**The Pennsylvania State University**
**The Graduate School**

# AIR POLLUTION ESTIMATION UNDER AIR STAGNATION

A Thesis in
Statistics
by
Ying Zhang



Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2021

The thesis of Ying Zhang was reviewed and approved by the following:

Le Bao
Associate Professor of Statistics
Thesis Advisor

Xiaoyue Niu
Associate Research Professor of Statistics

Murali Haran
Professor of Statistics
Head of the Department of Statistics

# Abstract

Air pollution continues to be a major environmental concern in China. The wind-driven transmission poses difficulties for understanding the air pollution patterns at the local level. The main objective of this study is to provide an easy approach to investigate the temporal trends and meteorological effects on the air pollutant concentrations during the generation process rather than incorporating the complex wind-driven transmission effect. We focus on the hourly data of the three most common air pollutants: PM2.5, $NO_2$, CO under air stagnations in Beijing, China during 2014-2017. For associations study, we apply the restricted spatio-temporal regression, an extension of restricted spatial regression [1], to alleviate the confounding between fixed effects and spatio-temporal random effects. The major findings show that the local pollution levels in Beijing have decreased over the years after we eliminate the wind effects. Other temporal trends of the air pollution levels under air stagnation reveal that winter is the severest month of the year, and Sunday is the clearest day of the week. Our model also interpolates the air pollutant concentrations at sites without monitoring stations and provides the map of air pollution concentrations under air stagnation. The results can potentially be used to identify locations that air pollutants easily accumulate.

# Table of Contents

# List of Figures

# List of Tables

# Acknowledgments

This material is partially supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number NIH/NIAID 5-R01-AI136664. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Institutes of Health.

I would like to thank my advisor, Dr. Le Bao, for his guidance and encouragement throughout the process of my research. Also, I would like to thank Dr. Song Xi Chen and Dr. Xiaoyue Niu for giving me invaluable insights and suggestions.

Chapter 1

# Introduction

In recent years, there is increasing awareness of the threat posed by air pollution among the international community. Several studies have observed associations between air pollution and respiratory diseases in the past [3, 4, 5, 6, 7, 8]. Epidemiological studies suggest that exposure to air pollution adversely affects human health, promoting the risk of asthma, lung cancer, heart diseases, and cardiovascular diseases [3]. Air pollution-related deaths account for around 17% of all deaths in China [9]. In the past few years, air pollution data in China became more readily available, including satellite imaging and hourly air pollution measurements from a ground monitoring network in China published by China National Environmental Monitoring Center (CNEMC, http://www.cnemc.cn/) since 2013. Particulate matter with aerodynamic diameters below 2.5 micrometers (PM2.5) serves as an indicator of air pollution level. These ultra-fine particles are fatal as they can translocate from the lungs and penetrate deep into the circulatory system [10, 11, 12]. The long-term exposure to PM2.5 contributes to an estimated 1.6-2.2 million premature deaths annually during 2013-2015 in China [13]. Besides, nitrogen dioxide ($NO_2$) and carbon monoxide (CO) are also of great concern. They are pervasive and primarily human-caused through fossil fuel consumption.

Covariates commonly used in the past studies include meteorological variables [14, 15, 16], land-use related [17, 18, 19, 20, 21], and road-traffic related variables [18, 21, 22]. More recently, satellite data became a popular information to model and estimate the air pollution concentrations [23, 24, 25, 26, 27]. [23] provides general methods to process satellite images for detecting and predicting air pollution. [24, 25, 26, 27] utilize remote sensing retrieved satellite retrieved aerosol optical depth data to predict the concentration

of air pollution concentrations. In this thesis, we include meteorological variables and nighttime light intensity from satellite images as covariates to help estimate the air pollution concentrations.

There exist several approaches to model the concentrations of air pollutants. The approaches include Bayesian hierarchical models [28, 29, 30, 31, 32], land-use regression [17, 18, 19, 20, 21], geographically weighted regression [33], non-parametric regression [34, 22, 35], and the kriging models [36, 37, 38, 39].

However, wind effects inevitably dominate the air pollution levels due to their strong transmission effects, and the effect varies across different land cover types. Many analysis results show that a higher wind speed can help disperse and dilute air pollutants and decrease their concentration [14, 40]. However, wind under different meteorological conditions could have different effects. There exist some cases when a higher wind speed can blow surface pollutants into the air and increase the air pollution levels [41, 42]. For Beijing, China, in particular, a study has shown that south wind brings pollutants while north wind guarantees clean air [35]. Therefore, in this thesis, we aim to study the air pollutants' levels under air stagnation, i.e., under calm-wind periods, to eliminate the noises caused by the wind effects and better estimate the temporal or meteorological effects. Our model considers the spatial heterogeneity and dependence at the grid cell level, the temporal pattern at different scales such as year, month, weekday, and hour. We also investigate the relationship between air pollutants and meteorological factors, including temperature, boundary layer height, pressure, dew point, evaporation, rain, and nighttime light intensity from satellite images. On the edge of the severest air-polluted region in China, we take Beijing as an illustrative example with the hourly PM2.5, $NO_2$, and CO data from 36 monitoring stations between February 1st, 2014 and May 31st, 2017.

The rest of the thesis is organized as follows. We present the descriptions of air pollution data and the air stagnation in chapter 2. In chapter 3, we build two spatio-temporal models for kriging and propose two evaluation metrics for their kriging performances. We also discuss the methods for inference of associations between air pollutants and other variables including meteorological variables and night-time light intensity. In chapter 4, we provide the kriging maps in different seasons and results of temporal and meteorological associations for three pollutants. In chapter 5, we conclude with discussions and future works.

Chapter 2

# Air Pollution and Air Stagnation Data

The hourly air pollution data in Beijing from February 1st, 2014 to May 31st, 2017 for this study were obtained from 36 monitoring stations, where 35 of them were from the Beijing Municipal Environmental Monitoring Center (BMEMC) and one from the US Embassy site. Figure 1 shows the locations of 36 monitoring stations in Beijing during our study period. The meteorological variables and the wind variables are the re-analysis data from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 hourly data (`https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/`

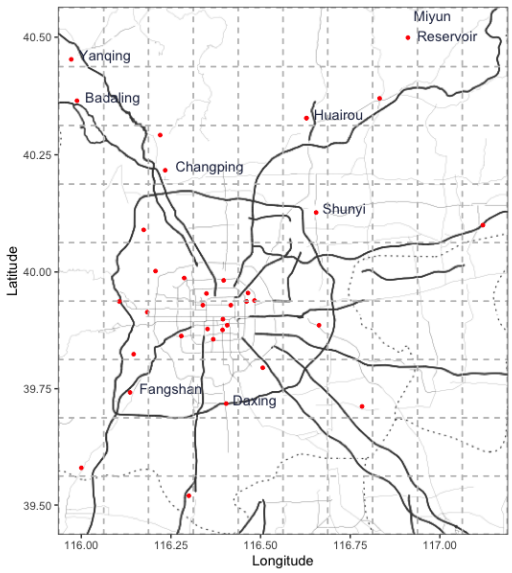



Figure 2.1: Map of monitoring sites for air pollution concentrations across Beijing, China from February 1st, 2014 to May 31st, 2017. The red points represent the monitoring sites. The grids with the dashed line are the spatial resolution of the meteorological variables from ECMWF ERA5 hourly data..

era5) with a grid resolution of $0.125° \times 0.125°$. The night-time light intensity data are obtained from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) produced by the Earth Observation Group, Payne Institute for Public Policy (https://payneinstitute.mines.edu/eog/nighttime-lights/). The night-time light intensity data resolution is 15 arc seconds (around 500 meters) in the geographic grid.

## 2.1 Data Description

We focus on three pollutants: PM2.5, $NO_2$, and CO. Table 2.1 summarizes the continuous variables being used in our analysis. The meteorological variables in Table 2.1 are mainly based on the past studies [15, 43, 35, 44, 45]. For example, temperatures at different levels are positively correlated with the air pollution level in [15, 43]. The pressure (PRES) is found to have associations with air pollution level in [35, 44, 45]. We also added a night-time light intensity as a variable, which reflected the population density across different locations [46, 47].

Table 2.1: Description for Continuous Variables

| Name | Type | Range | Mean | Description |
|---|---|---|---|---|
| PM2.5 | Pollutant | [3,696] | 79.65 | PM2.5 concentration level ($\mu g \cdot m^{-3}$) |
| $NO_2$ | Pollutant | [2,300] | 49 | $NO_2$ concentration level ($\mu g \cdot m^{-3}$) |
| CO | Pollutant | [100,10000] | 1,126 | CO concentration level ($\mu g \cdot m^{-3}$) |
| Longitude | Spatial | [116.0,117.1] | 116.4 | Longitude |
| Latitude | Spatial | [39.5,40.5] | 40.02 | Latitude |
| Temp at 2m | Meteorological | [261.7,314.1] | 293.6 | Temperature at 2 metres(unit:K) |
| t250 | Meteorological | [208.4,241.7] | 225.05 | Temperature at 250hPa (unit:K) |
| t850 | Meteorological | [261.5,300.1] | 286.7 | Temperature at 850hPa (unit:K) |
| blh | Meteorological | [10.27,3795.74] | 362.39 | Boundary layer height (unit:m) |
| evapor | Meteorological | [-6.67e-04,1.28e-05] | -7.48e-05 | Evaporation (unit:m) |
| DEWP | Meteorological | [265.4,298.3] | 280.9 | Dew point at 2 metres (unit:K) |
| HUMI | Meteorological | [18.30,81.17] | 54.51 | Relative humidity (unit: %) |
| Geopotential | Meteorological | [13249,15618] | 14,541 | Geopotential height: (unit:$m^2s^{-2}$) |
| PRES | Meteorological | [90978,103532] | 98,868 | Surface pressure (unit:Pa) |
| Radiation | Meteorological | [0,3839385] | 1276707 | Surface net solar radiation (unit:J/$m^2$) |
| IRAIN | Meteorological | [0,4.14e-03] | 9.94e-05 | Average precipitation over the past 24 hours (unit:m) |
| wind10m | Meteorological | [0,3.2] | 1.25 | Current wind speed at 10 metres (unit:m$\cdot s^{-1}$) |
| wind500 | Meteorological | [0,13] | 7.6 | Current wind speed at 500 hPa (unit:m$\cdot s^{-1}$) |
| iws10m | Meteorological | [0,3.2] | 1.49 | Integrated wind speed at 10 metres (unit:m$\cdot s^{-1}$) |
| IHour | Other | [1.39%,95.83%] | 37% | Integrated air stagnant hour (unit: %) |
| nightlight | Other | [39.5,40.5] | 40.02 | Night-time light intensity (unit: nanowatts $\cdot$ sr $^{-1}$ $\cdot$ cm $^{-2}$) |

Besides the instant measurements of wind and precipitation, their cumulative amounts are also expected to affect the air pollution level [35]. Therefore, we create the integrated (cumulative) version of those variables, IHour, IRAIN, iws10m: IHour (Integrated Stag-

nant Hour) stands for the percentage of hours under air stagnant status over the past 24 hours, where the air stagnation is defined in section 2.2; IRAIN stands for the average precipitation over the past 24 hours; iws10m stands for the integrated wind speed at the height of 10 meters, where the wind speed is accumulated if the wind direction remains the same as the wind direction in the previous hour.

