Atmospheric Environment 267 (2021) 118737

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Relative importance of meteorological variables on air quality and role of boundary layer height

Yaxuan Huang [a], Bin Guo [b,*], Haoxuan Sun [c], Huijie Liu [d], Song Xi Chen [e,**]

[a] Yuanpei College, Peking University, Beijing, 100871, China
[b] Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, 611130, China
[c] Center for Data Science, Peking University, Beijing, 100871, China
[d] The Experimental High School Attached To Beijing Normal University, Beijing, 100032, China
[e] School of Mathematical Sciences, Guanghua School of Management and Center for Statistical Science, Peking University, Beijing, 100871, China

HIGHLIGHTS

- Stable orders of meteorological effects on six air pollutants in North China.
- Boundary layer height's effects can be modeled by the surface meteorological variables.
- $PM_{2.5}$, PM_{10}, SO_2 and CO shared a common order of variable importance.

ARTICLE INFO

Keywords:
Air quality assessment
Statistical learning
Meteorological effects
Variable selection

ABSTRACT

To gain insight on the meteorological effects of air pollution, we study the relative importance of surface meteorological variables and boundary layer height (BLH) on six major air pollutants according to the orders of variables being selected in a forward variable selection algorithm. It is found that there was a strong agreement in the orders of relative importance for the major pollutants among six major cities in North China, which implies regularities in the meteorological processes of air pollution in that region. In particular, $PM_{2.5}$, PM_{10}, SO_2 and CO shared a common variable importance order and were mostly impacted by the dew point temperature and air pressure, while the NO_2 and O_3 were mostly influenced by the boundary layer height (BLH) and temperature, respectively. We evaluate the impacts of BLH on the pollution levels given the surface meteorological variables. It is found that BLH can be well modeled by the surface meteorological variables. Thus, air quality assessment without using BLH would also produce adequate results.

1. Introduction

Air pollution is an enduring challenge encountered by many countries including China. A substantial part of China has experienced severe air pollution in the last two decades. Epidemiological studies have shown a high correlation between exposure to air pollutants and human health (Donaldson et al., 1998; Schwartz, 2000; Pope III et al., 2002; Chen et al., 2013); research has also found that air pollution may cause social and economic problems (Xie et al., 2016; Zhang et al., 2017b; Feng et al., 2019).

Mitigation and control for air pollution are needed in these countries, which requires understanding the mechanisms of the air pollution with respect to its main drivers.

Air pollution is impacted by both the underlying emission and the meteorological condition. As one of the two main drivers, it is important to understand the meteorological aspects of air pollution based on the air quality and meteorological data which are quite available in this era of data.

The relationship between meteorological factors and the pollutant concentration has be considered in a range of studies. Tai et al. (2010) studied the correlation of $PM_{2.5}$ and its components with meteorological variables in the United States by applying multiple linear regression and

* Corresponding author. Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, 611130, China.
** Corresponding author. School of Mathematical Sciences and Guanghua School of Management, Peking University, Beijing, 100871, China.
E-mail addresses: huangyaxuan@pku.edu.cn (Y. Huang), guobin@swufe.edu.cn (B. Guo), hxsun@pku.edu.cn (H. Sun), liuhuijie16@pku.edu.cn (H. Liu), songxichen@pku.edu.cn (S.X. Chen).

https://doi.org/10.1016/j.atmosenv.2021.118737
Received 22 July 2021; Received in revised form 15 September 2021; Accepted 15 September 2021
Available online 2 October 2021
1352-2310/© 2021 Elsevier Ltd. All rights reserved.

found meteorological variables, can explain up to 50% of the variation of $PM_{2.5}$. Liang et al. (2015, 2016) revealed that $PM_{2.5}$ concentration was highly influenced by meteorological conditions in Beijing and found that 75% of the hourly variation of $PM_{2.5}$ was driven by meteorological factors. Shen et al. (2018) studied the monthly $PM_{2.5}$ levels in Beijing's winters during 2010–2017 and found that 81% of the variation can be explained by the first principal component of 850 hPa meridian wind velocity and the relative humidity. Vu et al. (2019) applied the random forest to assess the impact of clean air action on six major pollutants in Beijing by decoupling the impact of meteorology on ambient air quality.

Stafoggia et al. (2019) estimated the daily PM_{10} and $PM_{2.5}$ concentrations in Italy during 2013–2015 using the random forest with similar meteorological variables as well as the planetary boundary layer height (BLH). Choubin et al. (2020) studied the hazard prediction of PM_{10} in Barcelona Province, Spain and found seven critical features among a total of thirteen features, which included topographic wetness index and precipitation.

In an attempt to construct timely statistical measures to reflect the underlying emission based on the air quality data, Liang et al. (2015) and Zhang et al. (2020) proposed meteorologically adjusted pollution measures (average and quantile) under the meteorological baseline distribution constructed from the readily available surface meteorological data. They showed that the meteorologically adjusted pollution measures can reflect the underlying changes in the emission, while the unadjusted raw average pollution levels are subject to meteorological confounding and cannot objectively reflect the underlying emission.

