Evaluation and modeling of data from low-cost air quality sensors for accurate $PM_{2.5}$ estimation

Dinushani Senarathna[1], **Vijay Kumar**[1], **Supraja Gurajala**[2], **Suresh Dhaniyala**[3], **Shantanu Sur**[4], **Sumona Mondal**[1]

[1]Department of Mathematics, Clarkson University, [2]Department of Computer Science, Suny Potsdam University, [3]Department of Mechanical Engineering, Clarkson University, [4]Department of Biology, Clarkson University

Background

Motivation:

- Low-cost sensors are producing high resolution spatio-temporal data
- Low –cost sensors data can be noisy and low in accuracy[1][2] and sensitive to weather parameters
- Correction models are built to correct $PM_{2.5}$ form PurpleAir (PA) sensor data using EPA measurements as gold-standard[1][2]
- Existing models are applicable US wide and use relative humidity (RH) for correction [2]

Figure 1: PA sensor deployed for AQ measurement

Research objectives:

- Build and investigate PA $PM_{2.5}$ correction model performance as function of distance
- Evaluate the model performance based on multiple PA sensors vs single PA sensor

Methodology

- Data source: Cook county, Illinois, USA; 2019 August to 2020 July from EPA and PA
- Prediction models were built using PA-measured temperature (T) and relative humidity (RH) as correction factors
- Prediction accuracy of models built using single PA sensor and multiple PA sensor data were compared
- Determined the effect of distance between PA and EPA sites on model performance
- Identified the optimal number of PA sensors needed for accurate prediction

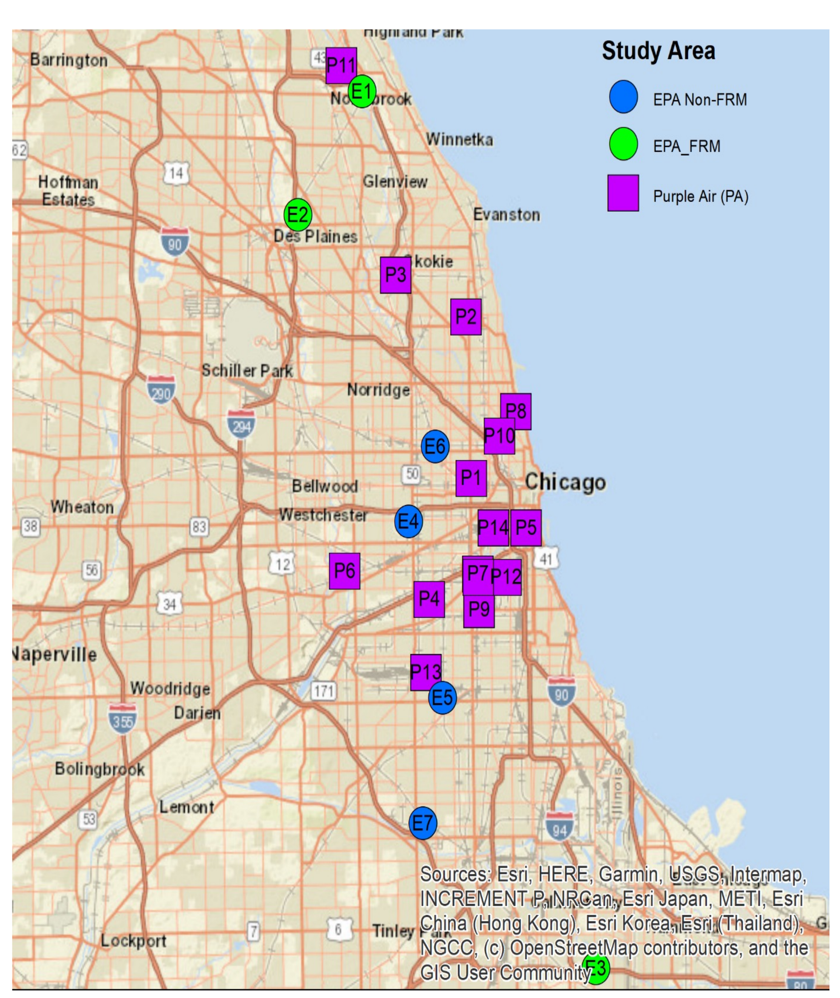


Figure 2: Location of 15 PA sensors (purple) and 7 EPA air monitoring sites including 3 FRM/FEM (green), and 4 non-FRM/FEM (blue) in Cook County, Chicago, Illinois, USA.