The spatial resolutions of monitoring stations, meteorological variables, and night-time light intensity are different. It is known as the change of support problem [48, 49]. Since most of our collection data are meteorological variables, we decide to use the centroids of their spatial grid as the standards for the location in modeling and inference. To match the spatial resolution of the meteorological variables, the hourly air pollution levels for each grid are set as the average air pollution levels at the monitoring stations within that grid. After the aggregation of monitoring stations, 23 grid cells with observed air pollutant concentrations are in total. Similarly, averaging night-time light intensity is taken in an 8-kilometer buffer zone within each grid to match the night-time light intensity with the spatial resolution of the meteorological variables.

Finally, in addition to variables described in Table 2.1, we include the categorical variables such as year, month, weekday, and hour, and a spatial indicator for being at the center of Beijing or not. Here we define the center of Beijing as the region within Beijing's Fifth Ring Road.

## 2.2 Air Stagnation

The air stagnation days were firstly defined by the National Oceanic and Atmospheric Administration (NOAA) in the United States [50]. The criteria of air stagnation days are determined by three factors: lower-air (10m) wind, upper-air (500hPa) wind, and precipitation. An air stagnation day is when 10m wind speed (lower-air wind) is less than 3.2m$\cdot s^{-1}$, the 500hPa wind (upper-air wind) is less than 13m$\cdot s^{-1}$, and it is precipitation-free.

NOAA air stagnation criteria are defined as daily criteria, but our data is obtained hourly. To accommodate for our data, we defined the criteria for air stagnant hours in a similar way. The lower-air wind is defined as 10m integrated wind speed, and the upper-air wind is defined as 500hPa wind speed. The daily wind speed is defined as the mean of wind speed (in all directions) over the past 24 hours. Our conditions for air stagnant hours

are:

- The current 10m integrated wind speed and daily 10m (lower-air) wind speed is less than 3.2 m$\cdot s^{-1}$.
- The current and daily 500hPa (upper-air) wind speed is less than 13 m$\cdot s^{-1}$.

Since we are also interested in the effects of rain on PM2.5, $NO_2$, and CO under air stagnation, precipitation is not used as the data exclusion criteria. Our air stagnant data's time scope is from February 1st, 2014, to May 31st, 2017. A total of 87,261 time-location combinations satisfy the air stagnation conditions and will be referred to as data points in our study. The proportions of air stagnation data takes up to 14% of collected data. Although it is often preferable to incorporate all existing data in a model, we believe the air stagnation data themselves are more suitable for revealing the spatial and temporal patterns of air pollution concentrations and investigating the associations. The large number of data points satisfying the air stagnation condition is also sufficient for deriving reliable estimates.

Chapter 3

# Spatio-temporal Modeling

## 3.1 Modeling of Air Pollution

We established a Gaussian spatio-temporal process model for modeling the air pollution concentrations. The model can be written as two parts: an unobservable process model and an observable data model[51].

The process model consists of two parts: one is the stationary mean effect explained by the meteorological, nightlight, spatial location and temporal related variables $\mathbf{x}(s,t)$ measured at the $t$ time point ($t = 1, \ldots, T$) and the $s$ site ($s = 1, \ldots, 25$). The other is the spatio-temporal random field $\delta_z(s,t)$ to better impute PM2.5, $NO_2$, and CO at the non-monitoring locations and also those that were missing or heavily affected by wind at the monitoring locations. The process model is

$$\mathbf{Z}(\mathbf{s},\mathbf{t}) = \mathbf{X}(\mathbf{s},\mathbf{t})\boldsymbol{\beta} + \boldsymbol{\delta_z(s,t)},$$

where $\mathbf{X}(\mathbf{s},\mathbf{t})$ is the $n \times p$ design matrix, where each row corresponds to the explanatory variables measured at location $\mathbf{s}$ and time $\mathbf{t}$. $\mathbf{X}(\mathbf{s},\mathbf{t})$ includes the potential main effects such as the temporal variables (time, year, month, weekday and hour), the meteorological variables (rain, temperatures at 2m, 250hPa, and 850hPa, pressure, dew point, boundary layer height, integrated rain, integrated air stagnant hour, radiation, dew point, geopotential height, and evaporation), nightlight and spatial location (indicator for being at the center of Beijing or not, latitude and longitude). We also consider two-way interactions between significant main effects. We take the grid-level air stagnation data as point referenced

data for modeling, using their centroids as the location of each grid cell to calculate the distance. $\boldsymbol{\beta}$ is the $p-$dimensional fixed effect vector, and the $\boldsymbol{\delta_z(s,t)}$ is the Gaussian spatio-temporal random effects with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma(s,t)}$. $Z(s,t)$, the element of $\mathbf{Z}(\mathbf{s},\mathbf{t})$, is a Gaussian process with mean $\mathbf{x}(s,t)^T\boldsymbol{\beta}$, where $\mathbf{x}(s,t)$ is a row of the design matrix $\mathbf{X}(\mathbf{s},\mathbf{t})$.

The data model is

$$\boldsymbol{y(s,t)} = \mathbf{Z}(\mathbf{s},\mathbf{t}) + \boldsymbol{\delta_y(s,t)},$$

where $y(s,t)$, the element of vector $\boldsymbol{y(s,t)}$, is the observed air pollution level measured at the $t$ time point and the $s$ site after cubic root transformation. The cubic root transformation provides better normality of the residuals that satisfy our assumption for the spatio-temporal random field. $\boldsymbol{\delta_y(s,t)}$ is the measurement error and it follows multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance $\sigma^2_{\text{nugget}}\mathbf{I}$, where the variance $\sigma^2_{\text{nugget}}$ is also known as the nugget variance. The observed data $y(s,t)$ given $Z(s,t)$ is an independent Gaussian process with mean $Z(s,t)$ and variance $\sigma^2_{\text{nugget}}$.

We combined two processes together, so we have

$$\boldsymbol{y(s,t)} = \mathbf{X}(\mathbf{s},\mathbf{t})\boldsymbol{\beta} + \boldsymbol{\epsilon(s,t)},$$

where $\boldsymbol{\epsilon(s,t)} \sim N(\mathbf{0}, \sigma^2_{\text{nugget}}\mathbf{I} + \boldsymbol{\Sigma(s,t)})$. Here $\sigma^2_{\text{nugget}}\mathbf{I} + \boldsymbol{\Sigma(s,t)}$ is assumed to follow either the Gneiting model structure [52] or the Product-Sum model structure [53]. These two variogram models both allow for the interactions of the space and time correlation, and they have been proved to be flexible to describe a wide range of variogram surfaces through simulations and empirical studies [54, 53, 55, 52].

## 3.2 Parameter Estimation

To expedite the computing time, we used a two-stage estimation procedure in the model selection process: the regression coefficients, $\boldsymbol{\beta}$, was estimated by the ordinary least squares method; the spatio-temporal covariance, $\boldsymbol{\Sigma}$, was estimated from the empirical variogram by using either the Gneiting model [52] or the Product-Sum model [53] using the ordinary least squares residuals. The variogram between a pair of $i^{th}$ and $j^{th}$data points is defined as

$$2\gamma_{i,j} = \text{Var}(\epsilon_i - \epsilon_j) = \text{Var}(\epsilon_i) + \text{Var}(\epsilon_j) - 2\text{Cov}(\epsilon_i, \epsilon_j),$$

and the half of the variogram is defined as semivariogram [51]

$$\gamma_{i,j} = \frac{1}{2}\text{Var}(\epsilon_i - \epsilon_j).$$

We assume that the residuals are isotropic and stationary under air stagnation. Under the isotropic assumption, i.e., that the semivariogram $\gamma$ depends on neither location nor time, the spatio-temporal semivariogram can be written as a function of distance $h_{ij}$ and time lag $k_{ij}$ between two points,

$$\gamma(h_{ij}, k_{ij}) = \frac{1}{2}\text{Var}(\epsilon_{s_i,t_i} - \epsilon_{s_j,t_j}),$$

where $h_{ij} = \text{Distance}(s_i, s_j)$ and $k_{ij} = \text{Time Lag}(t_i, t_j)$ (in hours). In addition, we refer the upper bound of the semivariogram as the global sill variance $\sigma^2$, and it is defined as

$$\sigma^2 = \text{Var}(\epsilon_i) = \sigma^2_{\text{nugget}} + \text{Var}(\delta_z(s_i, t_i))),$$

for any $i$.

To estimate this semivariogram, we utilize the observed process $\boldsymbol{y(s,t)}$ to obtain the empirical semivariogram of the residuals. The empirical semivariogram is

$$\hat{\gamma}(h, k) = \frac{1}{2|A_{h,k}|} \sum_{i,j \in A_{h,k}} (\hat{\epsilon}_i - \hat{\epsilon}_j)^2,$$

where $\hat{\epsilon}$ is the estimated residuals of the least square fit, $A_{h,k}$ is the set of pairs of data points that have distance $h$ kilometers and time lag $k$ hours respectively, and $|A_{h,k}|$ is the size of $A_{h,k}$. In practice, we only compute the empirical semivariogram at different values of time lag $k$ and distance $h$. In terms of distance $h$, we include all pairs of data points whose distances are within $(h - 0.5, h + 0.5]$ in $A_{h,k}$.

The Product-Sum model [53] and the Gneiting model [52] are two parametric spatio-temporal covariance models that fit to the above empirical spatio-temporal semivariograms (or correlations). Instead of estimating the spatio-temporal variogram at one time, their methods first utilize the information from empirical spatial semivariogram and temporal semivariogram to estimate the spatial variogram and temporal variogram respectively. The spatial semivariogram is the spatio-temporal semivariogram when the time lag $k = 0$, and the temporal semivariogram is the spatio-temporal semivariogram when the distance

$h = 0$. Estimating them separately can significantly simplify and expedite the optimization process. Given the estimated parameters for spatial and temporal semivariogram, we can estimate space and time interaction parameters with empirical spatio-temporal semivariogram. We describe two spatio-temporal covariance models as follows.

The Product-Sum model [53] assumes a spatio-temporal covariance structure:

$$\text{Cov}_{s,t}(h,k) = k_1\text{Cov}_s(h) + k_2\text{Cov}_s(h)\text{Cov}_t(k) + k_3\text{Cov}_t(k),$$

where $\text{Cov}_s(h)$ and $\text{Cov}_t(k)$ are valid temporal and spatial covariance structures, respectively. It can also be written in terms of the semivariograms:

$$\gamma_{s,t}(h,k) = \gamma_{s,t}(h,0) + \gamma_{s,t}(0,k) - m\gamma_{s,t}(h,0)\gamma_{s,t}(0,k),$$

where $\gamma_{s,t}(h,0)$ is the spatial semivariogram, and $\gamma_{s,t}(0,k)$ is the temporal semivariogram. Parameter $m$ controls the degree of space-time interaction. We use spherical $c_{\text{sill},1}\text{sph}(\frac{h}{r})$, and exponential semivariograms $c_{\text{sill},2}\text{exp}(\frac{h}{\alpha_{\text{scale}}})$ for $\gamma_{s,t}(h,0)$ and $\gamma_{s,t}(0,k)$ respectively. To be specific,

$$\gamma_{s,t}(h,0) = \begin{cases} \sigma^2_{\text{nugget, spatial}} + (c_{\text{sill},1} - \sigma^2_{\text{nugget}})(\frac{3}{2}h/r - \frac{1}{2}(h/r)^3), & \text{if } 0 < h \leq r \\ c_{\text{sill},1}, & \text{if } h > r \end{cases}$$

and

$$\gamma_{s,t}(0,k) = c_{\text{sill},2}\exp(-k/\alpha_{\text{scale}}).$$

The spatial semivariogram, $\gamma_{s,t}(h,0)$, and the temporal semivariogram, $\gamma_{s,t}(0,k)$, can be fitted separately using weighted least squares with their corresponding sample spatial or temporal semivariograms. Lastly, the estimate of $m$ can be then obtained by the weighted least squares with spatio-temporal semivariograms.

