A spatio-temporal model for the analysis and prediction of fine particulate matter concentration in Beijing

Yating Wan[1], Minya Xu[1] *, Hui Huang[2] and Song Xi Chen[1]

[1]Peking University and [2]Sun Yat-Sen University

Abstract

Effective air quality management and forecasting in Beijing is urgently needed as the region suffers from the worst air pollution in any standards. However, the statistical mechanism of the $PM_{2.5}$ formation with respect to various factors is under-explored in this region and China in general. Through an elaborate application with refinement of a spatio-temporal model with varying coefficients to the dynamics of $PM_{2.5}$ around Beijing based on a large data set, we provide a comprehensive interpretation for the dynamics of $PM_{2.5}$ concentration with respect to its gaseous precursors, meteorological conditions and geographical variables. Furthermore, we conduct multi-step temporal forecasts on a rolling basis for both the $PM_{2.5}$ concentration and the pollution levels. With the help of the expectation-maximization algorithm, the proposed models estimated for eight seasons

* *Corresponding author*: Minya Xu, E-mail: minyaxu@gsm.pku.edu.cn

from March 2015 to February 2017 around Beijing provide satisfactory in-sample fits and generate more accurate out-of-sample forecasts, compared with Finazzi and Fassò's original model as well as other alternative models. Valuable insights in tackling the excessive air pollution in Beijing are suggested from the comprehensive application of our model.

Keywords: Air pollution, expectation-maximization algorithm, random effects, rolling prediction.

1 Introduction

Air pollution has been a worldwide concern, and that in China is particularly pressing in terms of severity and spatial coverage in recent years. Beijing, the capital of China, is situated in a region that has endured the worst air pollution even in the Chinese standard. Air pollutants, especially the fine particulate matter $PM_{2.5}$ (fine particulate matter with a diameter of 2.5 micrometres or less), are known to cause harm to human health. $PM_{2.5}$ can penetrate deep into the lungs, blood vessels and other organs of human beings, and can aggravate heart and lung diseases when inhaled (Pope III et al., 2002; Yeatts et al., 2007). Besides, the adverse health outcomes of exposure in air pollution, e.g., respiratory diseases (Shao et al., 2010) and low birth weight (Berrocal et al., 2011), have been quantified by researchers.

Public attention and the adverse health effect of air pollution has motivated the studies on the triggers of $PM_{2.5}$ in China. Most studies are based on real-time measurements with a set of state-of-the-art instruments. Among many others, Guo et al. (2014) studied the haze episodes in 2013 fall and elucidated the dominat-

ing effect of meteorological conditions and secondary formation on haze pollution, while the effects from primary emissions and regional transport of $PM_{2.5}$ were found to be insignificant; Sun et al. (2015) conducted a year-long investigation of the different variations and formation processes among aerosol particle species, as well as uncovered impacts of several meteorological factors (relative humidity, temperature and wind) on aerosol loading. Previous environmental science research mainly studies the composition properties of aerosol particles and their interaction with meteorological factors through real-time measurements. Such research usually restricts the observations to a certain sampling site, and cannot untangle the different extents of impact among the various factors simultaneously.

As Beijing in China has been suffering from severe air pollution in these years, we seriously need an effective statistical model to interpret the influencing mechanism and predict the extent of the city's air pollution. In this work, we perform a comprehensive study with elaborate parametric statistical models to uncover the influences of a large number of environmental factors on $PM_{2.5}$ concentration, reveal space-time dependence structures, and provide predictions over time in Beijing. While Liang et al. (2015, 2016) and Chen et al. (2018) have conducted statistical analyses on air quality data in various parts of China, they mainly focused on air quality assessments using nonparametric models.

The present study aims at attaining a suitable model for studying $PM_{2.5}$ in Beijing, which can be used to offer environmental interpretation for policy purposes and to do predictions. We will build a spatio-temporal statistical model, since it can incorporate the spatiotemporal dependence structure (Cressie and Wikle, 2011, Chap. 1), as well as account for the influences of various covariates. The liter-

ature abounds with applications of spatio-temporal models to air quality data (i.e., the recent ones including Calculli et al. (2015), Cheam et al. (2017), Nicolis et al. (2019), Clifford et al. (2019), and Padilla et al. (2020)). For example, Cheam et al. (2017) applied an EM algorithm to the inference of a parametric spatio-temporal mixture model used for clustering the air quality data. Clifford et al. (2019) conducted Bayesian inference based on a semi-parametric spatio-temporal model for the airborne particulate matter concentration, with nonparametric temporal trends and the spatial random effect approximated using Gaussian Markov Random Fields (GMRF). These studies put more concerns on flexibility of models and computation, but they did not consider the environmental covariates which play an important role in triggering air pollution (Chen et al., 2015; Cai et al., 2017), and they did not apply the model to generate predictions. In addition, some studies developed spatio-temporal models with environmental covariates involved and used them for space-time predictions (Calculli et al., 2015; Nicolis et al., 2019; Padilla et al., 2020), however in contrast to our proposed model, they did not take varying coefficients into account.

Specifically, the model in our study is set up based on Fassò and Finazzi (2013) and Finazzi and Fassò (2014), which contains a spatially correlated random field and a separate latent temporal dynamic, and allows for stochastically varying coefficients to incorporate the interaction between spatial random effects and time-invariant covariates. In addition, the proposed model extends that of Finazzi and Fassò (2014), in that we incorporate the lag-one concentrations of $PM_{2.5}$ as well as other gaseous pollutants as covariates. We demonstrate that the improvement over the model framework in Finazzi and Fassò (2014) is substantial in terms of fitting

and prediction accuracy after including the lagged response as an autoregressive term. Such an adjustment is also suitable for situations in Beijing, as much heavier air pollution and the dense human socio-economic activities around Beijing make air pollution less likely to dilute quickly and thus more time-persistent. The inclusion of lagged gaseous pollutants better accounts for the secondary formation of $PM_{2.5}$.

By analyzing a quite large data set that consists of two years' hourly air quality data from 35 monitoring sites together with meteorological data from 15 weather observation stations and gridded geographical data, we establish the validity of our model through interpretation of the influencing mechanism of $PM_{2.5}$ dynamics as well as the temporal forecasting performance. The first part is aimed at investigating various driving factors of $PM_{2.5}$ formation, where we obtain satisfactory in-sample fits that outperform other alternatives using the proposed spatio-temporal model with the EM algorithm on a seasonal basis. In addition, we also demonstrate the model fitting performance if removing different model components including the lagged response and random effects, so that we can understand which component plays more important roles. In the second part, we perform an out-of-sample multi-step temporal prediction for both the $PM_{2.5}$ concentration and the severity levels for $PM_{2.5}$, with the model fitted on a rolling basis for each season. The out-of-sample forecasts of the proposed model outperform those obtained from the original model in Finazzi and Fassò (2014) as well as the simple $AR(1)$ model.

The rest part of the paper is organized as follows. Section 2 gives a description of data in the study region and provides the motivation for the proposed spatio-temporal model with an insightful data exploration. In Section 3, we establish

the spatio-temporal model, and introduce the EM algorithm for the maximum likelihood estimation. Section 4 shows the results of applying the model to our data in Beijing region. In Subsection 4.1 we fit the model on a seasonal basis, and provide the detailed interpretation of environmental covariates' contributions to the dynamics of $PM_{2.5}$ in Beijing. We further show the importance of different model components based on a comparative assessment of fitting performances. In Subsection 4.2, we illustrate the methodology and results of the out-of-sample rolling prediction. Section 5 provides the conclusion and discussion of our study.

2 Data And Pre-analysis Exploration

2.1 Data Description

The primary endpoint is the concentration of $PM_{2.5}$, which is collected hourly from 35 air-quality monitoring stations operated by Beijing Municipal Environmental Monitoring Center (BMEMC). The 35 air-quality monitoring stations are associated with specific latitudes and longitudes and denoted by $\mathcal{D} = \{s_1, \ldots, s_{35}\}$. The $PM_{2.5}$ concentration is measured in microgram per cubic meter of air ($\mu g/m^3$) and collected from March 1st, 2015 to February 28th, 2017, which encompasses two seasonal years with eight seasons.

For explanatory variables, first we include the $PM_{2.5}$ concentration from the previous hour, since $PM_{2.5}$ is temporally persistent. In addition, a portion of $PM_{2.5}$ can be attributed to the atmospheric chemical reactions involving gaseous pollutants. In line with the choice of other pollutants in Liang et al. (2015) and Chen et al. (2018), we consider four gaseous pollutants of SO_2, NO_2, O_3 and CO.

In particular, we use the lag-one (previous hour) concentrations of these pollutants as covariates, which would account for the secondary formation of $PM_{2.5}$. They are also collected hourly and obtained from 35 air-quality monitoring stations.

