Understanding the Source Components Captured by the Purple Air Network

Vijay Kumar

Department of Mathematics, Clarkson University, Potsdam-NY

Date: October 05, 2022

Co-authors:

Suresh Dhaniyala, Clarkson University

Supraja Gurajala, SUNY Potsdam

Shantanu Sur Clarkson University

Sumona Mondal, Clarkson University

Dinushani Senarathna, Clarkson University

CLARKSON UNIVERSITY 18 • TECHNOLOGIA • 96 ™

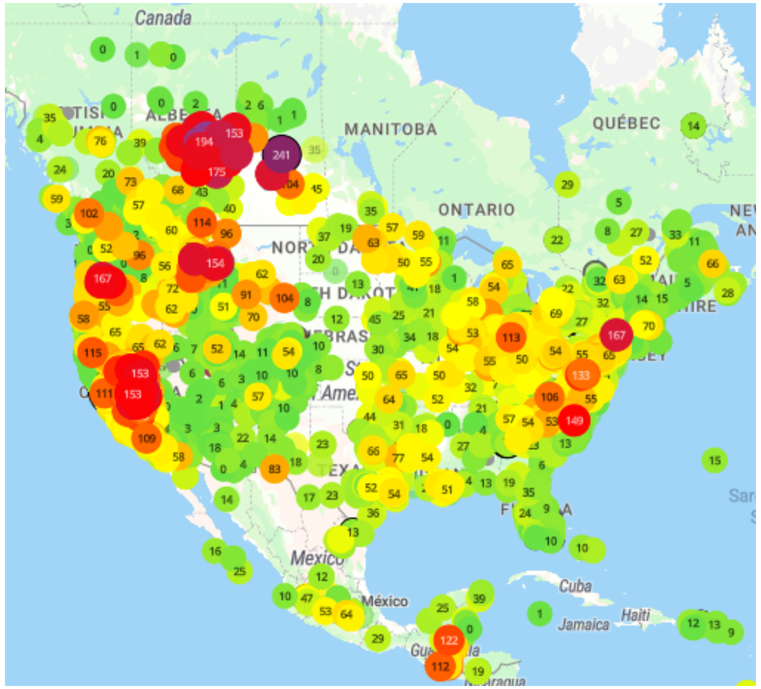
AIR LAB AEROSOL INSTRUMENTAL RESEARCH LAB

AAAR American Association for Aerosol Research

HEC HIGHER EDUCATION COMMISSION

Introduction

- Conventional instruments for air quality data are expensive and difficult to maintain
- Low-cost air sensor data increasingly used in health studies
- Sensitivity, noise, and accuracy of data of sensors remain a concern
- What is driving these differences, and can that be “calibrated” out?
- This study is focused on differences between $PM_{2.5}$ concentrations of EPA reference monitors and low-cost Purple Air (PA) sensors



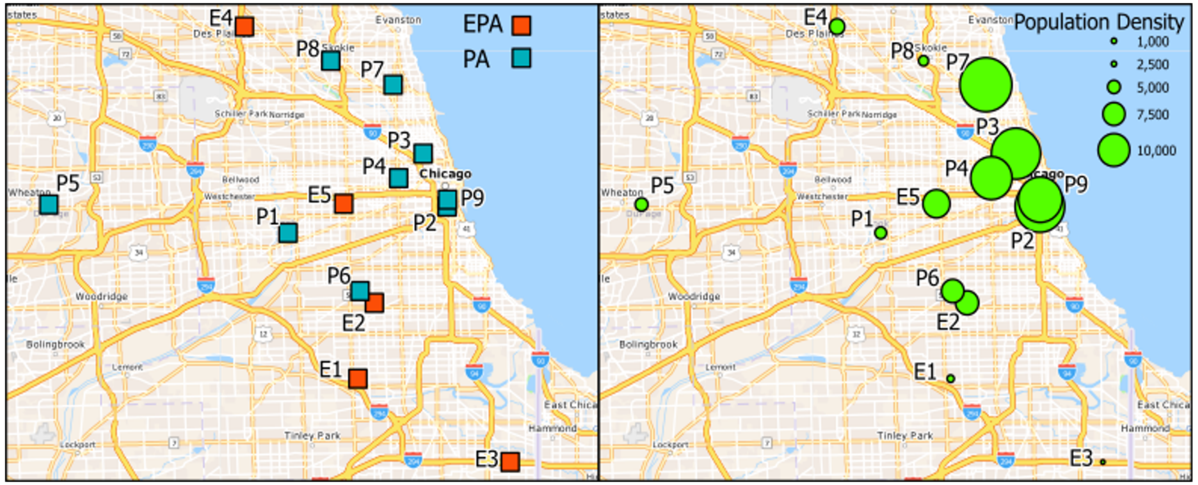
https://map.purpleair.com/1/mAQI/a10/p604800/cC0#2.33/42.94/-98.5