Despite the wide range of studies on the meteorological effects of the air quality motivated from different contexts, there is still a lack of in-depth understanding on the relative importance of the meteorological variables on different air pollutant species. Specifically, given the inter-dependence among the meteorological variables, is there an established order of relative meteorological importance for an air pollutant in a region? An established order of relative importance would imply a common mechanism in the meteorological process of the air pollutant, which would be useful for the modeling and prediction for the air pollutant.

This paper tries to answer the above question by analysing air quality data in six major northern China cities. The orders of the variable importance are determined by the ranks of variables being selected in a step-wise forward selection procedure (Hastie et al., 2009) in the regression of pollution concentration on the meteorological variables. The forward variable selection orders take into account the inter-dependence among the meteorological variables relative to the target pollutant. The study finds that there was a strong agreement in the orders of meteorological variables in affecting the pollution concentration among the six major cities in North China.

In particular, $PM_{2.5}$, PM_{10}, SO_2 and CO shared a common order of meteorological importance led by the dew point temperature, while NO_2 and O_3 were mostly influenced by the boundary layer height (BLH) and air temperature, respectively. These suggest there is regularity in the meteorological processes for the six species of air pollutants in North China.

The planetary boundary layer height (BLH) defines the vertical dispersion property and is known to be influential to the ground-level concentration of air pollutants. Studies have used BLH as a key factor in the formation and evolution of heavy air pollution (Liu et al., 2013; Quan et al., 2014; Tang et al., 2016; Miao and Liu, 2019). A negative correlation between BLH and pollution concentration from observed data was found in Zhang et al. (2009); Miao et al. (2017); Gui et al. (2019). Xiang et al. (2019) revealed a significant negative correlation between BLH and $PM_{2.5}$ in Beijing during the winter heavy pollution. Miao et al. (2021) found that the relationships between planetary boundary layer and $PM_{2.5}$ and O_3 in the summer of Beijing and Shanghai were different with heavy pollution being associated with lower BLH in Beijing, but higher BLH in Shanghai. That higher pollution being associated with higher BLH in Shanghai was due to confounding of transported pollution from warmer inland under westerly wind, which indicates the importance of treating the multivariate meteorological variables collectively.

However, unlike the surface meteorological variables which are readily available, there is a delay of around three months in the BLH data assimilated by the European Centre for Medium-Range Weather Forecasts (ECMWF). There are two pending questions regarding the role of BLH in the air pollution processes. One is how much information contained in BLH can be explained by the surface meteorological variables; and the other is the impacts of BLH on the meteorological adjusted average pollution levels. The study finds that, despite a significant negative correlation with the pollutants, BLH was not highly ranked for the five non-NO_2 species after considering the surface variables in the modeling.

Furthermore, more than 77% of the variation in the BLH can be explained by the surface variables in the six cities. Even for NO_2, where BLH was the most important variable, the effects of BLH on NO_2 was less than 1.24% of the meteorologically adjusted average concentration. Therefore, using only the surface meteorological variables for air quality assessment would produce adequate results.

The paper is organized as follows. Section 2 outlines the data and the study region. Research methods are outlined in Section 3, which include the nonparametric regression, the forward variable selection and the meteorological adjustment. Section 4 reports results of the analysis. Additional information, extra tables and figures are provided in the supplementary materials (SM).

2. Study region and data

The study region contains six major cities in the North China, which include two mega-cities of Beijing and Tianjin, and four cities: Shijiazhuang, Jinan, Zhengzhou and Taiyuan, which are the provincial capitals of Hebei, Shandong, Henan and Shanxi provinces, respectively. The region encompassed by Beijing, Tianjin and the four provinces has been the main battle ground for the air pollution mitigation campaign in China; see Fig. S1 in SM for the geographical location of the six cities. The time range of the study was from March 1, 2013 to February 28, 2021, which contained 8 seasonal years from spring to winter. The unit of our study is season, consisting of spring from March to May, summer from June to August, fall from September to November, and winter from December to February next year.

The air quality data consisted of hourly concentrations of air pollutants ($PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3) from 53 monitoring sites, which are directly managed by the Ministry of Ecology and Environment (MEE). For ozone, we only focused on the period from noon to 7pm (8-h O_3), when the concentrations tended to be the highest. The 53 air quality sites consisted of 11 sites in Beijing and Tianjin, 7 in Shijiazhuang, 8 from Jinan, Zhengzhou and Taiyuan, respectively.

The meteorological data contained hourly surface (2 m) meteorological observations of six variables: air temperature (TEMP) and pressure (PRES), dew point temperature (DEWP), wind direction (WD), cumulative wind speed (IWS), and cumulative precipitation (IPREC). The wind directions were grouped into five broad directions: Northwesterly (NW), Northeasterly (NE), Southwesterly (SW), Southeasterly (SE) and CV, where CV was for the calm (wind speed less than 0.5 m/s) and variable wind. The cumulative wind speed was the summation of the wind speed since a wind direction was established and was reset to zero when there was a change in the direction.

