



Spatio-temporal analysis of air pollution in North China Plain

Le Chang¹ · Tao Zou¹

Received: 15 November 2020 / Revised: 23 September 2021 / Accepted: 23 October 2021 /
Published online: 25 November 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Accompanying China's rapid industrialization, a vast area of the country, particularly the Beijing–Tianjin–Hebei (BTH) region, has significantly experienced concerning levels of air pollution over the past decade. Exposure to severe particulate matter (PM), $PM_{2.5}$ in particular, it raises a crucial public health concern, but quantifying $PM_{2.5}$ accurately across large geographic areas and across time poses a great challenge. To investigate $PM_{2.5}$ concentration in the BTH region, we utilize a spatio-temporal mixed effects model that includes geographic information system-based time-invariant spatial variables and time-varying meteorological covariates. Our kriging results find that $PM_{2.5}$ concentration is hazardous in the North China Plain (NCP), where major iron, steel, and cement industries are located. More importantly, our analysis of the impact of wind finds that the severity of air pollution highly depends on the direction of the wind. That is, a northerly wind can considerably reduce the level of $PM_{2.5}$ in the NCP, while a southerly wind generally does not alleviate air pollution and sometimes even increases it. Using prediction error as a proxy for the level of local emissions, we find that Shijiazhuang and Tangshan produce the most significant local emissions, which coincides with a heavier industry in these two cities. During the winter heating period, we find that the two densely populated cities of Beijing and Tianjin have dramatic increases in local emissions because of the massive coal consumption during this period.

Keywords Spatio-temporal model · Kriging · Air pollution · Meteorological condition · North China Plain · Winter heating

Handling Editor: Pierre R. L. Dutilleul

✉ Le Chang
le.chang@anu.edu.au
Tao Zou
tao.zou@anu.edu.au

¹ Research School of Finance, Actuarial Studies and Statistics, College of Business and Economics, Australian National University, Canberra, ACT, Australia

1 Introduction

During the period of rapid industrialization over the past decade, a vast area of China has experienced severe and chronic air pollution. The main pollutants are fine particulate matter (PM), $PM_{2.5}$ in particular, which consists of airborne particles with aerodynamic diameters of less than $2.5 \mu m$. Research has demonstrated that exposure to $PM_{2.5}$ can create severe public health concerns (Pope et al. 2009). Such exposure can cause lung morbidity, serious respiratory diseases, and even death (Donaldson et al. 1998). It can also negatively affect the climate, agriculture, ecosystems, and many other aspects of society (Zhao et al. 2013a).

The Beijing–Tianjin–Hebei (BTH) region is a region of North China where economic development is most active. It consists of two municipalities, Beijing and Tianjin, and other 11 cities in Hebei province. The North China Plain (NCP) that surrounds the BTH region has experienced China's most severe air pollution, with excessive $PM_{2.5}$ concentration recorded. According to China's Ministry of Environmental Protection, eight of the ten most polluted cities in China were in the BTH region, including Xingtai, Shijiazhuang, Baoding, Handan, Langfang, Hengshui, Tangshan, and Beijing (Wang et al. 2013). On average, 6.6 cities from the BTH region represented the 10 most polluted cities every month from 2013 to 2016. Considering that a large population in Beijing and in many other cities in Hebei has been exposed to a high level of $PM_{2.5}$ for a long period, it is vital to measure the severity of the $PM_{2.5}$ concentration in this region.

Extensive studies have been conducted to investigate the composition, characteristics, and sources of air pollutants in BTH (Zhao et al. 2013b; Li et al. 2017). As stated in Chen et al. (2018), the extremely high level of $PM_{2.5}$ concentration is widely considered to be causing the excessive local emission of pollutants from heavy industries, ranging from iron, steel, and cement production to the rapidly increasing number of motor vehicles being used on the roads. Apart from local emissions, the meteorological conditions, including temperature, relative humidity, and wind, can also strongly impact the level of $PM_{2.5}$ concentration (Wang et al. 2014; Liang et al. 2015; Yin et al. 2016). For example, Wang et al. (2014) stated that high humidity can contribute to aerosols and secondary transformation under high emissions, and resulted in severe $PM_{2.5}$ episodes in Beijing in January 2013. In addition, Liang et al. (2015) stated that the south and east of Beijing on the NCP are densely populated with heavy industries, which produce enormous amounts of air pollutants. These air pollutants originating in the NCP are likely to be transported to the north and influence the level of $PM_{2.5}$ in its downwind areas. Therefore, it is also crucial to evaluate the effect of the air quality in the neighboring areas of the NCP.

To capture both spatial and temporal effects, many spatio-temporal models for analyzing air pollution data have been developed under a hierarchical modeling framework over the last decade (Sahu et al. 2006; McMillan et al. 2010; Fassò 2013; Shaddick et al. 2013). For example, Sahu et al. (2006) proposed a random effect model for $PM_{2.5}$ concentration which introduces two random effects components to account for urban and rural differences in respect of both mean levels and variability. Shaddick et al. (2013) utilized a Bayesian hierarchical model to examine the effects of human activity on pollution in urban areas by isolating global effects on pollution concen-

trations. Moreover, Calculli et al. (2015) introduced a multivariate generalization of a well-known univariate hidden dynamic geostatistical (HDG) model and implemented the expectation maximization (EM) algorithm to obtain the parameter estimates (see Finazzi and Fassò (2014) for details of the algorithm). Fassò et al. (2016) further utilized this model to estimate the distribution of European population exposure to airborne pollutants. To study the level of $PM_{2.5}$ concentration in BTH region, we utilize a univariate spatio-temporal mixed effects model (Cressie and Wikle 2015) that includes both geographic information system (GIS)-based time-invariant variables and time-varying meteorological covariates to investigate and quantify the severity of $PM_{2.5}$ pollution in BTH. More specifically, using kriging results, we examine which areas of BTH on a particular day are heavily polluted and can be unhealthy to live in. We further analyze the impact of meteorological conditions (e.g., wind from different directions) based on universal kriging. By performing several quasi-experiments, we examine how the wind from different directions interacting with the location of mountains affects the level of $PM_{2.5}$ in BTH.

Importantly, our spatio-temporal model can provide greater insight into the local emissions produced in each city in BTH. The level of local emissions strongly influences the level of $PM_{2.5}$ but data for the level of local emissions are generally not publicly available. After removing the effect of geographic and meteorological conditions, the prediction error produced from the model can serve as a good proxy for the level of local emissions. We then perform an analysis of the impact of winter heating on local emissions. Our results reveal that two heavily industrialized cities—Shijiazhuang and Tangshan—produce the highest level in BTH of local emissions before the winter heating period, but during the winter heating period, we find that two densely populated cities—Beijing and Tianjin—have a dramatic increase in local emissions because of the massive coal consumption for winter heating.

2 Data

2.1 Data description

The BTH region is a coastal region that consists of 13 cities located in North China. Their locations are shown in Fig. 1. We collect $PM_{2.5}$ data from 90 monitoring sites located in the BTH region (see Fig. 2). The time-invariant geographic variables considered in this study are longitude (Lon), latitude (Lat), altitude (Alt), and distance to mountain (Dm). We add “distance to mountain” in our investigation given that “mountain” may play a very important role in blocking pollution transportation. We define “mountain” as areas with altitudes exceeding 500. The altitude of the BTH region and the locations of “mountain” in the region are presented in Fig. 3. This figure, along with Fig. 2, demonstrates that Beijing is surrounded by mountains: Taihang Mountains to the west and Yan Mountain to the north. The yellow area without mountains belongs to NCP. NCP is a quite large area that consists of some parts of Shandong province, Henan province, and most of Hebei province. We restrict this study to analyzing the Hebei region.

