Estimating the impact of ambient air pollution on overall mortality in the presence of preferential sampling and measurement error using electronic health record data

Tae Yoon (Harry) Lee
Banting Postdoctoral Fellow
Quantitative Sciences Unit
Stanford School of Medicine

JSM NASHVILLE TENNESSEE AUGUST 2–7, 2025
Statistics, Data Science, and AI Enriching Society

Learning goals

By the end of this talk, you should be able to:

- Recognize different sources of data and different types of models for air pollution exposure assessment
- Discuss the importance of addressing potential biases in environmental health studies:
  - Preferential sampling in exposure assessment
  - Measurement error
- Develop an appreciation for open science practices

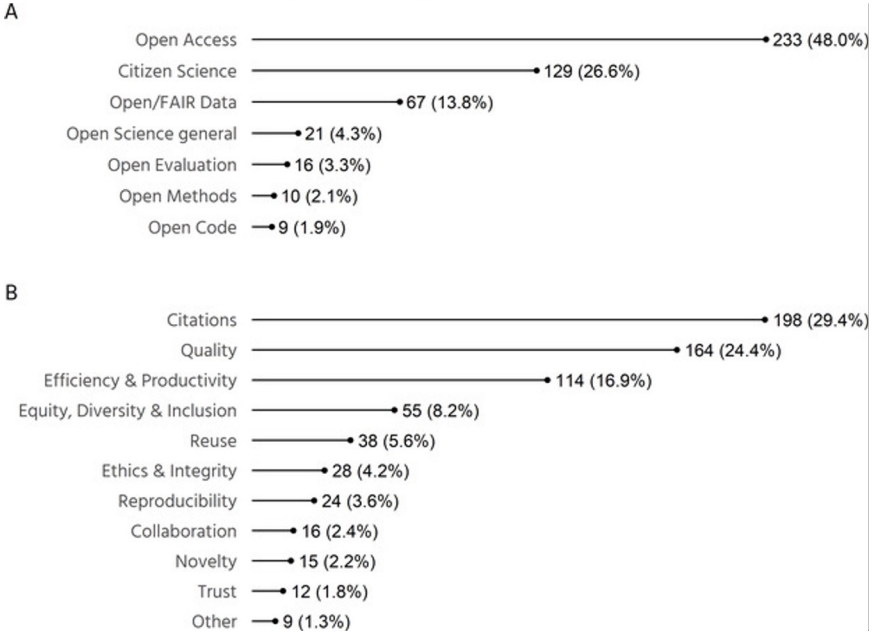
Open science practices

Review articles

The academic impact of Open Science: a scoping review

Thomas Klebel ✉, Vincent Traag, Ioanna Grypari, Lennart Stoy and Tony Ross-Hellauer

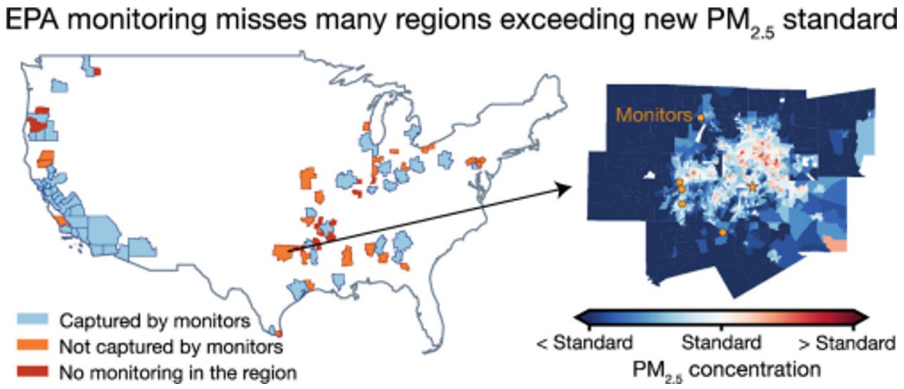
Published: 05 March 2025 https://doi.org/10.1098/rsos.241248



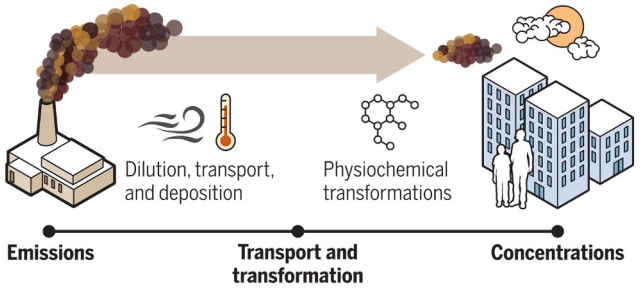
TQRCG

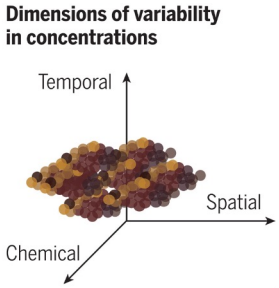
Air quality monitoring site

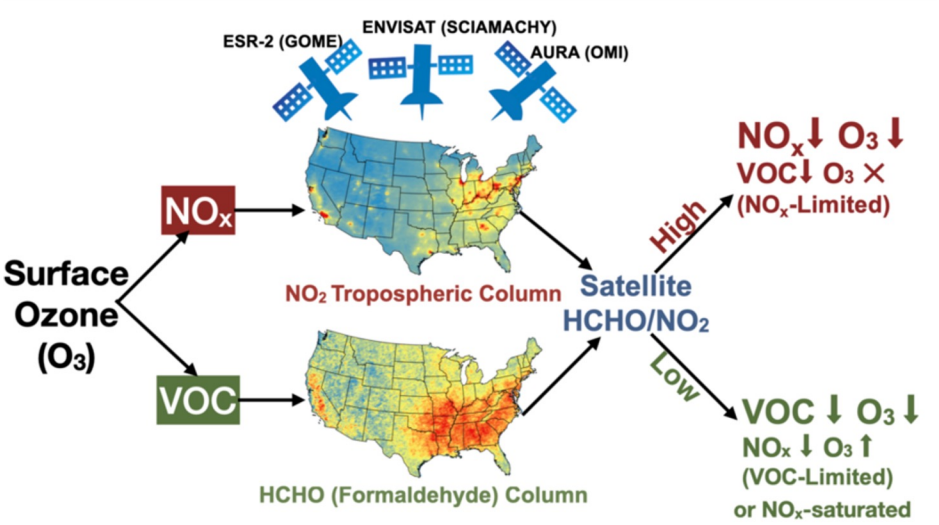
"FRM and FEM monitors are considered **the gold standard** for air quality monitoring."