Results

$PM_{2.5}$ data distributions:

- Median values of $PM_{2.5}$ from PA sensors (10.9) and EPA (8.7) are shown by the dashed-lines

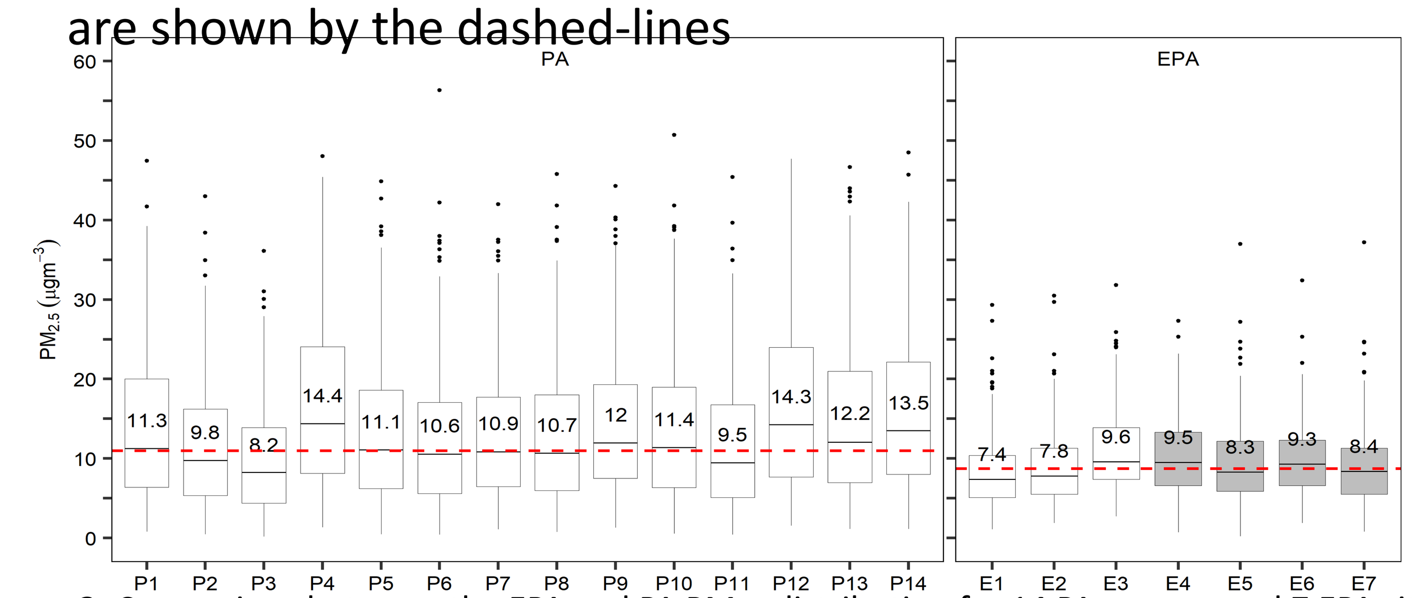


Figure 3: Comparison between the EPA and PA $PM_{2.5}$ distribution for 14 PA sensors and 7 EPA sites

Effects of temperature and relative humidity on model accuracy:

$PM_{2.5\,(EPA)} = \beta_0 + \beta_1 PM_{2.5\,(PA)} + \beta_2 RH_{(PA)} + \beta_3 T_{(PA)}$ where $\beta_2, \beta_3 = 0$ (None), $\beta_3 = 0$ (RH), $\beta_2 = 0$ (T)

- **RH** has a stronger influence in models when FRM data were used
- **T** has a stronger influence in models when non-FRM data were considered

