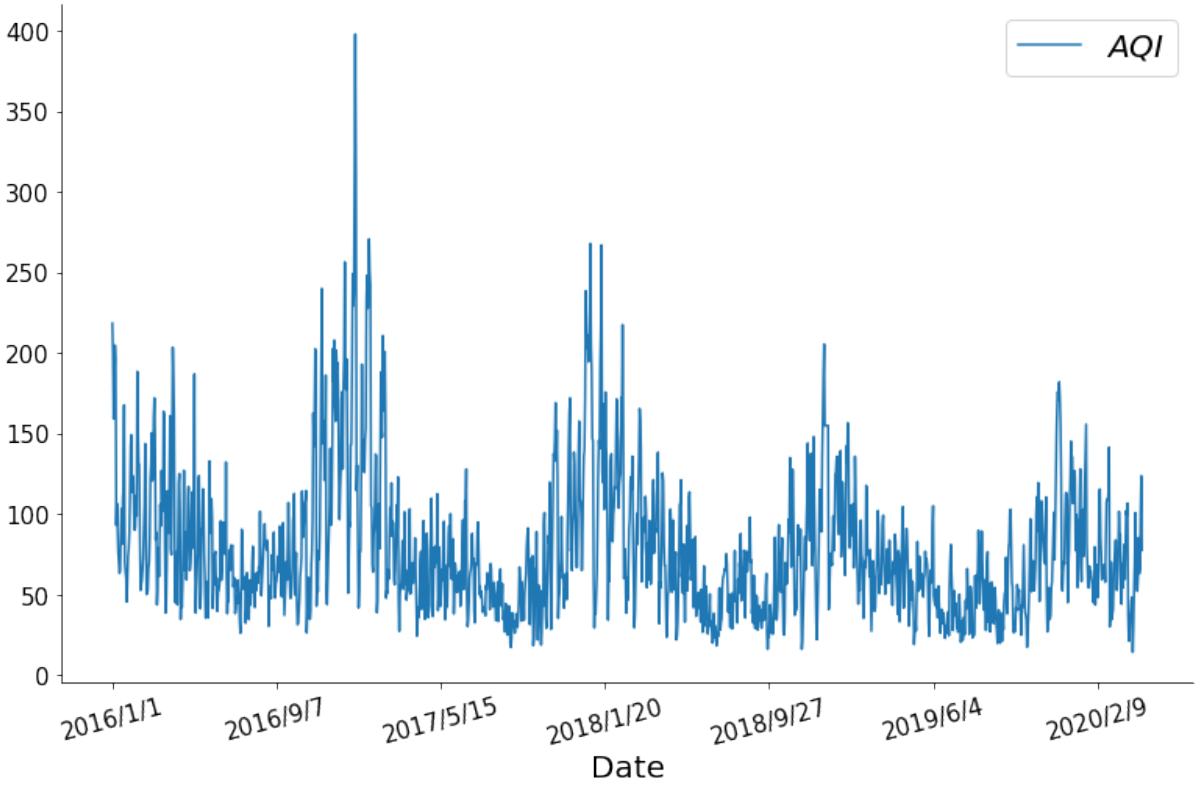

Figure 3: Daily series of AQI

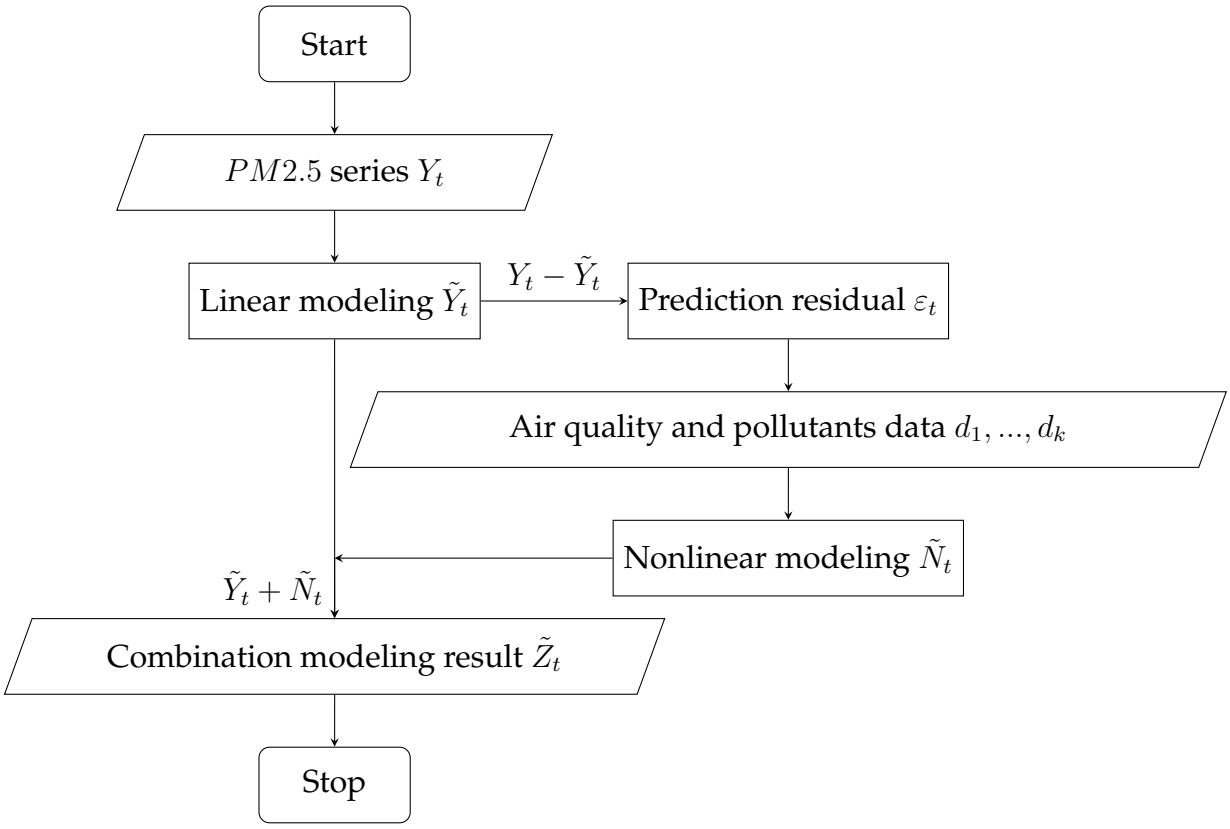