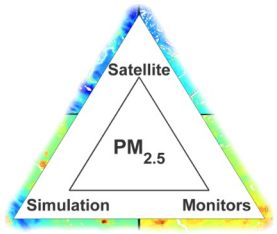
Other methods







SatPM2.5 (Satellite-derived PM2.5)



Global and regional $PM_{2.5}$ concentrations are estimated using information from satellite-, simulation- and monitor-based sources. Aerosol optical depth from multiple satellite instruments (MODIS/Terra, MODIS/Aqua, MISR/Terra, SeaWiFS/SeaStar, VIIRS/SNPP, and VIIRS/NOAA20) and their respective retrievals (Dark Target, Deep Blue, MAIAC) is combined with simulation (GEOS-Chem) based upon their relative uncertainties as determined using ground-based sun photometer (AERONET) observations to produce geophysical estimates that explain most of the variance in ground-based $PM_{2.5}$ measurements. A subsequent statistical fusion incorporates additional information from $PM_{2.5}$ measurements.

Why monitoring data?

Connecting Health Outcomes Research and Data Systems (CHORDS)

CAFE

Accelerating Climate & Health Research

CAFE Climate and Health Research Coordinating Center Collection

(Harvard University, Boston University)

Distance-based method

PLoS One. 2023 Feb 16;18(2):e0281499. doi: 10.1371/journal.pone.0281499

Short-term association between ambient air pollution and cardio-respiratory mortality in Rio de Janeiro, Brazil

Taísa Rodrigues Cortes [1,*], Ismael Henrique Silveira [2], Beatriz Fátima Alves de Oliveira [3], Michelle L Bell [4], Washington Leite Junger [1]

Exposure assessment

Individual-level exposures to air pollutants and meteorological variables were estimated using the inverse distance weighting (IDW) method. In this method, the values at unmeasured locations are estimated by averaging the weighted sum of values from the nearest neighbors (monitoring stations) within a search radius. The weights assigned to each neighboring value are expressed as a function of the inverse distance (between the measured and unmeasured locations) raised to a non-negative power [18].

J Atheroscler Thromb. 2025 Jan 25. doi: 10.5551/jat.65424. Online ahead of print.

All-Cause and Cause-Specific Mortality Associated with Long-Term Exposure to Fine Particulate Matter in Japan: The Ibaraki Prefectural Health Study

Takehiro Michikawa [1,2], Yuji Nishiwaki [1], Keiko Asakura [3], Tomonori Okamura [4], Toru Takebayashi [4], Shuichi Hasegawa [5], Ai Milojevic [6], Mihoko Minami [7], Masataka Taguri [8], Ayano Takeuchi [9], Kayo Ueda [10], Toshimi Sairenchi [2,11], Kazumasa Yamagishi [2,12], Hiroyasu Iso [13], Fujiko Irie [14], Hiroshi Nitta [15]

Affiliations + expand

PMID: 39864858 DOI: 10.5551/jat.65424

Methods: We used data of 46,974 participants (19,707 men; 27,267 women), who were enrolled in 2009 and followed up until 2019, in a community-based prospective cohort study (the second cohort of the Ibaraki Prefectural Health Study). We estimated $PM_{2.5}$ concentrations using the inverse distance weighing methods based on ambient air monitoring data, and assigned each participant to administrative area level concentrations. A Cox proportional hazard model was applied to estimate hazard ratios (HRs) and 95% confidence intervals (CIs) of mortality.

Kriging

> Am J Respir Crit Care Med. 2022 Oct 15;206(8):1008-1018. doi: 10.1164/rccm.202107-1770OC.

Traffic-related Air Pollution and Lung Cancer Incidence: The California Multiethnic Cohort Study

Iona Cheng [1 2], Juan Yang [1], Chiuchen Tseng [3], Jun Wu [4], Salma Shariff-Marco [1 2],

We used established approaches to estimate air pollutant concentrations at residential locations across the study period (1993–2013) as previously described (23, 24). For gaseous traffic-related pollutants, based on empirical Bayesian kriging interpolation, largely exposures from regional emission sources (25) were estimated using air monitoring data routinely collected by the U.S. Environmental Protection Agency for NO_x, NO_2, PM_{10}, CO, and ozone (O_3) (1993–2013) and $PM_{2.5}$ (2000–2013). $PM_{2.5}$ concentrations for 1993–1999

ARTICLES · Volume 22, 100500, June 2023
Open Access

Download Full Issue

Assessing socioeconomic bias of exposure to urban air pollution: an autopsy-based study in São Paulo, Brazil

Julio da Motta Singer [a] · Carmen Diva Saldiva de André [a] · Paulo Afonso de André [b,c] · Francisco Marcelo Monteiro Rocha [d] · Dunia Waked [b] · Aline Macedo Vaz [b] · et al. Show more

In parallel, we created a project in a Geographic Information System (ArcMap 10.8.1) to estimate PM_{10} concentrations at the residence of the deceased by interpolating an ordinary kriging model of the concentrations of particles determined by the network with 25 monitoring stations of the MASP. Kriging is a geostatistical method for interpolation based on statistical models that include autocorrelation. Kriging is most appropriate when there is a spatially correlated distance or directional bias in the data. Thus, we performed an autocorrelation analysis before kriging.[25]

Problem statement

- A network of monitoring sites is constructed for various purposes:
  - Compliance with the regulation
  - Evaluation of interventions (e.g., emission control strategies)
  - Forecasting
  - *Validating air quality models in rural regions*
- **Preferentially siting monitors:**

“Real monitors are clustered in high concentration areas.”

– Air Quality Monitoring Network Plan (2023)

Cohort: Stanford Health Care

- Subset of 4,391 patients diagnosed with lung cancer from 2000 to 2021
- Geolocation based on mailing address in EHR

