Spatio-temporal modeling of $PM_{2.5}$ concentrations with missing data problem: a case study in Beijing, China

Qiang Pu and EunHye Enki Yoo*

04/06/2019

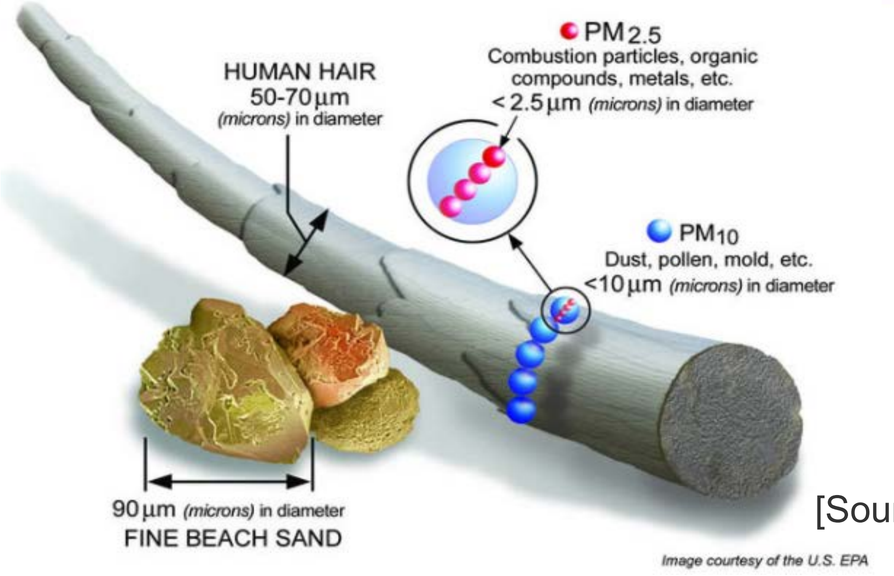
Department of Geography,

University at Buffalo, SUNY, USA

University at Buffalo The State University of New York

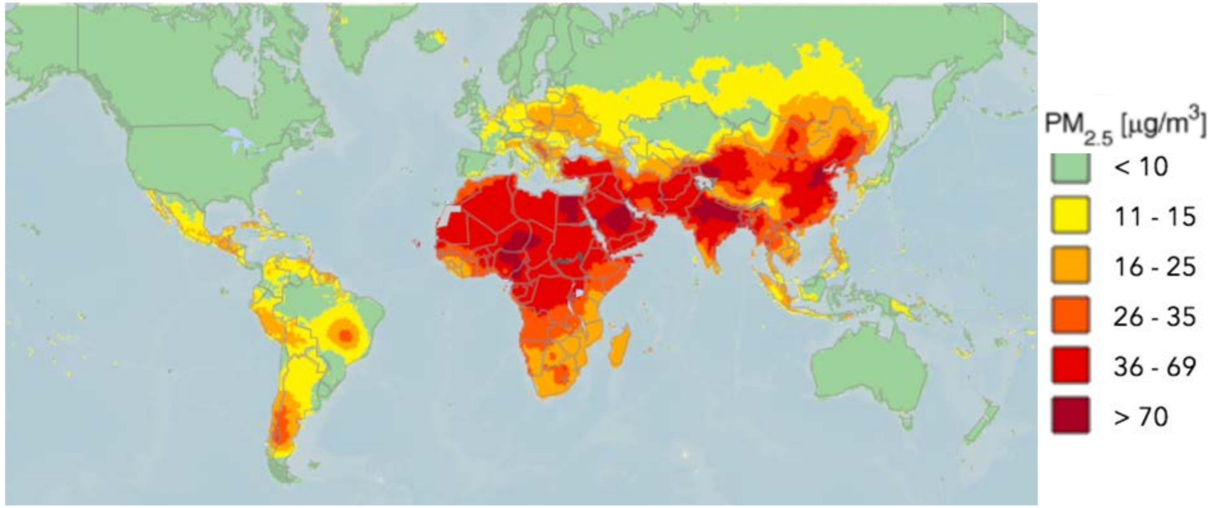
- ❑ INTRODUCTION
- ❑ DATA AND METHODS
- ❑ RESULTS
- ❑ DISCUSSIONS

$PM_{2.5}$ Pollution & Health Impact



[Source: EPA]

Global annual mean $PM_{2.5}$ for 2016 [Source: WHO, 2016]