Previous studies have found significant meteorological influence on $PM_{2.5}$. For example, Liang et al. (2015) showed that meteorological variables could explain 78% of $PM_{2.5}$ variation on average in their nonparametric model fitting. In this study, according to previous studies on Chinese air quality in Zhang et al. (2017a) and Chen et al. (2018), we consider the following meteorological variables: air temperature in degree Celsius, air pressure in hectopascal, dew point temperature in degree Celsius, integrated precipitation in millimeter, and cumulative wind speed in meter per second on four combined wind directions ("SE" for southeast, "SW" for southwest, "NE" for northeast and "NW" for northwest). They are obtained from 15 weather stations belonging to China Meteorological Administration (CMA) and are matched to the 35 monitoring sites according to Table S1 in Zhang et al. (2017b). In addition to the variables used in the previous studies, we also consider boundary layer height in kilometer, as it mainly influences the vertical dissipation of particulate matter (Tang et al., 2016; Miao et al., 2015). It is collected from ERA-Interim of the European Centre for Medium-Range Weather Forecasts (ECMWF) at a grid size of $0.125° \times 0.125°$, with a detailed description of the dataset found in Xu et al. (2020). We depict the spatial locations of the 35 air-quality monitoring stations along with the 15 weather stations in Figure 1. The altitudes of the monitoring sites are obtained from Google Maps over the region (39.4°N–40.7°N, 115.8°E–117.4°E) at a grid size of $0.01° \times 0.01°$.

Due to Beijing's the semi-open geographical configuration with mountains in

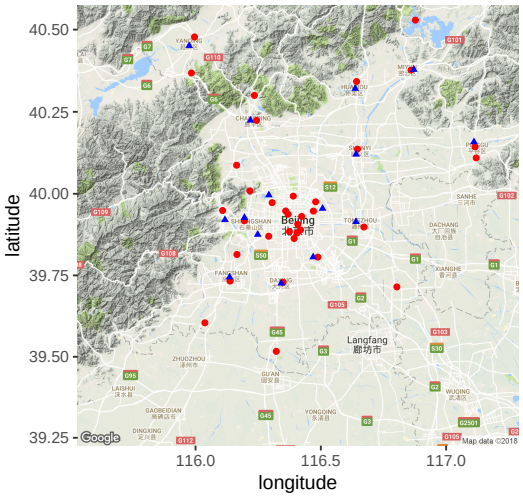


Figure 1: Locations of the 35 air-quality monitoring stations (red dots) along with the 15 weather observing stations (blue triangles) around Beijing. West and north of Beijing are mountains, while south and east of Beijing are open plains.

the west and north, in addition we consider two geographical variables: altitude in kilometer and distance to mountains in kilometer. As shown in Figure 1, the areas near mountains may suffer from more serious air pollution, as the mountains can easily trap pollutants (Sun et al., 2006). By defining the locations with altitude higher than 500 meters as mountains, we attain the distance from each monitoring station to its nearest mountain.

Since the distributions of $PM_{2.5}$, four gaseous pollutants, wind speed and boundary layer height tend to be skewed, we take the logarithm transformations for these variables. Particularly, the log-transformation of wind speed is made after adding 1, as it contains a proportion of zero values. After that, all variables are standardized with their seasonal means and standard deviations.

As 5.1% of the $PM_{2.5}$ concentration, 4.8% of the four gaseous precursors and

3.5% of meteorological variables are missing, for each monitoring site, we conduct linear imputation for those variables with no more than 12 continuous hours missing, otherwise we impute with the weighted mean of the corresponding observations using data of the nearest 20% stations with the weights being inversely proportional to the distances.

2.2 Pre-analysis Exploration

In this section, we do some data exploration and demonstrate the motivation for adopting the spatio-temporal Model (1), which we will set up in Section 3.1. First we explore the distributions of $PM_{2.5}$ concentration for the eight seasons in the two seasonal years. The box plots are presented in Figure 2. From the figure, we can see that there exist consistently strong seasonal patterns, where winters suffer most severe pollution with large mean and variability, and summers are relatively relieved with smaller mean and variation. This motivates us to construct distinct models for the eight seasons from March 2015 to February 2017.

For each season studied, we retrieve the variance inflation factors (VIF) of all covariates mentioned above from the diagonal elements of the inverse sample correlation matrix (Belsley et al., 2005). The VIFs with respect to the covariates are not exceeding 5 after taking average over the eight seasons, which relaxes our concern of collinearity.

In addition, to see the importance of modeling with time-dependent dynamics, we plot sample partial autocorrelations (PAC) in Figure 3, for the raw log-$PM_{2.5}$ concentrations in Panels (a), (c) and residuals after we fit a linear $AR(1)$ Model (3) with the covariates we consider in Panels (b), (d), respectively. Each box-plot

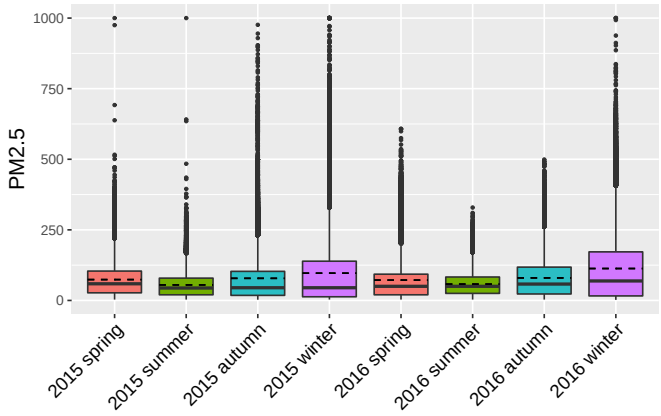
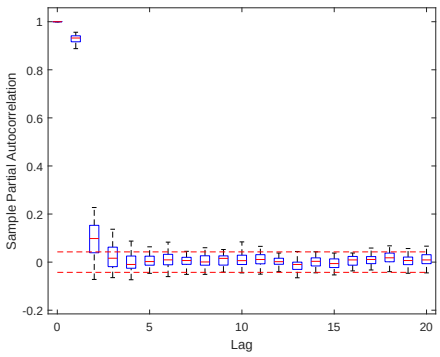


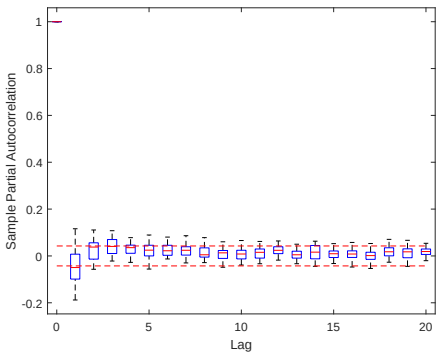
Figure 2: Box plots of $PM_{2.5}$ concentration ($\mu g/m^3$) in Beijing for the eight seasonal periods from March 2015 to February 2017. Black bars inside the boxes are the medians, and the dashed bars are the means.

in Figure 3 is constructed based on the PACs of the 35 time series from different monitoring sites. We demonstrate results from the first and the last seasonal period in the two years from 2015 to 2016, respectively. Other seasonal periods have similar patterns. First, Panels (a), (c) show a significantly large lag-one PAC for the raw data for all 35 monitoring sites, suggesting an autoregressive component in the model. Next, Panels (b), (d) indicate that there exists time-dependent random effect, which possesses a relatively pronounced first-order autocorrelation, contributing to the variabilities of $PM_{2.5}$ which can not be fully explained by its autoregressive component as well as other covariates.