Meeting Abstract: 2025 ASCO Annual Meeting I

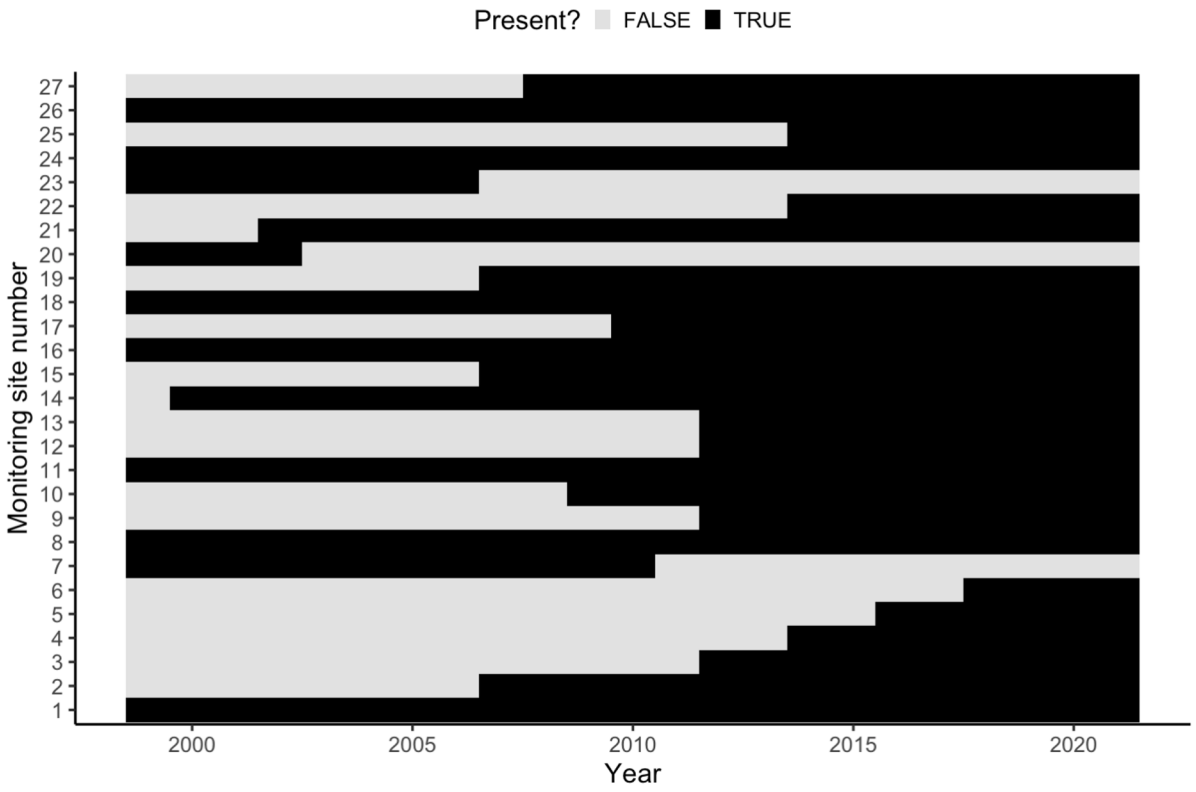
FREE ACCESS | Quality Care/Health Services Research | May 28, 2025

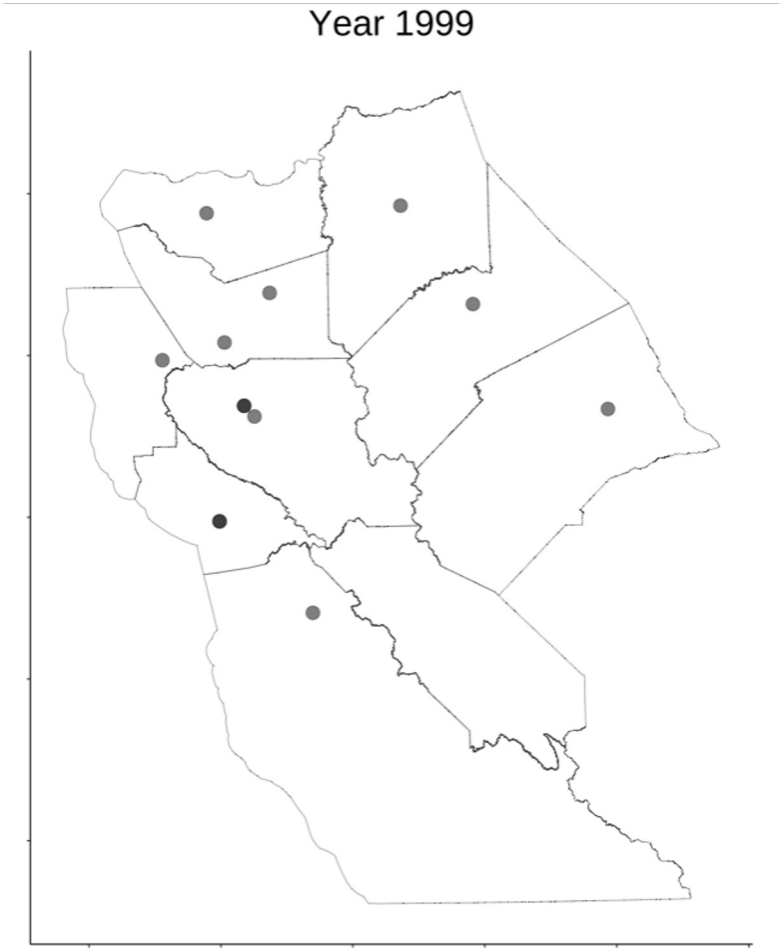
Oncoshare-Lung: Novel three-way linkage of neighboring academic and community medical centers to state cancer registry for lung cancer.

Authors: Chloe Su, Eunji Choi, Mina Satoyoshi, Archana Bhat, Tony Chen, Victoria Ding, Aparajita Khan, Tae Yoon Lee, Julie Tsu-yu Wu, Solomon Henry, Manisha Desai, Joel W. Neal, Leah Monique Backhus, Curtis Langlotz, Ann Leung, Scarlett L. Gomez, Heather A. Wakelee, Allison W. Kurian, Su-Ying Liang, and Summer Han SHOW FEWER | AUTHORS INFO & AFFILIATIONS

Publication: Journal of Clinical Oncology • Volume 43, Number 16_suppl
https://doi.org/10.1200/JCO.2025.43.16_suppl.e23292