The Gneiting model [52] assumes an alternative spatio-temporal correlation structure:

$$C_{s,t}(h,k) = \begin{cases} (ak^{2\alpha}+1)^{-(\delta+\eta)}, & \text{if } h = 0, \\ p(ak^{2\alpha}+1)^{-(\delta+\eta)} \times \exp(-\frac{bh}{(ak^{2\alpha}+1)^{\frac{\eta}{2}}}), & \text{otherwise.} \end{cases}$$

The corresponding semivariogram structure is

$$\sigma^2 - C_{s,t}(h,k) = \gamma_{s,t}(h,k),$$

where $\sigma^2 = \text{Var}(\epsilon)$.

The similar weighted least squares procedure can be used to estimate parameters: $a$ and $b$, non-negative scaling parameters of time and space respectively; $\alpha$, a smoothness parameter; $\eta$, a space-time interaction parameter; and $\sigma^2$, the variance of residuals (or global sill variance).

For association studies, we make inference using either non-spatio-temporal regression (ordinary least squares) or restricted spatial regression, which we will describe in section 3.4. For the point estimates of associations, the restricted spatial regression is the same as ordinary least squares estimates [1, 56]. For variance estimation in restricted spatial regression, we need the estimates of nugget effects. The nugget effect in our product-sum models [53] is spatial nugget effect $\sigma^2_{\text{nugget, spatial}}$, and it can be directly estimated. While the nugget effect in Geniting model [52] here can be implicitly obtained by the estimates of $(1-p)\sigma^2$.

## 3.3 Kriging

Kriging is first proposed by [57] and is used for the interpolation based on Gaussian process. The interpolation of air pollution concentration levels at location $s'$ and time $t'$ requires estimating two components, the global mean, $\mathbf{x}^T(s',t')\boldsymbol{\beta}$, and the Gaussian spatio-temporal random field, $\epsilon(s',t')$. $\boldsymbol{\beta}$ can be estimated by ordinary least squares, and the distribution of $\epsilon(s',t')$ given $\boldsymbol{\epsilon}$, the residuals of observed data points, is

$$\epsilon(s',t') \mid \epsilon \sim N(\text{Cov}(\boldsymbol{\epsilon}, \epsilon(s',t'))^T(\sigma^2_{\text{nugget}} I + \boldsymbol{\Sigma})^{-1}\boldsymbol{\epsilon}, \Sigma^*),$$

where

$$\Sigma^* = \text{Var}(\epsilon(s',t')) - \text{Cov}(\boldsymbol{\epsilon}, \epsilon(s',t'))^T(\sigma^2_{\text{nugget}} I + \boldsymbol{\Sigma})^{-1}\text{Cov}(\boldsymbol{\epsilon}, \epsilon(s',t')).$$

We used kriging to recover the maps of air pollution concentrations at different times. We showed the spatial maps of air pollution levels in four seasons for three pollutants in the result.

## 3.4 Inference for Associations

Our primary goal is to estimate each pollutant's yearly, monthly, weekly, and hourly trends. In addition, we want to study each meteorological variable's effect one by one after controlling for the temporal effects to avoid the collinearities. The association inference result would guide people better to understand the air pollution concentrations under varying weather conditions. We deploy a linear model to learn the association between air pollution and those covariates of interest. We use the same model structure as the one for kriging,

$$\mathbf{y}(\mathbf{s},\mathbf{t}) = \mathbf{X}(\mathbf{s},\mathbf{t})\boldsymbol{\beta} + \boldsymbol{\epsilon}(\mathbf{s},\mathbf{t}),$$

where $\boldsymbol{\epsilon}(\mathbf{s},\mathbf{t}) \sim \mathrm{N}(0, \sigma^2_{\text{nugget}} I + \Sigma)$.

To give an example of the mean function $\mathbf{X}(\mathbf{s},\mathbf{t})^T\boldsymbol{\beta}$ in our model with just the temporal effects, it can be expanded as

$$\boldsymbol{E}y(s,t) = \beta_0 + \beta_{\text{year},y_t} + \beta_{\text{month},m_t} + \beta_{\text{week},w_t} + \beta_{\text{hour},h_t},$$

where $y_t$, $m_t$, $w_t$, $h_t$ are the year, month, weekday, and hour that time $t$ is in.

The covariance $\sigma^2_{\text{nugget}} I + \Sigma$ is chosen in subsection 3.5.2 between the Gneiting model [52] and the Product-Sum model [53]. To estimate the regression coefficients of temporal and other meteorological effects, one could use the generalized least squares estimate, accounting for the spatio-temporal covariance:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T(\sigma^2_{\text{nugget}} I + \Sigma)^{-1}\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2_{\text{nugget}} I + \Sigma)^{-1}\mathbf{y},$$

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T(\sigma^2_{\text{nugget}} I + \Sigma)^{-1}\mathbf{X})^{-1}.$$

However, studies have shown that the direct use of estimates incorporating spatial effects would mask the estimates of associations between response and those fixed effects due to the confounding with spatial random effects [58, 1]. The spatial confounding issues could produce misleading inference or interpretation for the fixed effects, and it has not been addressed enough in the past studies about air pollution. We can avoid this issue by deploying restricted spatial regression to estimate the fixed effects [1, 56]. The idea is to restrict the spatial random effect to be orthogonal to the fixed effects. In our case, adding the spatio-temporal random effects will only exacerbate the confounding problem

in the association inference. We extend the spatial restricted regression to spatio-temporal restricted regression and show the closed-form solutions for Gneiting model [52], and the Product-Sum model [53].

We consider the following regression model with spatio-temporal random effects,

$$y(s,t) = \mathbf{x}(s,t)^T\boldsymbol{\beta} + \epsilon(s,t).$$

We restrict the spatio-temporal random effects in $\epsilon(\mathbf{s},\mathbf{t})$ to be orthogonal to the fixed effects. The sampling error part does not need orthogonal adjustment since it will not cause confounding problems. In practice, the spatio-temporal random effects can be restricted by multiplying the $(I-P)$ projection matrix where $P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, so that it is orthogonal to the fixed effects at observed locations and times. The covariance of the spatio-temporal error term after the restriction is

$$\boldsymbol{\epsilon}(\mathbf{s},\mathbf{t}) \sim \mathrm{N}(0, (I-P)\Sigma(I-P) + \sigma^2_{\text{nugget}}I),$$

and $\Sigma$ is the covariance of spatio-temporal random effects at the locations and times with observations. With this restricted model, our estimates using Generalized Least Squares become

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2_{\text{nugget}}(\mathbf{X}^T\mathbf{X})^{-1},$$

The point estimates of fixed effects are the same as the Ordinary Least Squares (OLS), and the form of the the variance is exactly the same as the restricted spatial regression in [1]. Note that the nugget effect here $\sigma^2_{\text{nugget}}$ is referred to spatio-temporal nugget effect instead of spatial nugget effect in [1]. For product-sum model [53], the spatio-temporal nugget effect comprises the spatial nugget effects $\sigma^2_{\text{nugget,spatial}}$ and temporal nugget effects $\sigma^2_{\text{nugget,spatial}}$. It can be derived as

$$\begin{aligned}
\sigma^2_{\text{nugget}} &= \lim_{(h,k)\to(0^+,0^+)} \gamma_{s,t}(h,k) \\
&= \lim_{h\to 0^+} \gamma_{s,t}(h,0) + \lim_{k\to 0^+} \gamma_{s,t}(0,k) - m \lim_{h\to 0^+} \gamma_{s,t}(h,0) \lim_{k\to 0^+} \gamma_{s,t}(0,k) \\
&= \sigma^2_{\text{nugget,spatial}} + \sigma^2_{\text{nugget,temporal}} - m\sigma^2_{\text{nugget,spatial}}\sigma^2_{\text{nugget,temporal}}.
\end{aligned}$$

In our specific case, $\sigma^2_{\text{nugget}} = \sigma^2_{\text{nugget,spatial}}$ since $\sigma^2_{\text{nugget,temporal}} = 0$.

For Gneiting model [52], the spatio-temporal nugget effects is

$$\sigma^2_{\text{nugget}} = \sigma^2 \lim_{(h,k)\to(0^+,0^+)} (C_{s,t}(h,k)) - C_{s,t}(0,0)).$$

In our specific case, $\sigma^2_{\text{nugget}} = \sigma^2(1-p)$

As is pointed out in [56], the variance estimate using restricted spatial regression is usually smaller than that in OLS, and could possibly result in higher chance of making type I error for fixed effect significant test than OLS. Therefore, in our case, we report both the OLS and restricted spatio-temporal estimates for associations and variance components $\hat{\sigma}^2$ and $\hat{\sigma}^2_{\text{nugget}}$ for inference in our association study.

## 3.5 Model Selection

### 3.5.1 Evaluation Metrics

We assess the prediction accuracy of different proposed models via cross-validation. The prediction accuracy is stratified by monitoring station is available at a grid cell:

- At grid cells with monitoring stations, we impute the PM2.5, $NO_2$, and CO that are missing due to either not being collected or not qualifying the air stagnant condition. The time-series data at a grid cell are randomly divided into $P$ equal-sized subsamples, and we set $P = 5$ in this paper. We evaluate the prediction accuracy by repeatedly withholding one subsample at a grid cell as the test data. We refer to this evaluation approach as the Leave-Partial-Time-Out (LPTO) cross-validation. It assesses the interpolation accuracy at a location with historical air pollutant data.

- To evaluate the prediction accuracy for grid cells without any monitoring stations, we take one observed grid cell as the test data repeatedly for all 25 observed grid cells. This evaluation approach will be referred to as the Leave-One-Location-Out cross-validation (LOLO). Such a target-oriented cross-validation design was recommended by [59] for space-time data to prevent over-fitting – the model very well predicts on subsets of the time series of the locations used for training but fails in the extrapolation of unknown locations.