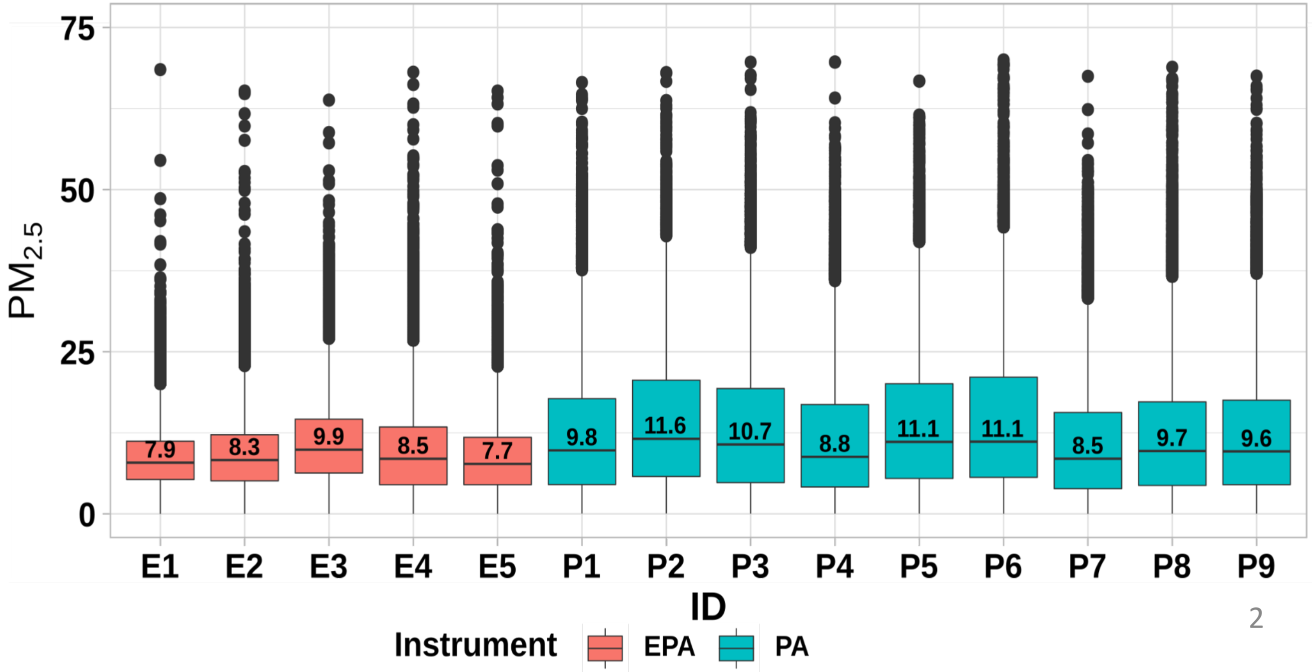
Monitoring data of $PM_{2.5}$

Data preprocessing

- $PM_{2.5}$ data from Sep 2019 to Dec 2021
- Removed sensors > 20% missing observations
- 05 EPA and 09 PA sensors in the final analysis
- $PM_{2.5}$ range $[0,70]\ \frac{\mu g}{m^3}$ to exclude outliers