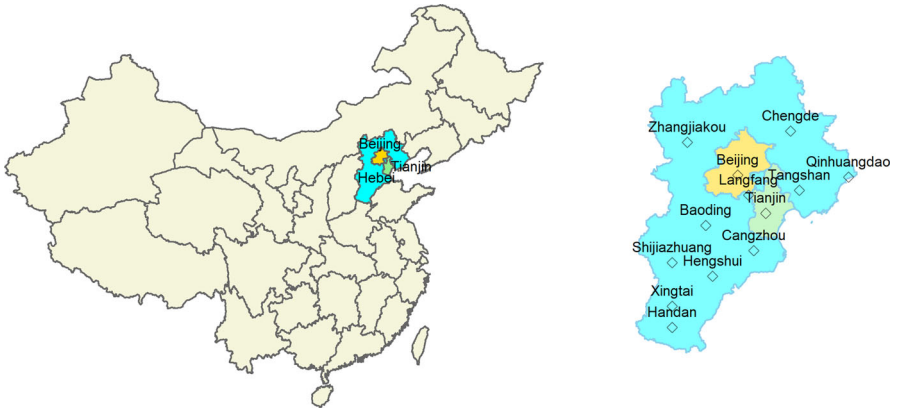


Fig. 1 Location of BTH region in China (Mainland) and location of cities in BTH region

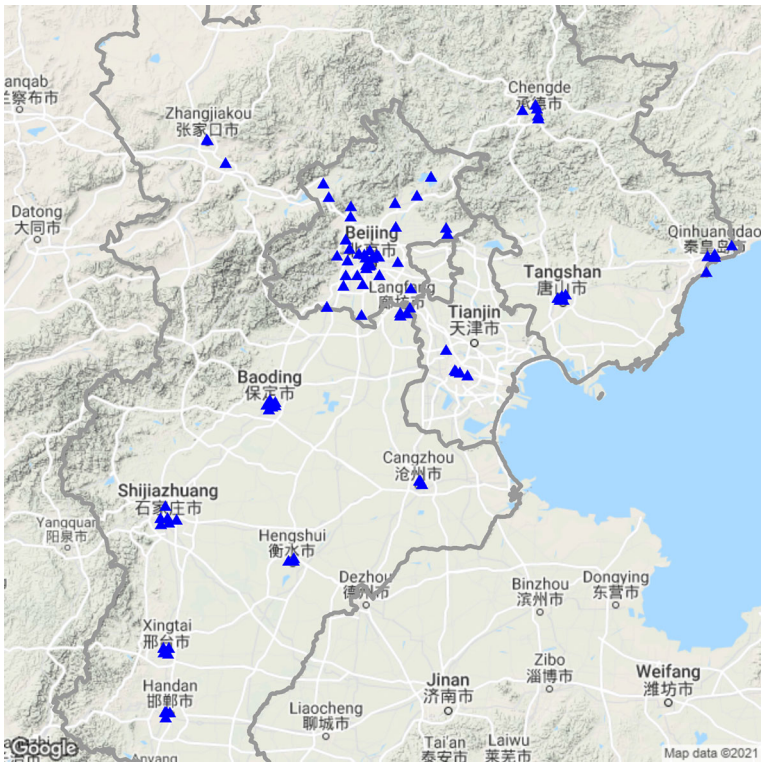


Fig. 2 Locations of the monitoring sites in the BTH region

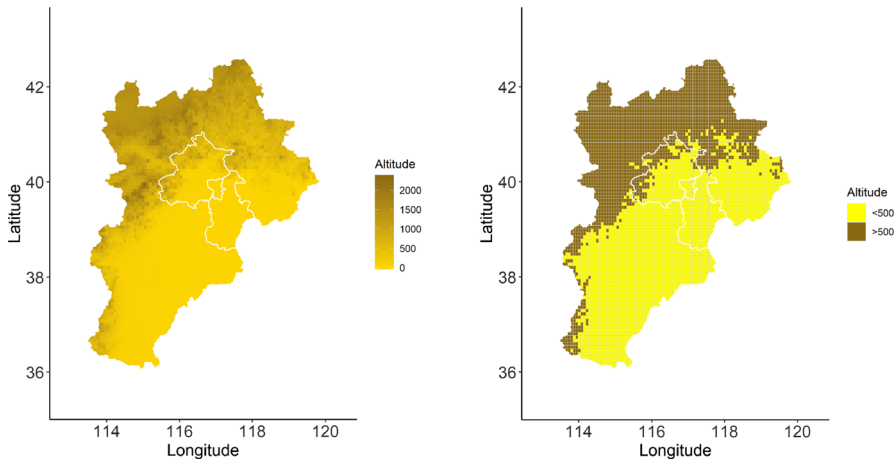


Fig. 3 The altitude of the BTH region (left). Areas with altitudes exceeding 500 are defined as “mountain”(right)

Many previous studies have found that meteorological conditions have a significant impact on the levels of $PM_{2.5}$ (Paciorek et al. 2009; Liang et al. 2015, 2016; Zhang et al. 2017). Similar to past studies on Beijing’s $PM_{2.5}$ (Liang et al. 2015), the time-varying meteorological variables included in this study are air pressure (Pres), temperature (Temp), dew point (Dewp), and wind—north-west (NW), north-east (NE), south-east (SE), south-west (SW), and calm and variable (CV), where CV is a category of the wind such that the wind direction is not well-established or the wind speed is less than 0.5 meter per second (0.5m/s). Following Alduchov and Eskridge (1996) and Liang et al. (2015), we do not consider relative humidity in our model because it can be determined by temperature and dew point based on a physical relationship. The original wind data are given as wind speed and direction. To ease the analysis, we create five variables (NW, NE, SE, SW, CV) based on the direction of wind, with the associated variable values being wind speed.

2.2 Data aggregation and data overview

The original data of the $PM_{2.5}$ and other meteorological variables are collected hourly. However, the hourly data are too noisy and difficult to predict. Given the variability of the hourly data, we aggregate hourly data into daily data in our analysis. It is also beneficial to investigate daily data given the issue of missing data. Figure 4 demonstrates that if we aggregate hourly data into daily data, there will be no missing explanatory variables, and the proportion of “missings” for $PM_{2.5}$ drops from 6% to 3%. After our preliminary investigation, we find that the way of missing is not related to the levels of $PM_{2.5}$, and there is no evidence of any patterns in the missing values. Thus we believe that the missing values can be considered missing at random. For simplicity, we impute the missing data using multivariate imputation by chained equations built in R package *mice*.

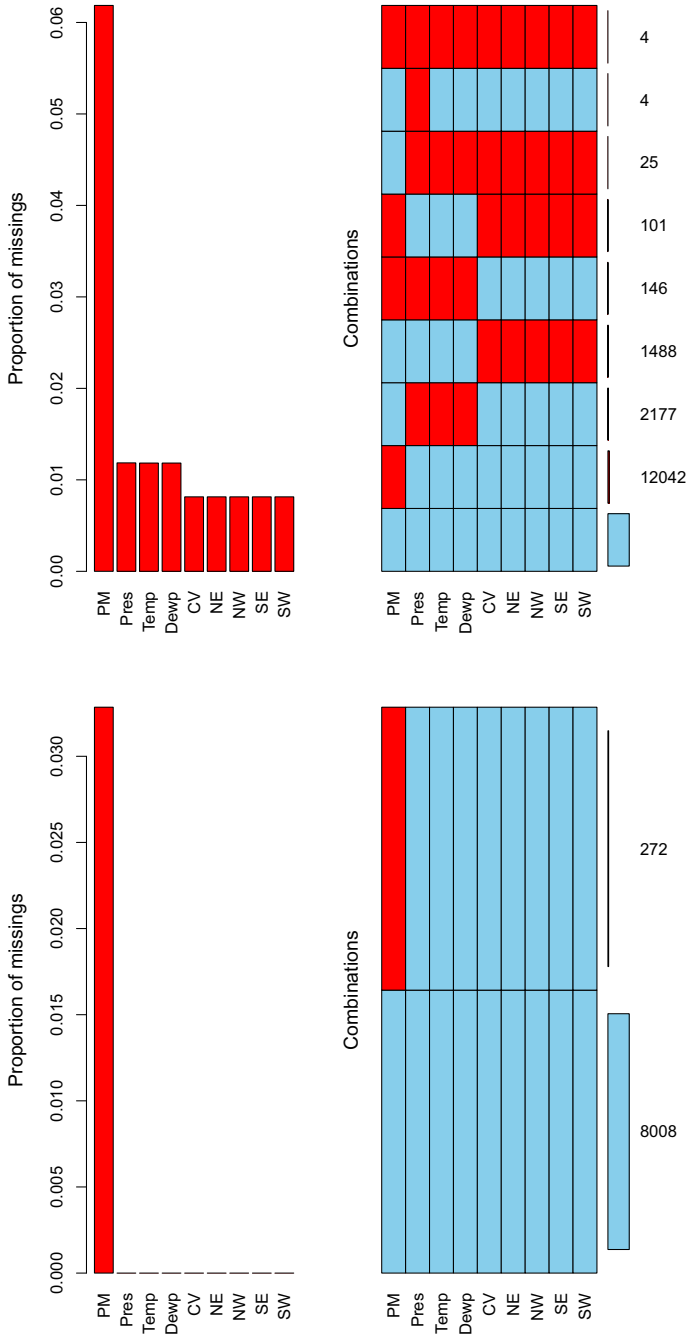


Fig. 4 Proportion and combinations of missings for hourly (top) and daily (bottom) data from October 1, 2016 to December 31, 2016

We analyze the period with a high level of $PM_{2.5}$ from October 2016 to December 2016. We are more interested in pollution episodes because these can be harmful to human health. A brief discussion on the choice of the period in this study, along with a graph showing the $PM_{2.5}$ concentration over the whole year of 2016, is provided in Sect. 1 of online supplementary materials. To facilitate visualization of the data, we averaged $PM_{2.5}$ observed at multiple monitoring sites for each city. Figure 5 presents daily $PM_{2.5}$ from October 1, 2016 to December 31, 2016 for all 13 cities in the BTH region. We observe that towards the end of year 2016, $PM_{2.5}$ can reach $600 \mu\text{g}/\text{m}^3$ for some cities such as Shijiazhuang and Handan. According to the standard of the United States Environmental Protection Agency (EPA), $PM_{2.5}$ concentration of approximately $35 \mu\text{g}/\text{m}^3$ is considered moderately polluted or acceptable air quality, and levels exceeding $150 \mu\text{g}/\text{m}^3$ are considered very unhealthy and even hazardous. The boxplot of $PM_{2.5}$ for each city demonstrates that for most of the cities in this study, $PM_{2.5}$ is above the moderate pollution level most of the time. $PM_{2.5}$ in Shijiazhuang is even above the unhealthy level half of the time. These results indicate very severe air pollution in the BTH region.