The LOLO cross-validation selects more variables that capture the spatial variation, while the LPTO cross-validation favors variables that capture temporal variation. The MAE of LOLO cross-validation is

$$MAE_{\text{LOLO cross-validation}} = \frac{1}{\sum_{s=1}^{S} \mathcal{T}_s} \sum_{s=1}^{S} \sum_{t \in \mathcal{T}_s} |y_{st} - \hat{y}_{st}|,$$

for each removed location $s$, $\{y_{st}\}$ are the observed response from test data, and $\hat{y}_{st}$ are the predicted value of test data with training data excluding location $s$. $\mathcal{T}_s$ is the observed time points in location $s$. And the MAE is averaged over the entire spatial location $s$. The MAE of LPTO cross-validation is

$$MAE_{\text{LPTO cross-validation}} = \frac{1}{\sum_{s=1}^{S} \mathcal{T}_s} \sum_{s=1}^{S} \sum_{p=1}^{P} \sum_{t \in \mathcal{T}_{sp}} |y_{st} - \hat{y}_{st}|,$$

for each removed location $s$ and partial time $p$, $\{y_{st}\}$ are the observed response from test data, and $\hat{y}_{st}$ are the predicted value of test data with training data excluding the partial time $p$ at location $s$. And the MAE is averaged over the entire spatial location $s$ and time $p$. Other metrics (MAE, RMSE, $R^2$) are calculated in a similar way.

We calculate the mean absolute value (MAE) and the square root of the mean square error (RMSE), with the data points being equally weighted for each approach. LPTO RMSE and LOLO RMSE are both aggregated across all locations and time points. We provide these two types of cross-validation metrics to show our kriging performance for locations with partial observations and new locations. We selected the predictors and covariance structures based on the mean absolute error (MAE) by combining these two cross-validation strategies.

The covariance structures of the final models were selected first based on the average of LPTO MAE and LOLO MAE for each pollutant. After we fixed the covariance structure, we obtained the variables in the final models by combining the predictors selected by both LOLO and LPTO under selected covariance structures. We also considered the interactions between selected main effects, but no interaction term has been selected in the forward selection process.

### 3.5.2 Variable Selection

The candidate predictors are first screened by the Bayesian Information Criterion (BIC) [60] to avoid the high computational cost and the forward selection procedure without taking the spatio-temporal effect into account. Within the candidate predictors for each pollutant, we use the forward selection cross-validation to determine the best set of predictors, including both main effects and two-way interactions of candidate predictors. This cross-validation variable selection procedures are conducted for each proposed covariance structures: product-sum [53] and Gneiting [52].

Chapter 4

# Results

In this section, we first analyze the temporal trends of three pollutants under air stagnation and their associations between meteorological and nightlight variables. We also discuss the differences of inference between OLS and restricted spatio-temporal regression. Then we summarize the model selection results and performance in section 4.3. Lastly, we present the spatial distributions of three air pollutants in four seasons after interpolations using the final selected models.

## 4.1 Temporal Trends of Air Pollutants

Figure 4.1 shows the year effects (2014 as the baseline) and the month effects (January as the baseline) of $PM2.5^{1/3}$, $NO_2^{1/3}$ and $CO^{1/3}$. The mean estimates of all three pollutants decrease over the years in general except for a small bump in 2016. The monthly trends suggest the severest air pollution is in winter. One of the reasons is the heating effect during winter, given that the central heating in Beijing usually turns on from November 15 to March 15. Comparing three pollutants, CO stays at a low concentration level much longer, from April to September. In contrast, PM2.5 remains at a low concentration level only from August to September. For three pollutants, Sunday is the cleanest day in Beijing, as more people stay at home and fewer human activities occur during that day. The weekly peaks are different for the three pollutants, but they are all concentrated from Wednesday to Friday as it approaches the end of the weekday. PM2.5 remains higher on Thursday and Friday, while it can stay lower from Sunday to Wednesday. $NO_2$ remains low from Friday to Sunday. The different weekly patterns could be attributed to different characteristics

of three pollutants. PM2.5 and CO are more stable and could accumulate over days while $NO_2$ is not. The hourly trend of the three pollutants all show a dip around 3pm-5pm during the afternoon, which could be attributed to the temperatures, sunlight, and human activities during that time.

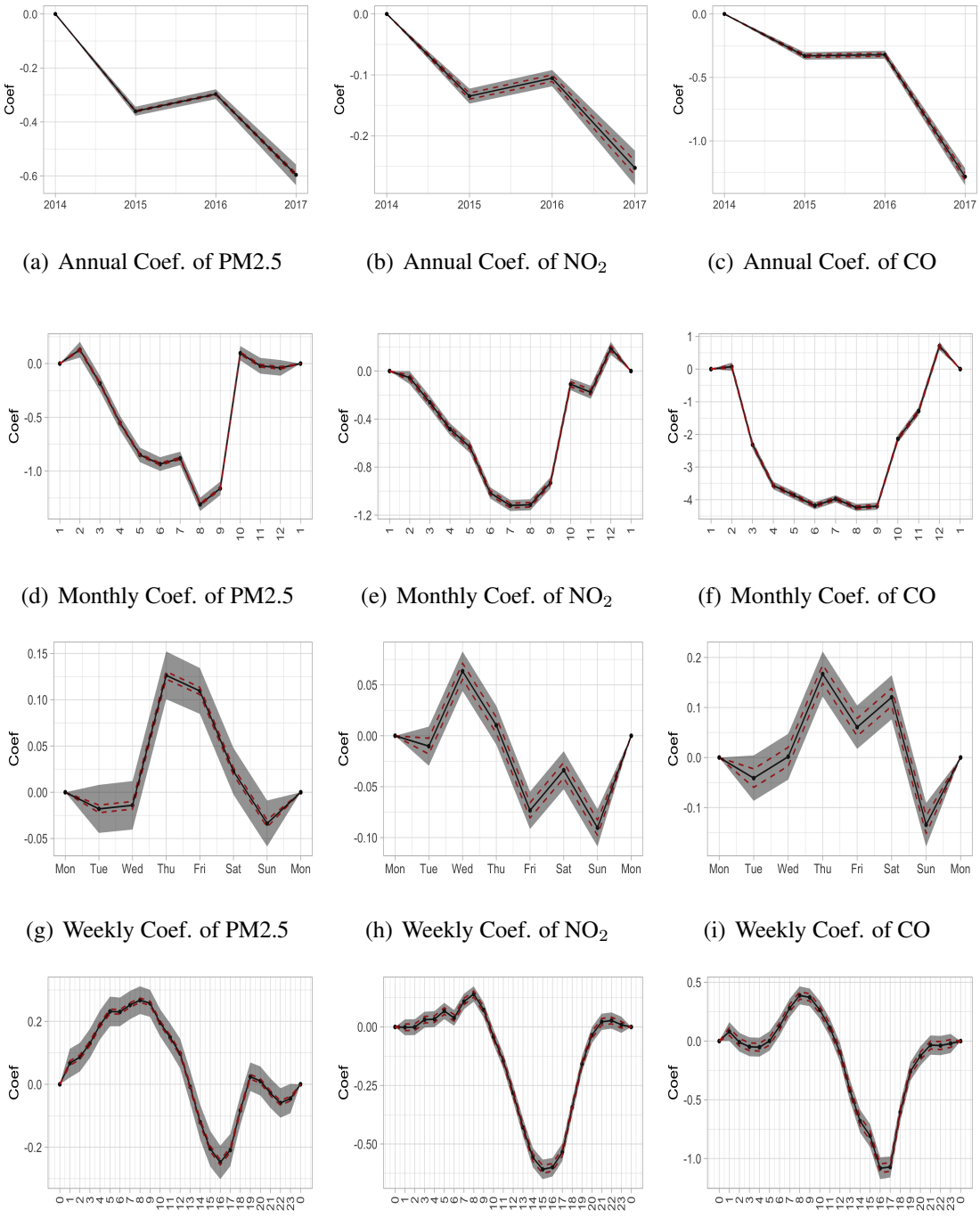
All temporal trends are tested significantly at 95% confidence level, using either OLS method or restricted spatio-temporal with the product-sum method. The estimated standard error $\sigma$ of OLS method is 1.015, 0.752, and 3.154 for PM2.5, $NO_2$, and CO, respectively. While the square root of the nugget effect $\sigma_{\text{nugget}}$ with product-sum is 0.141, 0.305, and 0.521. The estimated nugget effect with the Gneiting model is 0.144, 0.461, and 0.638. The nugget effects estimated from the product-sum model are slightly smaller than those from the Gneiting model. Here, we only plot the confidence bands estimated using product-sum. From Figure 4.1, the confidence bands of restricted spatio-temporal regression using product-sum are much narrower than OLS. The confidence bands of the Gneiting model are somewhere in between product-sum and OLS, and closer to product-sum. The restricted spatio-temporal regression did not alleviate the spatio-temporal random effects confounding with fixed effects in our case since the estimated nugget variances $\hat{\sigma}^2_{\text{nugget}}$ from Gneiting model and product-sum are larger than the OLS variance $\hat{\sigma}^2$.

## 4.2 Associations of Meteorological Variables/Nightlight

Table 4.1, Table 4.2, and Table 4.3 summarize the regression coefficients of meteorological variables and night-time light intensity after controlling for the temporal trends (year, month, week, and hour) for PM2.5, $NO_2$, and CO respectively. It can be interpreted as the associations between meteorological variables/nightlight and three pollutants respectively when controlling for all temporal effects.

For the PM2.5 model, the IHour (integrated air stagnant hour) shows the most significant association with PM2.5. It is positively correlated with PM2.5, that is, the longer the air stagnant hours it has been through, the higher the PM2.5 is. Consider the extreme case, when there is no wind over the past 24 hours (i.e., IHour = 100%), PM2.5 would likely accumulate over time. We quantify that with a 10% increase of air stagnation hour, the $\text{PM2.5}^{1/3}$ will increase 0.136 on average.

For $NO_2$ and CO model, nightlight density shows the most significant association. These two pollutants are expected to be higher on average if the night-time light intensity




(a) Annual Coef. of PM2.5 (b) Annual Coef. of $NO_2$ (c) Annual Coef. of CO

(d) Monthly Coef. of PM2.5 (e) Monthly Coef. of $NO_2$ (f) Monthly Coef. of CO

(g) Weekly Coef. of PM2.5 (h) Weekly Coef. of $NO_2$ (i) Weekly Coef. of CO

(j) Hourly Coef. of PM2.5 (k) Hourly Coef. of $NO_2$ (l) Hourly Coef. of CO

Figure 4.1: Temporal trends (regression coefficients) of PM2.5, $NO_2$, CO in Beijing under air stagnation from Feb 1st, 2014 to May 31st, 2017 with 95% confidence intervals. The gray bands are the 95% confidence interval using OLS inference. The red dashed lines are the 95% confidence interval using restricted spatio-temporal regression with the product-sum covariance model. The concentration levels are under the cubic root transformation. The first row presents the annual trends of (a) PM2.5, (b) $NO_2$ and (c) CO; the second row presents the monthly trends of (d) PM2.5, (e) $NO_2$, and (f) CO; the third row shows the weekly patterns of (g) PM2.5, (h) $NO_2$, and (i) CO; the last row shows the hourly patterns of (j) PM2.5, (k) $NO_2$, and (l) CO.

Table 4.1: Coefficients of Meteorological and Nightlight Variables for PM2.5 under ordinary least squares (OLS). The standard errors of restricted spatio-temporal regression (RSR) for the product-sum(PS) and Gneiting(GN) model are included. The coefficients are controlled for temporal effects, and sorted by their absolute t-values.