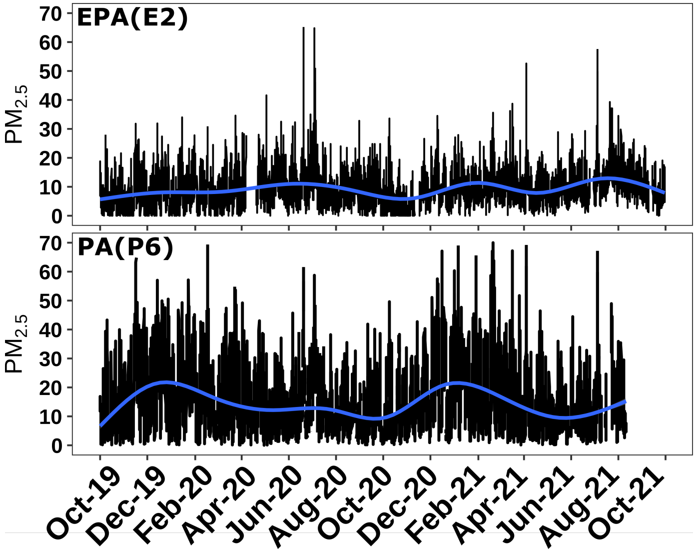
EPA & PA sites in Cook County, IL

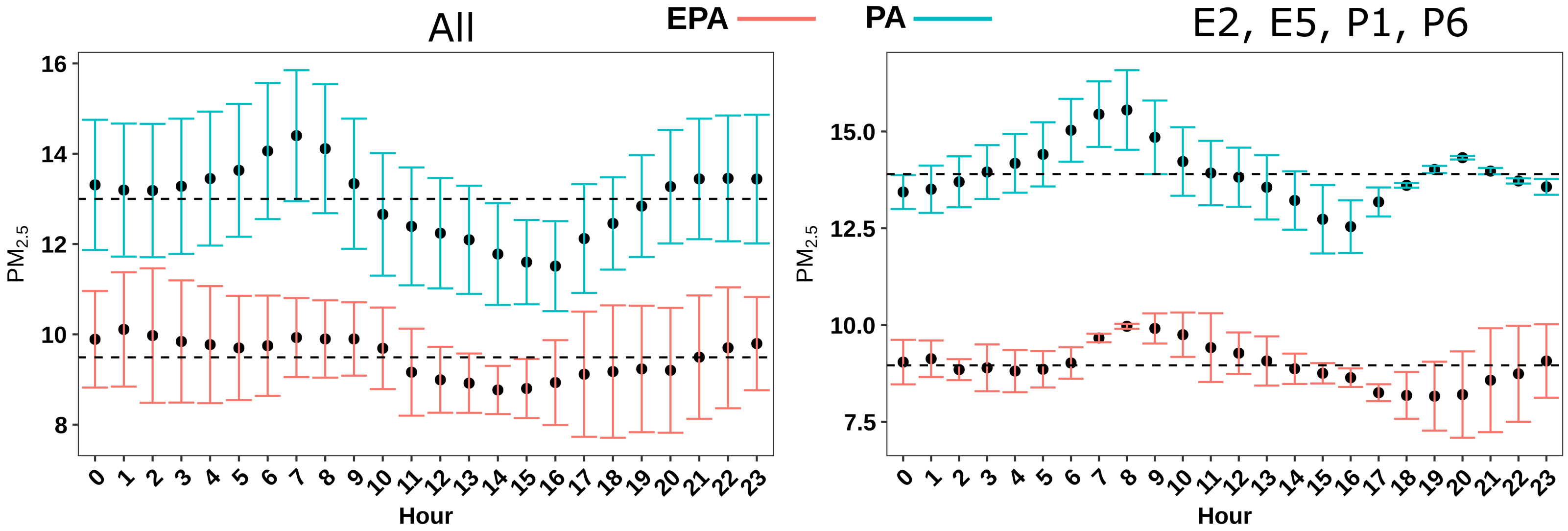




Current state in low-cost sensors: Data accuracy and noise

- PA data is noisy and overestimate the $PM_{2.5}$ measurements [1,2,3]
- Correction models accounting for environmental conditions on sensor performance [1,2]
- Only relative humidity and temperature were used as correction factors
- Correlation based methods have been used to correct the data





Objective: Characterizing the reasons for differences in low-cost and regulatory monitoring data

[1] Giordano, 2021 [2] Barkjohn, 2021 [3] Ardon-Dryer, 2020

Time series data (spectral analysis)

- Spectral analysis used to study daily and weekly variations in aerosols [1,2,3,4]
- Low-cost sensors show better agreement with reference monitors in low traffic locations [5]

$The\ power\ spectral\ density\ (PSD)\ for\ a\ finite\ time\ is\ the\ squared\ magnitude\ of\ X(k)$