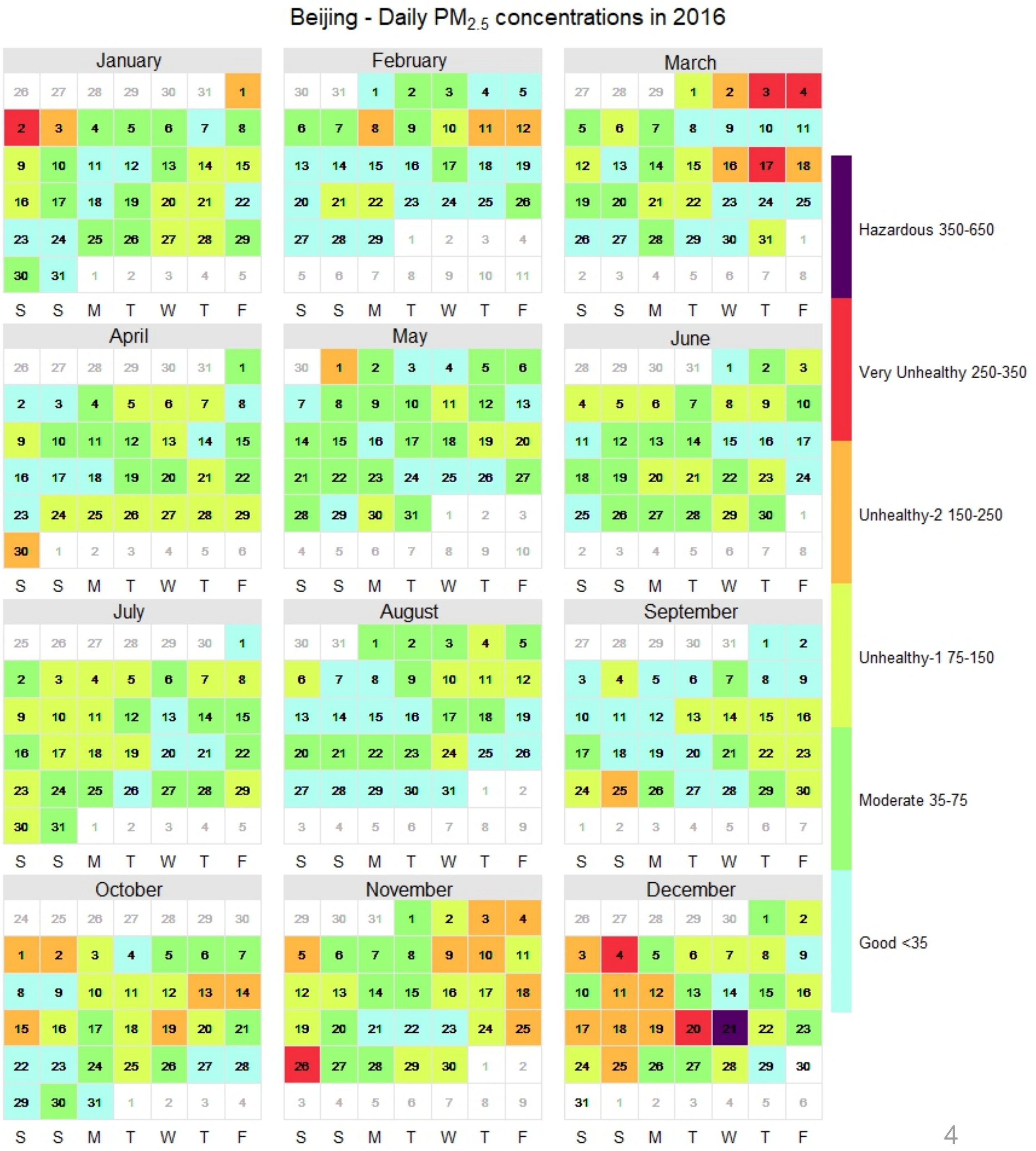
- $PM_{2.5}$ is associated elevated risk of mortality and cardiopulmonary diseases.
- In 2016, about 92% of world's population breathes unsafe air according to WHO.

$PM_{2.5}$ Pollution in China

China is one of the most populated and polluted counties.

Spatially and temporally varying $PM_{2.5}$ distribution

[Source: CCTV]

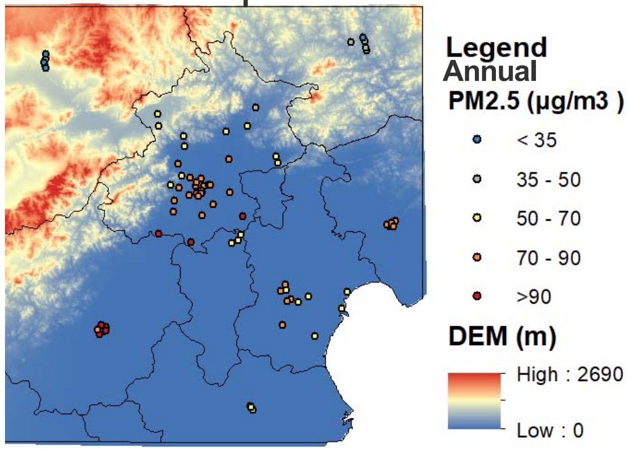


Air Pollution Monitoring

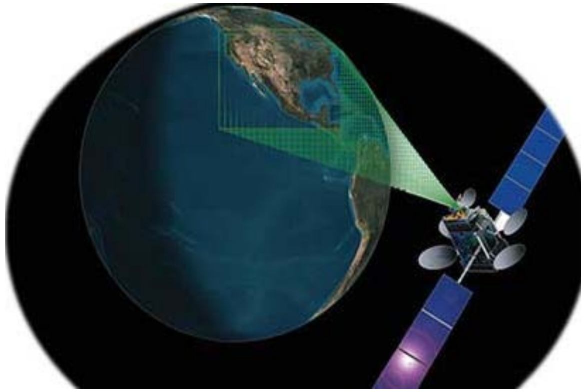
- Ground monitored data are insufficient for predicting spatially and temporally varying $PM_{2.5}$ concentrations at fine resolution.
- Satellite aerosol optical depth (AOD) with broad spatial coverage can be used to supplement sparse monitoring data.