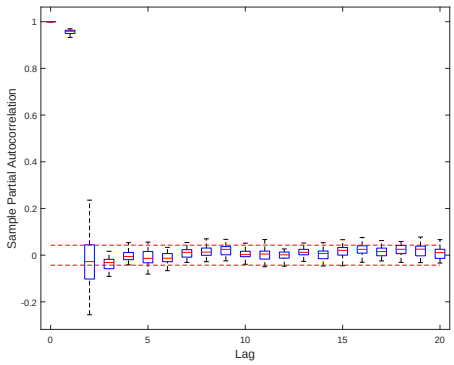
Besides, we present the spatial sample correlation of the fitted residuals of Model (3) in Figure 4. It is easily found that the correlation between residuals at two monitoring sites declines as their distance increases. It thus motivates us to incorporate a spatially-dependent random effect in the model to account for the



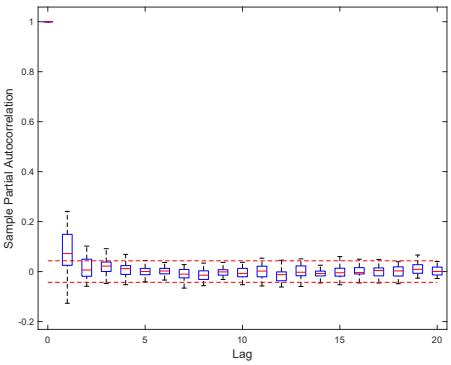
(a) Raw data, 2015 spring



(b) Residuals, 2015 spring



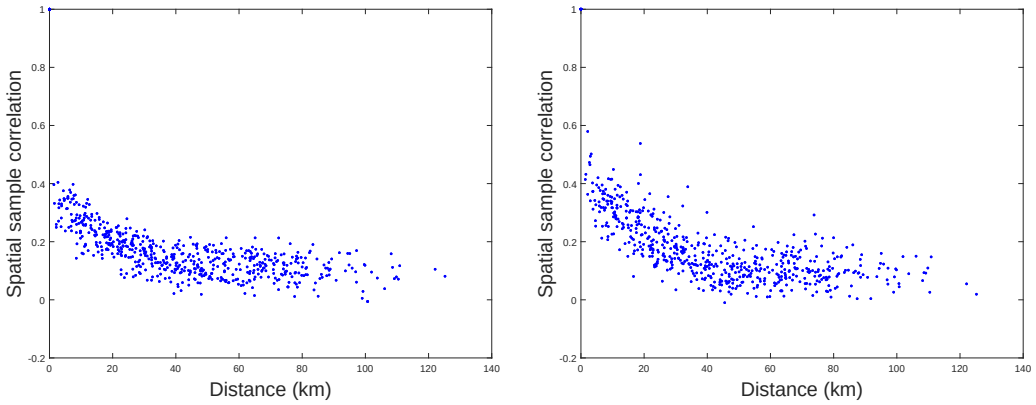
(c) Raw data, 2016 winter



(d) Residuals, 2016 winter

Figure 3: Sample partial autocorrelation (PAC) box-plots of (a), (c) raw log- $PM_{2.5}$ concentrations, and (b), (d) residuals after fitting a linear $AR(1)$ Model (3) with covariates. Panels at the top (bottom) show the results for the spring (winter) of seasonal year 2015 (2016), with similar patterns for other seasons. Each box-plot displays PACs obtained for each of the 35 monitoring sites.

spatial correlation structure of estimates in different locations.



(a) Residuals, 2015 spring

(b) Residuals, 2016 winter

Figure 4: Spatial sample correlation of the fitted residuals of a linear $AR(1)$ Model (3) with covariates between two monitoring sites with respect to their distance. Here we only show the model fitting result for the spring of seasonal year 2015 and the winter of seasonal year 2016, as other seasons have similar patterns.

3 Method

3.1 The Spatio-temporal Model

For each season in Beijing, we build a spatio-temporal model with varying coefficients, which can incorporate the dependencies among space and time, as well as the interaction between spatial random effects and time-invariant covariates.

Specifically, denoting $y(s,t)$ the log-$PM_{2.5}$ concentration at site $s \in \mathcal{D}$ and time $t \in \{1,\ldots,T\}$, we assume an $AR(1)$ model with covariates and latent random

effects:

$$
\begin{aligned}
y(s,t) = \quad & y(s,t-1)\beta_0 + \mathbf{x}_g'(s,t-1)\boldsymbol{\beta}_g + \mathbf{x}_m'(s,t)\boldsymbol{\beta}_m \\
& + x_{Alt}(s)(\beta_{Alt} + \alpha_1 w_1(s,t)) + x_{Dist}(s)(\beta_{Dist} + \alpha_2 w_2(s,t)) \\
& + z(t) + \varepsilon(s,t), \qquad (1) \\
z(t) = \quad & g z(t-1) + \eta(t), \quad z(0) \sim N(\mu_0, \sigma_0^2),
\end{aligned}
$$

where $\mathbf{x}_g(s,t-1)$ is a vector of four lag-one gaseous pollutants accounting for the secondary formation of $PM_{2.5}$, and $\mathbf{x}_m(s,t)$ is a vector of nine meteorological variables at time t. Note that the covariates are random in nature, and we consider the conditional inference by conditioning on the random covariates, which is similar to treat the covariates as fixed in standard regression. Moreover, according to the previous work which uses lagged responses as covariates in time series models (see, e.g., Li (1994) and Brumback et al. (2000)), we add the lagged log-$PM_{2.5}$ $y(s,t-1)$ as an autoregressive term to account for the temporal persistence of $PM_{2.5}$ generation process.

In the model, $x_{Alt}(s)$ and $x_{Dist}(s)$ are respectively altitude and distance to mountains, for which we use β_{Alt} and β_{Dist} to represent their global (space and time independent) effect, and use $w_j(s,t)$, $j = 1,2$ as location-specific (spatial) random effects. We assume that $w_j(s,t)$ with $j = 1,2$ are mutually independent and temporally uncorrelated Gaussian random fields with zero mean, unit variance and an exponential spatial correlation function in the form of $\rho(\|s - s'\|; \theta_j) = \exp(-\|s - s'\|/\theta_j)$, with $\|s - s'\|$ denoting the distance between two locations s and s'. Such simple exponential correlation function as a special form of the Matérn correlation family is commonly adopted in applications of spatially-

dependent models (see, e.g., Sahu (2012) and Finazzi and Fassò (2014)). Scale parameters α_j and θ_j respectively control variability and correlation decaying rate of the random effects $w_j(s,t)$. We introduce $w_j(s,t)$ into Model (1) to account for location-specific effects, so that it is flexible to model additional variations in data. Since we have non-replicated spatio-temporal data, "borrowing strength" from neighboring sample points is more efficient in parameter estimation, given that the spatial correlation structure is pre-determined in certain forms.

The time-dependent random variable $z(t)$ represents the other time-dependent influential factors which are not considered in the model. We assume it to be a Markovian dynamic with time-independent innovation $\eta(t) \sim N(0, \sigma_\eta^2)$. The $AR(1)$ coefficient g quantifies the persistency in $z(t)$'s variability overtime, and we require that $|g| < 1$ for stationarity.

Lastly, the measurement error $\varepsilon(s,t)$ is assumed to be a zero-mean Gaussian white noise with a variance σ_ε^2. It can be seen as a random effect describing uncertainties that cannot be fully explained by the other components of the model.

Under the model setup, the parameters to be estimated are

$$\Psi = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\varepsilon^2; \boldsymbol{\theta}; g, \sigma_\eta^2; \mu_0, \sigma_0^2),$$

with scale parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$, $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and regression coefficients $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_g', \boldsymbol{\beta}_m', \beta_{Alt}, \beta_{Dist})'$. Our main interest lies in the estimation and inference of $\boldsymbol{\beta}$.

3.2 The Expectation Maximization Algorithm

In the model fitting procedure, the expectation maximization (EM) algorithm (Dempster et al., 1977) is employed to obtain the maximum likelihood estimate (MLE) of the parameter vector Ψ. The expectation step (E-step) of the mth EM iteration requires computing the expectation of "complete-data" log-likelihood $l(\Psi; \mathbf{Y}, \mathbf{W}, \mathbf{Z})$ conditioned on the observed data $\{\mathbf{Y}, \mathbf{X}\}$ based on the current estimate $\hat{\Psi}^{(m-1)}$, that is,

$$Q(\Psi, \hat{\Psi}^{(m-1)}) = E_{\hat{\Psi}^{(m-1)}} \left[l(\Psi; \mathbf{Y}, \mathbf{W}, \mathbf{Z}) \mid \mathbf{Y}, \mathbf{X} \right],$$

where $\mathbf{Y}$, $\mathbf{W}$, $\mathbf{Z}$ and $\mathbf{X}$ represent the responses, spatial random effects, temporal random effects and covariates in Model (1), respectively. We emphasize that, though we add the lagged response to the covariates $\mathbf{X}$ in Model (1), $Q(\Psi, \hat{\Psi}^{(m-1)})$ takes the same expression with respect to $\{\mathbf{Y}, \mathbf{X}\}$ as that obtained from Finazzi and Fassò (2014). Thus our refinement of the model makes no difference from the perspective of estimation with the EM algorithm. To be specific, the computation of $Q(\Psi, \hat{\Psi}^{(m-1)})$ requires the conditional distribution of the random effects $\mathbf{Z}$ and $\mathbf{W}$ given observed data, which are obtained respectively by the Kalman smoother and the well-known formulas of multivariate normal distribution respectively, as detailed in Fassò and Finazzi (2011).

Then at the maximization step (M-step), we aim at finding the mth update of MLE $\hat{\Psi}$:

$$\hat{\Psi}^{(m)} = \arg\max_{\Psi} Q(\Psi, \hat{\Psi}^{(m-1)}).$$

The EM algorithm stops when the update of parameter vector is significantly small. The model fitting procedures are implemented based on the D-STEM

software (Finazzi and Fassò, 2014).

