# Real-Time Estimation of Population Exposure to PM2.5 Using Mobile- and Station-Based Big Data

Latthawat Pakornvanitcha, MD.

Data Science in Healthcare and Clinical Informatics

# Outline

- Proxy for personal exposure
- Overview
- Methodology
- Results
- Discusstion

# Proxy for personal exposure

- Outdoor and personal exposure:
  - Exposure misclassification
- Residential-based air pollution underestimated exposure by around 13% and up to 61% when the commute distance exceeded 30 km.

Lu Y. Beyond air pollution at home: Assessment of personal exposure to PM2.5 using activity-based travel demand model and low-cost air sensor network data. Environmental Research. 2021;201:111549.

# Real-Time Estimation of Population Exposure to PM2.5 Using Mobile- and Station-Based Big Data

# Real-Time Estimation of Population Exposure to PM2.5 Using Mobile- and Station-Based Big Data

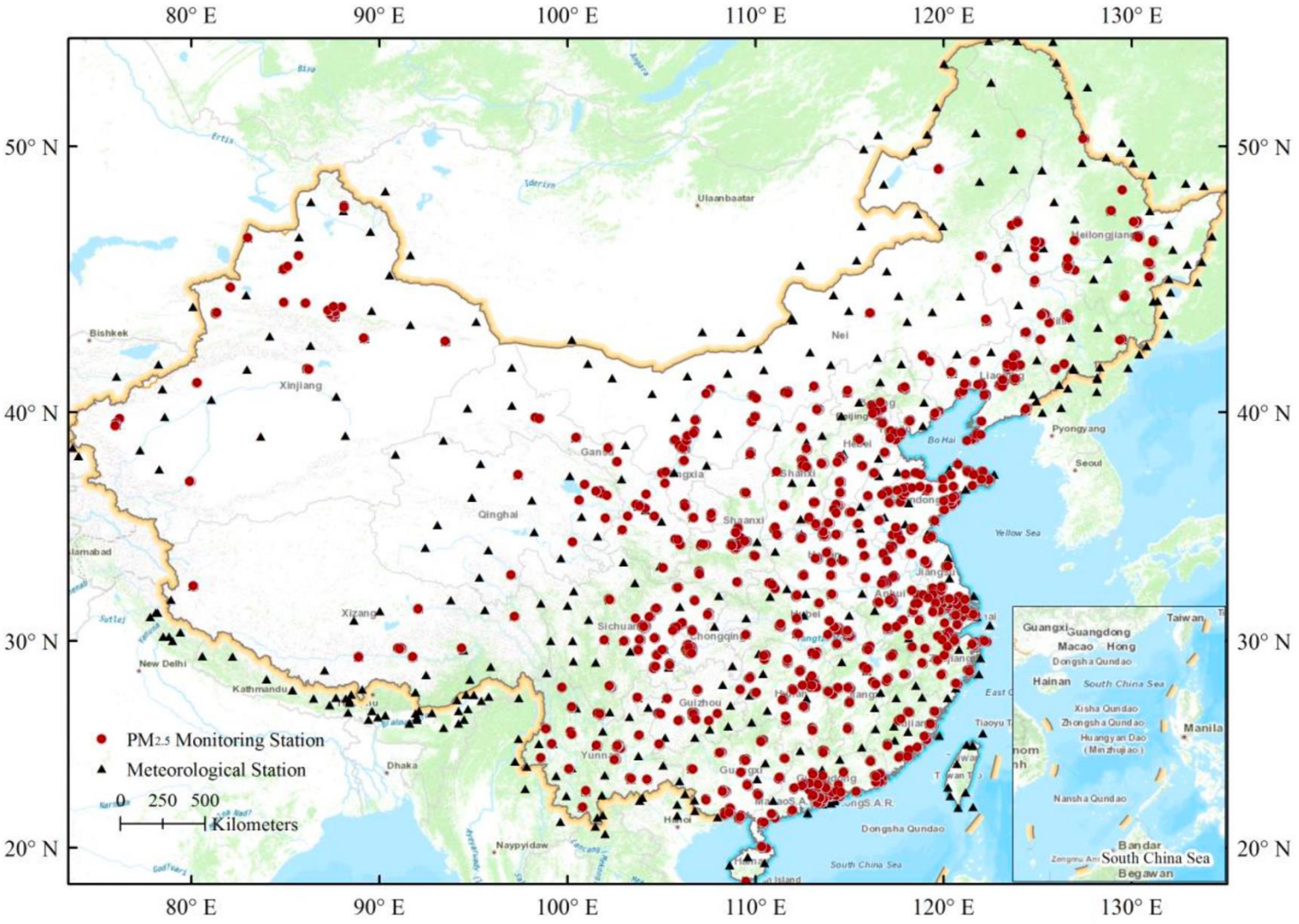
- Using the mobile-phone locating requests that are near-time data could gather data in more detail, such as commute, morning or evening peaks, to improve on the residential-based approach.
- Reduce exposure misclassification by only using the survey to update census data.