PM2.5 model

| Variables | Estimate | S.E. | RSR S.E. (PS) | RSR S.E. (GN) | t-value | p-value |
|---|---|---|---|---|---|---|
| IHour | 1.36 | 0.019 | 2.70e-03 | 3.31e-03 | 71.48 | <0.0001 |
| TEMP | 0.069 | 1.13e-03 | 1.73e-04 | 1.36e-04 | 61.00 | <0.0001 |
| t850 | 0.071 | 1.17e-03 | 1.68e-04 | 1.70e-04 | 60.91 | <0.0001 |
| t250 | 0.047 | 9.00e-04 | 1.30e-04 | 1.23e-04 | 51.78 | <0.0001 |
| PRES | 4.21e-05 | 1.53e-06 | 2.35e-07 | 1.99e-07 | 27.42 | <0.0001 |
| blh | -2.5e-04 | 1.22e-05 | 1.68e-06 | 2.02e-06 | -20.58 | <0.0001 |
| nightlight | 5.49e-03 | 3.14e-04 | 4.66e-05 | 4.45e-05 | 17.52 | <0.0001 |
| evapor | 823.22 | 55.29 | 7.70 | 7.80 | 14.89 | <0.0001 |
| IRAIN | -209.67 | 14.45 | 2.01 | 2.04 | -14.51 | <0.0001 |
| geopotential | -3.09e-05 | 1.30e-05 | 1.80e-06 | 1.83e-06 | -2.39 | 0.017 |
| HUMI | 2.49e-04 | 2.27e-04 | 3.16e-05 | 3.21e-05 | 1.09 | 0.274 |
| solar | 1.54e-09 | 3.30e-09 | 4.59e-10 | 4.67e-10 | 0.47 | 0.641 |
| DEWP | 3.49e-05 | 2.55e-04 | 3.54e-05 | 3.60e-05 | 0.14 | 0.891 |

is potent. Night-time light intensity reflects the population density at different locations [46, 47] and thus the spatial heterogeneity of human activities. The result implies that the high concentrations of these two pollutants are associated with more intensive human activities.

At least two levels of temperature and PRES (pressure) are among the Top 5 most significant associations for three pollutants, and they all show a positive correlation. As temperature or pressure goes up, air pollution concentrations increase. While HUMI, solar, and DEWP are the three least significant variables for all three pollutants, without significant associations (p-value are all greater than 0.1). We consider that these three variables are not strongly correlated with air pollutants after controlling for the temporal trends.

For IRAIN (integrated rain), the larger the integrated rain is, the lower the concentrations of PM2.5 and $NO_2$ will be on average. On the contrary, the larger the integrated rain is, the higher the concentration of CO will be on average, and this effect is less significant. The reason might be that PM2.5 and $NO_2$ can be dissolved in water while CO is not dissolvable. When there is a heavy rain, it usually washes out PM2.5 and $NO_2$, but it does not decrease the CO concentration.

Table 4.2: Coefficients of Meteorological and Nightlight Variables for $NO_2$ under ordinary least squares (OLS). The standard errors of restricted spatio-temporal regression (RSR) for the product-sum(PS) and Gneiting(GN) model are included. The coefficients are controlled for temporal effects, and sorted by their absolute t-values.

$NO_2$ model

| Variables | Estimate | S.E. | RSR S.E. (PS) | RSR S.E. (GN) | t-value | p-value |
|---|---|---|---|---|---|---|
| nightlight | 0.032 | 2.07e-04 | 1.30e-04 | 1.26e-04 | 152.43 | <0.0001 |
| PRES | 9.93e-05 | 1.09e-06 | 4.90e-07 | 6.63e-07 | 91.04 | <0.0001 |
| TEMP | 0.034 | 8.48e-04 | 3.56e-04 | 5.17e-04 | 40.42 | <0.0001 |
| t850 | 0.035 | 8.79e-04 | 3.61e-04 | 5.43e-04 | 39.57 | <0.0001 |
| IRAIN | -375.94 | 10.64 | 4.35 | 6.57 | -35.25 | <0.0001 |
| evapor | 1317.01 | 40.75 | 16.78 | 25.02 | 32.32 | <0.0001 |
| blh | -2.50e-04 | 9.00e-06 | 3.66e-06 | 5.55e-06 | -27.79 | <0.0001 |
| t250 | 8.29e-03 | 6.76e-04 | 2.75e-04 | 4.15e-04 | 12.26 | <0.0001 |
| geopotential | 1.05e-04 | 9.59e-06 | 3.90e-06 | 5.88e-06 | 10.96 | <0.0001 |
| IHour | 6.55e-02 | 1.45e-02 | 5.88e-03 | 8.88e-03 | 4.52 | <0.0001 |
| HUMI | 1.39e-04 | 1.68e-04 | 6.84e-05 | 1.03e-04 | 0.83 | 0.408 |
| solar | 1.97e-09 | 2.45e-09 | 9.94e-10 | 1.50e-09 | 0.81 | 0.421 |
| DEWP | -3.81e-05 | 1.89e-04 | 7.66e-05 | 1.16e-04 | -0.20 | 0.840 |

Table 4.3: Coefficients of Meteorological and Nightlight Variables for CO under ordinary least squares (OLS). The standard errors of restricted spatio-temporal regression (RSR) for the product-sum(PS) and Gneiting(GN) model are included. The coefficients are controlled for temporal effects, and sorted by their absolute t-values.

CO model

| Variables | Estimate | S.E. | RSR S.E. (PS) | RSR S.E. (GN) | t-value | p-value |
|---|---|---|---|---|---|---|
| nightlight | 0.030 | 5.40e-04 | 2.33e-04 | 2.41e-04 | 55.00 | <0.0001 |
| PRES | 1.37e-04 | 2.66e-06 | 1.14e-06 | 1.19e-06 | 51.45 | <0.0001 |
| t850 | 0.081 | 2.07e-03 | 8.51e-04 | 9.48e-04 | 39.29 | <0.0001 |
| t250 | 0.058 | 1.59e-03 | 6.51e-04 | 7.17e-04 | 36.49 | <0.0001 |
| IHour | 1.20 | 3.40e-02 | 1.39e-02 | 1.55e-02 | 35.38 | <0.0001 |
| blh | -6.38e-04 | 2.12e-05 | 8.66e-06 | 9.60e-06 | -30.02 | <0.0001 |
| TEMP | 0.052 | 2.01e-03 | 8.31e-04 | 8.90e-04 | 26.07 | <0.0001 |
| evapor | 1682.80 | 96.67 | 39.43 | 43.43 | 17.41 | <0.0001 |
| geopotential | 2.37e-04 | 2.26e-05 | 9.21e-06 | 1.02e-05 | 10.46 | <0.0001 |
| IRAIN | 80.90 | 25.30 | 10.28 | 11.38 | 3.20 | 0.0014 |
| solar | 7.18e-09 | 5.78e-09 | 2.35e-09 | 2.60e-09 | 1.24 | 0.214 |
| HUMI | 4.17e-04 | 3.97e-04 | 1.62e-04 | 1.79e-04 | 1.05 | 0.294 |
| DEWP | -3.47e-04 | 4.46e-04 | 1.81e-04 | 2.00e-04 | -0.78 | 0.436 |

## 4.3 Model Selection and Prediction Accuracy

We select the predictors and covariance structures based on the mean absolute error (MAE) under overall cross-validation performances. Table 4.4 summarizes the cross-validation performance of our final models for three pollutants in MAE, RMSE, and $R^2$ at the transformed scale (cubic root). The cross-validation $R^2$ is defined as

$$R^2_{\text{cross-validation}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_{test})^2},$$

where $\{y_i\}_{i=1}^n$ are the observed response from test data and $\{\hat{y}_i\}_{i=1}^n$ are the predicted value of test data. $\bar{y}_{test}$ is the sample mean of the test data. The cross-validation $R^2$ does not depend on the scale of response, and thus it is more comparable among different air pollutants than MAE and RMSE.

The final models provide estimates of the air pollution levels at all locations and all the time points, regardless of the availability of the monitoring station. According to the LOLO MAE and LPTO MAE average, the product-sum model is used for PM2.5, $NO_2$, and CO. The intermediate results of model selection are provided in Appendix A.1. The covariance parameters estimates are shown in Appendix A.2.

The Leave-Partial-Time-Out (LPTO) cross-validation $R^2$'s are above 0.9 for all three pollutants. With the product-sum covariance, LPTO $R^2$'s are 0.948 for PM2.5, 0.912 for $NO_2$, and 0.927 for CO. It indicates that our proposed model could accurately impute the missing values at the monitoring stations. The Leave-One-Location-Out (LOLO) cross-validation $R^2$'s are lower than LPTO $R^2$'s in general. The LOLO cross-validation $R^2$'s are 0.884 for PM2.5, 0.579 for $NO_2$ and 0.751 for CO. It is due to the lack of historical air pollution data at withholding grid cells, and thus the imputation is entirely driven by the spatial dependence. For both cross-validation strategies, $NO_2$ has the lowest $R^2$ because it is more reactive than PM2.5 and CO in the atmosphere and hence has a weaker spatio-temporal dependence.

In terms of the variables selected for kriging accuracy, the month and geopotential height are selected for all three pollutants. The usage of nightlight intensity and temperature at different levels seem to improve the prediction accuracy for $NO_2$ and CO but not PM2.5. Residuals of three pollutants models are all tested stationary using Dickey-Fuller test [61].

We also use the diagnostic plots to check the validity of model assumptions. The resid-

uals of the PM2.5 model are well in line with their theoretical distributions. In contrast, the residuals of CO model show a heavier tail due to heteroscedasticity because the variance of residuals increases with the wind speed. $NO_2$ also shows slight heteroscedasticity. See details of those diagnoses in Appendix B. One possibility is to define a more strict air stagnation condition, but a better solution would be to include the wind-driven transmission effects in the model.

Table 4.4: Evaluation Metrics under two types of Cross-validation Methods for the Final Models: Leave-One-Location-Out (LOLO) and Leave-Partial-Time-Out (LPTO)

| Pollutant | Covariance | Selected Variables | Type of CV | MAE | RMSE | $R^2$ |
|---|---|---|---|---|---|---|
| PM2.5 | Product-Sum | Month, Latitude, Hour, Geopotential, HUMI | LPTO | 0.168 | 0.257 | 0.948 |
| | | | LOLO | 0.268 | 0.385 | 0.884 |
| $NO_2$ | Product-Sum | Month, Weekday, Hour, Nightlight, Temp at 2m, t850, Geopotential, Latitude, Longitude | LPTO | 0.178 | 0.259 | 0.912 |
| | | | LOLO | 0.427 | 0.566 | 0.579 |
| CO | Product-Sum | Month, Nightlight, t850, t250, PRES, Geopotential | LPTO | 0.368 | 0.595 | 0.927 |
| | | | LOLO | 0.766 | 1.098 | 0.751 |

## 4.4 Spatial and Seasonal Distributions of Air Pollutants

The resulting maps were rendered using ggmap package in R[2]. Geo-location variables such as Latitude, Longitude, and Geopotential frequently appear in PM2.5 and $N0_2$ models, but not in CO models. It is probably because CO is stable and quickly travels for a long time and distance. As a result, its spatial heterogeneity is not as substantial as PM2.5 and $N0_2$. PM2.5 and $N0_2$ concentration levels decrease from south to north and from west to east. There were several cities with heavy industries in the south of Beijing like Baoding and Langfang in Hebei province, which explained why air pollution was severer in the south. The central area of Beijing lies slightly west within our studying range. Hence, the negative coefficient of longitude in $NO_2$ model captures Beijing's central effects, as the central area of Beijing showed higher air pollution levels than the rural area. Figure 4.2 presented the spatial distributions of PM2.5, $NO_2$, and CO concentration levels under air stagnation at their original scales and averaged by season. Darker colors indicate higher concentration levels. All three pollutants have the lowest concentration levels in the summer and are severest in the winter under air stagnant. Besides, the pollution level of $NO_2$ is worse in the central area of Beijing City than in rural areas. The reason is likely attributed

to the traffic emission on the roads of intensive transportation in the urban area.