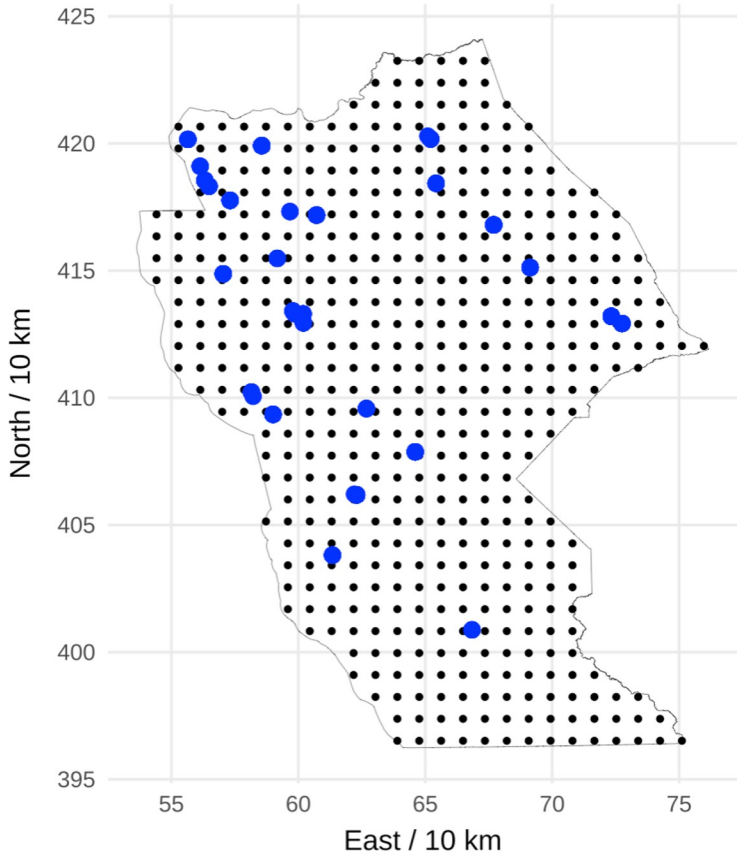
Monitoring sites for PM 2.5



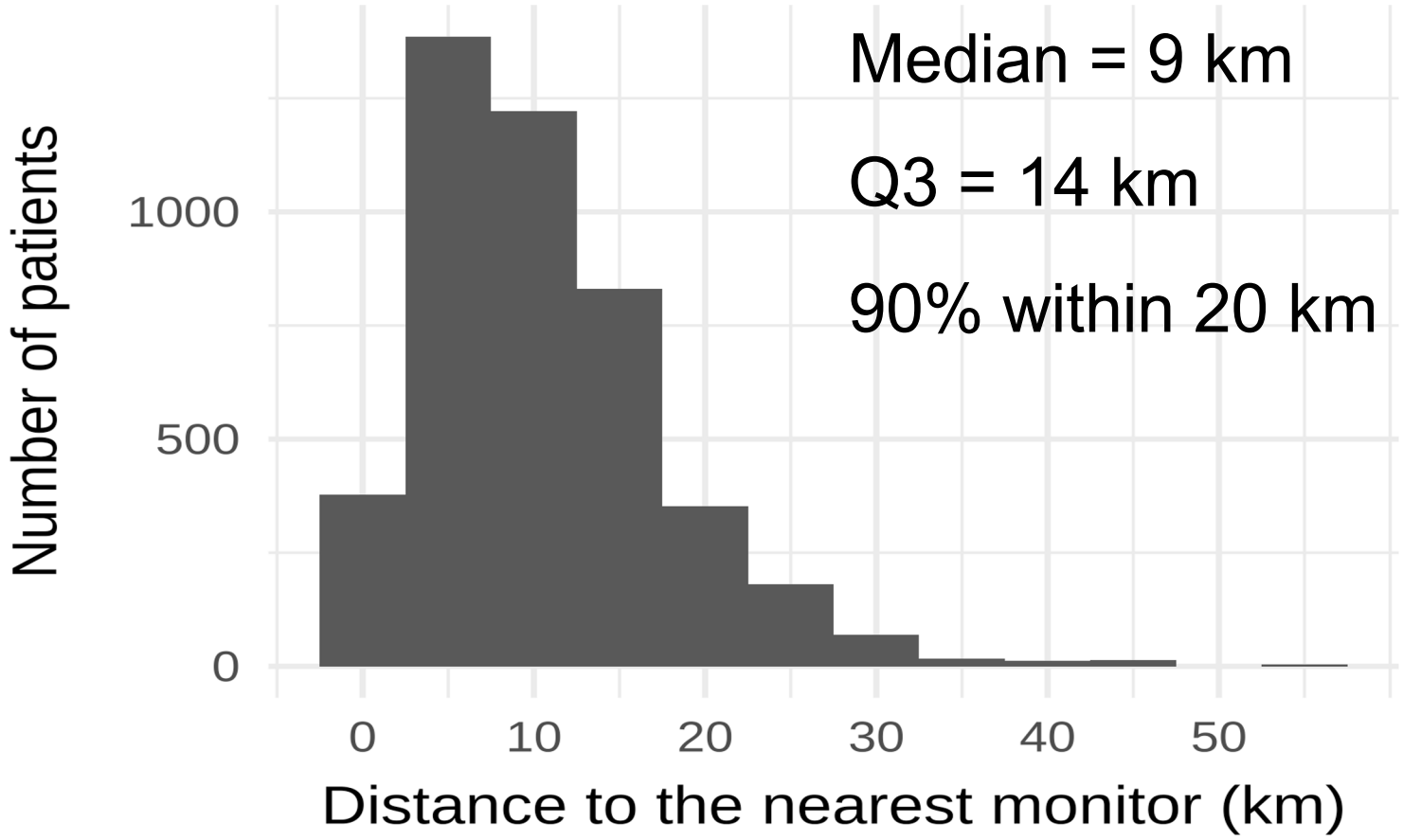


Potential preferential sampling?

- People's opinion, politics (e.g., budget cuts to environmental programs)
- Example: UK black smoke
  - Clean Air Act in 1956
  - 1235 sites in 1976
  - 563 sites in 1986
  - 225 sites in 1996
  - 65 sites in 2006



Distance to the nearest monitor



Acknowledgement - UBC Statisticians

James V Zidek

Joe Watson

Adrian Jones

Xinglong Li

December 2019

A general theory for preferential sampling in environmental networks

Joe Watson, James V. Zidek, Gavin Shaddick

Ann. Appl. Stat. 13(4): 2662-2700 (December 2019). DOI: 10.1214/19-AOAS1288

Preferential siting air pollution monitors

Detecting and adjusting for preferentially located air pollution monitoring sites: A case study with novel R packages

*

Adrian Jones[1], Xinglong Li[4], James V Zidek[2] and Joe Watson[3]

Observation process

$Z_{i,j}$ is the observed PM 2.5 on the log scale at site i (located at s_i) at time t_j.