# Real-Time Estimation of Population Exposure to PM2.5 Using Mobile- and Station-Based Big Data

## Objective

- To develop a framework to estimate population exposure to $PM_{2.5}$ in near-real-time by:
  - Mapping dynamic population via mobile-phone locating requests.
  - Interpolating (Ordinary Kriging) station-based $PM_{2.5}$ data to a spatial grid.
  - Calculating exposure metrics every 3 hours.

# Ground-Station PM2.5 and Meteorological* Measurements




***Meteorological variables:** air temperature, surface wind speed and horizontal visibility

# Mobile Phone Locating-Request Big data

- Retrieving the mobile locating-request (MPL) data from Tencent big data platform in China
- MPL data from active smartphone users using apps (e.g., WeChat, QQ, Tencent Map, etc.) from 1 Mar to 31 Mar 2016
- Recording by aggregating the real-time locations every 5 mins within a mesh grid 1.2 km

# Population Census Data

- Using the latest **city-level population** census in 2014 that includes permanent residents, registered residents at the country level by gender and age group since 2004

# Methodology

- Estimation of Spatiotemporal Continuous PM2.5 Concentrations
- Estimation of real-time population distribution by integrating MPL and Census Data
- Real-Time estimation of population exposure to PM2.5
- Estimation of cumulative inhaled PM2.5

# Estimation of Spatiotemporal Continuous PM2.5 Concentrations

- The ordinary Kriging method was performed using meteorological variables to retrieve data that covers the entire study area at a resolution of 1.2 km.
- To reduce the biases, all of the meteorological observations in 1.2 km were averaged around the PM2.5 monitoring station
- Using the Geographically Weighted Regression (GWR) model with adaptive Gaussian bandwidth to construct the statistical relationship between meteorological variables and PM2.5 concentrations. There are 8 groups for each time point (e.g., 2:00, 5.00,…, and 23:00)

$$PM_{2.5,i,t} = \beta_{0,i,t} + \beta_{1,i,t} VIS_{i,t} + \beta_{2,i,t} AT_{i,t} + \beta_{3,i,t} WS_{i,t}$$

$PM_{2.5,i,t}$: the PM2.5 concentration at the location i at time t at location i at time
$VIS_{i,t}$: the visibility (m) at location i at time
$AT_{i,t}$: air temperature ( C) at location i at time
$WS_{i,t}$: surface wind speed (m/s) t.
$\beta_{0,i,t}$, $\beta_{1,i,t}$, $\beta_{2,i,t}$,$\beta_{3,i,t}$: corresponding regression coefficients at location i at time t.

# Estimation of real-time population distribution by integrating MPL and Census Data

- Aggregating MPL data every 5 mins to 3 hours to be consistent with PM 2.5 monitoring stations.
- Using only MPL data can introduce a bias since physical environment and socio-economic development are different in various areas.
- Thus, Calibrate MPL with census data is performed to reduce the bias.

$$W_{ij} = \frac{p_{ij}}{\sum_{i=1}^{n} p_{ij}}$$

$$pop_{ij} = TR \times W_{ij}$$

$p_{i,j}$: amount of locating-request times within the i-th pixel at the hour j
n : the total number of pixels within a city
$W_{i,j}$ : the weight for redistributing population
TR : total population in the city from the census data