4 Results

4.1 Model Fitting and Interpretation

4.1.1 Model Fitting Results

The model parameter estimates together with their levels of significance are presented in Tables 1 and 2. Figure 5 provides a contrast of the magnitudes of estimated coefficients of gaseous pollutants, meteorological and geographical variables.

[Table 1 about here.]

[Table 2 about here.]

Generally speaking, almost all predictors in the proposed model are significant, showing the the formation of $PM_{2.5}$ is affected by multi-factors. Note that all the regressors have been standardized, contributions of each predictor are hence comparable. First, Tables 1 and 2 show that the $PM_{2.5}$ concentration from last hour contribute the most to the current $PM_{2.5}$ level. The average $AR(1)$ coefficient for eight seasons is 0.674 with a standard deviation of 0.048, indicating a strong autocorrelation. This is not surprising for hourly concentration, and echoes findings from other studies, e.g. Liang et al. (2015).

Second, Figure 5 shows that among the four gaseous precursors, CO is the most influential one affecting $PM_{2.5}$. CO is usually from traffic-related emissions

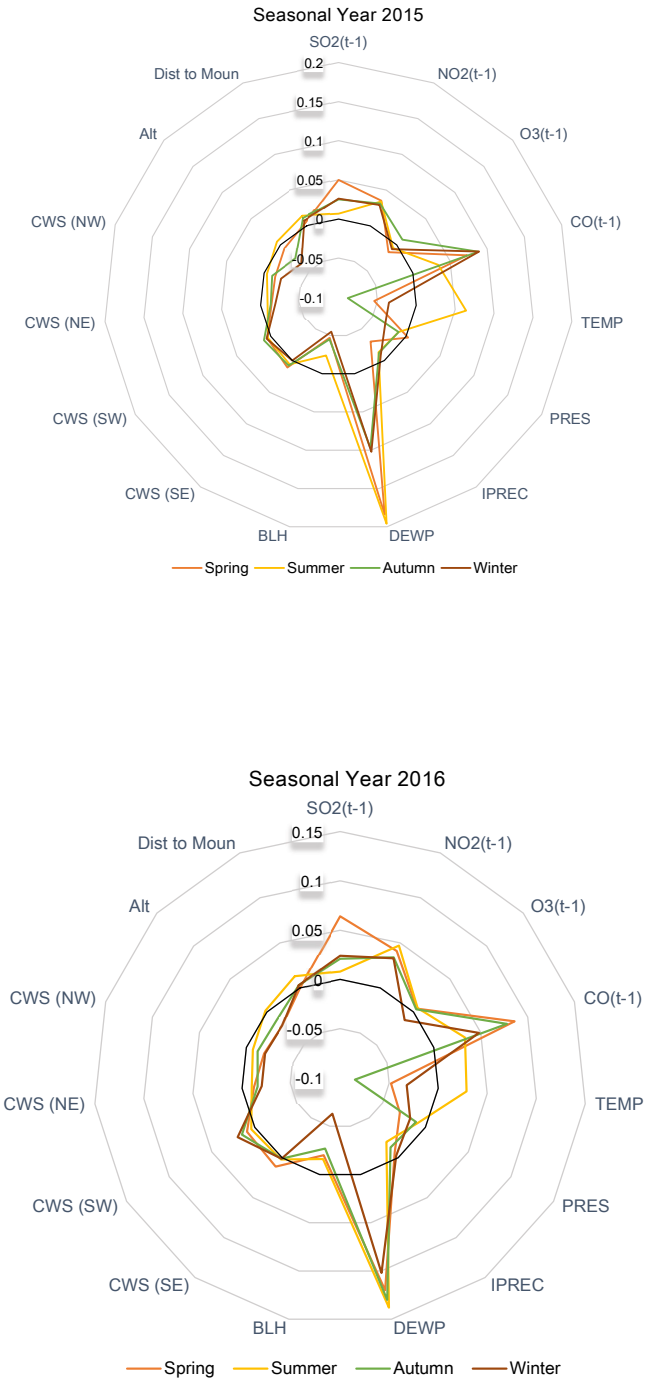


Figure 5: Radar plots of coefficient estimates of gaseous pollutants, meteorological and geographical variables in seasonal years 2015 and 2016, with the grid lines of zero highlighted in bold black.

and fossil fuel combustion, which makes significant contribution to the particulate matter pollution (Guo et al., 2014; Liu et al., 2015). More importantly, CO is a strong indicator of transported pollutant from outside of Beijing area, as it has quite long life span— it can stay in the air for almost a month (Chen et al., 2003; Weinstock, 1969). Between NO_2 and SO_2, the increase in concentration of NO_2 leads to a larger increase in $PM_{2.5}$ concentration than that of SO_2 in all seasons except spring which includes the burning period in March. As NO_2 is mainly generated from motor vehicle emission while SO_2 is associated with coal combustion in power plants and winter heating (Duan et al., 2006), the result indicates that motor vehicle emission has overtaken the coal combustion to become a more dominant emission source since 2015. The inconsistent signs of coefficients of O_3 between the two studied years indicate that O_3 displays a more complex relationship with $PM_{2.5}$, which may depend on other factors such as photochemical processes driven by solar radiation (Meng et al., 1997).

Third, Figure 5 shows the meteorological variables explaining a large part of $PM_{2.5}$ variation too. Among them the dew point temperature has the strongest and positive effect, which is indicated by a significantly larger magnitude of coefficients in comparison with other factors. This can be explained by the fact that dew point is a meteorological variable highly dependent on humidity and the temperature, both of which promote the formation of the secondary aerosols, and thus particulate matters (Yang et al., 2011). Besides, precipitation has a prominent effect on scavenging particulate matters, in that it is conducive to not only depositing particulate matter but also reducing the suspension of fugitive dust (Hu et al., 2006; Li et al., 2015). The correlation between temperature and air pol-

lution has a seasonal pattern. In summer, $PM_{2.5}$ concentration rises with higher temperature, as the summer in Beijing is humid and high temperature enhances the efficiency of secondary fine particulate matter formations (Chen et al., 2003). But in the other seasons, higher temperature (usually in the afternoon) increases the boundary layer height which in turn increases the vertical dispersion of air pollutants (Ito et al., 2007; Lin et al., 2009). The surface air pressure is found to have a negative effect on $PM_{2.5}$ concentration in general, which is similar to the findings in Liang et al. (2015). In addition, we find that the air pressure is not always significant and has the weakest meteorological effect on $PM_{2.5}$.

Furthermore, the boundary layer height (BLH) has a significant negative effect on $PM_{2.5}$ concentration. This supports the claim that relatively higher BLH leads to better vertical dispersion conditions of pollutants, while the lower BLH exacerbates air pollution in all seasons (Miao et al., 2015). Northerly (NW and NE) and southerly (SE and SW) winds have distinct effects on the pollution level. As Beijing's major sources of industrial pollution are from the south where a great many heavy industries are situated, a southerly wind can exacerbate air pollution by bringing the polluted air, and the pollution may remain trapped within the city because of the mountain ranging in the north and west of Beijing. Fortunately, the cleaner and drier wind from the north can help dilute $PM_{2.5}$ (Liang et al., 2015; Zhang et al., 2017a). Hence, lowering emission loading from industrialized southern regions would help to lower Beijing's air pollution level.

We note that the two geographical variables— altitude and distance to mountains, also have significant impacts on $PM_{2.5}$ concentrations. Locations with higher altitude tend to have better dispersion condition and thus have lower pollution lev-

el as compared to those locations with lower altitude, which are usually featured by stable atmosphere and strong temperature inversions (Chan et al., 2005). At the mean time, the mountainous topology in the north and west brings about north-northeasterly mountain-valley breezes, which makes the pollutant emissions from Beijing prone to disperse towards the southeastern plain (Zhao et al., 2009). In addition, the heavy emissions to the south of Beijing further exacerbate the accumulation of air pollutants (Zhang et al., 2013). Hence the southern districts far from mountains suffer more severe pollution than those near mountains.

The estimation results also support the necessity of random effects in Model (1). The average estimate of g among eight seasons is 0.877 with a standard deviation of 0.038, indicating a strong temporal persistency of the latent Markovian dynamic in $z(t)$. Estimates of the scale parameters θ_j, $j = 1,2$, are all above 50. Compared with the average distance 41.2km between monitoring stations, it shows a slow decaying rate of spatial correlation of latent random fields $w_j(s,t)$ as the distance increases. Figure 6 supports this observation by showing the sample correlations between the fitted random components $\hat{w}_j(s,t)$ and $\hat{w}_j(s',t)$ over all pairs of stations $\{s, s'\}$. Recall that $w_j(s,t)$ are location-specific random effects for geographical variables, the strong spatial dependence displays a relatively steady nonlocal contribution from geographical variables to local $PM_{2.5}$ concentrations.