3 Methodology

3.1 Spatio-temporal mixed effects model

To investigate $PM_{2.5}$ concentration in the BTH region, we consider a real-valued spatio-temporal process (Cressie and Wikle 2015, p. 125):

$$Z(\mathbf{s}, t) = Y(\mathbf{s}, t) + \epsilon(\mathbf{s}, t), \quad (1)$$

where $Z(\mathbf{s}, t)$ is the data process $\{Z(\mathbf{s}, t) : \mathbf{s} \in D \subset \mathbb{R}^2, t \in \{1, 2, \dots\}\}$, which is the sum of a hidden process $Y(\mathbf{s}, t)$ and a measurement error process $\epsilon(\mathbf{s}, t)$. The error $\epsilon(\mathbf{s}, t)$ is assumed to be a white-noise Gaussian process with mean 0 and variance σ_ϵ^2 . We further assume that $\text{cov}(\epsilon(\mathbf{s}_1, t_1), \epsilon(\mathbf{s}_2, t_2)) = 0$, unless $\mathbf{s}_1 = \mathbf{s}_2$ and $t_1 = t_2$. That is, $\epsilon(\mathbf{s}, t)$ is both spatially and temporally uncorrelated.

The hidden process is modelled as:

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + \nu(\mathbf{s}, t). \quad (2)$$

where the mean, or trend, $\mu(\mathbf{s}, t)$ models large-scale variation and is assumed to follow the linear structure $\mu(\mathbf{s}, t) = \mathbf{x}(\mathbf{s}, t)^T \boldsymbol{\beta}$, where $\mathbf{x}(\mathbf{s}, t)$ is a $p \times 1$ vector that represents p different known covariates at location \mathbf{s} and time t . In this study, $\mathbf{x}(\mathbf{s}, t)$ include spatio-temporal covariates (e.g., meteorological variables) and geographic covariates (e.g., longitude, latitude, altitude, distance to mountain). The regression coefficient $\boldsymbol{\beta}$ is a $p \times 1$ vector that is generally unknown. $\nu(\mathbf{s}, t)$ is a spatio-temporal random effect that captures small-scale behavior, which is assumed to be a Gaussian process with mean 0 and covariance function $\text{cov}\{\nu(\mathbf{s}_1, t_1), \nu(\mathbf{s}_2, t_2)\} = C(\mathbf{s}_1 - \mathbf{s}_2, t_1 - t_2)$.

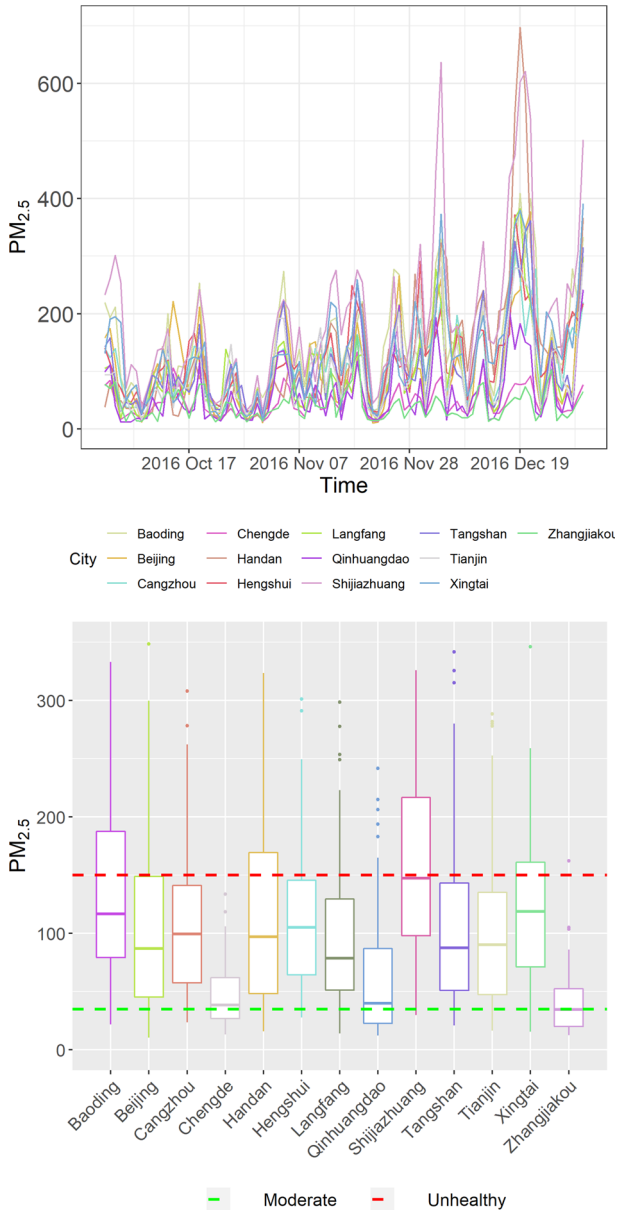


Fig. 5 Daily $PM_{2.5}$ ($\mu g/m^3$) from October 1, 2016 to December 31, 2016 for each city in the BTH region. Some extremely large values are truncated in the boxplot

Therefore, the covariance between two responses from the hidden process is

$$\text{cov} \{Y(\mathbf{s}_1, t_1), Y(\mathbf{s}_2, t_2)\} = C(\mathbf{s}_1 - \mathbf{s}_2, t_1 - t_2), \tag{3}$$

and the covariance between two responses from the data process is defined by

$$\text{cov} \{Z(\mathbf{s}_1, t_1), Z(\mathbf{s}_2, t_2)\} = C(\mathbf{s}_1 - \mathbf{s}_2, t_1 - t_2) + \sigma_\epsilon^2 I(\mathbf{s}_1 - \mathbf{s}_2, t_1 - t_2), \tag{4}$$

where I is an indicator function that returns 1 when $\mathbf{s}_1 = \mathbf{s}_2$ and $t_1 = t_2$, and 0 otherwise.

In this work, we utilize an isotropic spatio-temporal covariance function with a separable form to model $C(\mathbf{s}_1 - \mathbf{s}_2, t_1 - t_2)$. More detailed discussion on the choice of the covariance function is provided in Sect. 4 of online supplementary materials. Consider

$$C(\mathbf{s}_1 - \mathbf{s}_2, t_1 - t_2) = \sigma_v^2 \frac{2^{1-\theta}}{\Gamma(\theta)} \left(\frac{\|\mathbf{s}_1 - \mathbf{s}_2\|}{\phi} \right)^\theta K_\theta \left(\frac{\|\mathbf{s}_1 - \mathbf{s}_2\|}{\phi} \right) \rho^{|t_1 - t_2|} + \sigma_0^2 I(\mathbf{s}_1 - \mathbf{s}_2, t_1 - t_2), \tag{5}$$

where σ_v^2 is the variance parameter, θ measures the smoothness, K_θ is a modified Bessel function of the second kind of order θ , $\|\cdot\|$ denotes the Euclidean spatial distance, ϕ and ρ determine the strength of spatial and temporal dependence, respectively, and σ_0^2 is microscale variance of the hidden process. That is, the covariance function is constructed by multiplying the Matérn covariance function (Cressie and Wikle 2015, p. 126) of a spatial process and the correlation function of an AR(1) temporal process. Consequently, we have

$$\text{var} (Z(\mathbf{s}, t)) = \text{var} (Y(\mathbf{s}, t)) + \sigma_\epsilon^2 = \sigma_v^2 + \sigma_0^2 + \sigma_\epsilon^2, \tag{6}$$

which is referred to as the “sill” in a great deal of geostatistics literature. In addition, the term $\sigma_0^2 + \sigma_\epsilon^2$ is also known as the nugget effect. Because we cannot identify both parameters σ_0^2 and σ_ϵ^2 , without loss of generality, we set $\sigma_0^2 = 0$ in this study and the nugget effect is simply the measurement error variance σ_ϵ^2 .