$R_{i,j}$ is the site selection indicator for site i at time t_j.

$$
\begin{aligned}
Z_{i,j}|R_{i,j}=1 &\sim N(\mu_{i,j},\sigma^2_\epsilon) \\
\mu_{i,j} &= \gamma_0+\gamma_1 t_j+\gamma_2 t_j^2+ \\
&\quad b_{0,i}+b_{1,i}(t_j)+ \\
&\quad \beta_0(s_i) \\
[\beta_0(s_1),\ldots,\beta_0(s_M)]^T &\sim N(0,\Sigma(\xi_0)) \\
\Sigma(\xi_0) &= \text{Matern}(\xi_0) \\
[b_{0,i},b_{1,i}]^T &\sim N(0,\Sigma_b),\ \Sigma_b=\begin{bmatrix}\sigma^2_{b,0} & \rho_b \\ \rho_b & \sigma^2_{b,1}\end{bmatrix} \\
(\sigma^2_\epsilon,\gamma_0,\gamma_1,\gamma_2,\xi_0,\sigma^2_{b,0},\sigma^2_{b,1},\rho_b) &\sim \text{Priors}
\end{aligned}
$$

global variation

site-specific variation

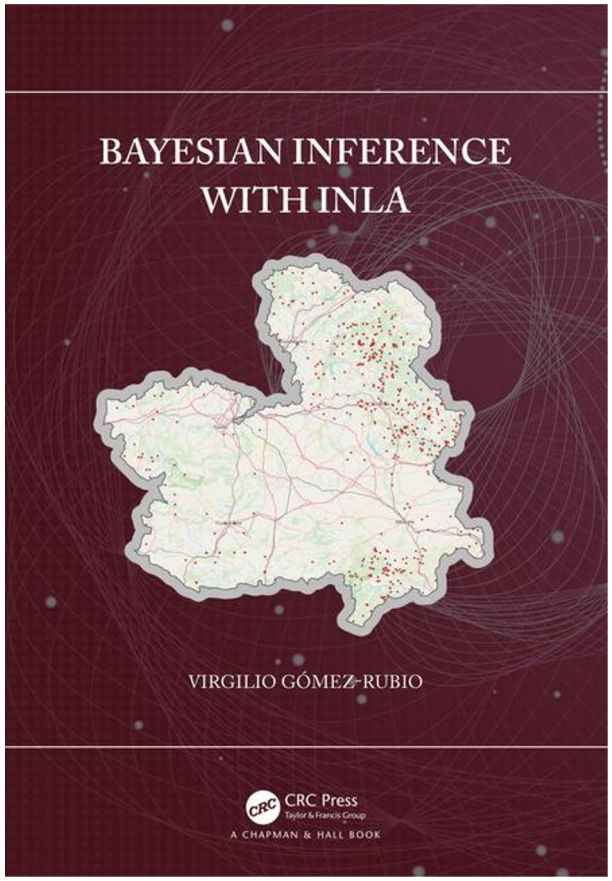
spatial variation

Site-selection process

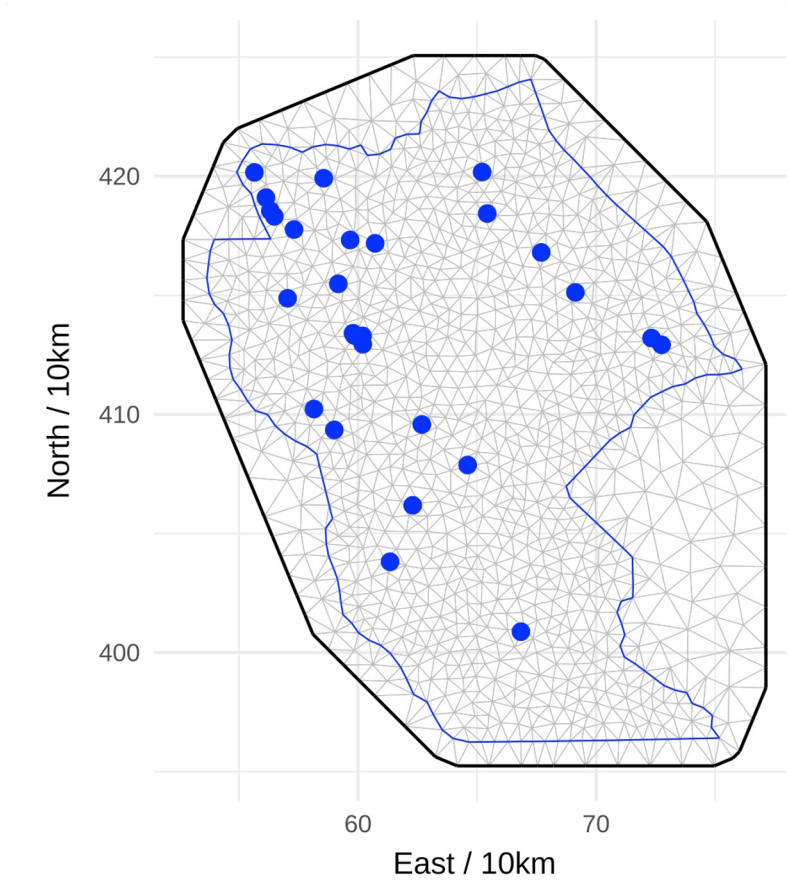
$$
\begin{aligned}
R_{i,j} &\sim \text{Bern}(p_{i,j}) \\
\text{logit}(p_{i,j}) &= \alpha_0 + \alpha_1 t_j + \alpha_2 t_j^2 + \beta_1^R(t_j) + && \text{global variation} \\
&\quad \beta_0^R(s_i) + && \text{spatial variation} \\
&\quad \alpha_{retention} R_{i,j-1} + && \text{retention effect} \\
&\quad d_b[b_{0,i} + b_{1,i} t_{j-1}] + && \text{“small-scale” PS} \\
&\quad d_\beta[\beta_0(s_i)]; && \text{“large-scale” PS} \\
[\beta_1^R(t_1), \ldots, \beta_1^R(t_T)]^T &= AR1(\rho_R, \sigma_R^2) \\
[\beta_0^R(s_1), \ldots, \beta_0^R(s_M)]^T &\sim N(0, \Sigma(\xi_R)) \\
\Sigma(\xi_R) &= \text{Matern}(\xi_R) \\
[\alpha_0, \alpha_1, \alpha_2, d_b, d_\beta, \rho_R, \sigma_R^2, \xi_R] &\sim \text{Priors}
\end{aligned}
$$