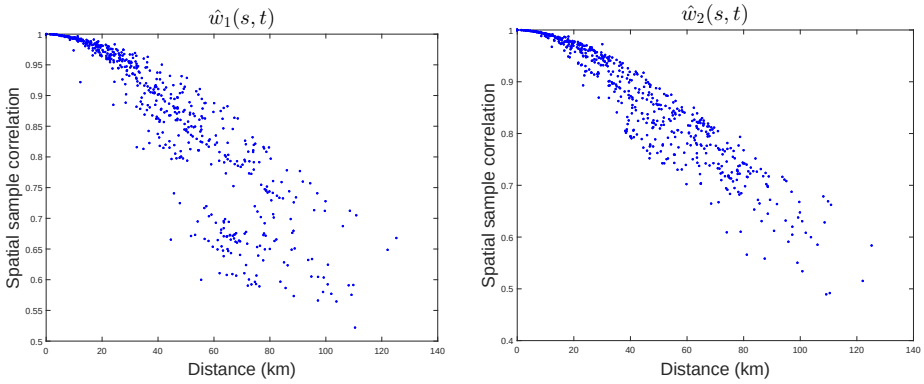


Figure 6: Spatial sample correlation of the fitted random fields $\hat{w}_1(s,t)$ and $\hat{w}_2(s,t)$ between two monitoring sites with respect to their distance. Here we only show the model fitting result for the spring of seasonal year 2015, as other seasons have similar patterns.

4.1.2 Model Assessment

To further demonstrate the benefits of the lagged $PM_{2.5}$ term and random effect components of Model (1), we compare it with four other models:

$$\begin{aligned} y(s,t) = \quad & \mathbf{x}_g'(s,t-1)\boldsymbol{\beta}_g + \mathbf{x}_m'(s,t)\boldsymbol{\beta}_m + x_{Alt}(s)(\beta_{Alt} + \alpha_1 w_1(s,t)) \\ & + x_{Dist}(s)(\beta_{Dist} + \alpha_2 w_2(s,t)) + z(t) + \varepsilon(s,t), \end{aligned} \tag{2}$$

$$\begin{aligned} y(s,t) = \quad & y(s,t-1)\beta_0 + \mathbf{x}_g'(s,t-1)\boldsymbol{\beta}_g + \mathbf{x}_m'(s,t)\boldsymbol{\beta}_m \\ & + x_{Alt}(s)\beta_{Alt} + x_{Dist}(s)\beta_{Dist} + \varepsilon(s,t), \end{aligned} \tag{3}$$

$$\begin{aligned} y(s,t) = \quad & y(s,t-1)\beta_0 + \mathbf{x}_g'(s,t-1)\boldsymbol{\beta}_g + \mathbf{x}_m'(s,t)\boldsymbol{\beta}_m \\ & + x_{Alt}(s)\beta_{Alt} + x_{Dist}(s)\beta_{Dist} + z(t) + \varepsilon(s,t), \end{aligned} \tag{4}$$

$$\begin{aligned} y(s,t) = \quad & y(s,t-1)\beta_0 + \mathbf{x}_g'(s,t-1)\boldsymbol{\beta}_g + \mathbf{x}_m'(s,t)\boldsymbol{\beta}_m \\ & + x_{Alt}(s)(\beta_{Alt} + \alpha_1 w_1(s,t)) + x_{Dist}(s)(\beta_{Dist} + \alpha_2 w_2(s,t)) + \varepsilon(s,t). \end{aligned} \tag{5}$$

Here Model (2) is the original model in Finazzi and Fassò (2014) with the lagged $PM_{2.5}$ term in Model (1) excluded from the covariates. Model (3) – (5) are relatively parsimonious in terms of the random effects compared to Model (1). Specifically, Model (3) includes no random effects, Model (4) includes no spatially correlated random component $w_j(s,t)$ for $j = 1, 2$, and Model (5) includes no temporal random effect $z(t)$.

We assess the fitting ability of these models with the root mean square error (RMSE),

$$\text{RMSE}_{\text{fit}} = \left\{ \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \left(Y(s_i,t) - \hat{Y}(s_i,t) \right)^2 \right\}^{\frac{1}{2}}, \tag{6}$$

where $n = 35$. The fitted value $\hat{Y}(s_i,t)$ of $PM_{2.5}$ concentration at site s_i and time $t \in \{1, \ldots, T\}$ is obtained from the log-standardized fitting given by

$$\hat{y}(s_i,t) = \mathbf{x}'(s_i,t)\hat{\boldsymbol{\beta}} + \hat{z}_t^T + \hat{\alpha}_1 x_{Alt}(s_i)\hat{E}(w_1(s_i,t)|\mathbf{Y}) + \hat{\alpha}_2 x_{Dist}(s_i)\hat{E}(w_2(s_i,t)|\mathbf{Y}),$$

with $\mathbf{x}(s,t)$ being the standardized covariate vector (including the lagged response variable), $\hat{\boldsymbol{\beta}}$, $\hat{\alpha}_1$, $\hat{\alpha}_2$ being the estimated parameters, $\hat{z}_t^T = E_{\hat{\Psi}}(z(t)|\mathbf{Y})$ being the Kalman-smoothed state of the temporal random effect $z(t)$, and $\hat{E}(w_j(s_i,t)|\mathbf{Y}) = E_{\hat{\Psi}}(w_j(s_i,t)|\mathbf{Y})$ being the estimated state of the jth spatial random effect, which is obtained from the properties of multivariate normal distribution.

The RMSEs of Models (1) – (5) are obtained from (6) and shown in Table 3, along with the standard deviations of the raw data which are the fitting RMSEs without any model used. Though the five models all present small fitting errors compared to the raw standard deviations, the proposed Model (1) demonstrates the best fitting performance. Note that Model (4) displays smaller fitting RMSEs than Model (5), which indicates that the temporal random effect $z(t)$ contributes more

to fitting accuracy than the location-specific one. Moreover, Model (2) is inferior in fitting accuracy to the alternative models, which demonstrates the necessity of including the lagged $PM_{2.5}$ concentration as a covariate in our case.

Furthermore, to gain a better understanding of the improvement in fitting after allowing for the lagged $PM_{2.5}$ concentration as a covariate, we conduct a comparative study of the parameter estimates between Model (1) and Model (2). Firstly, the estimates of g from the latter model, which represents the $AR(1)$ coefficient of the temporal random effect $z(t)$, get notably larger and even approaching 1. The average estimate among eight seasons is 0.975 with a standard deviation of 0.01. This indicates that modeling the time persistence of $PM_{2.5}$ concentration with only a latent autoregressive dynamic $z(t)$ may not be adequate for such extensive data in our study, as such large $AR(1)$ coefficient makes $z(t)$ approach a unit-root process which is highly unstationary with a long memory. Besides, without the lagged $PM_{2.5}$ concentration as a covariate, the estimated standard errors of most parameters get larger, in particular for the β coefficients of covariates, which suggests that Model (2) is less stable. In the following section, we will show that our proposed model also has better out-of-sample prediction ability than this model.

[Table 3 about here.]

4.2 Out-of-sample Temporal Prediction

In this section, we look into the out-of-sample forecasting performance for the proposed spatio-temporal Model (1). For each season, we use the data of the last 30 days (with $T_1 = 720$ hours) as the testing set, and the rest as the training set

(with $T_0 = 1488$ hours in springs and summers, 1464 hours in autumns and 2015 winter, and 1440 hours in 2016 winter).

We conduct a rolling prediction in an iterative way. Each time we move the training set forward by one hour while keeping a fixed time window of training observations (T_0 hours). We fit the proposed Model (1) as well as Model (2) in Finazzi and Fassò's framework, and then predict $PM_{2.5}$ values for 1, 2 and 3 hours forward in the out-of-sample. We also use a simple $AR(1)$ model as a benchmark for comparison, where we assume that the dynamic of log-standardized $PM_{2.5}$ concentration $y(s,t)$ at site s and time t follows an $AR(1)$ process

$$y(s,t) = \beta_0 y(s,t-1) + \varepsilon(s,t),$$

with $\varepsilon(s,t)$ a Gaussian white noise with zero mean.