Similarly, the cumulative precipitation referred to the sum of precipitation since the hour when it rained and was reset to zero when there was an hour without precipitation. Data of the above variables were from 12 weather observing stations from China Meteorological Administration (CMA), where 5 of them were located in Beijing, 3 in Tianjin and 1 each in other four cities.

We also considered the reanalyzed BLH data from the ERA5 data set of ECMWF, which contained hourly BLH at a grid size of $0.25^\circ \times 0.25^\circ$.

We matched each CMA site to the nearest grid of the ERA5 data for BLH as well as to the nearest air quality site. The time range of surface meteorological and BLH data was from March 1, 2011 to February 28, 2021. Here, two more years of meteorological data were employed to construct the meteorological baseline in the meteorological adjustment outlined in Section 3.3. The locations of the 53 air quality sites and the meteorological stations are displayed in Fig. S2 in SM.

In additional to the reanalyzed BLH data from ECMWF, there are other approaches to estimate BLH, which include the radiosonde approach by NOAA (Durre and Yin, 2008) and China's CMA (Guo et al., 2016). The BLH from the radiosonde observations have usually two (sometimes four in summer) observations per day. Seidel et al. (2012) analysed BLH measurements from four approaches over the United States and Europe, and found it was suitable to estimate BLH based on the bulk Richardson number (Vogelezang and Holtslag, 1996), which is the commonly used approach adopted by both the radiosonde observation and the reanalysis approach that produced the ECMWF's ERA5 reanalysis BLH dataset. Based on the radiosonde data of CMA, the properties of planetary boundary layer and its relationship with other meteorological variables in China have been studied. Guo et al. (2019) observed a shift in the temporal trend of BLH in China and found BLH was negatively associated with relative humidity, while positively associated with the near-surface temperature. Zhang et al. (2018) studied the thermodynamic stability of planetary boundary layer in China in summer. They found that convective boundary layer (CBL) dominated in summer throughout China (70%), with sometimes neutral boundary layer (NBL, 26%) and stable boundary layer (SBL, 4%) and BLH of CBL and NBL was positively (negatively) associated with near-surface temperature (humidity), whereas no apparent relationship was found for SBL.

Studies (Seidel et al., 2012; Guo et al., 2016) had shown that the BLH estimates from the radiosonde and the ERA Interim, the predecessor of the ERA5, had consistent pattern. Seidel et al. (2012) evaluated the BLH data at the United States and Europe and found that the radiosonde and reanalysis datasets show similar patterns of spatial and seasonal variability though with biases that vary spatially, seasonally and diurnally. Guo et al. (2016) compared the BLH measurements from the radiosonde measurement from China's CMA and the ERA Interim and found good agreement between the two sources. Another aspect is that the BLH data from the ERA5 is at the hourly frequency which matches the frequency of the air quality data. Therefore, we choose the BLH reanalysis data from the ERA5 of ECMWF in our analysis.

3. Methods

We introduce the regression model used in the study, the forward variable selection that determines the order of the meteorological variables' importance and the meteorological adjustment approach.

3.1. Nonparametric regression models

To avoid potential model mis-specification, we consider nonparametric regression modeling of each air pollutant with respect to the meteorological variables. The nonparametric aspect of the model makes it more adaptive to the complex meteorological processes than the linear or nonlinear parametric regression models.

Let $Y_{ijt}(s)$ denote the concentration of a pollutant, say $PM_{2.5}$, at hour t in season j and year i of a monitoring site s, $\mathbf{X}_{ijt}(s)$ be a d-dimensional vector collecting d meteorological variables, and $U_{ijt}(s)$ be the underlying emission leading to the pollution. The following nonparametric model quantifies the relationship between the concentration $Y_{ijt}(s)$, the emission $U_{ijt}(s)$ and the meteorological $\mathbf{X}_{ijt}(s)$:

$$( ) = \mathsf{e}\,(\quad ( ); \quad ( )) + \quad ( ); \qquad = 1; 2; \ldots; \quad ; \tag{3.1}$$

where $m_{ij}(u(s);\ \ (s)) = E(Y_{ijt}(s)|U_{ijt}(s) = u(s);\ \ _{ijt}(s) = \ \ (s))$ is the regression function, $e_{ijt}(s)$ is the residual satisfying $E(e_{ijt}(s)|U_{ijt}(s),\ \mathbf{X}_{ijt}(s)) = 0$ and has the finite second moment, and n_{ij} is the total number of observations in season j of year i at the site s.