3.2 Universal kriging

We are interested in making inferences on the hidden process $Y(\mathbf{s}, t)$ over a set of prediction regions. Kriging, or spatial best linear unbiased prediction (BLUP), is a predominant approach to spatial prediction in the spatial statistics literature. When the mean function $\mu(\mathbf{s}, t)$ is unknown, the regression parameter β must be estimated. In this situation, the unbiased linear predictor of $Y(\mathbf{s}, t)$ that minimizes mean squared prediction error is termed the “universal kriging predictor”. In this study, we wish to make point prediction of $Y(\mathbf{s}, t)$ at a location \mathbf{s}_0 , $\mathbf{s}_0 \in D$ (regardless of whether \mathbf{s}_0 is an observation location) and at a future time t_0 .

Suppose the data process $Z(\mathbf{s}, t)$ is observed at a set of locations $\mathbf{s}_1, \dots, \mathbf{s}_n$ and time t_1, \dots, t_T . Let the observed data $\mathbf{z} = \{Z(\mathbf{s}_1, t_1), \dots, Z(\mathbf{s}_n, t_T)\}$, an $nT \times 1$ vector. Let \mathbf{X} define the $nT \times p$ covariates matrix at observed locations and time, \mathbf{x}_0 denote the $p \times 1$ vector of spatio-temporal covariates at location \mathbf{s}_0 and time t_0 , Σ denote the $nT \times nT$ covariance matrix for the observed data \mathbf{z} , and define

$$\mathbf{c}_{0z} = \{C(\mathbf{s}_0 - \mathbf{s}_1, t_0 - t_1), \dots, C(\mathbf{s}_0 - \mathbf{s}_n, t_0 - t_T)\}^T$$

the $nT \times 1$ vector of covariances between $Y(\mathbf{s}_0, t_0)$ and the observed process \mathbf{z} . The universal kriging predictor for $Y(\mathbf{s}_0, t_0)$ is given by the following formula (Cressie 2015; Cressie and Wikle 2015):

$$\widehat{Y}(\mathbf{s}_0, t_0) = \mathbf{x}_0^T \widehat{\boldsymbol{\beta}}_{gls} + \mathbf{c}_{0z}^T \Sigma^{-1} (\mathbf{z} - \mathbf{X} \widehat{\boldsymbol{\beta}}_{gls}), \quad (7)$$

where $\widehat{\boldsymbol{\beta}}_{gls} = (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma^{-1} \mathbf{z}$ is known as the generalized least squares estimator of $\boldsymbol{\beta}$. The corresponding universal kriging variance is the mean squared prediction error of $\widehat{Y}(\mathbf{s}_0, t_0)$, which is shown to be

$$\begin{aligned} \sigma_{uk}^2(\mathbf{s}_0, t_0) &= E [Y(\mathbf{s}_0, t_0) - \widehat{Y}(\mathbf{s}_0, t_0)]^2 \\ &= \sigma_v^2 - \mathbf{c}_{0z}^T \Sigma^{-1} \mathbf{c}_{0z} \\ &\quad + (\mathbf{x}_0 - \mathbf{c}_{0z}^T \Sigma^{-1} \mathbf{X})(\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} (\mathbf{x}_0 - \mathbf{c}_{0z}^T \Sigma^{-1} \mathbf{X})^T. \end{aligned} \quad (8)$$

Recall that $\sigma_v^2 = \text{var}(Y(\mathbf{s}_0, t_0))$.

In this study, the covariance parameters are unknown. To find the kriging predictor and kriging variance, we must first estimate the covariance parameters. Define the covariance parameter as $\Psi = \{\sigma_\epsilon^2, \sigma_v^2, \theta, \phi, \rho\}$. We first estimate the covariance parameter Ψ using the restricted maximum likelihood (REML). Then, assuming Ψ is known, we estimate the regression parameter $\boldsymbol{\beta}$ according to the formula for $\widehat{\boldsymbol{\beta}}_{gls}$.

Generally, our prediction task is to investigate the behavior of the hidden process of $Y(\mathbf{s}, t)$ over a continuous region D^P at a given time t_P . This can be achieved by discretizing the region D^P into M adequately constructed finite grids or basic areal unit (BAU) as discussed in Nguyen et al. (2012). Define $\{\widetilde{\mathbf{s}}_1, \dots, \widetilde{\mathbf{s}}_M\}$ a set of centers for these M grids. Then, the kriging predictors and kriging variances at the locations $\widetilde{\mathbf{s}}_1, \dots, \widetilde{\mathbf{s}}_M$ and time t_P can be computed by equations (7) and (8), respectively.

3.3 Lognormal kriging

Lognormal transformation of data is commonly used to model many variables that take only positive values and have positively skewed distributions. This method has been widely employed to analyze air-pollution data (e.g., $PM_{2.5}$, NO_x ; see Fanshawe et al. (2008); Paciorek et al. (2009); Sampson et al. (2011)). In particular, the associated spatial statistical method for prediction in lognormal random fields is known as “lognormal kriging”. Given that our response of interest is $PM_{2.5}$,

which takes only positive values, we assume it follows a lognormal random process $\{W(\mathbf{s}, t) : \mathbf{s} \in D \subset \mathbb{R}^2, t \in \{1, 2, \dots\}\}$, such that $Z(\mathbf{s}, t) = \log W(\mathbf{s}, t)$, where $Z(\mathbf{s}, t)$ is a Gaussian random process as in (1). In this study, the actual observed response is defined as \mathbf{w} , such that $\mathbf{w} = \exp(\mathbf{z})$.

Providing the kriging predictor for the log scale according to equation (7), the predicted value can simply be exponentiated to predict on the original scale. However, it is widely known that the back-transformed value $\exp \widehat{Y}(\mathbf{s}_0, t_0)$ is a biased predictor (Cressie 2006; De Oliveira 2006; Paciorek et al. 2009). Define $U(\mathbf{s}, t)$ the hidden process of $W(\mathbf{s}, t)$, and m_{0Y} the Lagrange multiplier corresponding to the kriging predictor in equation (7)

$$m_{0Y} = (\mathbf{x}_0 - \mathbf{c}_{0z}^T \Sigma^{-1} \mathbf{X})(\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} \mathbf{x}_0^T. \tag{9}$$

It is not difficult to demonstrate that under the normal assumption on $Z(\mathbf{s}_0, t_0)$, the unbiased predictor for $U(\mathbf{s}_0, t_0)$ for the case with unknown mean function (Cressie 2006; De Oliveira 2006) is given by

$$\widehat{U}(\mathbf{s}_0, t_0) = \exp \left\{ \widehat{Y}(\mathbf{s}_0, t_0) + \frac{1}{2} \sigma_{uk}^2(\mathbf{s}_0, t_0) - m_{0Y} \right\}. \tag{10}$$

The case with a known mean function is also discussed in De Oliveira (2006) and Cressie (2015). Combining the calculations in equations (7), (8) and (9), the lognormal kriging predictors at locations $\tilde{\mathbf{s}}_1, \dots, \tilde{\mathbf{s}}_M$ and time t_P can be computed by equation (10).

4 Analyzing $PM_{2.5}$ in the North China Plain

4.1 Kriging results

In this study, the response of interest, $PM_{2.5}$ concentration, is assumed to be lognormal distributed and its log transformation $Z(\mathbf{s}, t)$ follows a spatio-temporal process as in (1).

We investigate the $PM_{2.5}$ readings at the beginning of the winter of 2016. Data at each observation location from November 19, 2016 to November 25, 2016 were used to estimate the covariance and the regression parameters. Therefore, the total number of observations is $90 \times 7 = 630$. We then perform the kriging on the next day, November 26, 2016. We find that one week of data is sufficient to capture the temporal structure and that using data with a longer period may not improve the prediction accuracy (see Sect. 2 of online supplementary materials for more details). The reason we analyze the $PM_{2.5}$ readings over this period is that the winter heating period usually runs from the middle of November in North China. The massive amount of fuel used to generate heating generally leads to poor air quality, and we are more interested in examining the period that has levels of $PM_{2.5}$ that are high enough to be harmful to human health.

According to the range of longitude and latitude of Hebei province, we construct a spatial bounding box for Hebei and discretize it into a 100×100 grid of pixels. The unit

measure of each grid is 0.07 degree of longitude times 0.07 degree of latitude (or 7.77 kilometers \times 7.77 kilometers). Thereafter, the spatial domain of interest is constructed by the set of pixels for which the center points fall in the spatial polygon of Hebei. It then results in a spatial domain with 4620 grids. When we perform the kriging, we require the covariates at the center of each of these 4620 grids. The geographic covariates can be easily obtained according to the location (longitude and latitude) of the pixel center. However, the meteorological variables are not known at the grid of interest, but only observed at the monitoring sites. Considering the difficulty of collecting the meteorological data at each grid, we define them by the corresponding ones at the nearest monitoring site from that grid. The distance is simply measured under the two-dimensional domain of interest D . Finally, kriging is performed at each of these 4160 locations to produce a continuous map of $PM_{2.5}$ concentration on November 26, 2016, following (7) and (10), as presented in Fig. 6.