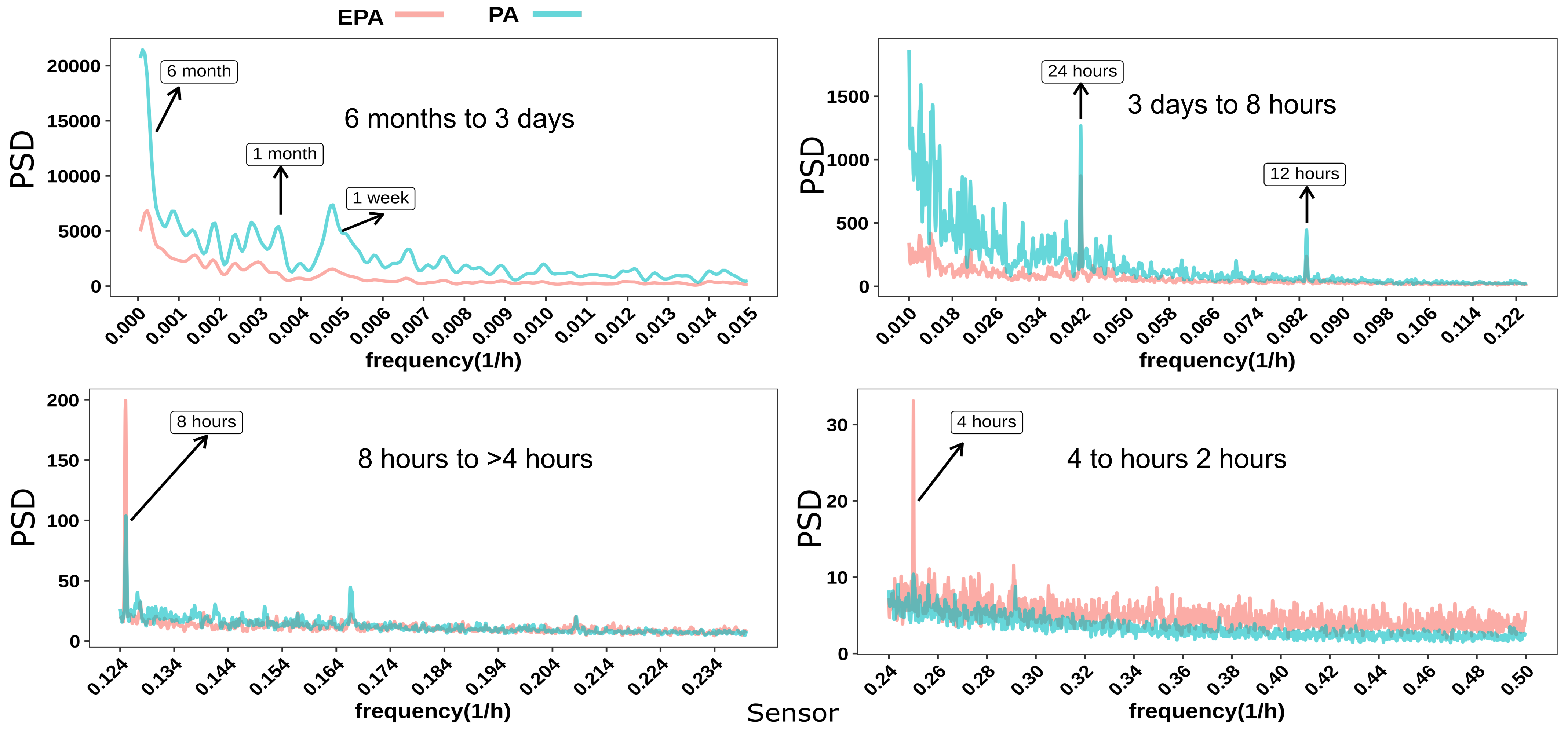
$$\Phi(v_k) = |X(k)|^2 = \left| \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} X_t e^{-2\pi i v_k t} \right|^2$$

$where\ k = 0, 1, \ldots, (N-1).\ N\ is\ the\ number\ of\ observations\ and\ v_k = \frac{k}{N}$
$and\ X(k)\ is\ discrete\ Fourier\ transform\ of\ hourly\ time\ series\ X_t$

[1] Hies, 2000 [2] Marr and Harley 2002 [3] Choi, 2008 [4] Tchepel and Borrego, 2010 [5] Castell, 2017