# Real-Time estimation of population exposure to PM2.5

$$PWP = \sum_{i=1}^{N} (pop_i \cdot pm_i) / \sum_{i=1}^{N} pop_i$$

**$pop_i$**: the population in the i-th pixel

**$pm_i$**: PM2.5 concentration level in the i-th pixel

**N**: total number of pixels within the corresponding administrative unit.

**PWP**: the population-weighted PM2.5 concentration level for the targeted administrative unit.

# Estimation of cumulative inhaled PM2.5

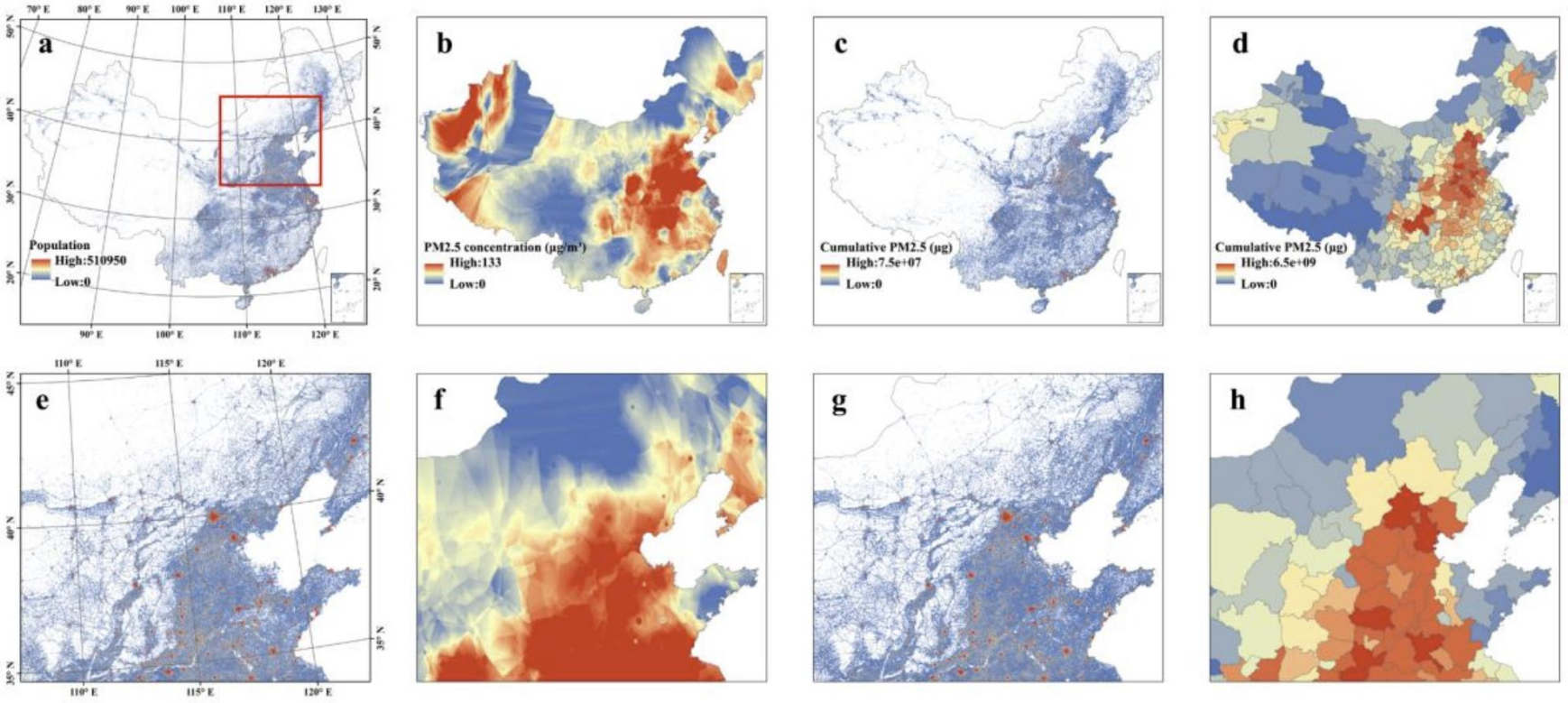
$$InPM'_{2.5} = \sum_{t=1}^{T} p_i(t) \cdot h \cdot m(t)$$

InPM2.5: the cumulative inhaled PM2.5 mass from the model
$p_i$: the number of population present in grid cell i at time t
h: the empirical inhaled volume of air around 15 $m^3$ / day (setting or resting in average)
m(t): the PM2.5 concentration level at time t

# Results

- The extracted example on 1 Mar 2016 (11:00)
- Temporal dynamics of population exposure to PM2.5 for 359 cities in China with every 3 h from 1 March to 31 March 2016
- Comparison of exposure assessment methods between the MPL-based estimations and the census-based estimations