(Gupta et al. 2006, Hoff and Christopher 2009, Van Donkelaar et al. 2010)

Ground air pollution monitoring



Satellite observations



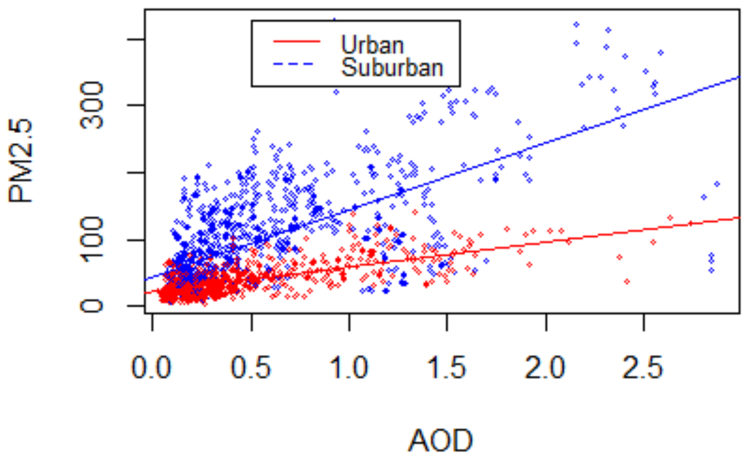
Source: NCAR

Major Challenges

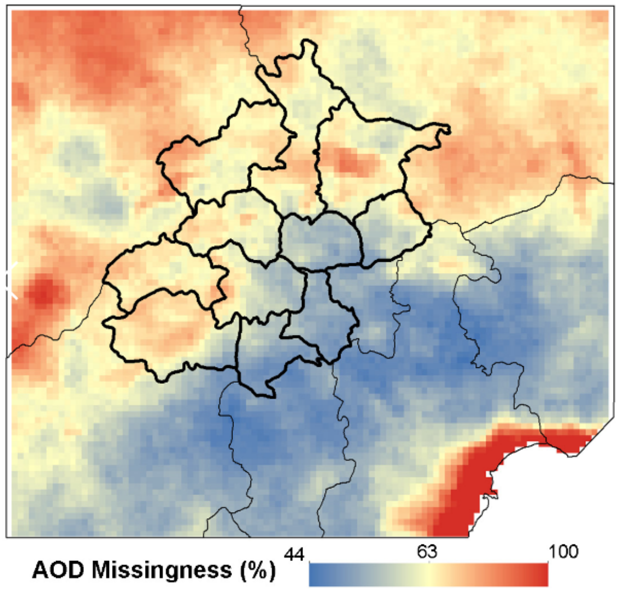
- **in predicting fine scale spatio-temporal $PM_{2.5}$ concentrations using satellite AOD**

  1. Spatial and temporal **heterogeneity** in the associations between $PM_{2.5}$ and AOD;
  2. **Missing AOD** issue may lead to biased $PM_{2.5}$-AOD relationships and incomplete $PM_{2.5}$ prediction;

AOD-$PM_{2.5}$ relationship



Missing AOD Data in 2016 (%)



Related Studies

❑ Modeling associations between $PM_{2.5}$ and AOD

- Account for spatially and temporally variable relationships using: Linear mixed effect models (LME) (Lee et al. 2011, Kloog et al. 2012)

❑ Adjusting sampling bias from missing data in AOD

- Using inverse probability weighting (IPW) (Wooldridge 2007).

❑ Obtaining full spatial coverage for $PM_{2.5}$ concentration

- Employing stochastic partial differential equations under integrated nested Laplace approximation (INLA-SPDE) (Cameletti *et al*., 2013)

Objective

We develop a multi-stage spatio-temporal $PM_{2.5}$ prediction model to estimate $PM_{2.5}$ values at fine resolutions in space and time, while accounting for

- **the spatially and temporally varying associations between measured $PM_{2.5}$ and satellite AOD**
- **the missingness of satellite-derived AOD**