As the emission information is largely not available at the hourly frequency, we have to condition on the observable meteorological variable $\mathbf{X}_{ijt}(s)$ to attain the following observable model

$$( ) = \quad (\quad ( )) + \quad ( ); \qquad = 1; 2; \ldots; \quad ; \tag{3.2}$$

where $m_{ij}(\mathbf{x}(s)) = E(Y_{ijt}(s)|\mathbf{X}_{ijt}(s) = \mathbf{x}(s))$ is the regression function by taking the conditional expectation on the observable meteorological data only, $_{ijt}(s)$ is the residual satisfying $E(\ _{ijt}(s)|\mathbf{X}_{ijt}(s)) = 0$, $\mathrm{var}(\ _{ijt}(s)|\ _{ijt}(s)) = \ ^{2}_{ij}(\ \ _{ijt}(s))$. Despite $U_{ijt}(s)$ is latent, its information has been reflected in the regression function $m_{ij}(\cdot)$ via the yearly index i which was not appeared in Model (3.1). See Liang et al. (2015) and Zhang et al. (2020) for more details on the linkage between Models (3.1) and (3.2).

In Model (3.2), specific parametric form of the regression function $m_{ij}(\cdot)$ is not assumed to allow model flexibility and non-linear pattern in the pollution generation processes. The yearly seasonal regression function $m_{ij}(\cdot)$ can be estimated by the nonparametric kernel estimator (Hardle, 1990). Specifically, let $k(\cdot)$ be a univariate kernel function which is a symmetric probability density function, for instance the Gaussian kernel $k(u) = (2\ \ )^{-1/2}\exp(-u^2/2)$, and define the multivariate product kernel

$$( ) = (\ _1\ _2 \cdots \quad)^{-1}\ (\ _1 = \ _1)\ (\ _2 = \ _2) \cdots (\ \ = \ \ );$$

where $\ = (x_1; x_2; \ldots; x_d)'$ and $\ = (h_1; h_2; \ldots; h_d)'$ is a vector of smoothing bandwidths. The kernel estimator of $m_{ij}(\mathbf{x})$ is

$$\mathsf{b}\ ( ) = \frac{\sum_{=1}^{\mathsf{P}} (\ - \quad ( ))\ \ ( )}{\sum_{=1}^{\mathsf{P}} (\ - \quad ( ))}. \tag{3.3}$$

The smoothing bandwidths $\mathbf{h}$ can be selected by the cross validation method (Hardle, 1990; Liang et al., 2015).

3.2. Forward variable selection

An important question in the air quality assessment is the relative importance of the meteorological variables on the average pollution level through the regression function $m_{ij}(\cdot)$. We consider the forward variable selection procedure for the nonparametric regression model, which selects the most influential variable once at a time that achieves the best fitting performance (Hastie et al., 2009).

To find the most important (the first one to be selected) variable for a pollutant, we start with the univariate model

$$( ) = \ ^{(1)}(\ \ ^{(1)}( )) + \ ^{(1)}( ); \qquad = 1; 2; \ldots; \quad ; \tag{3.4}$$

where $X^{(1)}_{ijt}(s)$ is a candidate covariate, say DEWP or BLH, $m^{(1)}_{ij}(X^{(1)}_{ijt}(s)) = E(Y_{ijt}(s)|X^{(1)}_{ijt}(s))$ is the conditional mean given the covariate and $\ ^{(1)}_{ijt}(s)$ is the residual. It is noted here and throughout this subsection that the function $m^{(1)}_{ij}(\cdot)$ changes with respect to different covariate $X^{(1)}_{ijt}(s)$ and target cities.

Let $\hat{m}^{(1)}_{ij}(\cdot)$ be the kernel estimate of $m^{(1)}_{ij}(\cdot)$ according to the generic estimator (3.3). The fitting Mean Square Error (MSE) of each variable $X^{(1)}_{ijt}(s)$ at site s for year i and season j is

$$\mathrm{MSE}^{(1)}(\ \ ^{(1)}( );\ \ ) = \frac{1}{\ } \sum_{=1}^{\mathsf{X}} \{\ \ ( ) - \mathsf{b}^{(1)}(\ \ ^{(1)}( ))\}^2. \tag{3.5}$$

To evaluate the explanation power of the meteorological variables, we use the coefficient of determination R^2 to assess fitting performance of the covariate. For nonparametric regression, Doksum and Samarov (1995) proposed to use

$$^{2}(1;\ ) = \frac{\left[\sum_{=1}\{\ \ (\ ) - \overline{\ }(\ )\}\{\mathrm{b}^{(1)}(\ ^{(1)}(\ )) - \overline{\ }(\ )\}\right]^{2}}{\sum_{=1}\{\ \ (\ ) - \overline{\ }(\ )\}^{2}\sum_{=1}\{\mathrm{b}^{(1)}(\ ^{(1)}(\ )) - \overline{\ }(\ )\}^{2}}; \tag{3.6}$$

where $\overline{Y}_{ij}(s) = \sum_{t=1}^{n_{ij}} Y_{ijt}(s)/n_{ij}$ is the average of $Y_{ijt}(s)$ in season j of year i.