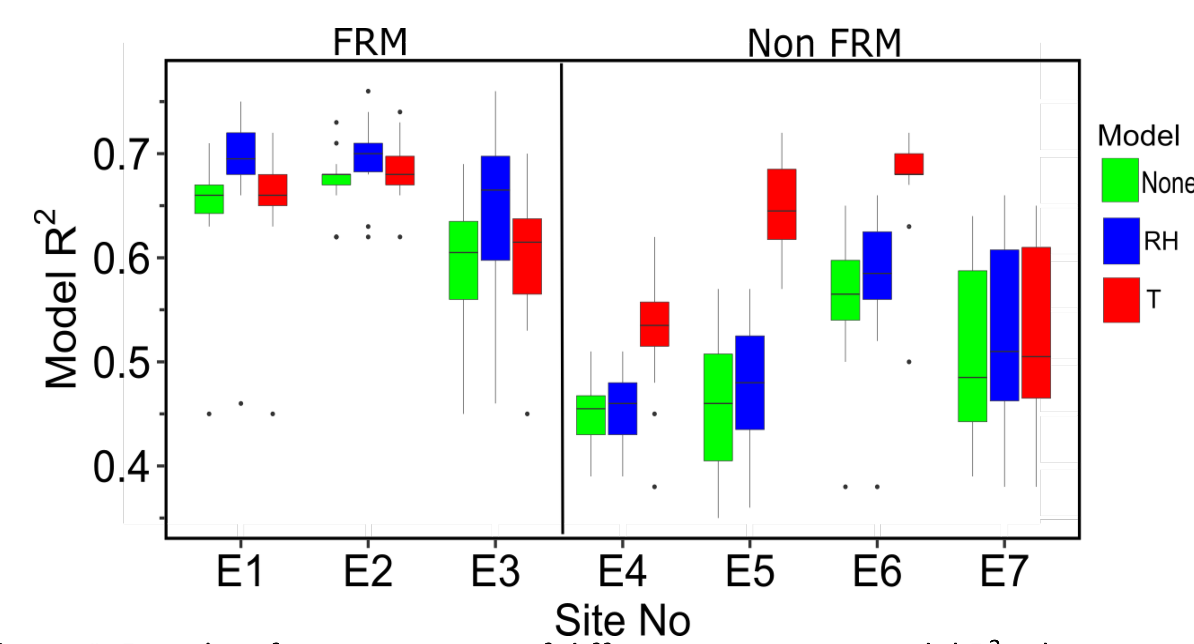


Figure 4: Box plots for comparisons of different correction model R^2 values

Effect of distance on model prediction accuracy:

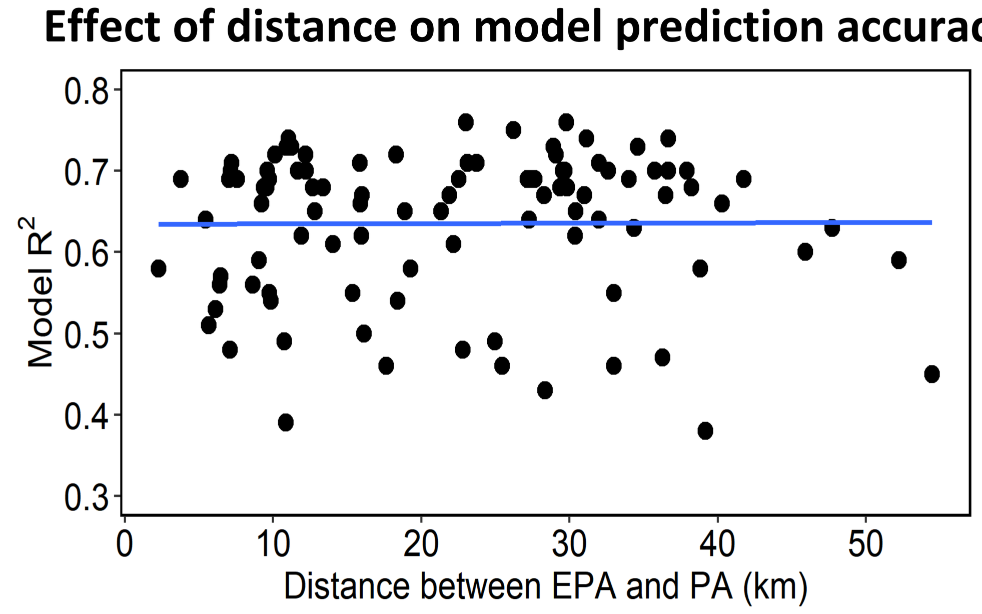


Figure 5: Model R^2 values plotted against distances between PA and EPA sites

- When consider overall, no effect from distance for the model R2 values

Changes of model accuracy across various distance groups:

- Distances were divided into 4 groups (0-10, 10-20, 20- 30 and > 30 km)
- Order restricted inference[10] applied for model R^2 with distance groups

Hypotheses

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs. $H_1: \mu_1 \geq \mu_2 \geq \mu_3 \geq \mu_4$

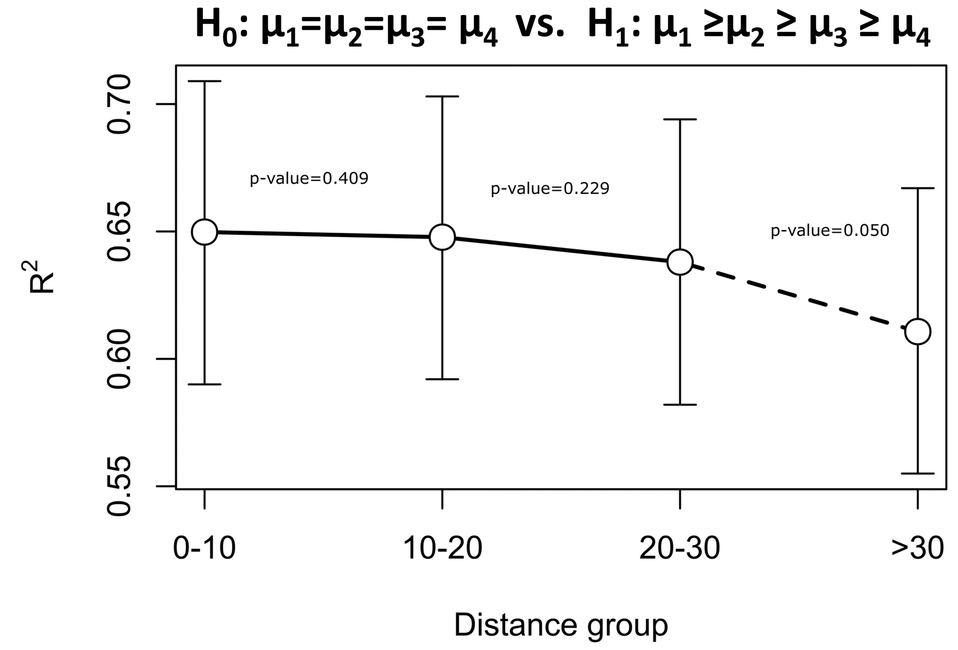


Figure 6: Plot of mean model R^2 under each distance groups

- Global test: no deceasing trend in model R^2 values were observed with increasing distances
- Pairwise analysis: model accuracy decreased for groups when distance is > 30 km

Results

Correction models built with multiple PA sensors:

$PM_{2.5\,(EPA)} = \beta_0 + \beta_1 T_{(PA)} + \beta_2 RH_{(PA)} + \sum_{i=3}^{7} \beta_i PM_{2.5\,(PAj)}$, where j = 1,..,5

- PA sensors within 30 km for each EPA site with all possible combination of 2, 3, 4, 5 PA sensors were considered
- Considering > 3 PA sensors in the model results in minimal improvement of model performance.