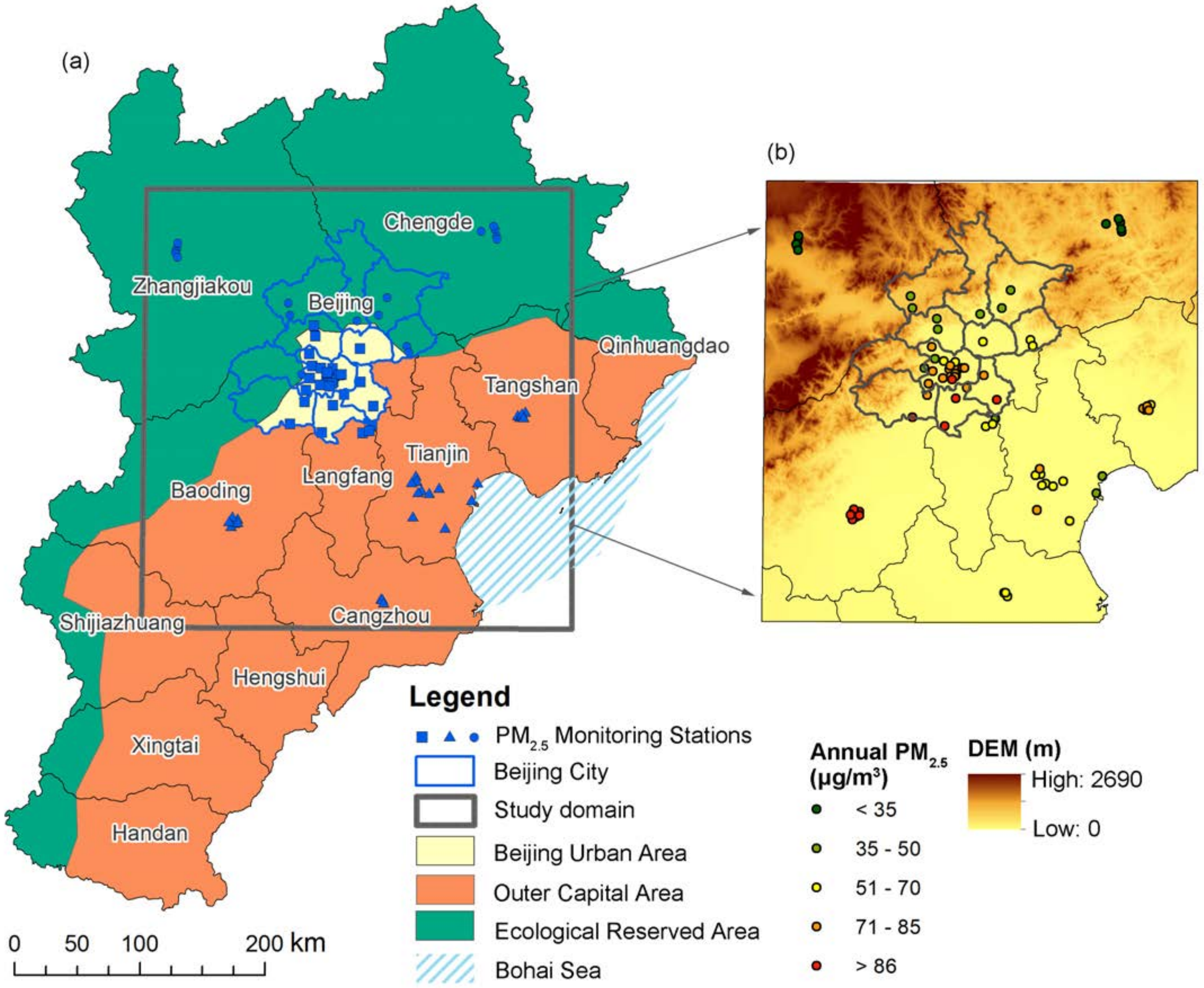
- ❑ INTRODUCTION
- ❑ **DATA AND METHODS**
- ❑ RESULTS
- ❑ DISCUSSIONS

Study Area

Beijing City
35 air pollution monitoring stations

Surrounding Areas
39 stations from 7 cities: Tianjin, Tangshan, Baoding, Chengde, Langfang, Cangzhou, Zhangjiakou;

74 monitoring stations in total in 2016



In accordance with governmental urban planning: 2016-2035

Details of Data Used

University at Buffalo The State University of New York

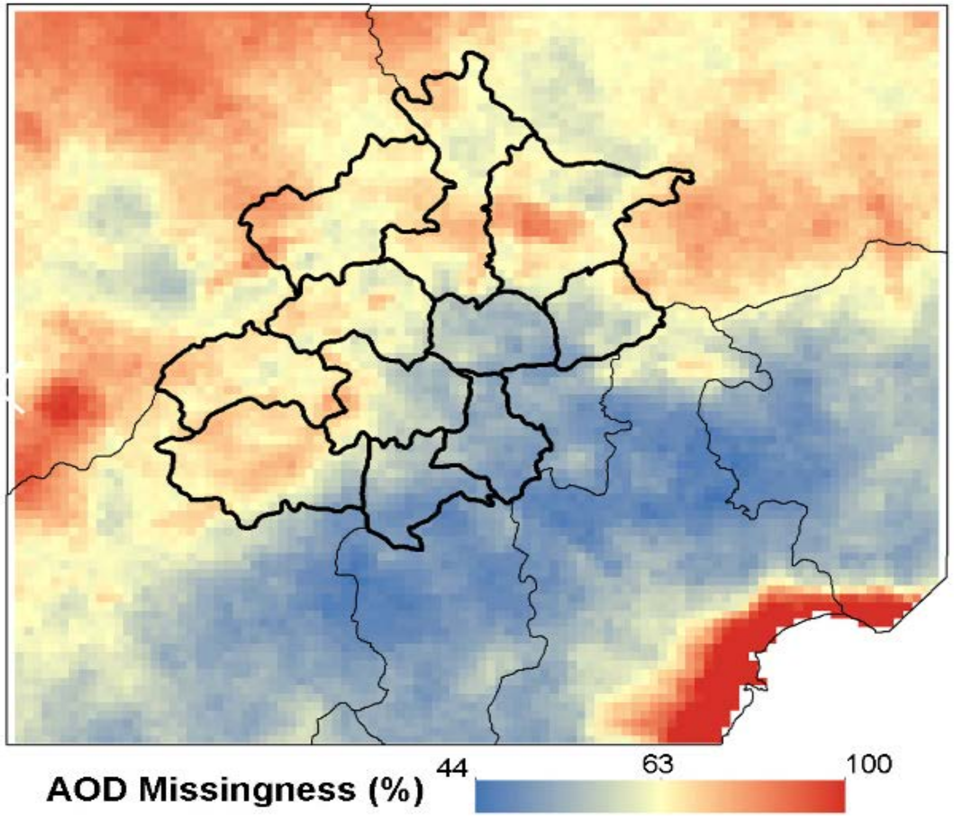
Category	Variables	Unit	Positive/Negative Effect	Granularity		Description
				Space	Time	
Air pollutant observations	$PM_{2.5}$	$\mu g/m^3$	N/A	Point	Daily	Hourly monitoring from ground stations
Satellite observations	AOD	N/A	possitive	3 km*3 km	Daily	MYD/MOD04_3K (DT and DB product) from MODIS/Aqua & Terra
	NDVI	N/A	Negative	1 km*1 km	Monthly	MOD13A3 from MODIS/Terra
Meteorological conditions	Boundary Layer Height	meter	Negative	14 km*14 km	Daily	ECMWF ERA-Interim global reanalysis dataset that has 8 time slots per day (3 hour interval from 0:00-12:00)
	Temperature	°C	Negative	Point	Daily	Daily observations from meteorological stations
	Wind Speed	m/s	Negative	Point	Daily	
	Relative Humidity	%	Negative	Point	Daily	
	Precipitation	mm	Negative	Point	Daily	
	Ground Air Pressure	hPa	Positive	Point	Daily	
Land use type	Major Road Length	meter	Positive	Line	N/A	Include expressway, national, provincial, county roads and major urban roads
	Build-up Land	%	Positive	30 m*30 m	N/A	Land use type classification from 30-m resolution Landsat image
	Farm Land	%	Positive	30 m*30 m	N/A	
	Forest Land	%	Negative	30 m*30 m	N/A	
	Grass Land	%	Negative	30 m*30 m	N/A	
	Water Body	%	Negative	30 m*30 m	N/A	
	Bare Land	%	Positive	30 m*30 m	N/A	
	Elevation	meter	Negative	90 m*90 m	N/A	
	Total Population	person/km^2	Positive	1 km*1 km	N/A	