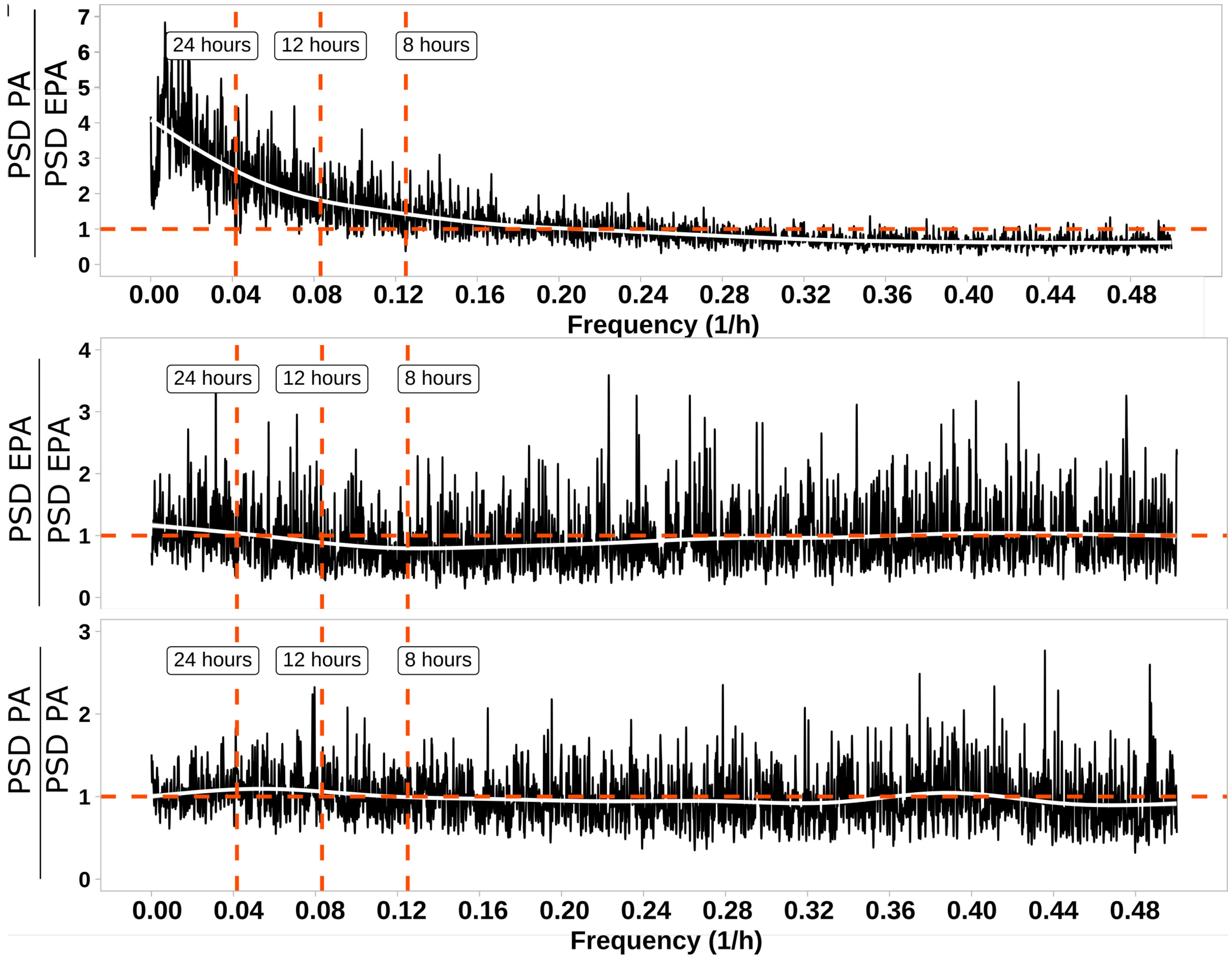
Comparison of EPA and PA power spectral density (PSD)

- The PSD of EPA and PA data from each site was averaged and plotted
- The PSD were divided into different frequency ranges for easier comparison



EPA and PA power spectral density (PSD) ratio

- EPA sites show larger PSD peaks for variations in the time-scale of 8 hours and less



Decomposition of data: Kolmogorov–Zurbenko (KZ) filter

Time series data can be decomposed into different components using the KZ filter [1]

Let $\{\mathrm{X(t)}\}, \mathrm{t} = 0, \pm 1, \pm 2, \ldots$ be a real − valued time series
The KZ filter with parameters moving **window** $(\mathbf{m})$ and **iterations** $(\mathbf{p})$ is defined as

$$\mathrm{KZ}_{\mathrm{m},p}[X(t)] = \sum_{\mathrm{j}=\frac{-p(m-1)}{2}}^{\frac{p(m-1)}{2}} X(t+j)\ \mathrm{a}_{\mathrm{j}}^{m,p}$$

Where coefficients $\mathrm{a}_{\mathrm{j}}^{m,p} = \frac{\mathrm{c}_{\mathrm{j}}^{m,p}}{m^p}, j = \frac{-p(m-1)}{2}, \ldots, \frac{p(m-1)}{2}$ are given by the polynomial coefficients obtained

$$\sum_{\mathrm{r}=0}^{p(m-1)} \mathrm{z}^{\mathrm{r}} \mathrm{c}_{\mathrm{r-p(m-1)}}^{m,p} = (1 + z + \ldots + z^{m-1})^p$$