Specifically, let $\hat{Y}_{T_0+r}(s_i, T_0+r+k)$ be the k-step forward prediction for the r-th rolling at site s_i. The k-step rolling prediction root mean square error (RMSE) is obtained after taking average on the testing set:

$$\begin{aligned} \mathrm{RMSE}_{\mathrm{roll}}(k) &= \left\{ \sum_{i=1}^{n} \sum_{r=0}^{T_1-k} \left(Y(s_i, T_0+r+k) - \hat{Y}_{T_0+r}(s_i, T_0+r+k) \right)^2 \right\}^{\frac{1}{2}} \\ &\quad \times (n(T_1-k+1))^{-\frac{1}{2}} . \end{aligned} \tag{7}$$

In the above equation, $\hat{Y}_{T_0+r}(s_i, T_0+r+k)$ is obtained with the r-th k-step rolling prediction of the log-standardized $PM_{2.5}$ $\hat{y}_{T_0+r}(s_i, T_0+r+k)$, with respect to time T_0+r+k and station s_i. The prediction is based on the model fitted to data with a time span from $r+1$ to T_0+r. For Model (1) prediction it is given iteratively by

$$\hat{y}_{T_0+r}(s_i, T_0+r+k) = \hat{\beta}_{0,(r)} \hat{y}_{T_0+r}(s_i, T_0+r+k-1) + \mathbf{x}_c'(s_i, T_0+r+k) \hat{\boldsymbol{\beta}}_{c,(r)} + \hat{g}^k_{(r)} \hat{z}^{T_0+r}_{T_0+r}, \tag{8}$$

where $\mathbf{x}_c(s_i,t)$ is the standardized covariate vector at time t apart from the lagged log-transformed $PM_{2.5}$, $\hat{z}_{T_0+r}^{T_0+r}$ is the Kalman filter output of the r-th rolling fit, and the parameter estimate for the r-th rolling procedure is denoted as $\hat{\Psi}_{(r)} = (\hat{\boldsymbol{\alpha}}_{(r)}, \hat{\boldsymbol{\beta}}_{(r)}, \hat{\sigma}^2_{\varepsilon,(r)}; \hat{\boldsymbol{\theta}}_{(r)}; \hat{g}_{(r)}, \hat{\sigma}^2_{\eta,(r)};\ \hat{\mu}_{0,(r)}, \hat{\sigma}^2_{0,(r)})$, with $\hat{\boldsymbol{\beta}}_{(r)} = (\hat{\beta}_{0,(r)}, \hat{\boldsymbol{\beta}}'_{c,(r)})'$. For predicting Model (2), the term $\hat{\beta}_{0,(r)}\hat{y}_{T_0+r}(s_i, T_0+r+k-1)$ in (8) contributed by predicted lagged log-standardized $PM_{2.5}$ is leaved out. The predicted log-standardized $PM_{2.5}$ from $AR(1)$ model is simply obtained as following:

$$\hat{y}_{T_0+r}(s_i, T_0+r+k) = \hat{\beta}_{0,(r)}\hat{y}_{T_0+r}(s_i, T_0+r+k-1).$$

Table 4 displays the one- to three-step forward rolling prediction RMSEs of each season obtained from (7) comparing the three methods: Model (1), Model (2) and the $AR(1)$ model. The raw standard deviations (SD) of responses in the testing sets of eight seasons are shown as a benchmark, and the average RMSEs and SD over the eight seasons are also provided as a summary. We can easily see that the proposed Model (1) produces overall smaller RMSEs for out-of-sample temporal prediction up to three steps forward, compared to both Model (2) without allowing for the lagged $PM_{2.5}$ as a covariate and the simple $AR(1)$ model without other covariates as well as random effects. More specifically, compared to Model (2), the proposed model improves the prediction accuracy in particular for the near future after allowing for the $AR(1)$ dependence of $PM_{2.5}$. Besides, comparing the two spatio-temporal Models (1) and (2) with the simple $AR(1)$ model, we find that the space-time dependence and the use of covariates in models help to improve the accuracy of forecasts for relatively far forward. From Table 4, we can see that the simple $AR(1)$ model produces overall larger root mean square errors than those obtained from Model (1) for two and three steps forward in particular, as well as

in comparison with Model (2) for three steps forward.

Though the relatively large prediction RMSEs for the autumns and winters is reconciled with the large variation and level of $PM_{2.5}$ concentration in those periods as shown in Table 4 and Figure 2, they sometimes exceed the ranges of some $PM_{2.5}$ pollution levels and thus become less informative. In view of this, we also consider the out-of-sample prediction for the $PM_{2.5}$ pollution severity in terms of 7 different levels according to the Air Quality Standard in China, with Levels 1–7 standing for the $PM_{2.5}$ concentration in $\mu g/m^3$ falling inside the intervals $[0, 35)$, $[35, 75)$, $[75, 115)$, $[115, 150)$, $[150, 250)$, $[250, 500)$ and $[500, +\infty)$, respectively. The out-of-sample prediction for the $PM_{2.5}$ levels is also studied in Gao et al. (2019)), whereas they adopted a vector autoregressive model with no other covariates and is thus lacking in interpretation for the dynamics of $PM_{2.5}$ concentration. Table 5 presents the percentages of correct one- to three-step forward rolling predictions at each of the 7 levels as well as for overall levels over the two seasonal years consisting of eight seasons, comparing Models (1), (2) and the simple $AR(1)$ model. We can obtain similar findings from Table 5 as above. The proposed Model (1) still produces overall more accurate out-of-sample predictions for the $PM_{2.5}$ pollution levels in comparison with Model (2) and the $AR(1)$ model.

The above comparisons among various models indicate the efficacy of the proposed model in out-of-sample temporal prediction, and confirms the utility of taking both $AR(1)$ dynamics and space-time dependence of $PM_{2.5}$ concentration into consideration.

[Table 4 about here.]

[Table 5 about here.]

5 Discussion

Our study shows that the spatio-temporal model with varying coefficients provides us with a comprehensive description of the formation of $PM_{2.5}$ in Beijing region with satisfactory in-sample fitting. The estimates from the model offer valuable insights in designing effective management strategies to alleviate air pollution in Beijing. For example, we can draw focus on controlling motor vehicle emission as we found that NO_2 leads to a great increase in $PM_{2.5}$ concentration in most seasons. CO was also found to have large influence on $PM_{2.5}$ concentration, which has quite long life span and can be transported from outside of Beijing area. In addition, the southerly wind was discovered to be mostly relevant to an increase in air pollution. Hence, decreasing emission loadings from industrialized southern regions will help to lower Beijing's air pollution level.

We further establish the adequacy of the proposed model by demonstrating its out-of-sample temporal predictions for both the $PM_{2.5}$ concentration and the pollution levels. Based on a rolling prediction for each studied season, the model shows a decent forecasting performance, with the multi-step prediction accuracy notably outperforming that of other existing methods including the original Finazzi and Fassò (2014) model framework and the simple $AR(1)$ model.

There also exist some limitations in the present study, which need to be cautious about and potentially can be overcome in future studies. We note that, though the proportion of the missing data is small, the imputation for missing covariates (including the missing response of $PM_{2.5}$ concentration) before modelling may result in bias and sensitivity issues for inference. For potentially tackling the missing data in the future, we suggest possible avenues are through making use of

the space-time dependence structure of the missing response (Calculli et al., 2015; Boaz et al., 2019; Padilla et al., 2020), or building a multiple regression model for the missing covariates (Yi et al., 2011).

Acknowledgments

The study was supported by China's National Key Research Special Program (2016YFC0207701, 2016YFC0207702). We would also like to thank the support from Center for Statistical Science in Peking University, and the Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University), Ministry of Education. We thank Shuyi Zhang, Ziping Xu and Huijie Liu from Peking University, Hengfang Wang from Iowa State University and Wenqing Wang for their generous helps with data and domain knowledge. The boundary layer height data are obtained from `http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc`, and the altitude data are extracted from Google Maps. Other data used in the paper are available upon request from S.X.C. (csx@gsm.pku.edu.cn). We have no competing interests.

References

Belsley, D. A., Kuh, E., and Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons, Hoboken, New Jersey.

Berrocal, V. J., Gelfand, A. E., Holland, D. M., Burke, J., and Miranda, M. L.

(2011). On the use of a PM2.5 exposure simulator to explain birthweight. *Environmetrics*, 22(4):553–571.

Boaz, R., Lawson, A., and Pearce, J. (2019). Multivariate air pollution prediction modeling with partial missingness. *Environmetrics*, 30(7):e2592.

Brumback, B. A., Ryan, L. M., Schwartz, J. D., Neas, L. M., Stark, P. C., and Burge, H. A. (2000). Transitional regression models, with application to environmental time series. *Journal of the American Statistical Association*, 95(449):16–27.

Cai, W., Li, K., Liao, H., Wang, H., and Wu, L. (2017). Weather conditions conducive to Beijing severe haze more frequent under climate change. *Nature Climate Change*, 7(4):257–262.

Calculli, C., Fassò, A., Finazzi, F., Pollice, A., and Turnone, A. (2015). Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in Apulia, Italy. *Environmetrics*, 26(6):406–417.