*AOD: Merged product based on Dark Target and Deep Blue algorithms both Terra and Aqua

Satellite AOD

Merged AOD with a two-step data merging scheme

Percentage of missing AOD Data



- Merged Dark Target and Deep Blue AOD from Terra and Aqua using Simplified Merge Scheme (SMS) (Bilal et al., 2017);
- A domain wide linear regression model against merged-AOD from Aqua and Terra to combine AODs from both satellites.

* Average missing rate is 61.40% for the study area.

Statistical Prediction Model

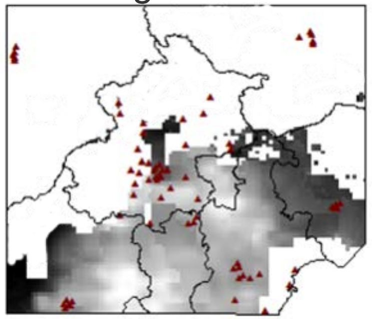
University at Buffalo The State University of New York

First Stage:
Use Inverse probability weighting to alleviate sampling bias caused by the missing AOD values;

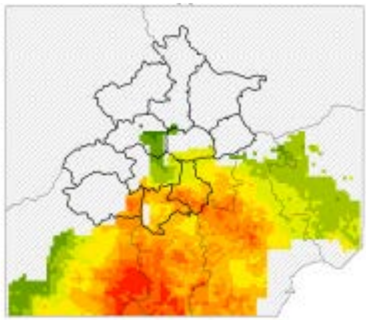
Second Stage: (adopt weights from stage-1)
Build Linear Mixed Effect Model using $PM_{2.5}$-AOD collocation pairs and predict $PM_{2.5}$ levels over spatial grids with AOD values;

Third Stage: (gap-filling, based on stage-1&2)
Utilize INLA-SPDE to predict $PM_{2.5}$ levels over areas without AOD retrievals.

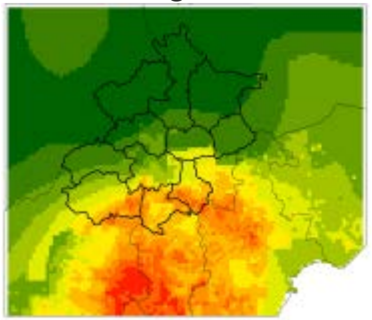
Fist Stage



Second Stage



Third Stage



First Stage: Inverse probability weighting

$$\begin{aligned} \ln \frac{p(i,j)}{1-p(i,j)} &= \alpha_0 + \sum_{k=1}^{6} \alpha_k W_k(i,j) \\ IPW(i,j) &= \frac{1}{p(i,j)} \end{aligned} \quad (1)$$

$\{W_k(i,j), k = 1, ..., 6\}$ denotes six predictors at grid cell *i* and day *j*;

(elevation, BLH, temperature, air pressure, forest cover, subregion class)

Second Stage: Linear Mixed Effect (LME) model

$$Y(i,j) = \tilde{\beta}_0(r,j) + \tilde{\beta}_1(r,j)A(i,j) + \sum_{m=2}^{8} \beta_m X_m(i,j) + \sum_{n=9}^{17} \beta_n Z_n(i) + \epsilon(i,j) \qquad (2)$$

$Y(i,j)$ is observed $PM_{2.5}$ on station *i* and day *j;*

$A(i,j)$ denotes DTB_3K AOD value at grid cell *i* and day *j*;

$X_m(i,j)$ and $Z_n(i,j)$ are spatio-temporal and spatial predictors, respectively;

$\tilde{\beta}_0(r,j)$ and $\tilde{\beta}_1(r,j)$ are intercept and slope that assumed to be region(r)- and day(j)- specific;

Second level linear model:

$$\begin{aligned} \tilde{\beta}_0(r,j) &= \beta_0 + \beta_0(r) + \beta_0(j) \\ \tilde{\beta}_1(r,j) &= \beta_1 + \beta_1(r) + \beta_1(j) \end{aligned}$$

[Pu and Yoo (2019), under revision]

Third Stage: INLA-SPDE

University at Buffalo The State University of New York

$$y(i,j) = c_0 + c_1\tilde{y}(i,j) + \xi(i,j) + \upsilon(i,j) \qquad (3)$$

$y(i,j)$ denotes both observed and predicted $PM_{2.5}$ from previous stages at cell *i* on day *j*.

$\tilde{y}(i,j)$ is the spatial average (105 km buffer) of $PM_{2.5}$ values from either ground observation or the LME predictions surrounding grid cell *i* on day *j*;

$\xi(i,j) = a\xi(i,j-1) + \omega(i,j)$ is a spatio-temporal process (first order autoregressive in time):

$\omega(i,j)$ captures spatial autocorrelation and is temporally independent, and it follows a Matérn spatial covariance function.

Model Validation

University at Buffalo The State University of New York

10-Fold Cross Validation (LME model):

R^2, RMSE, and MAE

Simple Spatial Validation (INLA-SPDE):

1. For each day, randomly select 20% of collocated grid cells with ground monitors for validation purpose;
2. Build INLA-SPDE model for each day and predict at $PM_{2.5}$ concentrations validation stations;
3. Calculate R^2, RMSE, and MAE.