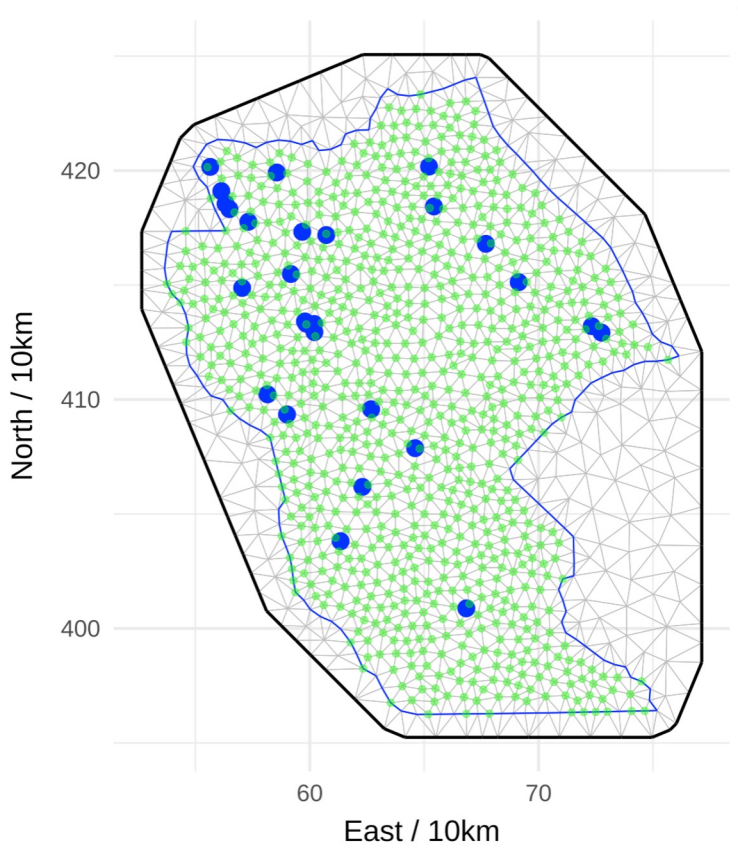
Integrated Nested Laplace Approximation (INLA)

- Bayesian hierarchical model
- Main interests:
  - Posterior for the latent field
  - Posterior for parameters
- Big “N” problem
- Fast, scalable approximation method
- Memory intensive



Mesh: presence only vs. pseudo-site





Evidence for preferential sampling

$$d_b[b_{0,i} + b_{1,i}t_{j-1}] + \quad \text{“small-scale” PS}$$
$$d_\beta[\beta_0(s_i)]; \quad \text{“large-scale” PS}$$

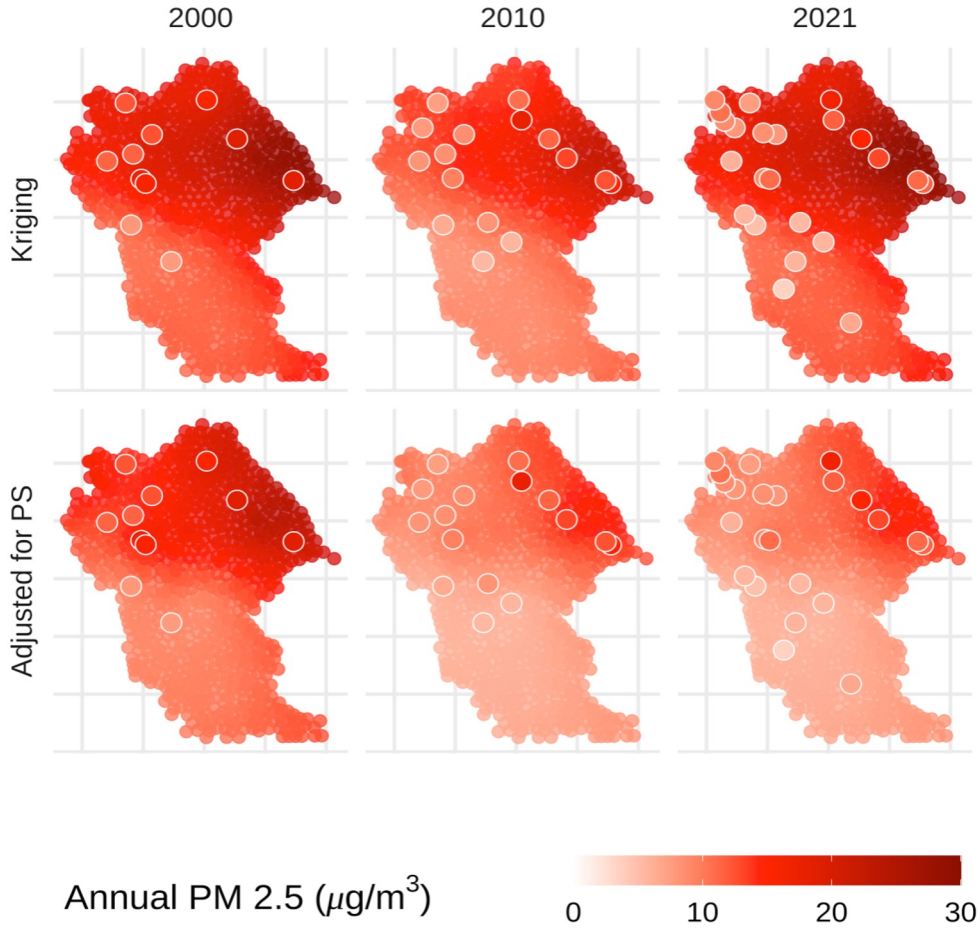
Observation process

Parameter	Kriging	Presence only	Pseudo-site
γ_0	0.28 (0.10)	0.27 (0.07)	0.28 (0.06)
γ_1	-1.54 (0.15)	-1.53 (0.13)	-1.53 (0.14)
γ_2	1.03 (0.13)	1.03 (0.12)	1.03 (0.12)
ρ_b	0.00 (0.21)	-0.02 (0.11)	-0.00 (0.02)
range_{ξ_0}	20.42 (13.42)	15.99 (3.5)	12.34 (0.70)
σ_{ξ_0}	0.34 (0.12)	0.29 (0.05)	0.24 (0.01)