[1] Rao and Zurbenko, 1994; Rao, 1997

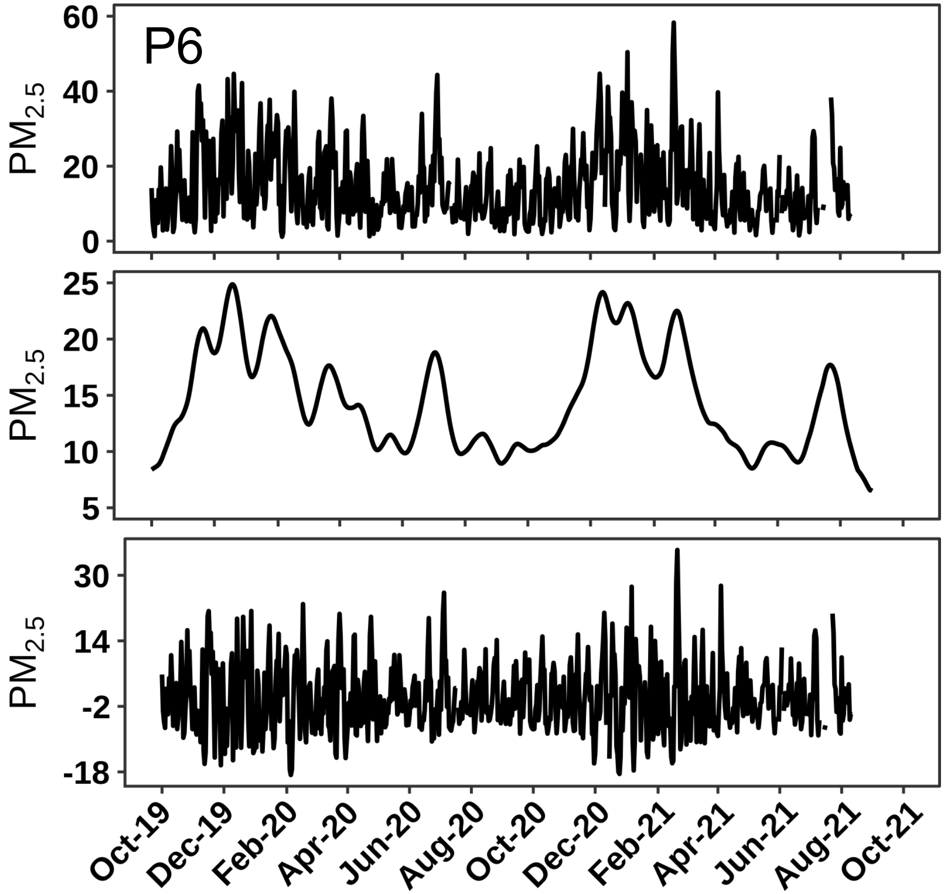
Decomposition of data: Kolmogorov–Zurbenko (KZ) filter

- For our study we have used $KZ_{15,3}$ which was designed by [1] and used for $PM_{2.5}$ [2]
- The short-term fluctuations are due to local temporal sources; traffic, and weather [1,2,3]
- Baseline component is the effect of long-term emissions and climate changes [1,2,3]

Total time series

Baseline component
$PM_{Baseline}(t) = KZ_{15,3}PM_{2.5}(t)$

Short-term component
$PM_{2.5,\,Short-term}(t) =$
$PM_{2.5}(t) - PM_{2.5,\,Baseline}(t)$



[1] Rao, 1994, 1997, [2] *Bai*, 2022, [3] Wise and Comrie, 2005

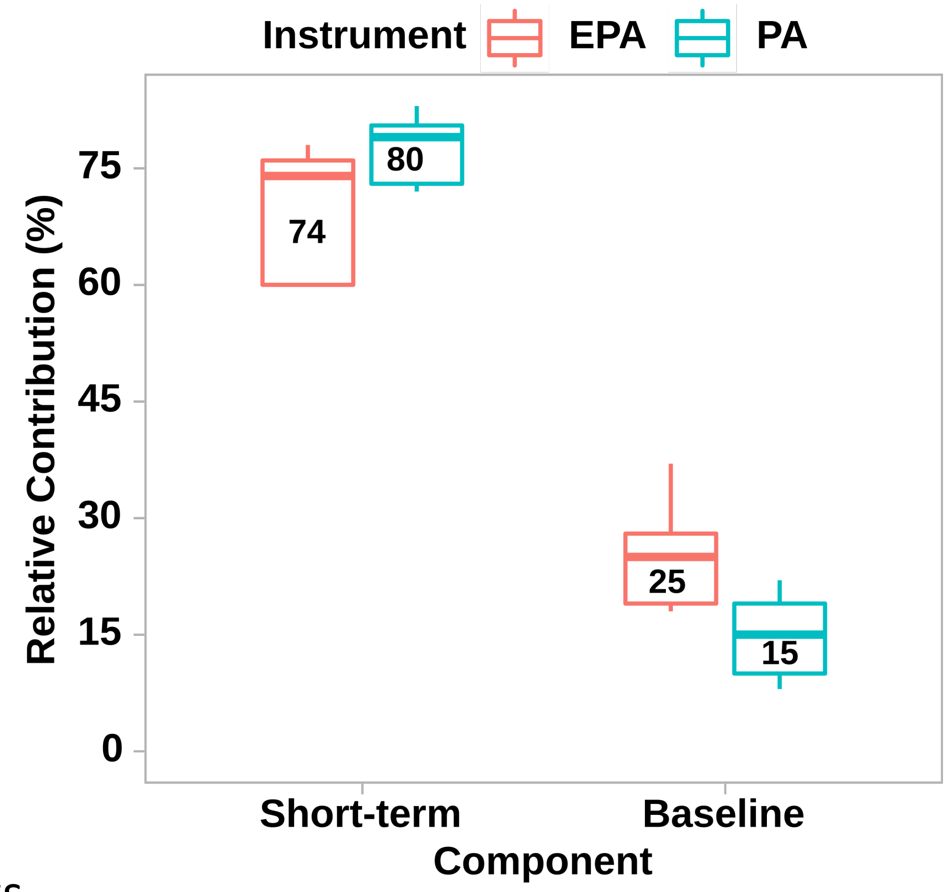
Relative contribution of temporal components on $PM_{2.5}$ data