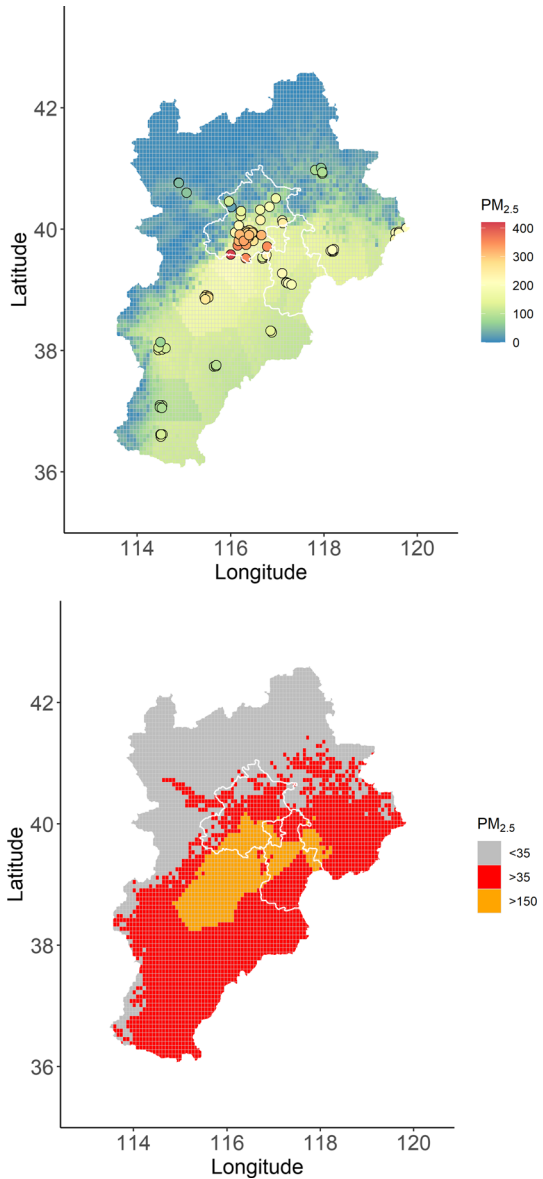
Figure 6(top) clearly demonstrates that the northern area of Hebei province has a much lower level of predicted $PM_{2.5}$ than the NCP, which coincides with what we actually observe. Within the NCP, the northern area, including the south of Beijing, seems to have the highest predicted and observed $PM_{2.5}$ concentration. It is widely known that major iron, steel, and cement industries are located in the south of Hebei province, but not in Beijing. The high level of $PM_{2.5}$ in the south of Beijing is widely attributed to pollutant transportation through wind blow from the industrial Hebei province. A further investigation into the impact of the wind on $PM_{2.5}$ concentration will be presented in the following section. Overall, Fig. 6(top) demonstrates that the kriging mainly agrees with the observed level of $PM_{2.5}$ in Hebei on November 26, 2016, apart from several locations in the south of Beijing with extremely high $PM_{2.5}$ concentrations. More discussion on the model assessment can be found in Sect. 3 of online supplementary materials.

It is widely known that exposure to a high level of $PM_{2.5}$ can be very harmful for human health and cause serious respiratory diseases. As discussed in Liang et al. (2015, 2016) and Zhang et al. (2017), $PM_{2.5}$ levels reaching $150 \mu\text{g}/\text{m}^3$ are widely considered very harmful and even hazardous to human health. Regions with kriginings of $PM_{2.5}$ greater than $35 \mu\text{g}/\text{m}^3$ (moderately polluted) and $150 \mu\text{g}/\text{m}^3$ (unhealthy) are plotted in Fig. 6(bottom). The region in orange color in the north of the NCP indicates the area where the kriging predictor of $PM_{2.5}$ exceeds $150 \mu\text{g}/\text{m}^3$. People living in this region should avoid prolonged outdoor activities at this high level of $PM_{2.5}$. The grey coloring indicates the region where the kriging of the $PM_{2.5}$ level is below $35 \mu\text{g}/\text{m}^3$, which coincides with the locations of “mountain” as presented in Fig. 3. Therefore, we can clearly see that “mountain” plays a very important role in blocking pollution transportation from the south of Hebei. Based on this finding, people should seek to conduct outdoor exercises in the mountainous north of Hebei to avoid high levels of $PM_{2.5}$ on November 26, 2016.

4.2 The impact of wind

It is widely accepted by the residents of Hebei that strong wind tends to alleviate the air pollution. Studies on air quality in Beijing (Liang et al. 2015; Zhang et al. 2017)

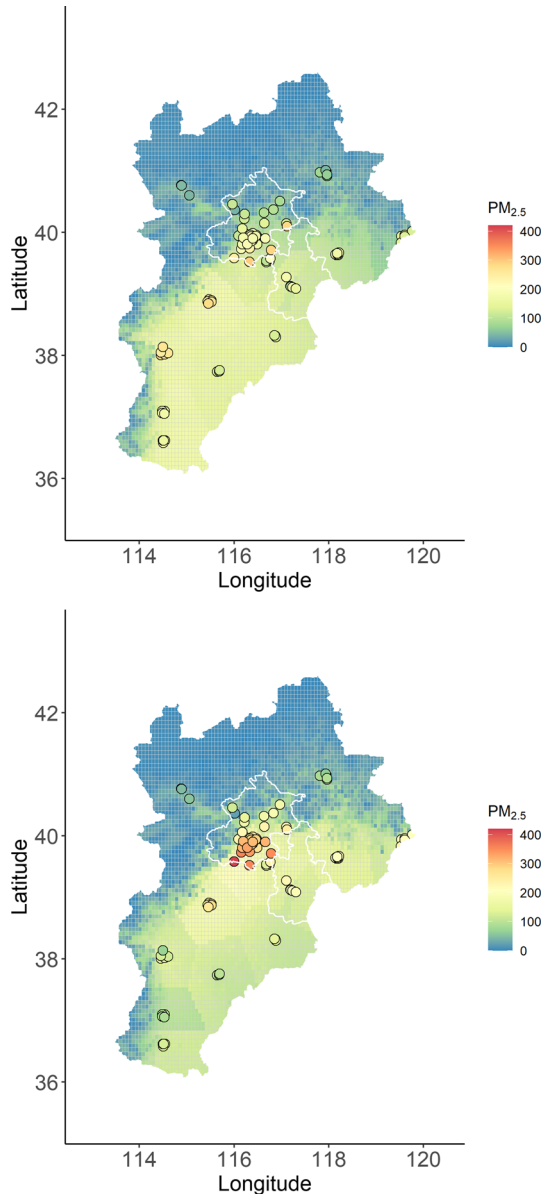
Fig. 6 Predicted (grid) and observed (circle) $PM_{2.5}$ concentration (top) and regions with krigings of $PM_{2.5}$ greater than $50 \mu\text{g}/\text{m}^3$ and $150 \mu\text{g}/\text{m}^3$ (bottom) in Hebei province on November 26, 2016



have also found that a lack of wind contributes to high levels of $PM_{2.5}$ more often than anthropogenic activities. In this section, we examine the impact of wind on $PM_{2.5}$ concentration in Hebei province.

We first investigate the predicted and the observed $PM_{2.5}$ concentration in Hebei province on November 25, 2016 and November 26, 2016, as presented in Fig. 7. The kriging on November 25, 2016 is performed by using the training data from November 17, 2016 to November 24, 2016. Figure 7 reveals that south-west of the NCP (Shiji-

Fig. 7 Predicted (grid) and observed (circle) $PM_{2.5}$ concentration in Hebei province on November 25, 2016 (top) and November 26, 2016 (bottom)



azhuang, Xintai, and Handan) is the region with the most severe predicted (observed) $PM_{2.5}$ concentration on November 25, 2016. However, on November 26, 2016, the highest predicted (observed) $PM_{2.5}$ concentration appears in north of the NCP (Beijing, Baoding, and Tangshan), while on this date, the $PM_{2.5}$ level is substantially reduced in the south-west of the NCP. Figure 7 clearly reveals a transmission of $PM_{2.5}$ concentration from the south-west of the NCP to the north of the NCP. As presented in Fig. 8a, the wind in Hebei on November 26, 2016 is predominantly SW