Site-selection process

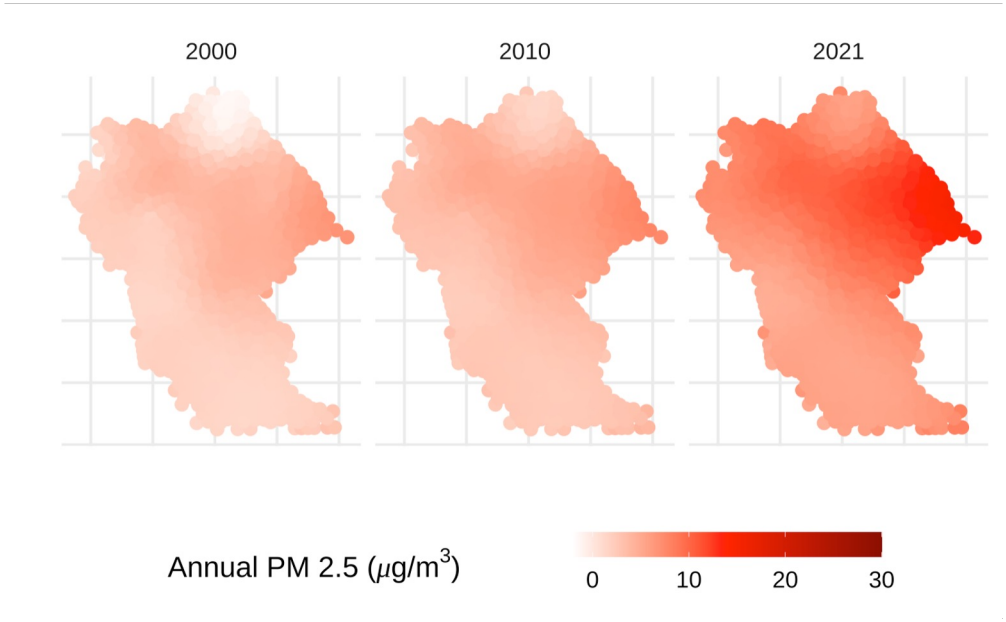
Parameter	Kriging	Presence only	Pseudo-site
α_0	-2.54 (0.61)	-1.84 (0.93)	-5.75 (0.45)
α_1	-2.54 (3.03)	-3.97 (4.36)	-5.21 (2.34)
α_2	3.61 (3.15)	4.47 (4.44)	3.75 (2.42)
$\alpha_{retention}$	8.14 (0.69)	7.74 (0.65)	11.94 (0.64)
ρ_R	-0.45 (0.35)	0.62 (0.07)	0.82 (0.01)
$1/\sigma_R^2$	1.24 (0.86)	1.24 (0.24)	14.43 (0.78)
range_{ξ_R}	11.18 (15.23)	4.72 (1.33)	5.78 (0.41)
σ_{ξ_R}	0.27 (0.27)	0.12 (0.03)	0.68 (0.03)
d_b	NA	1.04 (0.18)	0.97 (0.06)
d_β	NA	0.87 (0.16)	0.15 (0.05)

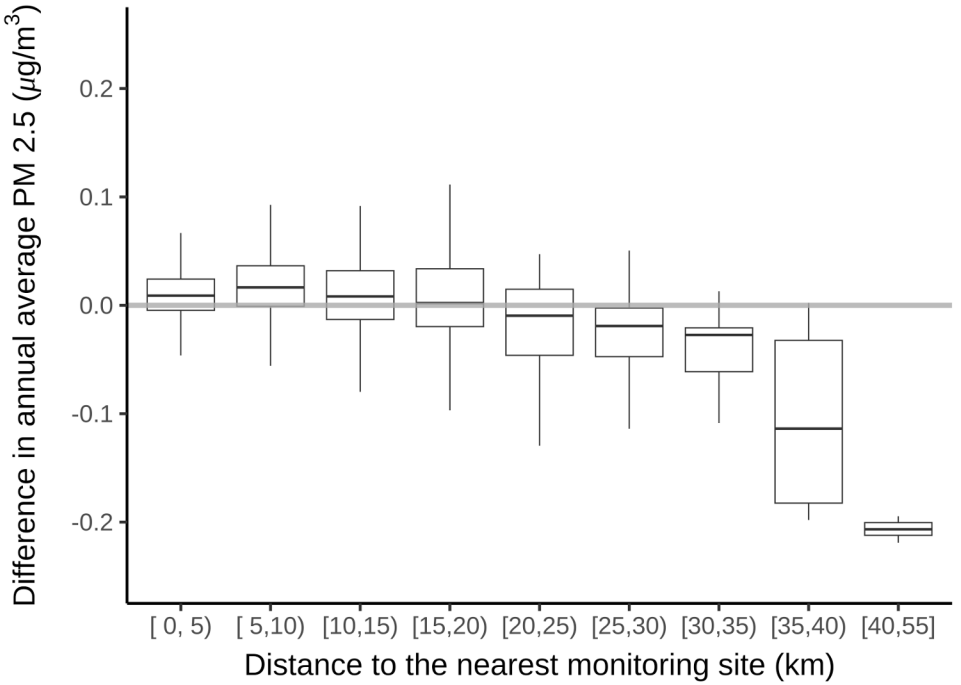
Estimated air pollution field



Difference in the estimated field

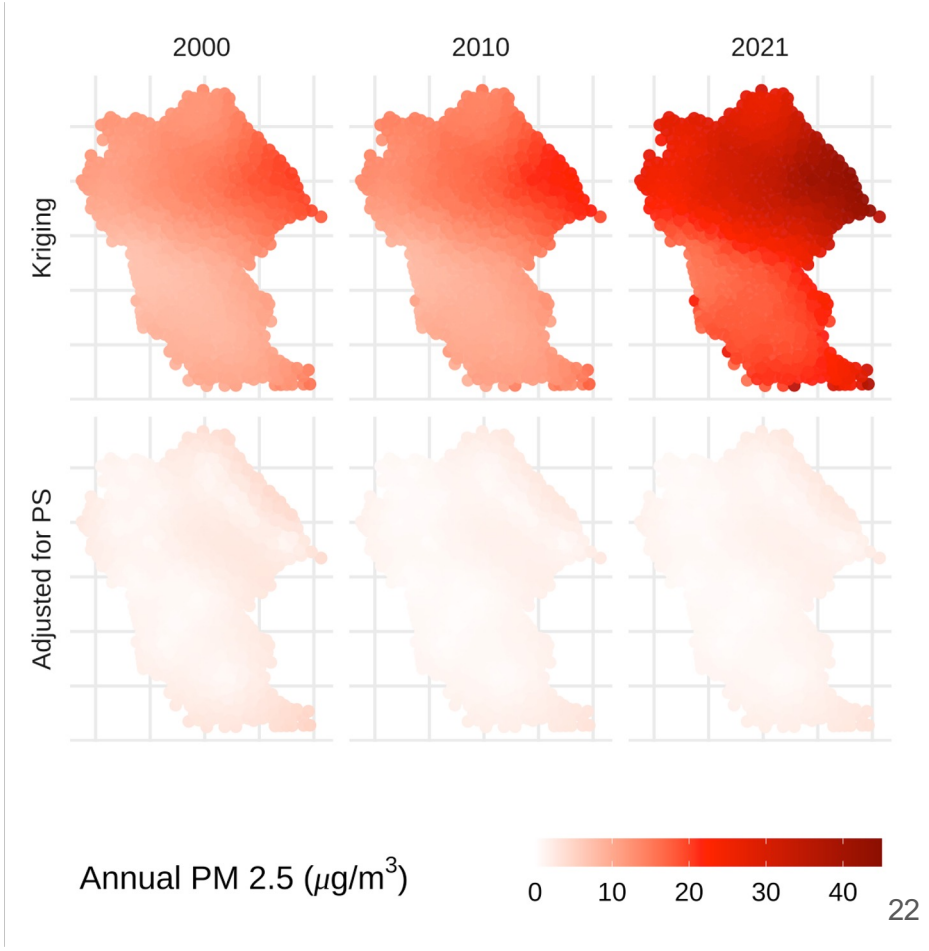
Difference = Kriging - Adjusted for PS





Downstream implication in health policy

- Without adjusting for PS, the air pollution exposure assessment could be biased.
- This bias could lead to misleading results in environmental health studies.
- Measurement error is often ignored in applied research.