# The extracted example on 1 Mar 2016 (11:00)



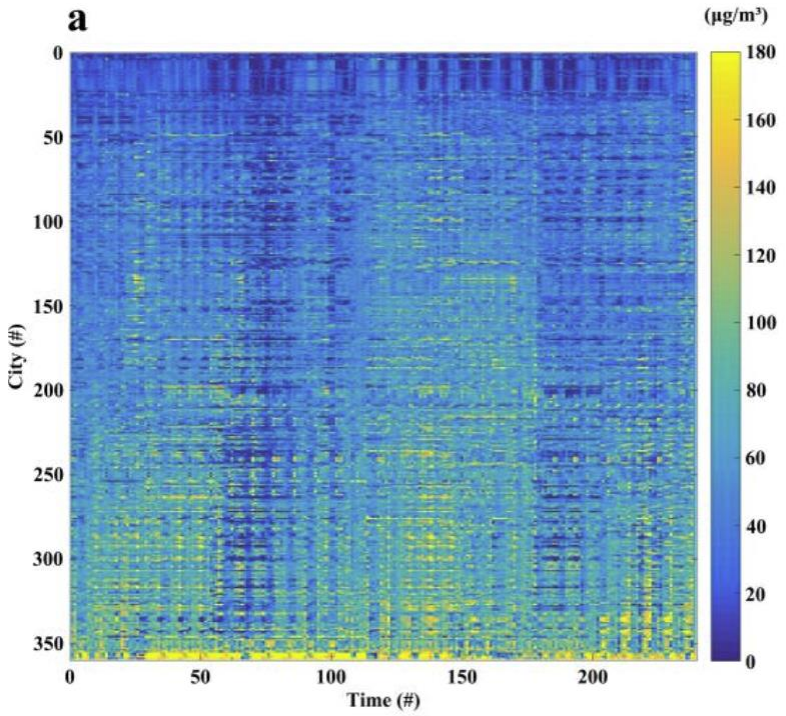

Population distribution

PM2.5 concentration level

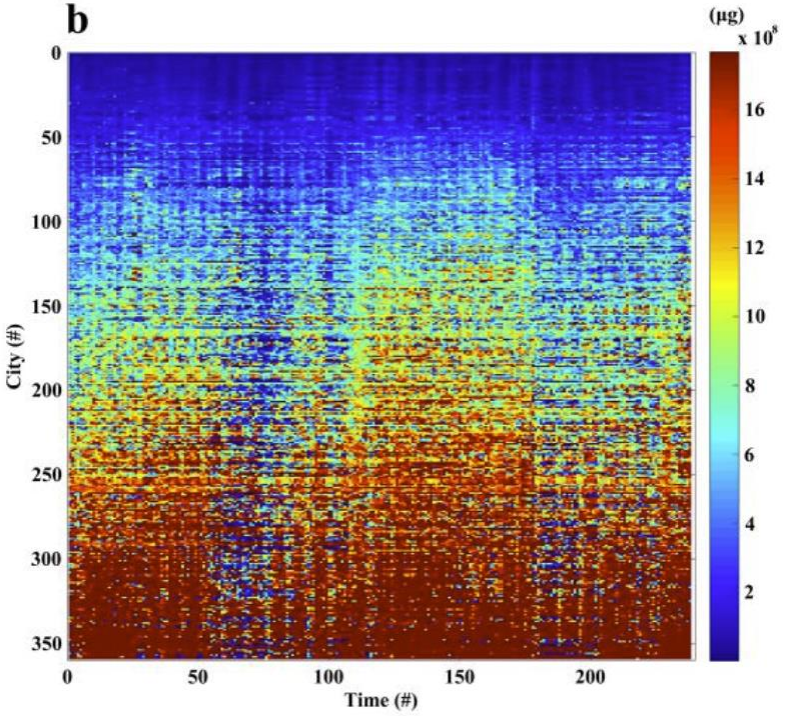
Cumulative inhaled PM2.5 based on MPL data