- INTRODUCTION
- DATA AND METHODS
- **RESULTS**
- DISCUSSIONS

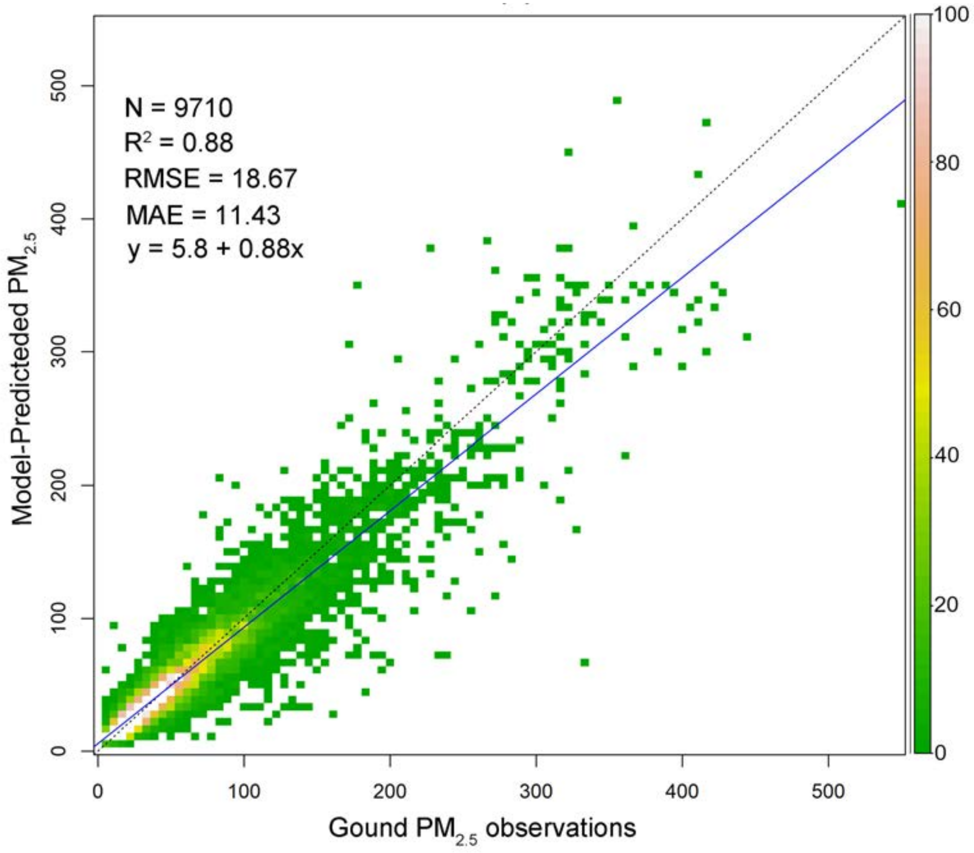
LME: On AOD-retrieved cases

University at Buffalo The State University of New York

Summary of the fixed effects for LME model

Fixed Effect	β (2.5%, 97.5%)	t value	p value	Correlation with $PM_{2.5} (P < 0.01)$	VIF
Interept	3.51 (3.12, 3.90)	17.57	< 0.001	-	-
AOD	1.23 (0.86, 1.59)	6.58	< 0.001	0.59	1.42
BLH	-0.24 (-0.27, -0.21)	-15.95	< 0.001	-0.38	2.43
temperature	0.03 (-0.02, 0.07)	15.83	< 0.001	-0.10	2.36
humidity	0.16 (0.08, 0.24)	3.89	< 0.001	0.32	2.13
wind speed	-0.42 (-0.48, -0.36)	-14.44	< 0.001	-0.22	1.66
forest	-0.07 (-0.08, -0.06)	-16.15	< 0.001	-0.23	1.71
build-up	0.08 (-0.09, -0.06)	-11.80	< 0.001	-0.08	1.59

Cross Validation:



Role of IPW:

- Mean $PM_{2.5}$ predictions with IPW = 60.43 μg/m^3 versus without IPW = 57.96 μg/m^3);
- LME model with IPW reduced the CV-RMSE by 1.75 μg/m^3.

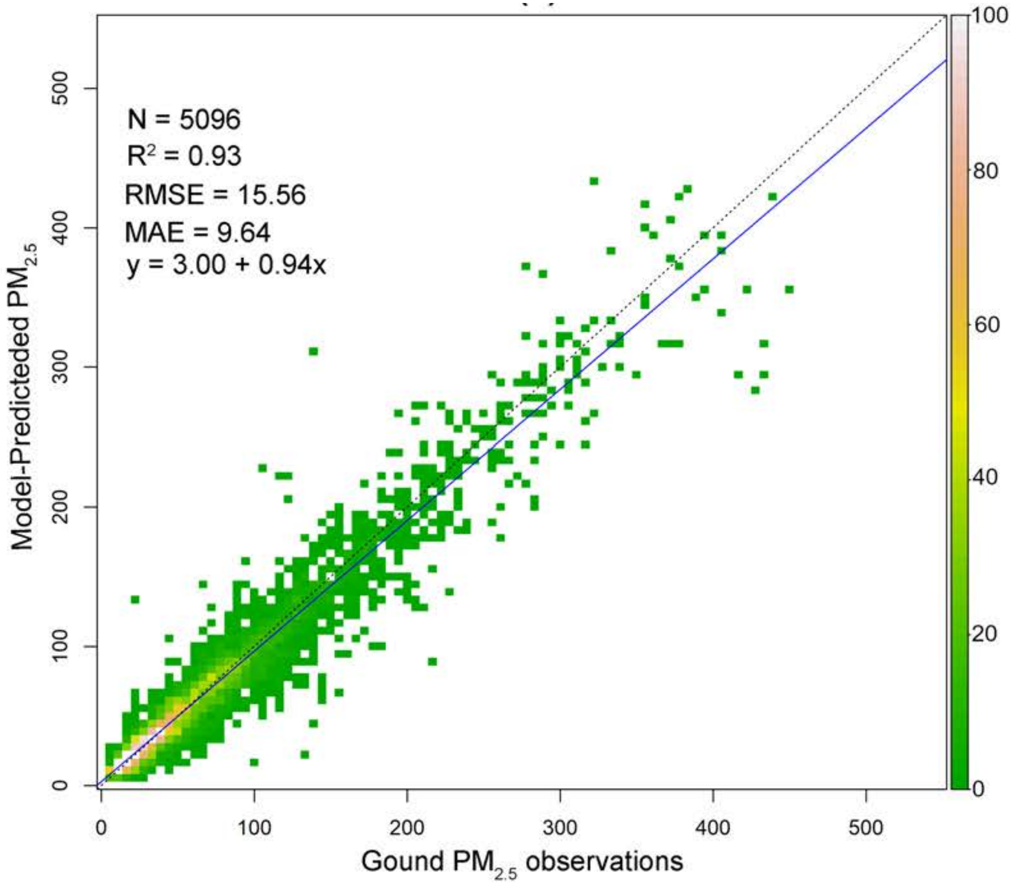
INLA-SPDE: On AOD-missing cases (Gap-filling)