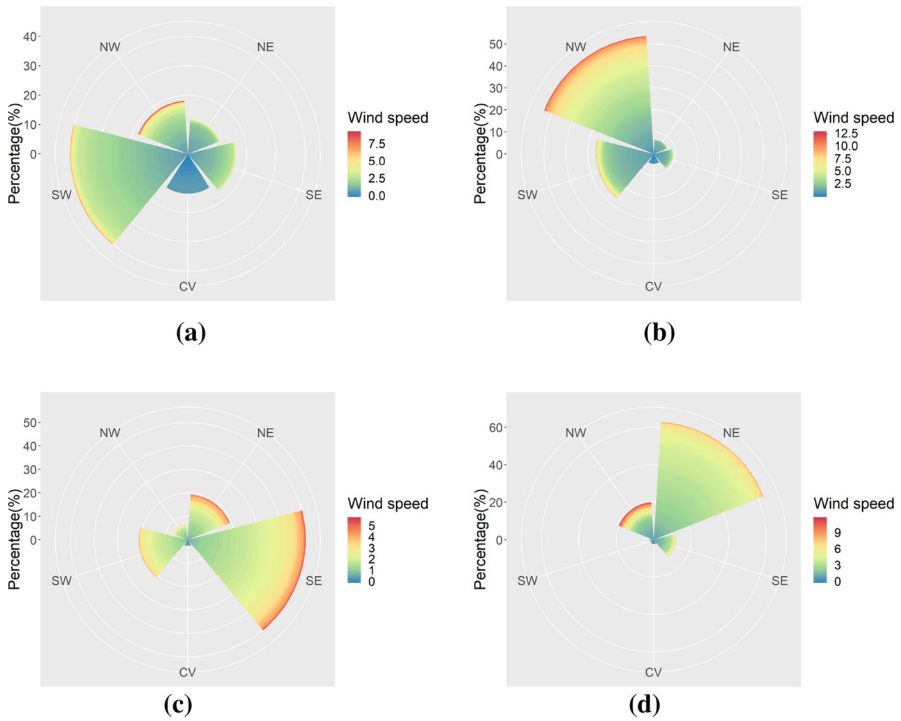


Fig. 8 Hourly wind speed (ms^{-1}) and wind frequency (%) of five grouped directions (SW, SE, NW, NE, CV) at all observation locations on four different days. **a** November 26, 2016, dominated by SW, **b** December 2, 2015, dominated by NW, **c** November 19, 2015, dominated by SE, **d** November 6, 2015, dominated by NE

wind. As a consequence, it is reasonable to believe that the SW wind blew the high level of $PM_{2.5}$ concentration from the south-west of the NCP to the north of the NCP on November 26, 2016.

By controlling the other meteorological variables, we now perform three quasi-experiments to further examine the impact of wind on $PM_{2.5}$ concentration in Hebei. We consider three different days with wind conditions that are dominated by NW, SE, and NE winds, respectively. For illustrative purposes, hourly wind speed and frequency on these days (note that daily wind data are still used to fit the model) are shown in 8 (b), (c), and (d). Days at the beginning of winter in 2015 are chosen so that their corresponding wind conditions are likely to occur on November 26, 2016 because each season usually has a homogeneous wind condition. The wind data on November 26, 2016 are then replaced with the wind data on these three days and kriging is performed under each of these three wind conditions. The kriging-based contour plots of $PM_{2.5}$ concentration using the wind data on November 26, 2016 and the three quasi-wind data are presented in Fig. 9.

Figure 8b demonstrates the scenario under which more than 50% of the wind comes from the north-west in all the observation locations. In addition, a certain percentage of NW wind achieves a wind speed that reaches 10 ms^{-1} . Along with Fig. 9b, we note

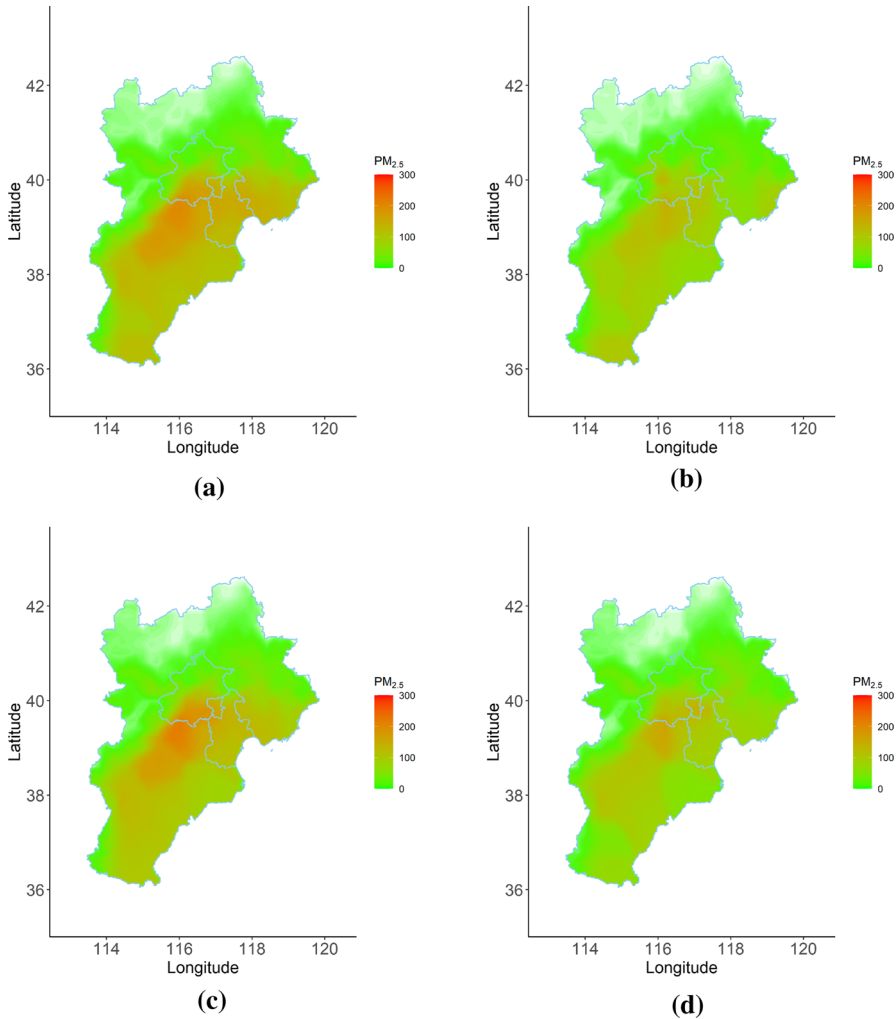


Fig. 9 Kriging-based contour plots of $PM_{2.5}$ concentration in Hebei province on November 26, 2016 using the wind data on November 26, 2016 and the quasi-wind data on three different days in 2015. **a** November 26, 2016, dominated by SW, **b** December 2, 2015, dominated by NW, **c** November 19, 2015, dominated by SE, **d** November 6, 2015, dominated by NE

that strong wind from the north-west substantially reduces the $PM_{2.5}$ concentration in the entire NCP. In the absence of high-polluting industries in the region to the north of the NCP, the wind from the north-west of Hebei does not travel with high $PM_{2.5}$ concentration, but rather blows $PM_{2.5}$ away from the NCP. In contrast, in Figs. 8c and 9c, we observe that wind at a moderate speed from the south-east does not alleviate the $PM_{2.5}$ concentration; interestingly, it even slightly increases the level of $PM_{2.5}$ in north-west of the NCP. The north-west of the NCP is bordered by Taihang Mountains to the west and Yan Mountain to the north. Seeing that the mountains block the wind

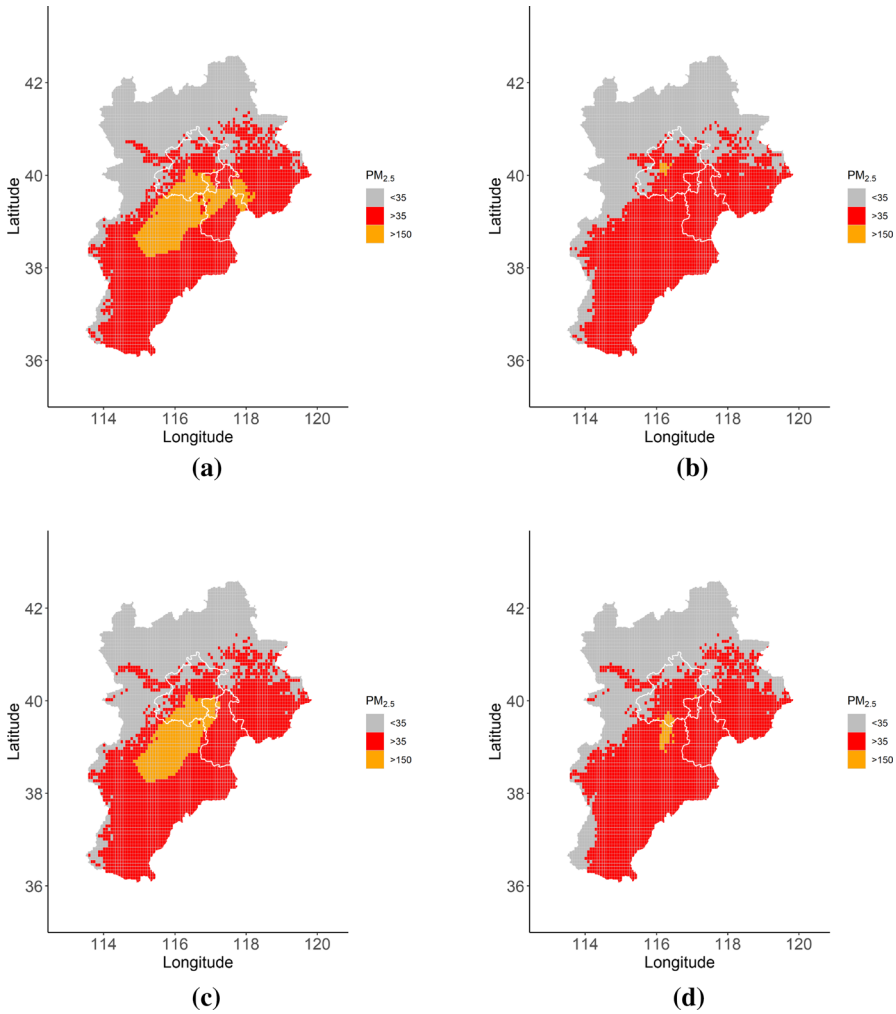


Fig. 10 Regions with krigings of $PM_{2.5}$ greater than $50\ \mu\text{g}/\text{m}^3$ and $150\ \mu\text{g}/\text{m}^3$ on November 26, 2016 using the wind data on November 26, 2016 and the quasi-wind data