Case study: short-term effects of PM 2.5 on 2-year overall mortality

- **Cohort:** 4,391 patients diagnosed with lung cancer
- **Outcome**: 2-year overall mortality (N=1,659)
- **Exposure**: 1-year past annual average exposure to PM 2.5 at the time of lung cancer diagnosis
- **Censoring:** 14% (N=378) among those who did not die
- **Covariates**: sex, age, race, ethnicity, marital status, histology, stage, smoking status, personal history of lung cancer, surgery, radiation, immunotherapy, chemotherapy
- **Model:** Bayesian logistic regression model using the brms package in R.

Measurement error model

- Joint modeling of the health and exposure models can be computationally infeasible.
- Two-stage Bayesian approaches
  - **Independent normal prior** (brms: `me()`)
  - Multiple imputation
  - Kernel density estimation
  - Scalable MVN approach

> Biostatistics. 2023 Dec 15;25(1):20-39. doi: 10.1093/biostatistics/kxac034.

A Bayesian framework for incorporating exposure uncertainty into health analyses with application to air pollution and stillbirth

Saskia Comess [1], Howard H Chang [2], Joshua L Warren [3]

▸ Biostatistics. 2024 Oct 4;26(1):kxae038. doi: 10.1093/biostatistics/kxae038

A scalable two-stage Bayesian approach accounting for exposure measurement error in environmental epidemiology

Changwoo J Lee [1], Elaine Symanski [2,3], Amal Rammah [4], Dong Hun Kang [5], Philip K Hopke [6], Eun Sug Park [7,✉]

Impact of PS and ME on inference

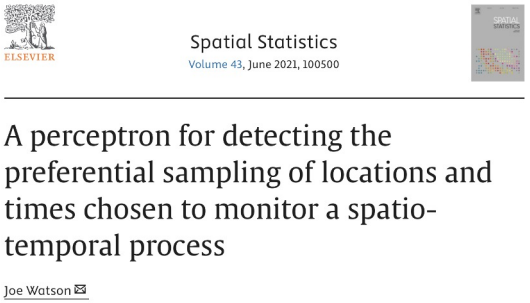
Adjusted for PS	Adjusted for ME: No	Adjusted for ME: Yes
No	1.41 (0.97-2.03)	**1.52 (0.99-2.33)**
Yes	1.41 (0.97-2.04)	**1.51 (0.99-2.34)**

Odds ratio per 10 unit increase in PM 2.5

Posterior mean (posterior 95% CI: Q2.5 and Q97.5)

Discussion: methods

- "Quick" test for PS
  - PStestR not maintained
- Spatially-varying PS
- Spatial data fusion under PS
  - Monitoring data
  - Satellite data
- Extension to temporal
- Extension to multivariate exposures
- Validation using monitoring data?



Spatial Statistics
Volume 43, June 2021, 100500

A perceptron for detecting the preferential sampling of locations and times chosen to monitor a spatio-temporal process

Joe Watson

Model-based geostatistics under spatially varying preferential sampling

AVR Amaral, ET Krainski, R Zhong, P Moraga

Journal of Agricultural, Biological and Environmental Statistics, 2024 • Springer

Spatial data fusion adjusting for preferential sampling using integrated nested Laplace approximation and stochastic partial differential equation

Ruiman Zhong , André Victor Ribeiro Amaral , Paula Moraga Author Notes

Journal of the Royal Statistical Society Series A: Statistics in Society, Volume 188, Issue 1, January 2025, Pages 140–157, https://doi.org/10.1093/jrsssa/qnae058

Published: 27 June 2024 **Article history**

Biostatistics. 2024 Dec 31;26(1):kxae038. doi: 10.1093/biostatistics/kxae038.

A scalable two-stage Bayesian approach accounting for exposure measurement error in environmental epidemiology

Changwoo J Lee [1], Elaine Symanski [2 3], Amal Rammah [2], Dong Hun Kang [4], Philip K Hopke [5], Eun Sug Park [4]

Discussion: applications

- Novel exposurement assessment models
- “Daunting task” for an applied healthcare researcher
- Lack of open research practices
- Open repositories
- Open-models for addressing
  - Preferential sampling
  - Measurement error
  - …

Connecting Health Outcomes Research and Data Systems (CHORDS)

CAFE
Accelerating Climate & Health Research

CAFE Climate and Health Research Coordinating Center Collection

(Harvard University, Boston University)

› Account Res. 2023 Jan;30(1):34-62. doi: 10.1080/08989621.2021.1962713. Epub 2021 Aug 17.

Open science, the replication crisis, and environmental public health

Daniel J Hicks [1]

Acknowledgement

Summer S Han

Chloe C Su

Stanford MEDICINE | Quantitative Sciences Unit

NIH

KISS KOREAN INTERNATIONAL STATISTICAL SOCIETY

UBC STATISTICS

CIHR IRSC

James V Zidek

Joe Watson

Adrian Jones

Xinglong Li

Thank you!

Back-up slides

Test for preferential sampling (Watson 2021)

Idea: preferentially sampled monitoring sites more/less clustered in regions with above/below-average levels

Implication: the nearest distances between sites will be negatively/positively correlated with the observed concentration at each site.

Test statistic: non-parametric Spearman's Rho correlation between ranked nearest neighbor distances and ranked exposure levels at the sites

Procedures

1. Fit a model (e.g., kriging) to the observed exposure data under the null hypothesis of no PS
2. Simulate a network of sites using that model.
3. Estimate the exposure field using the sampled network.
4. Compute the average of the K-nearest neighbor distances at each site in the sampled network.
5. Compute the rank correlation test statistic between the average distance and the estimated concentration.
6. Repeat Steps 2-5 for N times.

Weak evidence for preferential sampling