Parameter estimates of INLA-SPDE model

Model Parameter	Mean	SD	Quantiles		
			2.5%	50%	97.5%
Interept	3.63	0.83	1.96	3.63	5.28
Mean $PM_{2.5}$	0.66	0.04	0.59	0.66	0.73
σ_ϵ^2	-0.02	0.00	0.01	0.02	0.03
σ_ω^2	2.34	0.52	1.42	2.34	3.38
a	0.91	0.02	0.87	0.91	0.93
κ	314.71	21.98	263.95	302.15	346.79

- Presence of substantial temporal and spatial autocorrelation of $PM_{2.5}$ concentrations;
- INLA-SPDE was capable to accurately capture the complex spatio-temporal dynamics of $PM_{2.5}$.

Out-of-sample Validation



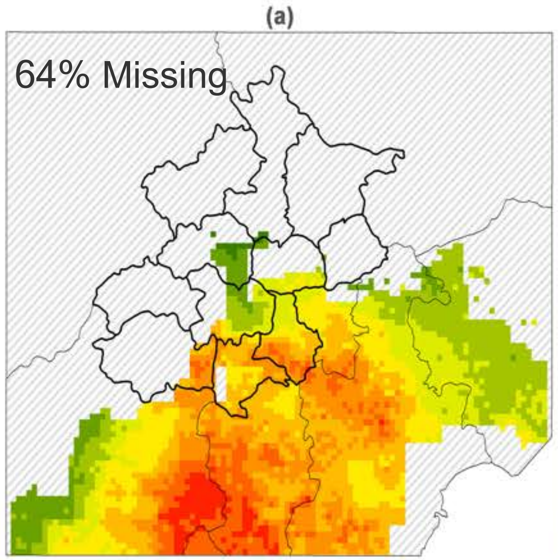
PM2.5 Estimates: (LME + INLA-SPDE)

University at Buffalo The State University of New York

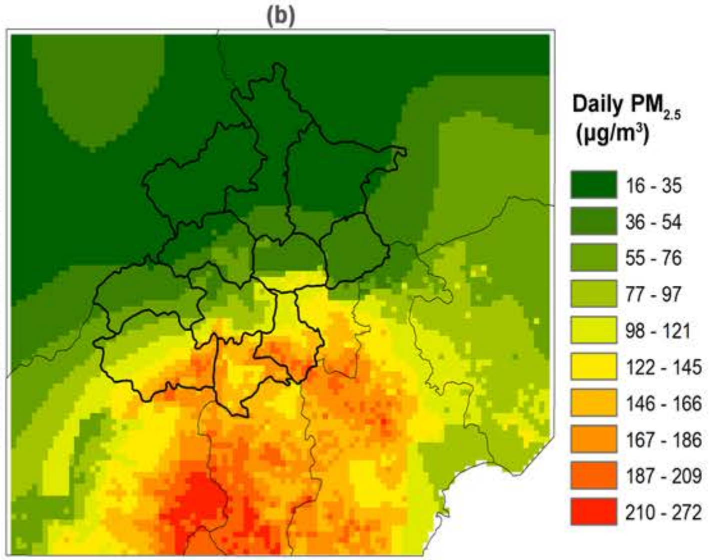
Daily prediction: (January 14th, 2016)

- Heavy pollution levels in the southern areas;
- Higher prediction uncertainty (> 80 μg/m^3) over areas farther away from monitoring stations;