Chan, C., Xu, X., Li, Y., Wong, K., Ding, G., Chan, L., and Cheng, X. (2005). Characteristics of vertical profiles and sources of PM2.5, PM10 and carbonaceous species in Beijing. *Atmospheric Environment*, 39(28):5113–5124.

Cheam, A., Marbac, M., and McNicholas, P. (2017). Model-based clustering for spatiotemporal data on air quality monitoring. *Environmetrics*, 28(3):e2437.

Chen, B., Lu, S., Li, S., and Wang, B. (2015). Impact of fine particulate fluctuation and other variables on Beijing's air quality index. *Environmental Science and Pollution Research*, 22(7):5139–5151.

Chen, L., Guo, B., Huang, J., He, J., Wang, H., Zhang, S., and Chen, S. X. (2018). Assessing air-quality in Beijing-Tianjin-Hebei region: The method and mixed tales of PM2.5 and O3. *Atmospheric Environment*, 193:290–301.

Chen, L.-W. A., Chow, J. C., Doddridge, B. G., Dickerson, R. R., Ryan, W. F., and Mueller, P. K. (2003). Analysis of a summertime PM2.5 and haze episode in the mid-Atlantic region. *Journal of the Air & Waste Management Association*, 53(8):946–956.

Clifford, S., Low-Choy, S., Mazaheri, M., Salimi, F., Morawska, L., and Mengersen, K. (2019). A Bayesian spatiotemporal model of panel design data: Airborne particle number concentration in Brisbane, Australia. *Environmetrics*, 30(7):e2597.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Duan, F., He, K., Ma, Y., Yang, F., Yu, X., Cadle, S., Chan, T., and Mulawa, P. (2006). Concentration and chemical characteristics of PM2.5 in Beijing, China: 2001–2002. *Science of the Total Environment*, 355(1-3):264–275.

Fassò, A. and Finazzi, F. (2011). Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics*, 22(6):735–748.

Fassò, A. and Finazzi, F. (2013). A varying coefficients space-time model for

ground and satellite air quality data over Europe. *Statistica & Applicazioni, Special Online Issue*, pages 45–56.

Finazzi, F. and Fassò, A. (2014). D-STEM: a software for the analysis and mapping of environmental space-time variables. *Journal of Statistical Software*, 62(6):1–29.

Gao, Z., Ma, Y., Wang, H., and Yao, Q. (2019). Banded spatio-temporal autoregressions. *Journal of Econometrics*, 208:211–230.

Guo, S., Hu, M., Zamora, M. L., Peng, J., Shang, D., Zheng, J., Du, Z., Wu, Z., Shao, M., Zeng, L., Molina, M. J., and Zhang, R. (2014). Elucidating severe urban haze formation in China. *Proceedings of the National Academy of Sciences*, 111(49):17373–17378.

Hu, M., Liu, S., Wu, Z.-J., Zhang, J., Zhao, Y.-L., Wehner, B., and Wiedensolher, A. (2006). Effects of high temperature, high relative humidity and rain process on particle size distributions in the summer of Beijing. *Huan jing ke xue= Huanjing kexue*, 27(11):2293–2298.

Ito, K., Thurston, G. D., and Silverman, R. A. (2007). Characterization of PM2.5, gaseous pollutants, and meteorological interactions in the context of time-series health effects models. *Journal of Exposure Science & Environmental Epidemiology*, 17(S2):S45–S60.

Li, W. K. (1994). Time series models based on generalized linear models: Some further results. *Biometrics*, 50(2):506–511.

Li, Y., Chen, Q., Zhao, H., Wang, L., and Tao, R. (2015). Variations in PM10, PM2.5 and PM1.0 in an urban area of the Sichuan Basin and their relation to meteorological factors. *Atmosphere*, 6(1):150–163.

Liang, X., Li, S., Zhang, S., Huang, H., and Chen, S. X. (2016). PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. *Journal of Geophysical Research: Atmospheres*, 121(17):10,220–10,236.

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., and Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257.

Lin, J., Liu, W., Li, Y., Bao, L., Li, Y., Wang, G., and Wu, W. (2009). Relationship between meteorological conditions and particle size distribution of atmospheric aerosols. *Journal of Meteorology and Environment*, 25(1):1–5.

Liu, Z., Hu, B., Wang, L., Wu, F., Gao, W., and Wang, Y. (2015). Seasonal and diurnal variation in particulate matter (PM10 and PM2.5) at an urban site of Beijing: analyses from a 9-year study. *Environmental Science and Pollution Research*, 22(1):627–642.

Meng, Z., Dabdub, D., and Seinfeld, J. H. (1997). Chemical coupling between atmospheric ozone and particulate matter. *Science*, 277(5322):116–119.

Miao, Y., Hu, X.-M., Liu, S., Qian, T., Xue, M., Zheng, Y., and Wang, S. (2015). Seasonal variation of local atmospheric circulations and boundary layer structure

in the Beijing-Tianjin-Hebei region and implications for air quality. *Journal of Advances in Modeling Earth Systems*, 7(4):1602–1626.

Nicolis, O., Díaz, M., Sahu, S. K., and Marín, J. C. (2019). Bayesian spatiotemporal modeling for estimating short-term exposure to air pollution in Santiago de Chile. *Environmetrics*, 30(7):e2574.

Padilla, L., Lagos-Álvarez, B., Mateu, J., and Porcu, E. (2020). Space-time autoregressive estimation and prediction with missing data based on Kalman filtering. *Environmetrics*, page e2627.

Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., and Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA*, 287(9):1132–1141.

Sahu, S. K. (2012). Hierarchical Bayesian models for space-time air pollution data. In Rao, T. S., Rao, S. S., and Rao, C., editors, *Time Series Analysis: Methods and Applications*, volume 30 of *Handbook of Statistics*, pages 477–495. Elsevier.

Shao, Q., Wong, H., Ip, W.-C., and Li, M. (2010). Effect of ambient air pollution on respiratory illness in Hong Kong: a regional study. *Environmetrics*, 21(2):173–188.

Sun, Y., Wang, Z., Du, W., Zhang, Q., Wang, Q., Fu, P., Pan, X., Li, J., Jayne, J., and Worsnop, D. (2015). Long-term real-time measurements of aerosol particle composition in Beijing, China: seasonal variations, meteorological effects, and source analysis. *Atmospheric Chemistry and Physics*, 15(17):10149–10165.

Sun, Y., Zhuang, G., Tang, A., Wang, Y., and An, Z. (2006). Chemical characteristics of PM2.5 and PM10 in haze-fog episodes in Beijing. *Environmental Science & Technology*, 40(10):3148–3155.

Tang, G., Zhang, J., Zhu, X., Song, T., Münkel, C., Hu, B., Schäfer, K., Liu, Z., Zhang, J., Wang, L., et al. (2016). Mixing layer height and its implications for air pollution over Beijing, China. *Atmospheric Chemistry and Physics*, 16(4):2459–2475.

Weinstock, B. (1969). Carbon monoxide: Residence time in the atmosphere. *Science*, 166(3902):224–225.

Xu, Z., Chen, S. X., and Wu, X. (2020). Meteorological change and impacts on air pollution – results from North China. *Journal of Geophysical Research: Atmospheres*, page e2020JD032423.

Yang, L., Wu, Y., Davis, J. M., and Hao, J. (2011). Estimating the effects of meteorology on PM2.5 reduction during the 2008 Summer Olympic Games in Beijing, China. *Frontiers of Environmental Science & Engineering in China*, 5(3):331–341.

Yeatts, K., Svendsen, E., Creason, J., Alexis, N., Herbst, M., Scott, J., Kupper, L., Williams, R., Neas, L., Cascio, W., Devlin, R. B., and Peden, D. B. (2007). Coarse particulate matter (PM2.5–10) affects heart rate variability, blood lipids, and circulating eosinophils in adults with asthma. *Environmental Health Perspectives*, 115(5):709–714.

Yi, G. Y., Liu, W., and Wu, L. (2011). Simultaneous inference and bias analysis

for longitudinal data with covariate measurement error and missing responses. *Biometrics*, 67(1):67–75.

Zhang, A., Qi, Q., Jiang, L., Zhou, F., and Wang, J. (2013). Population exposure to PM2.5 in the urban area of Beijing. *PloS one*, 8(5):e63486.

Zhang, S., Guo, B., Dong, A., He, J., Xu, Z., and Chen, S. X. (2017a). Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 473(2205):20170457.

Zhang, S., Guo, B., Dong, A., He, J., Xu, Z., and Chen, S. X. (2017b). Supplementary Material for Cautionary Tales on Air Quality Improvement in Beijing from Cautionary tales on air-quality improvement in Beijing.