Figure 4.3 shows the standard errors of cubic root transformed PM2.5, $NO_2$ and CO concentration levels at all grid cells (a darker color indicates a higher standard error), and the Leave-One-Location-Out RMSE at the monitoring stations (a larger dot indicates a larger RMSE). The central Beijing area has relatively lower standard errors and cross-validation RMSE because of the excellent coverage of monitoring stations. The standard errors are more prominent in spring and winter since there are fewer air stagnant hours, or in other words, more windy days in spring and winter. The large cross-validation RMSEs often appear on the map's boundary, and they are likely to be improved by enlarging the study area to include more monitoring stations.

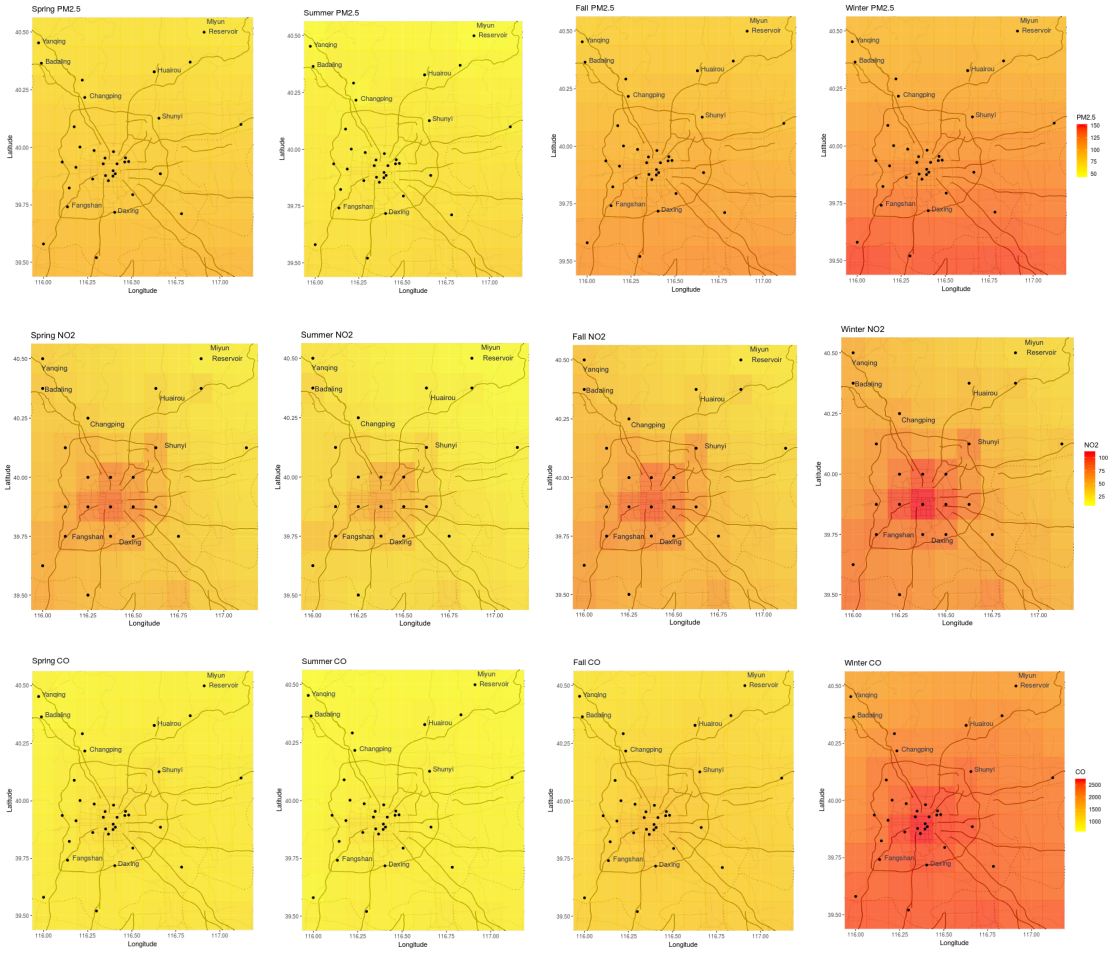



Figure 4.2: Maps of average pollutant concentrations for PM2.5 (first row), $NO_2$ (second row), CO (third row) in Beijing from spring (left) to winter (right) from Feb 1st, 2014 to May 31st, 2017. Concentrations are shown using color gradients from yellow to red, where yellow represents a lighter polluted level and red represents a more severe polluted level. Monitoring stations are located in the grids with black dots. Missing pollution concentrations were interpolated using available hourly air stagnation data from Feb 2014 to May 2017, and the hourly concentration fields were averaged over for each season. The resulting maps were rendered using the ggmap package in R[2].


Figure 4.3: Standard errors of estimated pollutant hourly concentrations for PM2.5 (first row), $NO_2$ (second row), CO (third row) in Beijing from spring (left) to winter (right) from Feb 1st, 2014 to May 31st, 2017. The standard errors at each grid are averaged over each season and shown using color gradients from white (low value) to purple (high value). Monitoring stations are located in the grids with black dots. The size of the dots represents the average leave-one-location-out RMSE. The resulting maps were rendered using the ggmap package in R[2].

Chapter 5

# Discussion

## 5.1 Conclusion

The behavior of air pollutants under air stagnant conditions has been studied as an important index for air pollution severity. [50] showed that air pollution is significantly severer during air stagnation days in the United States. [62] and [43] focused on creating the air stagnant index related to the heavily polluted period in Northern China and Beijing, China respectively. [63] further interpolated the mean air stagnant days in China. However, all the previous studies have not modeled or quantified the spatial and temporal trend of air pollutants under air stagnation status. We fill this gap by developing a spatio-temporal model for three air pollutants during the air stagnant period.

It is an initial attempt of separating the meteorological and temporal effects from the wind-driven transmission effects for the air pollution problem. When rich data under air stagnation are available (over 80,000 data points in our case), we believe that the air stagnation analysis leads to reliable and interpretable estimates of the associations between air pollution levels and auxiliary variables of interest. Our results indicate that Beijing's pollution levels decrease over the years which is consistent with China's emission control effort in recent years. The monthly trends suggest the severest air pollution is in winter which is likely due to the need for heating. Sunday is the cleanest day as more people stay at home and fewer human activities occur during that day. The hourly trend of three pollutants all show the dip around 3pm-5pm, which could be attributed to the temperatures, sun lights, together with human activities during that time.

In terms of different inference methods for fixed effects, in our model result, the fixed

effect associations' variance estimates using restricted spatio-temporal regression are usually smaller than that in OLS. It implies that the use of restricted spatio-temporal regression cannot alleviate the spatio-temporal random effects confounding in our model. It coincides with the findings and theorems of restricted spatial regressions in [1, 56].

We also compare our spatial mapping with the maps including windy days in [64]. Our maps show the same spatial distribution as [64] that the south part of Beijing has worse air quality than the north part. In addition, the spatio-temporal models uncover the relationships between the meteorological effects and air polluted levels under air stagnation. These association estimates are considered stable and reliable after eliminating the noises caused by winds. Most of our association results align with the existing studies [64, 35, 43, 15, 35], but with a much simpler model and stable data.

## 5.2 Future Work

Our results have the following limitations that are worth investigating in the future studies. The variable selection procedure is based on cross-validation error, which cannot imply any causal effects between the weather condition and the polluted levels. Another future research direction is to combine our air stagnation models with other wind-driven transmission models, which will provide estimates for both calm and windy days. The wind-driven transmission can be better estimated if we know the source of emissions in the surrounding areas or have monitoring data in larger geographical areas.

The prediction accuracy could be further improved by modeling all three pollutants jointly using a multivariate model with a more advanced model architecture (*e.g.*, linear coregionalization model [65, 66, 67, 68] or multi-task recurrent neural networks [69, 70, 71, 72, 73]). In addition, we acknowledge that there exists the change of support problem [48, 49] in our collected data, and we use a simple average to solve this problem. We can deploy more advanced approaches [74, 75, 76, 48, 49] to solve this problem more rigorously.

Appendix A

# Model Selection and Estimation Result

## A.1 Model Selection

We selected the predictors and covariance structures based on the mean absolute error (MAE) under different cross-validation strategies. Our intermediate model selection results are shown in Table A.1. The rows correspond to different pollutants, and covariance structures, and cross-validation strategies. The variable selection result column shows the set of variables that produced minimum cross-validation MAE under forward selection procedures. The variables are presented in the forward selection order.

The LOLO cross-validation selects more variables that capture the spatial variation, while the LPTO cross-validation favors variables that capture temporal variation. The MAE of LOLO cross-validation is

$$MAE_{\text{LOLO cross-validation}} = \frac{1}{\sum_{s=1}^{S} \mathcal{T}_s} \sum_{s=1}^{S} \sum_{t \in \mathcal{T}_s} |y_{st} - \hat{y}_{st}|,$$

for each removed location $s$, $\{y_{st}\}$ are the observed response from test data, and $\hat{y}_{st}$ are the predicted value of test data with training data excluding location $s$. $\mathcal{T}_s$ is the observed time points in location $s$. And the MAE is averaged over the entire spatial location $s$. The

MAE of LPTO cross-validation is

$$MAE_{\text{LPTO cross-validation}} = \frac{1}{\sum_{s=1}^{S} \mathcal{T}_s} \sum_{s=1}^{S} \sum_{p=1}^{P} \sum_{t \in \mathcal{T}_{sp}} |y_{st} - \hat{y}_{st}|,$$

for each removed location $s$ and partial time $p$, $\{y_{st}\}$ are the observed response from test data, and $\hat{y}_{st}$ are the predicted value of test data with training data excluding the partial time $p$ at location $s$. And the MAE is averaged over the entire spatial location $s$ and time $p$. Other metrics (MAE, RMSE, $R^2$) are calculated in a similar way.

In terms of covariance structures, all three metrics (MAE, RMSE, $R^2$) show a relatively small difference between the two choices of covariance structures: Gneiting [52], and product-sum [53]. The strong predictors are selected under both covariance structures, i.e., the first three selected variables for PM2.5 LPTO and all variables in PM2.5 LOLO. The most inconsistent variable selection results between two covariance structures are the LPTO cross-validation for CO. In that case, the prediction accuracy was mostly achieved by the spatial dependence and the time dependence. Different combinations of weak predictors provide similar prediction accuracy.

The covariance structures of the final models were selected firstly based on the average of LPTO MAE and LOLO MAE for each pollutant. Product-sum [53] outperforms Gneiting model [52] for all of these three pollutants.