We use the MSE to evaluate the importance of explanatory variables. By taking the average of $\mathrm{MSE}_{ij}^{(1)}(X_{ijt}^{(1)}(s);s)$ for all the monitoring sites in the city, we obtain the MSE of the city for using the univariate regressor $X_{ijt}^{(1)}$ at year i and season j. The most important explanatory variable for the city, denoted as $X_{ijt}^{(1)*}$, at year i and season j is the one that produced the smallest MSE of the city.

To select the second most important regressor, we consider the bivariate regression model

$$(\ ) = \ ^{(2)}(\ ^{(2)}(\ )) + \ ^{(2)}(\ ); \qquad = 1;2;\ldots; \quad ; \tag{3.7}$$

where $\ ^{(2)}_{ijt}(s) = (X_{ijt}^{(1)*}(s);X_{ijt}^{(2)}(s))'$ for a $X_{ijt}^{(2)}(s)$ other than the already selected $X_{ijt}^{(1)*}(s)$, $m_{ij}^{(2)}(\ ^{(2)}_{ijt}(s)) = E(Y_{ijt}(s)|\ ^{(2)}_{ijt}(s))$ is the conditional mean given $\ ^{(2)}_{ijt}(s)$ and $\ ^{(2)}_{ijt}(s)$ is the corresponding residual. Similar to selecting the first variable, the MSE with $X_{ijt}^{(2)}(s)$ as covariates is

$$\mathrm{MSE}^{(2)}(\ ^{(2)}(\ );\ ) = \frac{1}{\ }\sum_{=1}\{\ \ (\ ) - \mathrm{b}^{(2)}(\ ^{(2)}(\ ))\}^{2}; \tag{3.8}$$

where $\hat{m}_{ij}^{(2)}(\cdot)$ is also attained via (3.3). We select the second important variable by minimizing the average MSE in (3.8) over the sites in the city. The coefficient of determination $R^2(2, s)$ by using $\ ^{(2)}_{ijt}(s)$ is

$$^{2}(2;\ ) = \frac{\left[\sum_{=1}\{\ \ (\ ) - \overline{\ }(\ )\}\{\mathrm{b}^{(2)}(\ ^{(2)}(\ )) - \overline{\ }(\ )\}\right]^{2}}{\sum_{=1}\{\ \ (\ ) - \overline{\ }(\ )\}^{2}\sum_{=1}\{\mathrm{b}^{(2)}(\ ^{(2)}(\ )) - \overline{\ }(\ )\}^{2}}: \tag{3.9}$$

It is shown in Theorem S1 of SM that the variance of $\ ^{(2)}_{ijt}(s)$ in Model (3.7) is less than or equal to the variance of $\ ^{(1)}_{ijt}(s)$ in Model (3.4) with the $X_{ijt}^{(1)*}(s)$ as the regressor. This informs the benefits of adding more regressors to the regression model as it can reduce the variations of the residual term and hence makes the regression model having better fitting performance, although it is another matter for prediction as having more variables may not lead to better forecast. An implication of the result is that asymptotically $\mathrm{MSE}_{ij}^{(2)}(\ ^{(2)}_{ijt}(s);s) \le \mathrm{MSE}_{ij}^{(1)}(X_{ijt}^{(1)*}(s);s)$ and the R^2 monotonically increases with the nested regressors.

We proceed the nested forward procedure to select the third, the fourth important covariates and beyond until all the candidate covariates are considered. The order of the variables being selected contains useful information on the relative importance of the meteorological factors on the pollutant concentration and has taken into consideration of the inter-dependence of the meteorological variables.

To find out how much information in the BLH that can be explained by the surface meteorological variables, we carry on the same nested forward selection procedure by substituting the pollution concentration with BLH as the response variable and the surface variables as the regressors. The order of the variables being selected are used as the order of relative importance for BLH.

Furthermore, to avoid over-fitting of the nonparametric regression, we used a 10-fold cross validation (CV) procedure to gain information on the out-of-sample performance with the selected variables. The 10-fold CV randomly divides the data into 10 segments and alternatively uses any 9 segments at a time to fit the regression model and the remaining one to gain validation performance of the model. As the validation is made on the data set which is not used in the model fitting, it is more objective and is called the out-of-sample validation. Specifically, we randomly divided the data of a season at a station into 10 segments. Then, for each data segment (as the validation data set), we fit the nonparametric regression model on the remaining 9 segments of data (the training set) and evaluate its performance on the validation data set by calculating the out-of-sample $R^2_{CV;k}$

$$^{2}\ ; = \frac{\left[\sum_{=1}^{n}(\ ; - \overline{\ })(\mathrm{b}_{-}(\ ;) - \overline{\ })\right]^{2}}{\sum_{=1}(\ ; - \overline{\ })^{2}\sum_{=1}\{\mathrm{b}_{-}(\ ;) - \overline{\ }\}^{2}}; \qquad = 1;2;\ldots;10; \tag{3.10}$$