Zhao, X., Zhang, X., Xu, X., Xu, J., Meng, W., and Pu, W. (2009). Seasonal and diurnal variations of ambient PM2.5 concentration in urban and rural environments in Beijing. *Atmospheric Environment*, 43(18):2893–2900.

Table 1: Parameter estimates of Model (1) and their levels of significance (* for p-value <0.05; ** for p-value <0.01; *** for p-value <0.001) for four seasons in seasonal year 2015, along with the fitting root mean square errors (RMSE) and the standard deviations of $PM_{2.5}$ (Raw SD).

covariates	spring	summer	autumn	winter
$PM_{2.5}$(t–1)	0.589***	0.639***	0.708***	0.685***
SO_2(t–1)	0.050***	0.007***	0.025***	0.026***
NO_2(t–1)	0.035***	0.034***	0.031***	0.029***
O_3(t–1)	-0.014***	-0.006**	0.01***	-0.008***
CO(t–1)	0.073***	0.034***	0.088***	0.089***
Temperature	-0.054***	0.064***	-0.088***	-0.035***
Pressure	0.003	-0.011***	-0.011***	-0.031***
Precipitation	-0.030***	-0.013***	-0.012***	-0.008***
Dew point	0.184***	0.196***	0.097***	0.102***
Boundary layer height	-0.047***	-0.024***	-0.045***	-0.055***
Wind speed (SE)	0.011***	0.006***	0.008***	0.000
Wind speed (SW)	0.005*	0.004*	0.010***	0.006***
Wind speed (NE)	-0.014***	-0.009***	-0.013***	-0.019***
Wind speed (NW)	-0.015***	-0.004*	-0.011***	-0.023***
Altitude	-0.007*	0.006*	-0.026***	-0.036***
Distance to mountains	0.004*	0.014***	0.011***	0.007***
α_{Alt}	0.064***	0.060***	0.043***	0.036***
α_{Dist}	0.058***	0.041***	0.044***	0.031***
θ_{Alt}	51.105***	60.225***	51.936***	59.706***
θ_{Dist}	55.326***	62.430***	63.734***	71.859***
g	0.911***	0.932***	0.888***	0.826***
σ_η^2	0.008***	0.008***	0.005***	0.005***
σ_ε^2	0.088***	0.091***	0.059***	0.052***
RMSE	15.728	11.428	17.579	25.233
Raw SD	61.310	44.027	89.025	120.754

Table 2: Parameter estimates of Model (1) and their levels of significance (* for p-value <0.05; ** for p-value <0.01; *** for p-value <0.001) for four seasons in seasonal year 2016, along with the fitting root mean square errors (RMSE) and the standard deviations of $PM_{2.5}$ (Raw SD).

covariates	spring	summer	autumn	winter
$PM_{2.5}$(t–1)	0.637***	0.700***	0.735***	0.697***
SO_2(t–1)	0.064***	0.008***	0.021***	0.024***
NO_2(t–1)	0.041***	0.047***	0.034***	0.033***
O_3(t–1)	0.005*	0.006***	0.004*	-0.012***
CO(t–1)	0.086***	0.033***	0.078***	0.048***
Temperature	-0.048***	0.029***	-0.085***	-0.032***
Pressure	-0.030***	-0.009**	-0.011***	-0.018***
Precipitation	-0.006***	-0.020***	-0.013***	-0.003
Dew point	0.120***	0.138***	0.130***	0.102***
Boundary layer height	-0.020***	-0.016**	-0.027***	-0.063***
Wind speed (SE)	0.011**	0.002	0.001	0.001
Wind speed (SW)	0.009***	0.004*	0.015***	0.020***
Wind speed (NE)	-0.012***	-0.010***	-0.016***	-0.020***
Wind speed (NW)	-0.019***	-0.007***	-0.012***	-0.020***
Altitude	-0.021***	0.002	-0.013***	-0.021***
Distance to mountains	0.000	0.013***	0.003*	0.003**
α_{Alt}	0.050***	0.053***	0.056***	0.039***
α_{Dist}	0.049***	0.057***	0.027***	0.032***
θ_{Alt}	60.498***	67.863***	64.322***	61.482***
θ_{Dist}	66.365***	59.98***	59.903**	65.802***
g	0.849***	0.908***	0.846***	0.853***
σ_η^2	0.009***	0.008***	0.006***	0.006***
σ_ε^2	0.088***	0.077***	0.054***	0.054***
RMSE	14.695	9.573	13.350	25.897
Raw SD	74.122	41.333	72.197	123.046

Table 3: Root mean square fitting errors of the proposed Model (1), three of its parsimonious versions (3) – (5), and Model (2) without lagged $PM_{2.5}$ in covariates, along with the standard deviations of the raw $PM_{2.5}$ (Raw SD), for the eight seasons.

models	Seasonal Year 2015				Seasonal Year 2016			
	Spring	Summer	Autumn	Winter	Spring	Summer	Autumn	Winter
Model (1)	**15.728**	**11.428**	**17.579**	**25.233**	**14.695**	**9.573**	**13.350**	**25.897**
Model (4)	16.806	12.042	18.014	26.246	15.612	10.413	14.131	27.096
Model (5)	18.893	14.305	20.262	28.890	19.044	11.689	15.573	31.003
Model (3)	20.420	15.155	21.334	31.066	20.609	13.607	17.572	32.492
Model (2)	20.283	15.066	22.462	37.679	22.417	13.766	21.532	40.742
Raw SD	61.310	44.027	89.025	120.754	74.122	41.333	72.197	123.046

Table 4: Root mean square errors (RMSE) of one-step, two-step and three-step forward rolling predictions obtained with the proposed Model (1), Model (2) without lagged $PM_{2.5}$ in covariates, and the simple $AR(1)$ model, along with the standard deviations of the raw $PM_{2.5}$ (Raw SD). The RMSEs and Raw SDs are computed over the testing sets consisting of the last 30 days in each of the eight seasons respectively as well as after taking averages.

models		Seasonal Year 2015				Seasonal Year 2016				
		Spring	Summer	Autumn	Winter	Spring	Summer	Autumn	Winter	mean
Model (1)	1-step	15.447	10.082	25.529	25.922	15.586	8.587	20.134	21.714	17.875
	2-step	19.101	13.786	38.985	37.220	22.492	12.358	29.982	31.006	25.616
	3-step	21.092	16.054	47.978	43.697	26.902	14.875	37.188	37.131	30.615
Model (2)	1-step	20.250	15.097	41.543	36.137	22.853	12.775	31.447	37.440	27.193
	2-step	21.092	16.019	45.395	39.032	27.160	13.941	33.651	40.430	29.590
	3-step	21.907	17.031	49.388	41.697	30.652	15.168	35.955	43.743	31.943
$AR(1)$	1-step	16.823	11.239	24.893	26.820	16.857	9.533	21.010	21.275	18.556
	2-step	22.034	16.709	40.562	41.357	24.611	14.297	33.105	33.318	28.249
	3-step	25.551	20.504	52.171	51.285	29.491	17.376	41.806	41.823	35.001
Raw SD		41.727	37.196	107.02	79.366	41.454	33.607	83.064	86.229	63.708

Table 5: Percentages of correct one-step, two-step and three-step forward rolling predictions for $PM_{2.5}$ pollution levels obtained with the proposed Model (1), Model (2) without lagged $PM_{2.5}$ in covariates, and the simple $AR(1)$ model. The percentages are computed with respect to the 7 different pollution levels respectively as well as for overall levels, over the two seasonal years across the 35 stations.

models		Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	All
Model (1)	1-step	93.6	81.4	74.1	59.8	76.1	70.3	61.7	83.3
	2-step	90.0	74.7	63.4	44.8	64.0	55.4	44.5	76.1
	3-step	87.3	70.6	57.1	35.5	54.5	44.4	38.3	71.3
Model (2)	1-step	91.3	71.9	59.8	41.0	61.0	59.6	45.5	75.1
	2-step	90.2	69.7	56.4	37.8	57.3	56.5	40.7	72.9
	3-step	89.0	67.1	53.5	35.2	53.9	53.4	36.2	70.8
$AR(1)$	1-step	92.5	81.1	71.9	56.5	75.6	68.9	67.8	82.2
	2-step	88.0	73.5	57.2	35.9	60.7	46.1	46.8	72.9
	3-step	84.2	68.9	46.2	25.0	48.6	30.2	23.3	66.4

Notes: In this table, Levels 1–7 stand for the $PM_{2.5}$ concentration in $\mu g/m^3$ falling inside the intervals $[0, 35)$, $[35, 75)$, $[75, 115)$, $[115, 150)$, $[150, 250)$, $[250, 500)$ and $[500, +\infty)$, respectively.