After we fixed the covariance structure, we obtained the variables in the final models by combining the predictors selected by both LOLO and LPTO under selected covariance structures. We also considered the interactions between selected main effects, but no interaction term has been selected in the forward selection process.

Table A.1: Variable Selection Result and Performance

| Pollutant | Type of CV | Variable Selection Result | MAE | RMSE | $R^2$ |
|---|---|---|---|---|---|
| PM2.5 (Gneiting) | LPTO | Month, Latitude, Hour IRAIN, IHour | 0.170 | 0.259 | 0.944 |
| PM2.5 (Product-Sum) | LPTO | Month, Latitude, Hour | 0.167 | 0.255 | 0.946 |
| PM2.5 (Gneiting) | LOLO | Month, Latitude, Geopotential, HUMI | 0.269 | 0.385 | 0.875 |
| PM2.5 (Product-Sum) | LOLO | Month, Latitude, Geopotential, HUMI | 0.268 | 0.385 | 0.876 |
| $NO_2$ (Gneiting) | LPTO | t850, Nightlight, Month, Longitude, PRES, Geopotential | 0.192 | 0.272 | 0.903 |
| $NO_2$ (Product-Sum) | LPTO | Nightlight, Hour, Longitude, Month, Weekday, Geopotential | 0.177 | 0.257 | 0.913 |
| $NO_2$ (Gneiting) | LOLO | Month, evapor, t850, Hour, Year, Geopotential, IRAIN | 0.429 | 0.577 | 0.562 |
| $NO_2$ (Product-Sum) | LOLO | Nightlight, Latitude, Temp at 2m, Hour, t850 | 0.427 | 0.568 | 0.576 |
| CO (Gneiting) | LPTO | evapor, Year | 0.396 | 0.633 | 0.918 |
| CO (Product-Sum) | LPTO | PRES | 0.368 | 0.596 | 0.927 |
| CO (Gneiting) | LOLO | Month, Nightlight, t850, t250 | 0.755 | 1.093 | 0.753 |
| CO (Product-Sum) | LOLO | Nightlight, Month, t850, t250, Geopotential | 0.763 | 1.097 | 0.751 |

## A.2 Covariance Components Estimation Result

Table A.2: Covariance Components Estimation Result of Product-Sum

| Pollutant | m | spatial nuggets | spatial sill | temporal sill | temporal scale | range |
|---|---|---|---|---|---|---|
| PM2.5 | 0.59 | 0.020 | 0.35 | 0.58 | 9.50 | 150 |
| $NO_2$ | 0.63 | 0.093 | 0.62 | 0.23 | 3.64 | 100 |
| CO | 2.47 | 0.52 | 2.08 | 1.87 | 7.04 | 100 |

Appendix B

# Model Diagnosis

The residual plots in Figure B.1 for the three models provide a diagnosis of three models. The residual versus fitted plots is to verify the homoscedasticity assumption for error terms and detect non-linearity in our variables. The plots show that the homoscedasticity of the residuals is satisfied. Most of the trends were captured by the fitted values and there is no much non-linearity in our variables. Noted that there are some cuts in the lower half of residual versus fitted plots. The sources of these cuts came from the fact that some air pollution levels with low concentrations are frequently observed, especially for $CO$. The normal Q-Q plots of three models show that the normality assumptions of our models are satisfied.

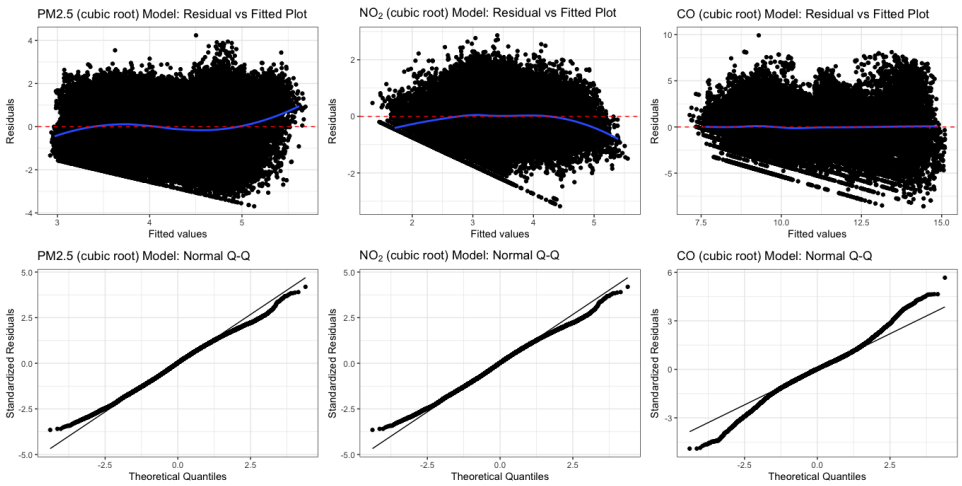



Figure B.1: Diagnostic Plot of PM2.5/$NO_2$/CO Model under Air Stagnation

# Bibliography


[1] HANKS, E. M., E. M. SCHLIEP, M. B. HOOTEN, and J. A. HOETING (2015) "Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification," *Environmetrics*, **26**(4), pp. 243–254.

[2] KAHLE, D. and H. WICKHAM (2013) "ggmap: Spatial Visualization with ggplot2," *The R journal*, **5**(1), pp. 144–161.

[3] KAMPA, M. and E. CASTANAS (2008) "Human health effects of air pollution." *Environmental pollution*, **151**(2), pp. 362–367.

[4] CHEN, R., H. KAN, B. CHEN, W. HUANG, Z. BAI, G. SONG, and G. PAN (2012) "Association of particulate air pollution with daily mortality: the China Air Pollution and Health Effects Study," *American journal of epidemiology*, **175**(11), pp. 1173–1181.

[5] LAI, H.-K., H. TSANG, and C.-M. WONG (2013) "Meta-analysis of adverse health effects due to air pollution in Chinese populations," *BMC Public Health*, **13**(1), pp. 1–12.

[6] XU, P., Y. CHEN, and X. YE (2013) "Haze, air pollution, and health in China," *The Lancet*, **382**(9910), p. 2067.

[7] SHANG, Y., Z. SUN, J. CAO, X. WANG, L. ZHONG, X. BI, H. LI, W. LIU, T. ZHU, and W. HUANG (2013) "Systematic review of Chinese studies of short-term exposure to air pollution and daily mortality," *Environment international*, **54**, pp. 100–111.

[8] FERKOL, T. and D. SCHRAUFNAGEL (2014) "The global burden of respiratory disease," *Annals of the American Thoracic Society*, **11**(3), pp. 404–406.

[9] ROHDE, R. A. and R. A. MULLER (2015) "Air pollution in China: mapping of concentrations and sources," *PloS one*, **10**(8), p. e0135749.

[10] XING, Y.-F., Y.-H. XU, M.-H. SHI, and Y.-X. LIAN (2016) "The impact of PM2. 5 on the human respiratory system," *Journal of thoracic disease*, **8**(1), p. E69.

[11] PUN, V. C., F. KAZEMIPARKOUHI, J. MANJOURIDES, and H. H. SUH (2017) "Long-term PM2. 5 exposure and respiratory, cancer, and cardiovascular mortality in older US adults," *American journal of epidemiology*, **186**(8), pp. 961–969.

[12] XU, Q., X. LI, S. WANG, C. WANG, F. HUANG, Q. GAO, L. WU, L. TAO, J. GUO, W. WANG, ET AL. (2016) "Fine particulate air pollution and hospital emergency room visits for respiratory disease in urban areas in Beijing, China, in 2013," *PLoS one*, **11**(4), p. e0153099.

[13] LI, J., H. LIU, Z. LV, R. ZHAO, F. DENG, C. WANG, A. QIN, and X. YANG (2018) "Estimation of PM 2.5 mortality burden in China with new exposure estimation and local concentration-response function." *Environmental Pollution*, **243**, pp. 1710–1718.

[14] LI, L., J. QIAN, C.-Q. OU, Y.-X. ZHOU, C. GUO, and Y. GUO (2014) "Spatial and temporal analysis of Air Pollution Index and its timescale-dependent relationship with meteorological factors in Guangzhou, China, 2001–2011," *Environmental Pollution*, **190**, pp. 75–81.

[15] HOLZWORTH, G. C. (1967) "Mixing depths, wind speeds and air pollution potential for selected locations in the United States," *Journal of applied Meteorology*, **6**(6), pp. 1039–1044.

[16] LEUNG, D. M., A. P. TAI, L. J. MICKLEY, J. M. MOCH, A. V. DONKELAAR, L. SHEN, and R. V. MARTIN (2018) "Synoptic meteorological modes of variability for fine particulate matter (PM 2.5) air quality in major metropolitan regions of China," *Atmospheric Chemistry and Physics*, **18**(9), pp. 6733–6748.

[17] HOEK, G., R. BEELEN, K. DE HOOGH, D. VIENNEAU, J. GULLIVER, P. FISCHER, and D. BRIGGS (2008) "A review of land-use regression models to assess spatial variation of outdoor air pollution," *Atmospheric environment*, **42**(33), pp. 7561–7578.

[18] HOCHADEL, M., J. HEINRICH, U. GEHRING, V. MORGENSTERN, T. KUHLBUSCH, E. LINK, H.-E. WICHMANN, and U. KRÄMER (2006) "Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information," *Atmospheric Environment*, **40**(3), pp. 542–553.

[19] ROSS, Z., M. JERRETT, K. ITO, B. TEMPALSKI, and G. D. THURSTON (2007) "A land use regression for predicting fine particulate matter concentrations in the New York City region," *Atmospheric Environment*, **41**(11), pp. 2255–2269.

[20] MOORE, D., M. JERRETT, W. MACK, and N. KÜNZLI (2007) "A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA," *Journal of Environmental Monitoring*, **9**(3), pp. 246–252.

[21] HENDERSON, S. B., B. BECKERMAN, M. JERRETT, and M. BRAUER (2007) "Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter," *Environmental science & technology*, **41**(7), pp. 2422–2428.

[22] MAYNARD, D., B. A. COULL, A. GRYPARIS, and J. SCHWARTZ (2007) "Mortality risk associated with short-term exposure to traffic particles and sulfates," *Environmental health perspectives*, **115**(5), pp. 751–755.

[23] PROCHÁZKA, A., M. KOLINOVA, J. FIALA, P. HAMPL, and K. HLAVATY (2000) "Satellite image processing and air pollution detection," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, vol. 4, IEEE, pp. 2282–2285.

[24] PENG, J., S. CHEN, H. LÜ, Y. LIU, and J. WU (2016) "Spatiotemporal patterns of remotely sensed PM2. 5 concentration in China from 1999 to 2011," *Remote Sensing of Environment*, **174**, pp. 109–121.

[25] MA, Z., X. HU, A. M. SAYER, R. LEVY, Q. ZHANG, Y. XUE, S. TONG, J. BI, L. HUANG, and Y. LIU (2016) "Satellite-based spatiotemporal trends in PM2. 5 concentrations: China, 2004–2013," *Environmental health perspectives*, **124**(2), pp. 184–192.

[26] KROTKOV, N. A., C. A. MCLINDEN, C. LI, L. N. LAMSAL, E. A. CELARIER, S. V. MARCHENKO, W. H. SWARTZ, E. J. BUCSELA, J. JOINER, B. N. DUNCAN, ET AL. (2016) "Aura OMI observations of regional SO 2 and NO 2 pollution changes from 2005 to 2015," *Atmospheric Chemistry and Physics*, **16**(7), pp. 4605–4629.