where $\hat{m}_{-k}(\cdot)$ is the fitted regression function without the kth segment, $\mathbf{X}_{k,t}$ and $Y_{k,t}$ denote the tth observation in the kth segment. The cross-validated coefficient of determination R^2_{CV} is

$$^{2} = \frac{1}{10}\sum_{=1}^{10}\ ^{2}\ ; \tag{3.11}$$

3.3. Meteorological adjustment

To gain information on the underlying emission, the meteorological effects on the pollution concentration has to be removed. There were methods to adjust for the meteorological variation, including the trend analysis approach proposed in Thompson et al. (2001) based on the linear regression and the three-year moving average method proposed by the US Environmental Protection Agency (EPA). Chen et al. (2018) showed that the latter method can not remove the meteorological confounding. Liang et al. (2015) proposed a general framework to adjust for the meteorological confounding, see Zhang et al. (2020) for comprehensive theoretical analysis.

For air quality evaluation based on the meteorological variables $\mathbf{X}_{ijt}(s)$, we consider Model (3.2) to quantify the meteorological relationship with the pollutant's concentration $Y_{ijt}(s)$. To remove the meteorological confounding, Liang et al. (2015) proposed a version of the average pollution concentration

$$(\ ) = (\ \ (\ )) = \{\ \ (\ \ (\ ))\} = \int (\ ;\ )(\ ;\ )\ \ ; \tag{3.12}$$

where $m_{ij}(\mathbf{X}_{ijt}(s))$ is the regression function that can be estimated by (3.3), and $f_j(\mathbf{x}, s)$ is a baseline probability density function for the meteorological variable $\mathbf{X}(s)$ in season j which can be constructed as

$$(\ ;\ ) = \ ^{-1}\sum_{=1}^{\ }(\ ;\ );$$

where $f_{aj}(\mathbf{x}, s)$ is the density function of meteorological variable $\mathbf{X}_{ajt}(s)$ in year a of season j, and A_j is the number of years that the meteorological data are available. In our study, A_j was 10 as we used 10 years' meteorological data from March 2011 to February 2021 to build the baseline density $f_j(\mathbf{x}, s)$.

An estimator for $\ _{ij}(s)$ which is free of the meteorological variation and can reflect the emission is

$$\mathrm{b}\ (\ ) = \sum_{=1}^{\ }\left[\ ^{!\ -1}\sum_{=1}^{\ }\sum_{=1}^{\ }\mathrm{b}\ (\ \ (\ ))\right]; \tag{3.13}$$

where n_{aj} is the number of meteorological sample size in season j of year a. The nonparametric regression estimator $\hat{m}_{ij}(\cdot)$ is (3.3). Unlike the raw averages, the temporally adjusted averages $\mathrm{b}_{ij}(s)$ are comparable for the specific season over different years as shown in Zhang et al. (2017a) and Chen et al. (2018).

The air quality measures $\mathrm{b}_{ij}(\mathcal{A})$ over a region $\mathcal{A}$ occupied by a city is an average of the adjusted $\mathrm{b}_{ij}(s)$ over the monitoring sites in the region:

$$\mathrm{b}\ (\mathcal{A}) = |\mathcal{A}|^{-1}\sum_{\in\mathcal{A}}\mathrm{b}\ (\ ); \tag{3.14}$$

where $|\mathcal{A}|$ denotes the number of monitoring sites in the region $\mathcal{A}$.

PM2.5 - Spring
PM2.5 - Summer
PM2.5 - Fall
PM2.5 - Winter
1st
2nd
3rd
4th
5th
6th
7th
DEWP (1.14)
PRES (2.13)
TEMP (2.99)
BLH (4.04)
IWS (5.07)
WD (5.63)
IPREC (7)
DEWP (1.28)
PRES (1.98)
TEMP (3.23)
BLH (3.96)
IWS (5.04)
WD (5.67)
IPREC (6.83)
DEWP (1.33)
PRES (2.06)
TEMP (2.89)
BLH (3.74)
IWS (5.28)
WD (5.77)
DEWP (1.03)
PRES (2.16)
TEMP (3.45)
BLH (3.51)
IWS (5.23)
WD (5.64)
Beijing
Jinan
Shijiazhuang
Taiyuan
Tianjin
Zhengzhou
Variables
DEWP
PRES
TEMP
BLH
IWS
WD
IPREC