that travels to the north-west, the polluted air from the south of the NCP tends to accumulate in the north-west of the NCP, which explains the slight increase of $PM_{2.5}$ revealed in Fig. 9 (c). In addition, the wind from the north-east also greatly reduces the $PM_{2.5}$ concentration in the NCP, as demonstrated in Figs. 8d and 9d. It is also worth noting that $PM_{2.5}$ in the northern area of Hebei (i.e., the areas with mountains) remains at a low level under all wind conditions, and thus, is not affected by the wind. Figure 10 presents the regions with krigings of $PM_{2.5}$ greater than $35\ \mu\text{g}/\text{m}^3$ and $150\ \mu\text{g}/\text{m}^3$ on November 26, 2016 using the wind data on November 26, 2016 and the quasi-wind data. We observe that when there is a northerly wind, people living in Hebei are relatively safe, but if there is a southerly wind, the people living in the north

Table 1 Estimated regression coefficients for wind from NE, NW, SE and SW, with their associated p-values (less than 5% p-values are in bold)

	Coef	p-value
NE	−0.064	0.021
NW	−0.054	0.025
SE	−0.002	0.964
SW	0.050	0.394

of NCP face a health risk harmed and should avoid outdoor activities, as demonstrated by the orange areas in Fig. 10a and c).

To conclude, our results demonstrate that a northerly wind can considerably reduce the level of $PM_{2.5}$ in the NCP, while a southerly wind generally does not alleviate the air pollution and sometimes even increases it. To confirm this finding, we also report the estimated regression coefficients for wind from NE, NW, SE and SW, with their associated p-values in Table 1. We observe that the northerly wind (NE and NW) has a negative coefficient that is significantly different from zero at 5%, implying that increasing the wind strength in this direction can reduce the level of $PM_{2.5}$. On the other hand, the estimated coefficient for the southerly wind (SE and SW) has a large p-value and SW even has a positive coefficient.

This finding is consistent with the finding for Beijing discussed in Liang et al. (2015). In addition, neither southerly nor northerly winds have a substantial impact on the level of $PM_{2.5}$ in the mountainous north of Hebei. In practice, the wind direction is difficult to predict and is unknown in advance. Given that the severity of air pollution greatly depends on the wind, it is worth presenting predicted $PM_{2.5}$ concentration in Hebei under different wind conditions.

4.3 Prediction errors and local emissions

As Fig. 7 demonstrates, predicted $PM_{2.5}$ can mainly capture the level of observed $PM_{2.5}$ in most areas of Hebei. However, extremely high levels of observed $PM_{2.5}$ in some areas tend to be underestimated (e.g., $PM_{2.5}$ in Beijing on November 26, 2016). The formation of $PM_{2.5}$ is generally caused by both meteorological conditions and the emission of pollutants. Given that our model has already considered the impact of meteorological and geographic variables on $PM_{2.5}$, local emissions can be a main factor in the measurement error $\epsilon(\mathbf{s}, t)$ as in (1). More specifically, the kriging predictor for a known location \mathbf{s}_1 and a future time t_0 can also be written as a weighted sum of observed values, where the weights are determined by the covariance function:

$$\widehat{Y}(\mathbf{s}_1, t_0) = k(\mathbf{s}_1, t_0) + \sum_{j=1}^T \sum_{i \neq 1}^N w_{ij} z(\mathbf{s}_i, t_j) + \sum_{j=1}^T w_{1j} z(\mathbf{s}_1, t_j),$$

where $k(\mathbf{s}_1, t_0) = \mathbf{x}_0^T \widehat{\boldsymbol{\beta}}_{gls} - \mathbf{c}_{0z}^T \boldsymbol{\Sigma}^{-1} \mathbf{X} \widehat{\boldsymbol{\beta}}_{gls}$ and $\mathbf{w}_1 = \mathbf{c}_{0z}^T \boldsymbol{\Sigma}^{-1}$. From this formalization, we observe that the kriging predictor for a location \mathbf{s} considers local meteorological conditions in the mean term, along with the impact from other locations in past times

and the impact of itself on past dates. Therefore, we reasonably believe that the prediction error or what is left unexplained, $Z(\mathbf{s}, t_0) - \widehat{Y}(\mathbf{s}, t_0)$, can serve as a proxy for the increment of the level of daily local emissions. Although the local emission data are generally not publicly available in China, the magnitude of prediction errors from our model may reveal the level of local emissions in different areas.

We investigate the $PM_{2.5}$ in early winter 2016, from October 1 to December 31. We produce one-step ahead forecast of $\log PM_{2.5}$ at each monitoring site on a rolling-window basis using a fixed window size of one week. That is, the training dataset shift towards with the passage of time (with the first set from October 1 to October 7) and one-step ahead forecast is performed until we obtain the predicted $\log PM_{2.5}$ as in (7) from October 8 to December 31 at each monitoring site. The prediction errors are then computed as the difference between the observed $\log PM_{2.5}$ and the predicted $\log PM_{2.5}$. As presented in Fig. 2, the locations of the monitoring sites in each city tend to be clustered together. Hence, we perform our analysis on a city level by simply averaging the prediction errors in each city so that we obtain 13 (which is the number of cities in the BTH region) prediction errors on each day from October 8 to December 31.

As stated, the winter heating period in China usually begins in the middle of November. Aside from the pollutants from heavy industries, the excessive coal burned to generate heating during the winter heating period greatly contributes to the local emissions. Considering this, we split the period under study into “before winter heating” and “after winter heating”. According to Fig. 11, there is a dramatic temperature drop around November 22 in the cities in Hebei. Consequently, “before winter heating” and “after winter heating” are defined as the study period before and after November 22, respectively.

We compute the average prediction errors before November 22 and after November 22 (inclusive) for each city in Hebei (see Table 2). We aim to test whether the prediction errors for each city separately are significantly different from zero before and after the winter heating period. To consider the serial correlation of the prediction errors, the corresponding p-value for each city presented in Table 2 is calculated based on a test statistic constructed as follows:

$$Z = \frac{\bar{e}}{\sqrt{\hat{\sigma}_{\infty}^2/T}} \xrightarrow{d} N(0, 1), \quad (11)$$

where \bar{e} is the average of the prediction errors, T is the number of days, and $\hat{\sigma}_{\infty}^2$ is a consistent estimator of the long-run covariance. Details of obtaining $\hat{\sigma}_{\infty}^2$ can be found in Andrews (1991) and Newey and West (1986).