[27] VERSTRAETEN, W. W., J. L. NEU, J. E. WILLIAMS, K. W. BOWMAN, J. R. WORDEN, and K. F. BOERSMA (2015) "Rapid increases in tropospheric ozone production and export from China," *Nature geoscience*, **8**(9), pp. 690–695.

[28] SAHU, S. K. and K. V. MARDIA (2005) "A Bayesian kriged Kalman model for short-term forecasting of air pollution levels," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(1), pp. 223–244.

[29] SZPIRO, A. A., P. D. SAMPSON, L. SHEPPARD, T. LUMLEY, S. D. ADAR, and J. D. KAUFMAN (2010) "Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies," *Environmetrics*, **21**(6), pp. 606–631.

[30] AL-AWADHI, F. A. and S. A. AL-AWADHI (2006) "Spatial-temporal model for ambient air pollutants in the state of Kuwait," *Environmetrics*, **17**(7), pp. 739–752.

[31] SHADDICK, G. and J. WAKEFIELD (2002) "Modelling daily multivariate pollutant data at multiple sites," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **51**(3), pp. 351–372.

[32] ZIDEK, J., L. SUN, N. LE, and H. ÖZKAYNAK (2002) "Contending with space–time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM10 field," *Environmetrics: The official journal of the International Environmetrics Society*, **13**(5-6), pp. 595–613.

[33] HU, X., L. A. WALLER, M. Z. AL-HAMDAN, W. L. CROSSON, M. G. ESTES JR, S. M. ESTES, D. A. QUATTROCHI, J. A. SARNAT, and Y. LIU (2013) "Estimating ground-level PM2. 5 concentrations in the southeastern US using geographically weighted regression," *Environmental research*, **121**, pp. 1–10.

[34] SMITH, R. L., S. KOLENIKOV, and L. H. COX (2003) "Spatiotemporal modeling of PM2. 5 data with missing values," *Journal of Geophysical Research: Atmospheres*, **108**(D24).

[35] LIANG, X., T. ZOU, B. GUO, S. LI, H. ZHANG, S. ZHANG, H. HUANG, and S. X. CHEN (2015) "Assessing Beijing's PM2. 5 pollution: severity, weather impact, APEC and winter heating," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **471**(2182), p. 20150257.

[36] SAMPSON, P. D., M. RICHARDS, A. A. SZPIRO, S. BERGEN, L. SHEPPARD, T. V. LARSON, and J. D. KAUFMAN (2013) "A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2. 5 concentrations in epidemiology," *Atmospheric environment*, **75**, pp. 383–392.

[37] CAMELETTI, M., R. IGNACCOLO, and S. BANDE (2011) "Comparing spatiotemporal models for particulate matter in Piemonte," *Environmetrics*, **22**(8), pp. 985–996.

[38] DE IACO, S., D. MYERS, and D. POSA (2002) "Space–time variograms and a functional form for total air pollution measurements," *Computational Statistics & Data Analysis*, **41**(2), pp. 311–328.

[39] CHAN, T.-C., M.-L. CHEN, I.-F. LIN, C.-H. LEE, P.-H. CHIANG, D.-W. WANG, and J.-H. CHUANG (2009) "Spatiotemporal analysis of air pollution and asthma patient visits in Taipei, Taiwan," *International journal of health geographics*, **8**(1), pp. 1–10.

[40] DAWSON, J., P. ADAMS, and S. PANDIS (2007) "Sensitivity of PM 2.5 to climate in the Eastern US: a modeling case study," *Atmospheric chemistry and physics*, **7**(16), pp. 4295–4309.

[41] KUKKONEN, J., M. POHJOLA, R. S. SOKHI, L. LUHANA, N. KITWIROON, L. FRAGKOU, M. RANTAMÄKI, E. BERGE, V. ØDEGAARD, L. H. SLØRDAL, ET AL. (2005) "Analysis and evaluation of selected local-scale PM10 air pollution episodes in four European cities: Helsinki, London, Milan and Oslo," *Atmospheric environment*, **39**(15), pp. 2759–2773.

[42] VARDOULAKIS, S. and P. KASSOMENOS (2008) "Sources and factors affecting PM10 levels in two European cities: Implications for local air quality management," *Atmospheric Environment*, **42**(17), pp. 3949–3963.

[43] CAI, W., K. LI, H. LIAO, H. WANG, and L. WU (2017) "Weather conditions conducive to Beijing severe haze more frequent under climate change," *Nature Climate Change*, **7**(4), pp. 257–262.

[44] PENDERGRASS, D., L. SHEN, D. JACOB, and L. MICKLEY (2019) "Predicting the impact of climate change on severe wintertime particulate pollution events in Beijing using extreme value theory," *Geophysical Research Letters*, **46**(3), pp. 1824–1830.

[45] ZHANG, S., B. GUO, A. DONG, J. HE, Z. XU, and S. X. CHEN (2017) "Cautionary tales on air-quality improvement in Beijing," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **473**(2205), p. 20170457.

[46] ZHUO, L., T. ICHINOSE, J. ZHENG, J. CHEN, P. SHI, and X. LI (2009) "Modelling the population density of China at the pixel level based on DMSP/OLS non-radiance-calibrated night-time light images," *International Journal of Remote Sensing*, **30**(4), pp. 1003–1018.

[47] TAN, M., X. LI, S. LI, L. XIN, X. WANG, Q. LI, W. LI, Y. LI, and W. XIANG (2018) "Modeling population density based on nighttime light images and land use data in China," *Applied Geography*, **90**, pp. 239–247.

[48] GOTWAY, C. A. and L. J. YOUNG (2002) "Combining incompatible spatial data," *Journal of the American Statistical Association*, **97**(458), pp. 632–648.

[49] GELFAND, A. E., L. ZHU, and B. P. CARLIN (2001) "On the change of support problem for spatio-temporal data," *Biostatistics*, **2**(1), pp. 31–45.

[50] WANG, J. X. and J. K. ANGELL (1999) "Air stagnation climatology for the United States," *NOAA/Air Resource Laboratory ATLAS*, **1**.

[51] CRESSIE, N. and C. K. WIKLE (2015) *Statistics for spatio-temporal data*, John Wiley & Sons.

[52] GNEITING, T. (2002) "Nonseparable, stationary covariance functions for space–time data," *Journal of the American Statistical Association*, **97**(458), pp. 590–600.

[53] DE IACO, S., D. E. MYERS, and D. POSA (2001) "Space–time analysis using a general product–sum model," *Statistics & Probability Letters*, **52**(1), pp. 21–28.

[54] DE IACO, S. (2010) "Space–time correlation analysis: a comparative study," *Journal of Applied Statistics*, **37**(6), pp. 1027–1041.

[55] DE IACO, S., D. POSA, and D. MYERS (2013) "Characteristics of some classes of space–time covariance functions," *Journal of Statistical Planning and Inference*, **143**(11), pp. 2002–2015.

[56] KHAN, K. and C. A. CALDER (2020) "Restricted Spatial Regression Methods: Implications for Inference," *Journal of the American Statistical Association*, pp. 1–13.

[57] MATHERON, G. (1963) "Principles of geostatistics," *Economic geology*, **58**(8), pp. 1246–1266.

[58] HODGES, J. S. and B. J. REICH (2010) "Adding spatially-correlated errors can mess up the fixed effect you love," *The American Statistician*, **64**(4), pp. 325–334.

[59] MEYER, H., C. REUDENBACH, T. HENGL, M. KATURJI, and T. NAUSS (2018) "Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation," *Environmental Modelling & Software*, **101**, pp. 1–9.

[60] SCHWARZ, G. ET AL. (1978) "Estimating the dimension of a model," *Annals of statistics*, **6**(2), pp. 461–464.

[61] DICKEY, D. A. and W. A. FULLER (1981) "Likelihood ratio statistics for autoregressive time series with a unit root," *Econometrica: journal of the Econometric Society*, pp. 1057–1072.

[62] FENG, J., J. QUAN, H. LIAO, Y. LI, and X. ZHAO (2018) "An air stagnation index to qualify extreme haze events in northern China," *Journal of the Atmospheric Sciences*, **75**(10), pp. 3489–3505.

[63] HUANG, Q., X. CAI, Y. SONG, and T. ZHU (2017) "Air stagnation in China (1985–2014): climatological mean features and trends," *Atmospheric Chemistry and Physics*, **17**(12), pp. 7793–7805.

[64] QI, J., B. ZHENG, M. LI, F. YU, C. CHEN, F. LIU, X. ZHOU, J. YUAN, Q. ZHANG, and K. HE (2017) "A high-resolution air pollutants emission inventory in 2013 for the Beijing-Tianjin-Hebei region, China," *Atmospheric environment*, **170**, pp. 156–168.

[65] GOULARD, M. and M. VOLTZ (1992) "Linear coregionalization model: tools for estimation and choice of cross-variogram matrix," *Mathematical Geology*, **24**(3), pp. 269–286.

[66] DE IACO, S., D. MYERS, and D. POSA (2003) "The linear coregionalization model and the product–sum space–time variogram," *Mathematical Geology*, **35**(1), pp. 25–38.

[67] DE IACO, S., D. MYERS, M. PALMA, and D. POSA (2013) "Using simultaneous diagonalization to identify a space–time linear coregionalization model," *Mathematical Geosciences*, **45**(1), pp. 69–86.

[68] SCHMIDT, A. M. and A. E. GELFAND (2003) "A Bayesian coregionalization approach for multivariate pollutant data," *Journal of Geophysical Research: Atmospheres*, **108**(D24).

[69] WANG, B., Z. YAN, J. LU, G. ZHANG, and T. LI (2018) "Deep multi-task learning for air quality prediction," in *International Conference on Neural Information Processing*, Springer, pp. 93–103.

[70] ZHANG, Q., S. WU, X. WANG, B. SUN, and H. LIU (2020) "A PM2. 5 concentration prediction model based on multi-task deep learning for intensive air quality monitoring stations," *Journal of Cleaner Production*, **275**, p. 122722.

[71] GUO, W., D. MU, X. XING, M. DU, and D. SONG (2019) "{DEEPVSA}: Facilitating Value-set Analysis with Deep Learning for Postmortem Program Analysis," in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 1787–1804.

[72] YUAN, Z., J. LIU, Y. LIU, Q. ZHANG, and R. W. LIU (2020) "A multi-task analysis and modelling paradigm using LSTM for multi-source monitoring data of inland vessels," *Ocean Engineering*, **213**, p. 107604.

[73] BAI, L., L. YAO, S. S. KANHERE, Z. YANG, J. CHU, and X. WANG (2019) "Passenger demand forecasting with multi-task convolutional recurrent neural networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

[74] STEIN, M. L. (2012) *Interpolation of spatial data: some theory for kriging*, Springer Science & Business Media.

[75] KYRIAKIDIS, P. C. (2004) "A geostatistical framework for area-to-point spatial interpolation," *Geographical Analysis*, **36**(3), pp. 259–289.

[76] CRESSIE, N. A. (1996) "Change of support and the modifiable areal unit problem," .