C & . * * - , 6 * - * - ' % 3 C 3
> C C
' 4 1 H 8 ' 5

& . 3 C / ! 9 5 . 5 6 !
3 ! 9
& . * ' $ % 3 & . *
3 1 ' ? A & . * 3 3 ! ! C H
C $' C : 0 8 ((0
% H % 0) (H 9 & * C 9
4 * - - 6 5 ' * 1 K ' % !
8 ! 3 H 3 3 7 3
9 3 3 '
3 & . * 3 1 ' , * 1 3 3 ! 9
, 6 3 ! 1 ' ? A 9 3
& . * ' !
% . > % 9
3 H
> / 3 1 ' , - , ,
3 . 6 3 3 H
$ < 9 . 6 3 3 % 3 !
3 ! & . * 3 % 3
> 3 1 ' ? A 3 H 4 K ' * 5 ! 3 & . * '
3 ! H & 3 3 !
> 9 3 1 6 ' 3 !
1 K * 9 '
3 3 9 % 3
' H 4 6 ' * 5 3
1 ' , K , ? ' % H 3 * H ! e
C 3 3 E
3 * 3 3 '
% 4 K ' , 5 9 4 K ' 6 5 ^ .:. e †‡ ; ^ ; ; ; : 4 K ' K 5
'
1 ' * 6) C 3
3 3 4 % 0) (5 3
4 : 0 8 (5 4 (0 5 !

C < 9 . 6 * - , 6 * - * - '% 3 C 3
> C C
' 4 1 H 8 ' 5

9 3 9 * 4 5 ( ) * ' +
C 3 4 5 C ! 4 5% C
! C C 4 1 H 8 ' 5

9 3 9 * 4 5 < 9 . 6
C 3 4 5 C ! 4 5% C
! C C 4 1 H 8 '5

% ! 4 5 ! 4 3 5 ...A† 4 K ' + 5
6 * * - , 6 * - * - '

(	G	>	%	%	@ >	/
( ) *'+	-'<64-',*5 *'+K	,'*;4-',+5 K'-*	-'<*4-',*5 *'AK	,'+A4-'*K5 +'+*	,'-*4-',65 *'<;	-'A64-'-;5 ,';<	,'-K4-',*5 6'*<
( ) , -	-'?+4-',*5 *'<*	,'-A4-',K5 K'-?	-'?A4-',-5 *'*A	,'6K4-'**5 K'6,	-';64-',,5 *',+	-'A-4-'-<5 ,'<-	-'<;4-',-5 *';-
. *	,'*-4-',65 *'AK	,'6A4-',;5 K'6<	,',;4-'*,5 ?'-?	*'6*4-'6?5 A';-	,',+4-',65 6'*+	-';?4-',K5 6'**	,'6A4-',<5 K'+;
& . *	-';K4-',65 *';?	,'+<4-',;5 K'*;	,'*;4-',+5 K'-?	,'6;4-',<5 6'AA	,',64-',65 *'?6	,',64-'*<5 <'++	,'*K4-'-<5 K'6A
# .	-'A<4-',K5 *'<?	,'*K4-',<5 K'-K	,',-4-',<5 6'A,	-';;4-',A5 K',?	-'<-4-',-5 *'6;	-'A*4-',+5 K'6A	-';K4-'-<5 6'+;
< 9 . 6	-'+?4-'-<5 ,'A?	*'A*4-'6<5 A',;	,'*;4-'*,5 +',6	6'A;4-'KA5 ,-'?*	-'A;4-',,5 *'*A	-'A64-',*5 *'6K	,'?+4-'K;5 K'<<
/	-'<64-'-;5 *'?,	,'++4-'*K5 K'??	,'-?4-',,5 6';;	,';-4-'K*5 ?'-+	-';A4-'-?5 *'?-	-'<64-'-A5 6'A,	,',;4-',-5 6';K

1 4 5 H 4 +L 5 ! 3 !
6 * * - , 6 * - * - '

(	G	>	%	%	@ >
( ) *'+	- 4 - 5	- 4 - 5	- 4 - 5	- 4 - 5	- 4 - 5	- 4 - 5
( ) , -	- 4 - 5	- 4 - 5	- 4 - 5	- 4 - 5	- 4 - 5	- 4 - 5
. *	- 4 - 5	- 4 - 5	- 4 - 5	* 4 ? ' * + L 5	- 4 - 5	- 4 - 5
& . *	- 4 - 5	- 4 - 5	- 4 - 5	6 4 ; ' 6 < L 5	- 4 - 5	* 4 ? ' * + L 5
# .	- 4 - 5	* 4 ? ' * + L 5	- 4 - 5	- 4 - 5	- 4 - 5	, 4 6 ', 6 L 5
< 9 . 6	- 4 - 5	* 4 ? ' * + L 5	- 4 - 5	K 4 , * ' + L 5	- 4 - 5	- 4 - 5

! " $ ( ) % 3 1 : ' %
>) 8
$ % 3 1 : 8 1 '
'& ' % # >) 3
1 : 8 $ %
3 1 : 8 '

% H
7 3 C
'

, - . "

% 3 & &
1 # ;*-K?-*,,*-A,-,6A,;A6--+
,,;A,6;- !)0J1 (C B ' C 3
1 1 # B
G N,<-?--*'