Cumulative inhaled PM2.5 based on census data

# Temporal dynamics of population exposure to PM2.5 for 359 cities in China with every 3 h from 1 March to 31 March 2016



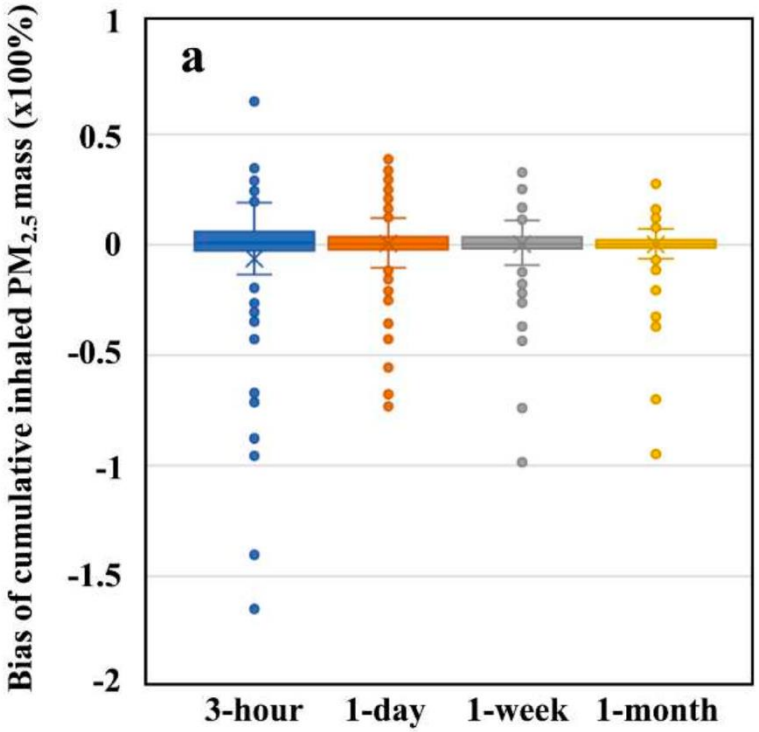

The estimated population-weighted PM2.5 concentrations




cumulative inhaled PM2.5 masses

*Note: x axis represents evey 3-h since 1 Mar to 31 Mar 2016 (8 time periods x 31 days)

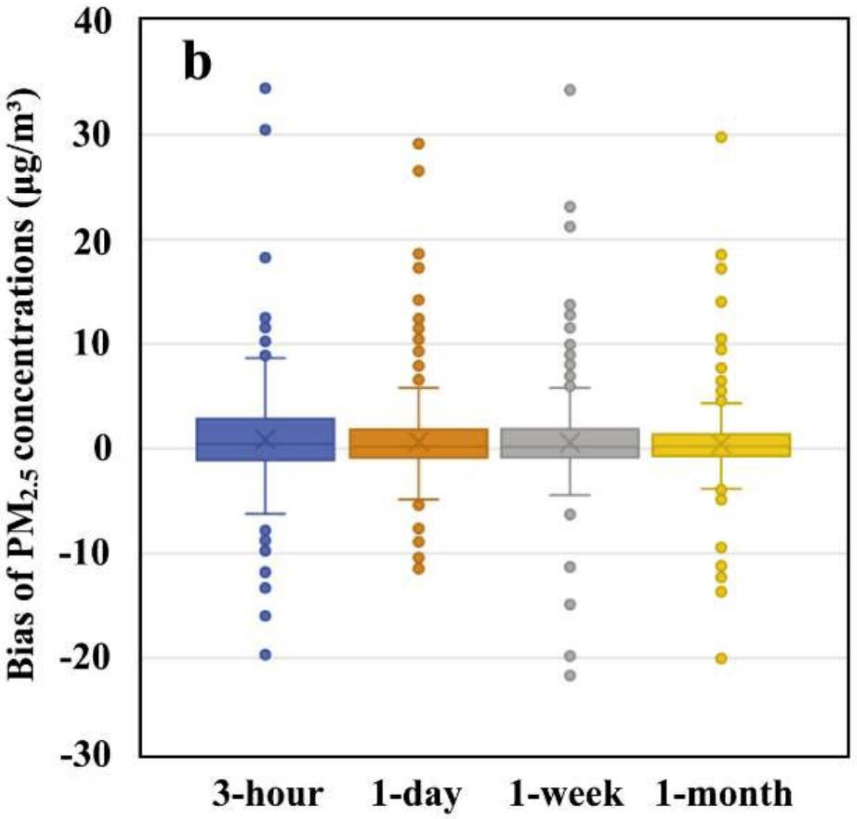
# Comparison of exposure assessment methods between the MPL-based estimations and the census-based estimations




The biases of cumulative inhaled PM2.5 mass (x 100%)

- The maximum biases are over 100% in some points
- On average, the biased percentage between the MPL-based and the census-based:
  - 14.9% (3-h),
  - 5.8% (1-day),
  - 4.7% (1-week),
  - 3.9% (1-month)

# Comparison of exposure assessment methods between the MPL-based estimations and the census-based estimations




The biases of the per capita PM2.5 exposure concentration

- The maximum biases are around 30 μg/m$^3$ on 3-hour
- The exact average of each period isn't mentioned in this paper

# Strengths

- **High spatiotemporal resolution**
  - Combines 3-hourly MPL-derived population maps with GWR-interpolated PM2.5 surfaces, enabling near-real-time exposure tracking.
- **Dynamic population weighting**
  - Using MPL data calibrated at the city level to capture daily mobility patterns (such as commutes, urban-rural shifts) to improve on static census approaches.

# Strengths

- **Captures daily/weekly cycles**
  - The results showed morning/evening peaks in population-weighted PM2.5 in inhaled mass that static or daily-average approaches would miss

# Limitations

- **Census calibration constraints**
  - Relies on the 2014 city-level census might be outdated.
- **Indoor exposure unaddressed**
- **Single-pollutant focus**
  - Only PM2.5 is used.
- **MPL proxy coverage may exclude offline users**


Thank
You