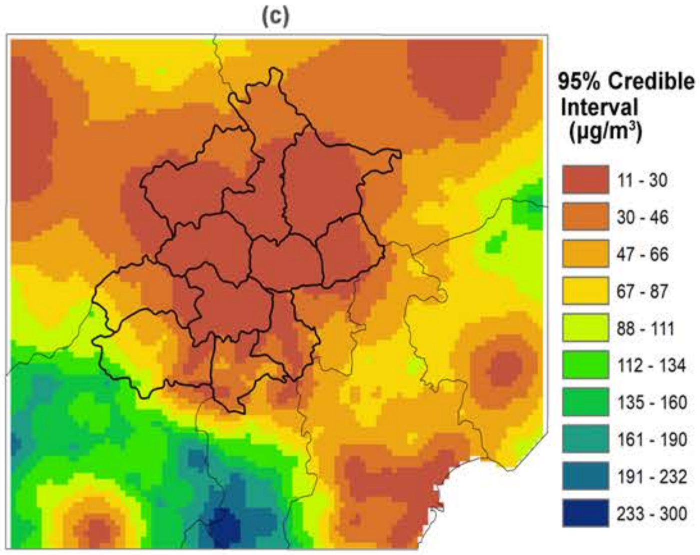
LME predictions



Gap-filling with INLA-SPDE



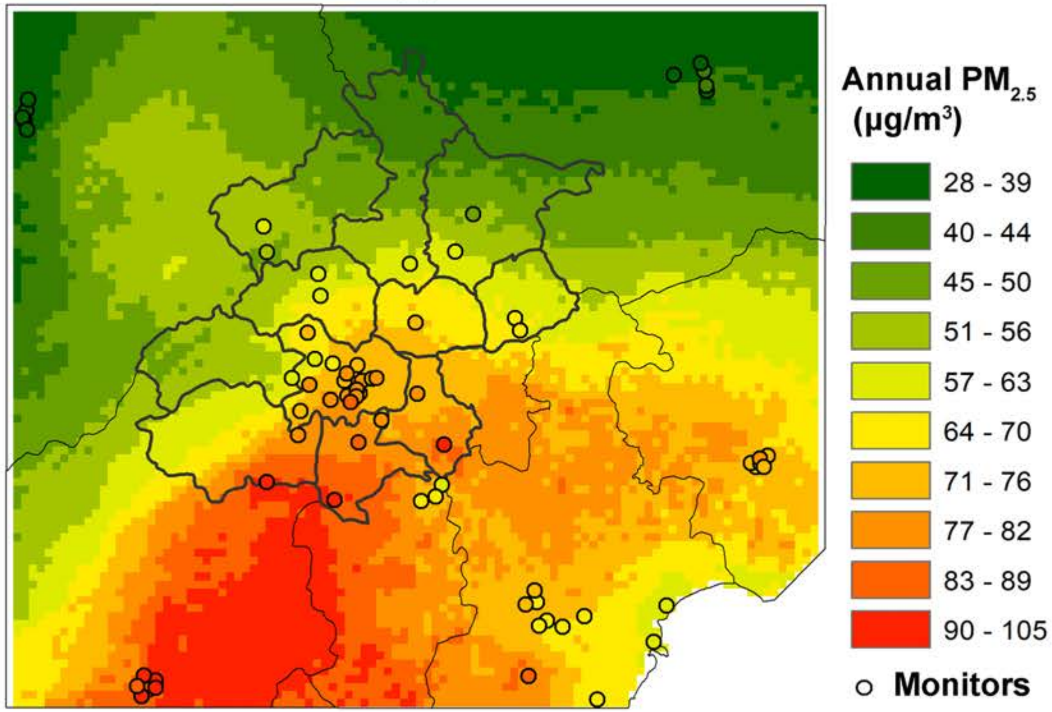
Prediction uncertainty (95% CI)



PM2.5 Estimates: (LME + INLA-SPDE)

University at Buffalo The State University of New York

Annual average:



- In the range of 28.63 to 104.30 $\mu g/m^3$ with the mean of 61.04 $\mu g/m^3$;
- Most of the study area (about 99 %) exceeded the annual Level-2 standard (35 $\mu g/m^3$) according to the Chinese National Ambient Air Quality Standard.

- ❑ INTRODUCTION
- ❑ DATA AND METHODS
- ❑ RESULTS
- ❑ **DISCUSSIONS**

Summary and Discussion

University at Buffalo The State University of New York

- The day- and region-specific LME model captures the spatially and temporally varying relationships between ground measured $PM_{2.5}$ and satellite AOD;

- IPW is a simple and effective method to adjust uneven sampling problems caused by missing data;

- INLA-SPDE effectively captures complex space-time dynamics of $PM_{2.5}$ while offers a computationally efficient support for model inference;

- The extensive daily $PM_{2.5}$ estimates with quantified uncertainty can be used to improve our understanding of the regional pollution processes.

Limitations

University at Buffalo The State University of New York

- Change-of-support problems in data aggregation process;
- Additional prediction uncertainty by using the multi-stage model;
- The data-intensive method has limited applicability.

Reference:

Bilal, M., Nichol, J. E., & Wang, L. (2017). New customized methods for improvement of the MODIS C6 Dark Target and Deep Blue merged aerosol product. Remote Sensing of Environment, *197*, 115-124.

Cameletti, M., Lindgren, F., Simpson, D., & Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. Asta-Advances in Statistical Analysis, 97(2), 109-131.

Gupta, P., Christopher, S. A., Wang, J., Gehrig, R., Lee, Y., & Kumar, N. (2006). Satellite remote sensing of particulate matter and air quality assessment over global cities. Atmospheric Environment, 40(30), 5880-5892.

Hoff, R. M., & Christopher, S. A. (2012). Remote Sensing of Particulate Pollution from Space: Have We Reached the Promised Land? Journal of the Air & Waste Management Association, 59(6), 645-675.

Kloog, I., Nordio, F., Coull, B. A., & Schwartz, J. (2012). Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2.5 exposures in the Mid-Atlantic states. Environ Sci Technol, 46(21), 11913-11921.

Lee, H. J., Liu, Y., Coull, B. A., Schwartz, J., & Koutrakis, P. (2011). A novel calibration approach of MODIS AOD data to predict PM2.5 concentrations. Atmospheric Chemistry and Physics Discussions, 11(3), 9769-9795.

van Donkelaar, A., Martin, R. V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., & Villeneuve, P. J. (2010). Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environ Health Perspect, 118(6), 847-855.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. Journal of Econometrics, 141(2), 1281-1301.

THANK YOU!

Q&A