Figure 4: Flow chart of methodology

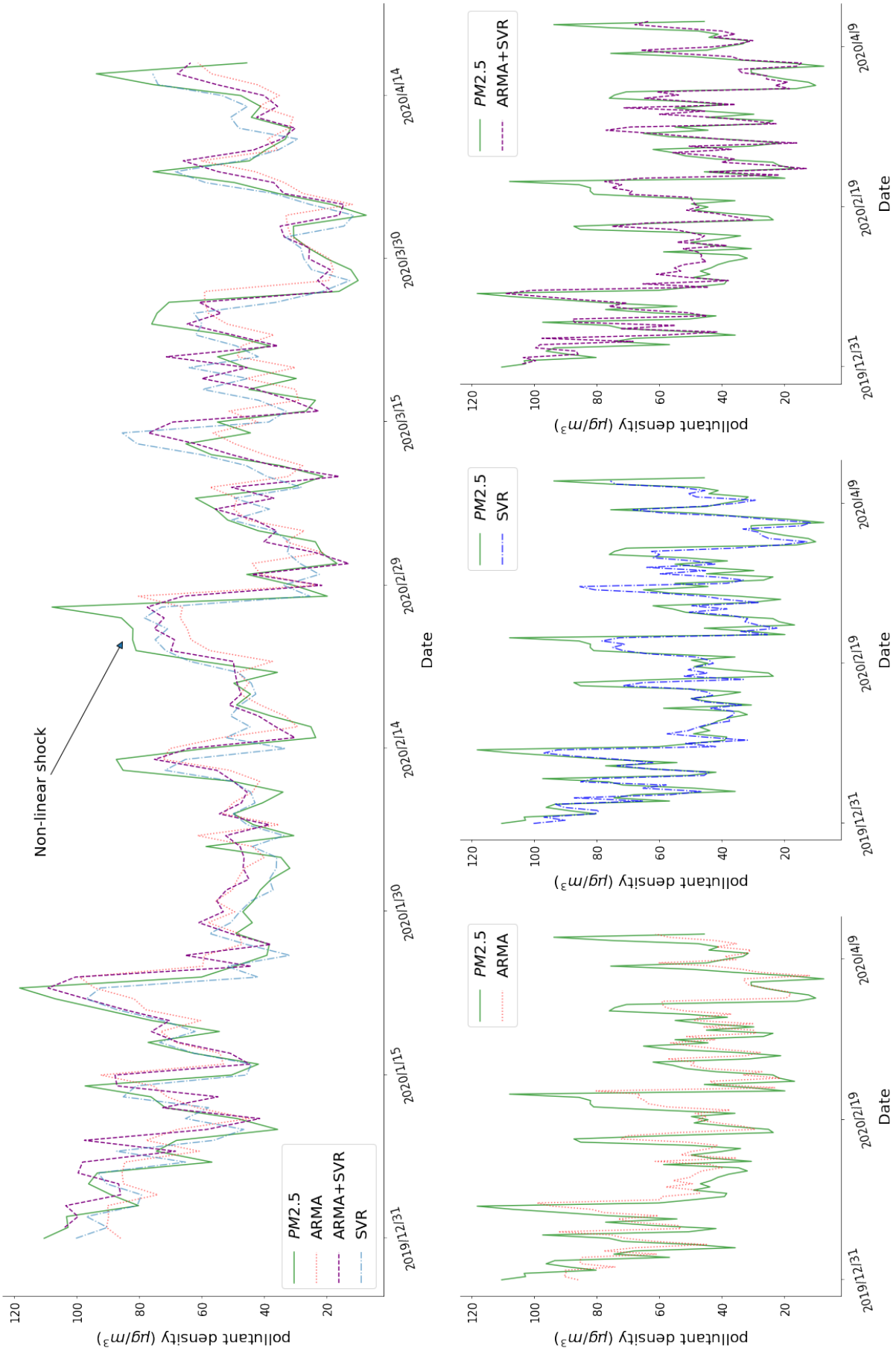

Figure 5: Out-of-sample prediction result