The relative contribution of temporal components

$$Relative\ contribution(\%) = \frac{Var(PM_{2.5,\,component}(t))}{Var(PM_{2.5}(t))} \times 100$$

where $PM_{2.5,\ component}$ (t) is either

short-term or baseline component of $PM_{2.5}$



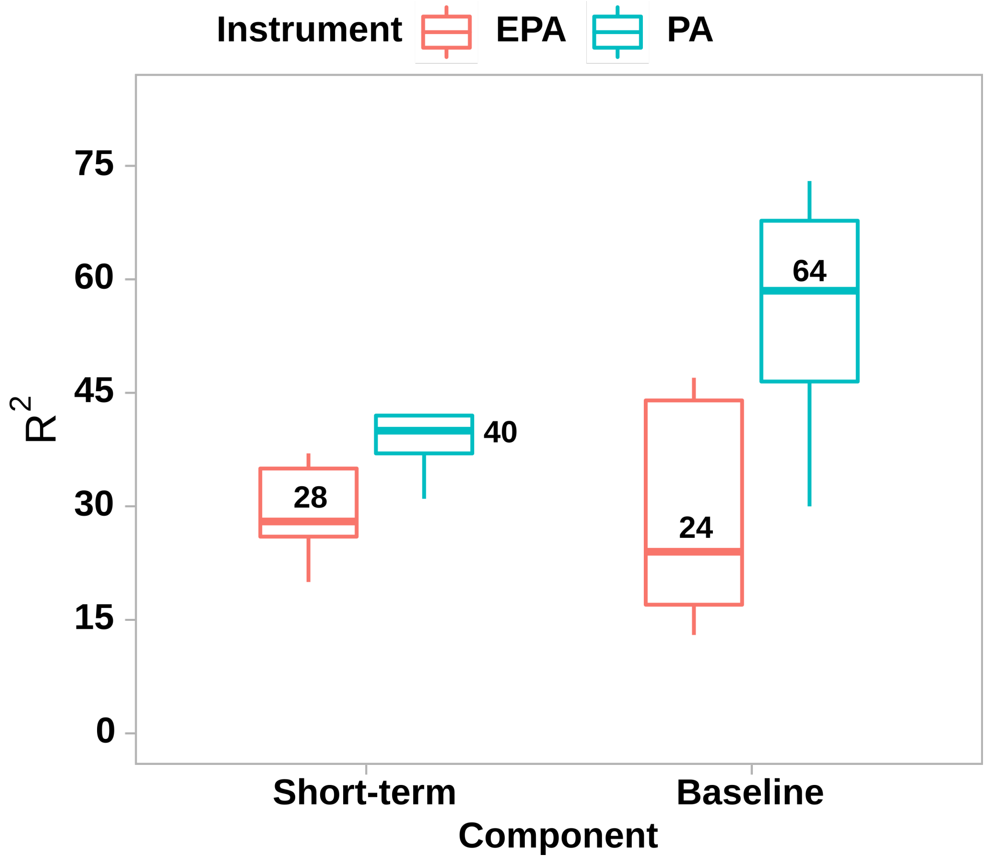
- PA sensors in short-term as well as baseline component have more relative contribution than EPA
- This indicates PA data has more variations due to source components