In Table 2, we observe that Shijiazhuang and Tangshan achieve exceedingly large average prediction errors before the winter heating period, indicating significant daily local emissions produced from heavy industries in these two cities. This can be expected because these two cities are densely populated with iron, steel, and cement industries that consume enormous amounts of coal and other fuels. In contrast, after winter heating period begins, the densely populated cities of Beijing, Langfang (a city very close to Beijing), and Tianjin show an increase in average prediction errors, which is probably because of the massive level of coal consumption by residents for winter

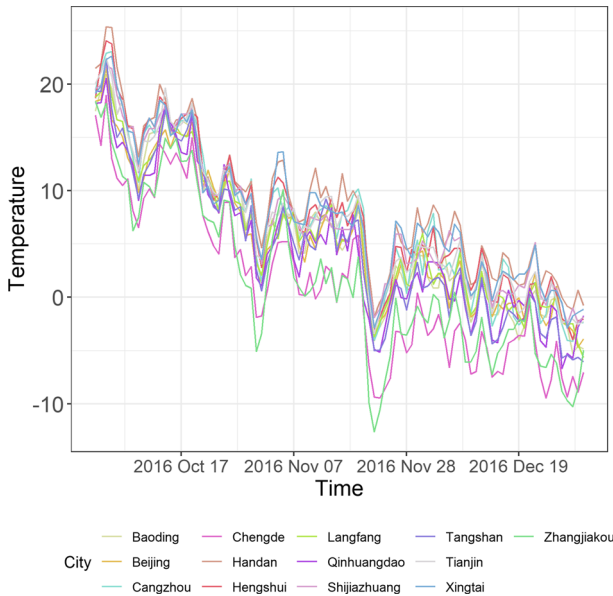


Fig. 11 Temperatures for different cities in Hebei from October 2016 to December 2016

Table 2 Average out-of-sample prediction errors before and after the winter heating period with associated p-values (less than 5% p-values are in bold)

	City	Before	p-value	After	p-value	No.
1	Baoding	0.061	0.379	0.083	0.162	6
2	Beijing	0.091	0.208	0.144	0.023	36
3	Cangzhou	0.064	0.211	0.033	0.084	3
4	Chengde	0.041	0.511	0.033	0.323	5
5	Handan	-0.036	0.481	-0.083	0.137	4
6	Hengshui	0.052	0.262	-0.033	0.411	3
7	Langfang	0.034	0.506	0.112	0.041	4
8	Qinhuangdao	0.044	0.508	0.025	0.736	5
9	Shijiazhuang	0.159	0.014	0.126	0.037	6
10	Tangshan	0.177	0.001	0.099	0.119	6
11	Tianjin	0.077	0.173	0.082	0.098	5
12	Xingtai	0.057	0.347	-0.096	0.129	4
13	Zhangjiakou	0.003	0.962	-0.051	0.137	3

heating. In particular, we observe that in the before winter heating period, Beijing and Langfang have a large p-value, but a p-value smaller than 5% after the winter heating period, implying that the winter heating period affects the level of $PM_{2.5}$ for densely populated cities.

5 Conclusion

Utilizing our spatio-temporal mixed effects model, we found that $PM_{2.5}$ can be hazardous in many cities in BTH, where major iron, steel, and cement industries are located. Our analysis of the impact of wind supports that the mountains block the wind traveling to the north-west, such that the polluted air from the south of the NCP tends to accumulate in the north-west of the NCP under certain meteorological conditions. This results in the interesting phenomenon that a northerly wind can considerably reduce the level of $PM_{2.5}$ in the NCP, while a southerly wind generally does not alleviate air pollution and sometimes even increases it. Note that such a conclusion might be limited to the winter season due to some seasonal variations in wind and emission patterns as discussed in Liang et al. (2015). The quasi-wind analysis on other seasons can be performed in a similar fashion and it is left to future research. Additionally, in our current analysis, we define the meteorological variables at each grid by the corresponding ones at the nearest monitoring site from that grid. In our future work, we will seek for some satellite atmospheric data to improve the quality and resolution of the meteorological data to be used in our analysis.

Moreover, we find that the prediction errors produced from the model can serve as a good proxy for the level of local emissions. By analyzing the prediction errors before and after the winter heating period separately, we discover that the coal burning that occurs during the winter heating period substantially contributes to the local emissions, particularly in densely populated cities.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10651-021-00521-4>.

Acknowledgements This research was undertaken with the assistance of data resources provided by Professor Song Xi Chen's research group based at Peking University (<https://www.songxichen.com>).

References

- Alduchov OA, Eskridge RE (1996) Improved magnus form approximation of saturation vapor pressure. *J Appl Meteorol* 35(4):601–609
- Andrews D (1991) Heteroskedasticity and autocorrelation consistent covariant matrix estimation. *Econometrica* 59(3):817–858
- Calculli C, Fassò A, Finazzi F, Pollice A, 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
- Chen L, Guo B, Huang J, He J, Wang H, Zhang S, Chen SX (2018) Assessing air-quality in beijing-tianjin-hebei region: The method and mixed tales of pm_{2.5} and o₃. *Atmos Environ* 193:290–301
- Cressie N (2006) Block kriging for lognormal spatial processes. *Math Geol* 38(4):413–443
- Cressie N (2015) *Statistics for spatial data*. Wiley, New York
- Cressie N, Wikle CK (2015) *Statistics for spatio-temporal data*. Wiley, New York
- De Oliveira V (2006) On optimal point and block prediction in log-gaussian random fields. *Scand J Stat* 33(3):523–540
- Donaldson K, Li X, MacNee W (1998) Ultrafine (nanometre) particle mediated lung injury. *J Aerosol Sci* 29(5–6):553–560

- Fanshawe TR, Diggle PJ, Rushton S, Sanderson R, Lurz P, Glinianaia SV, Pearce MS, Parker L, Charlton M, Pless-Mullooli T (2008) Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach. *Environmetrics* 19(6):549–566
- Fassò A (2013) Statistical assessment of air quality interventions. *Stoch Environ Res Risk Assess* 27(7):1651–1660
- Fassò A, Finazzi F, Ndongo F (2016) European population exposure to airborne pollutants based on a multivariate spatio-temporal model. *J Agric Biol Environ Stat* 21(3):492–511
- Finazzi F, Fassò A (2014) D-stem: a software for the analysis and mapping of environmental space-time variables. *J Stat Softw* 62(1):1–29
- Li H, Zhang Q, Zhang Q, Chen C, Wang L, Wei Z, Zhou S, Parworth C, Zheng B, Canonaco F et al (2017) Wintertime aerosol chemistry and haze evolution in an extremely polluted city of the north china plain: significant contribution from coal and biomass combustion. *Atmos Chem Phys* 17(7):4751–4768
- Liang X, Zou T, Guo B, Li S, Zhang H, Zhang S, Huang H, Chen SX (2015) Assessing beijing's pm_{2.5} pollution: severity, weather impact, APEC and winter heating. In: *Proc. R. Soc. A*, vol 471, p 20150257. The Royal Society
- Liang X, Li S, Zhang S, Huang H, Chen SX (2016) Pm_{2.5} data reliability, consistency, and air quality assessment in five chinese cities. *J Geophys Res* 121 (17)
- McMillan NJ, Holland DM, Morara M, Feng J (2010) Combining numerical model output and particulate data using bayesian space-time modeling. *Environmetrics* 21(1):48–65
- Newey WK, West KD (1986) A simple, positive semi-definite, heteroskedasticity and autocorrelation-consistent covariance matrix
- Nguyen H, Cressie N, Braverman A (2012) Spatial statistical data fusion for remote sensing applications. *J Am Stat Assoc* 107(499):1004–1018
- Paciorek CJ, Yanosky JD, Puett RC, Laden F, Suh HH (2009) Practical large-scale spatio-temporal modeling of particulate matter concentrations. *Ann Appl Stat* 370–397
- Pope CA, Ezzati M, Dockery DW (2009) Fine-particulate air pollution and life expectancy in the united states. *N Engl J Med* 360(4):376–386
- Sahu SK, Gelfand AE, Holland DM (2006) Spatio-temporal modeling of fine particulate matter. *J Agric Biol Environ Stat* 11(1):61–86
- Sampson PD, Szpiro AA, Sheppard L, Lindström J, Kaufman JD (2011) Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmos Environ* 45(36):6593–6606
- Shaddick G, Yan H, Vienneau D (2013) A bayesian hierarchical model for assessing the impact of human activity on nitrogen dioxide concentrations in Europe. *Environ Ecol Stat* 20(4):553–570
- Wang L, Wei Z, Yang J, Zhang Y, Zhang F, Su J, Meng C, Zhang Q (2013) The 2013 severe haze over the southern Hebei, China: model evaluation, source apportionment, and policy implications. *Atmos Chem Phys Discuss* 13(11)
- Wang L, Zhang N, Liu Z, Sun Y, Ji D, Wang Y (2014) The influence of climate factors, meteorological conditions, and boundary-layer structure on severe haze pollution in the Beijing-Tianjin-Hebei region during January 2013. *Adv Meteorol*
- Yin Q, Wang J, Hu M, Wong H (2016) Estimation of daily pm_{2.5} concentration and its relationship with meteorological conditions in Beijing. *J Environ Sci* 48:161–168
- Zhang S, Guo B, Dong A, He J, Xu Z, Chen SX (2017). In: *In Proc R, Soc A* (eds) Cautionary tales on air-quality improvement in Beijing, vol 473, p 20170457. The Royal Society
- Zhao H, Che H, Zhang X, Ma Y, Wang Y, Wang H, Wang Y (2013a) Characteristics of visibility and particulate matter (pm) in an urban area of northeast china. *Atmos Pollut Res* 4(4):427–434
- Zhao X, Zhao P, Xu J, Meng W, Pu W, Dong F, He D, Shi Q (2013b) Analysis of a winter regional haze event and its formation mechanism in the north china plain. *Atmos Chem Phys* 13(11):5685–5696

Le Chang Research School of Finance, Actuarial Studies and Statistics, College of Business and Economics, the Australia National University, Australia.

Tao Zou Research School of Finance, Actuarial Studies and Statistics, College of Business and Economics, the Australia National University, Australia.