,!! , !! "

EFF

F,-',-,?F ' '*-*,',;<A6A

+

/ %'8*--6' $) / 'G
8 & 3 C
!' ' G' G'8 ' @ ' # '"'*-,<'/ 9
9% 9 E ()*'+ .6'
/ '0 ;;6*;-6-,'
' @ @'N ' *-,6' # '/ '
G' '# #) ',<<,,A-,,A,'
' /)') 0'J "') /' '
(' *-*-' > (),-
' '% 0 A-,,6KKAK
: C N' /' ,;;+'&
3 /' ' *6
,KK6KA'6
: N'! "') & 8' ,;;<'B H4 5
G' '*; ++6+?'
: $' "'*--<'0 ' '
/ ') ' '<; ,*+A,*?*
1 @':) /' / /')' ('% ' 1 1'% 1'
/'(0'*-,;'0 >
'0 $;6,,-K;?'?
N'# ' 8 ' 8 ' @ !' @ ' @ ' %'@ "'
-,;' 9 ()'+ 0 # ,;;<
*-,?E 0 ('
KA,,+,,66
G'! ' # G' ! G'# :' " ' ! !' G' N'@ (' *-,;'
9
4,;A;*-,?5 ' ' ! 'K??-<-?-<'
G') ' @ ' ! ' ! @'@ 8' G'!)' ' !'
@ (' *-,?%
'/ ' # '(',?,66-;,66,;'
• 8',;;-'/ & # B '
%'% ' 1 G'*--;'% 0 ! E
) $ ('
! "'! ' @ ' ' # '"'*-,?'()*'+
H # G' ' 'E
/ ,*, ,-**-,-*6?
! "' @ %' ' ! ' @ ' @ ' ' # '"'*-,+'@ %'!
/ ()*'+ E 3 /(0# 3 /
(' '/KA,*-,+-*+'A ' *,+6-+6,?
" ' : ' :

! "' ' ! G'J ' %' !' G'# #' ' ! "' %'
@ ' % ')' *-,6'1
> E # '/ ' # '(',6
K+-,K+,'K
) ' # ' @ "' ! ' *-*,' 3
()*'+ .6 3
'0 (*?<,,+AA'+
) ' G'! ' ! ' ! @'@ 8' @ (' *-,A'# H
3
'/ ' # '(',A6-;A6,,-'
) ' ! ' *-,;'! C 3
' '% 0 ?+-*<<*;?
($$$'/' '%'%)'G'# 0'0'N 3 C:' $ N'% ':'
--'! 9 H
G'/') ' / '*<A,,6*,,K,'
J G'% "' @ J'! J'! "' ' @ :' *-,K'#
*-,**-,63 # '/ '
0 <<<6<;'
3 >G'*---'% 3 '
0 ,, 6*-6*?
:'G'@ ' /' >G'#G /' ') ' *-,*'
B
0 'G' ' 'E ,,A,A,-?
!' G :'G') C !'G'8 ' @ J'*-,<'$ H
3 > '/ '# '(',<
,AK<;AK;'?
)' %' ':)' N' : 1'
>>#'! /') >>('>)')' /'
2 ' N $' 3 >G'*-,;'0 ().- ()*'+
$ *-,69*-,+ 9 9
'0 $;*K,A-,A;'
% /'(') C !'G'G :'G'*-,-'# 3 H
4()*'+5 B E
() *'+ '/ '0 KK6;A?6;<K
% ' @ G'@ "' %')O C #' ' • N'! @'@ G'
8 !'" G' ('8 '*-,?')
'/ ' # '(',?
*K+;*KA'+
%)'!' G'# !' ' (' ('*--,'/ 3
> '/ '
0 6+?,A?6'
2 > :'(' /'/')' ,;;?'0
9 ' 9!) '<,*K+*?;'
2 %'2' @'# G'@ J' N'8 ' ')' *-,;'/

'/ ' # '(',; ,,6-6,,6,K
" ' @ %'! G'! !': ' # @*-,;'/
3 '
/ ' '*,+6-+6,?
" ' : ' : ' C %') %'*-,?'0 ()*'+
9 # E 9 0 ' '
% +-K<6?K<K6
@ J') "'% "')' @ #'*--;'2
3 E 9
'/ '0 K6++*?++6'+
@ ' # '"' ' 8 ' ! 8' *-*-' 9
' ') '+-++A++<%
0 E EFF ' FB F1 F(
F #)9*-,;9-6?< *P(' P '
@ ' ' : /' G'" @'# '"'*-,A #' 9
(' '/KA6*-,A-K+A
@ 8' G') ' ! ' ' 1 @' G'!)' ' ! '
@ ('#*-,<(',; ,,6-6,,600K '
%'! G'! !': ' # @*-,;'/
3 '
(' '/KA,*-,+-*+'A ' *,+6-+6,?
" ' : ' : ' C %') %'*-,?'0 ()