[1] Botlaguduru, 2018

$PM_{2.5}$ contributions from meteorology and anthropogenic emissions

M(t)=Meteorological variable

1. RH= Relative humidity (%)
2. T= Temperature (°C)
3. WS = Wind speed (miles per hour)
4. WD = Wind direction (degree from true north)
5. P= Air pressure (inches of mercury)
6. V= Visibility (horizontal distance an object in mile)



- PA sensors in short-term, as well as long-term component have more R^2 than EPA
- This shows PA sensors are more influenced by meteorological conditions

[1] Bai, 2022 [2] Li, 2019 [3] Zhai, 2019

Relative importance of predictors: Lindeman, Merenda, and Gold (LMG)

This method was first proposed by Lindeman, Merenda, and Gold (1980) known as LMG

Variance decomposition in linear regression

The formulas in this section are taken from [1]

$$E(Y \mid X) = X\beta \quad (1)$$

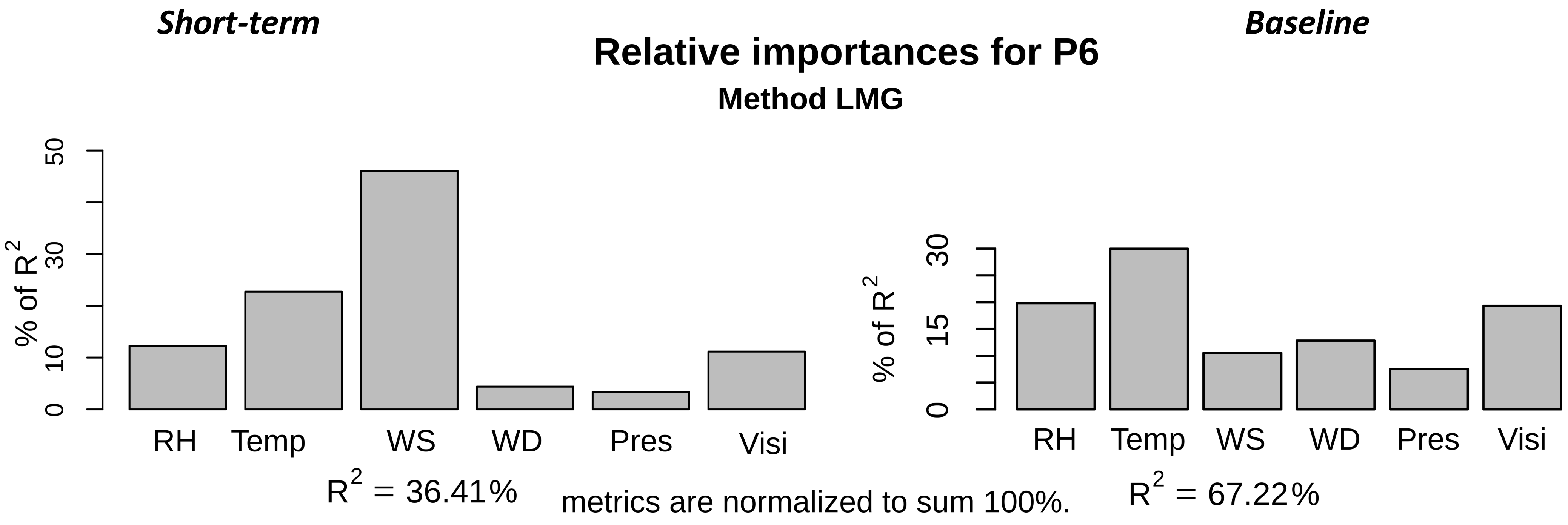
$$\mathrm{Var(Y)} = \sum_{j=1}^{p} \beta_j^2 v_j + 2 \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \beta_j \beta_k \sqrt{v_j v_k} \rho_{jk} + \sigma^2 \quad (2)$$

v_j is variance of each regressor , j=1,2,…,p, ρ_{jk}= covariance of regressors, k=j +1,…,p

$$LMG(x_k) = \frac{1}{p!} \sum_{r\ permutations} seq\ R^2\ (\{x_k\}|r) \quad (5)$$

[1] Grömping, 2015

Relative importance of predictors: $LMG(x_k)$



- The WS is the most influential factor in most PA sensors' short-term component
- In rest of PA sensors WS is 2nd influential factor where is first important factor is temperature or RH
- In baseline temperature/RH is consistently important factor of PA sensors and in EPA half sites as well
- Karoline K. Barkjohn, 2021, Iasonas Stavroulas,2020 found that RH is important correction factor for PA sensors

Summary

- The PA sensor data is sensitive to meteorological conditions
- Especially wind speed, temperature, and relative humidity
- Wind speed must be used for correction of PA data
- Corrections should be performed in different components; short-term, and baseline

Thank You!