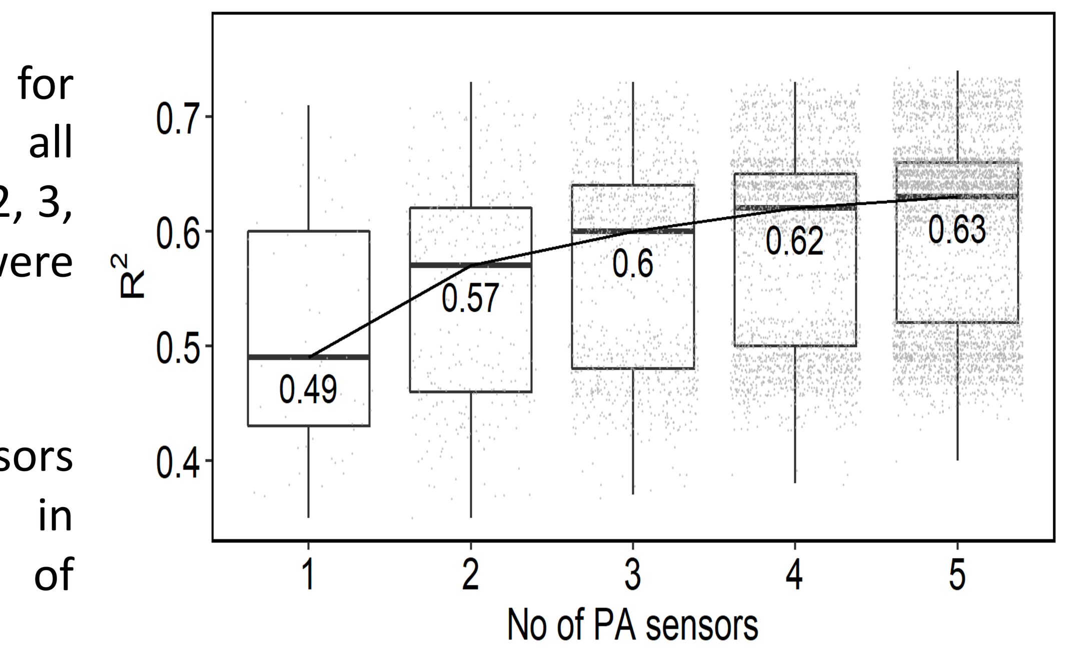
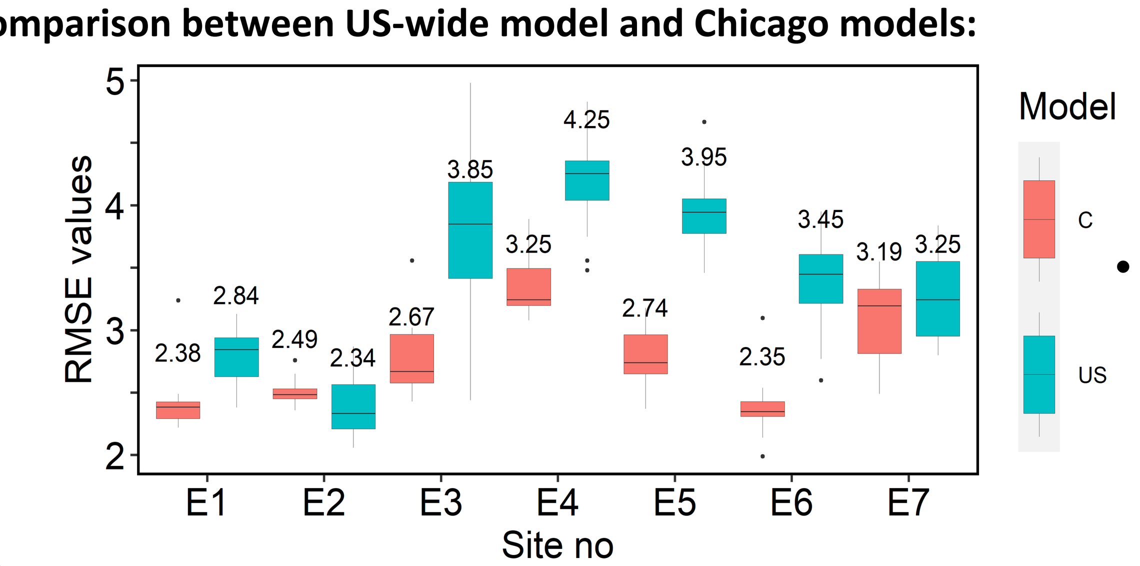


Figure 7: Box plots for multiple PA sensor model R^2 values

Comparison between US-wide model and Chicago models:



- Local models have a lower RMSE values compared with the US-wide model

Figure 8: Box plots of RMSE values for local Chicago models (C) and US-wide models (U)

Conclusion & Future work

- Relative humidity and temperature provide more accurate prediction for FRM and non-FRM data, respectively
- Model R^2 values decrease significantly when the distance between EPA and PA sensors are > 30 km
- Models using multiple PA sensors performed better than using a single PA sensor, however, improvement was minimal for more than >3 PA sensors
- Consideration of additional parameters such as wind speed and wind direction might help to obtain higher model accuracy

References

1. Ardon-Dryer, K. (2020), Measurements of $PM_{2.5}$ with Purple Air under atmospheric conditions, Atmos. Meas. Tech., 13, 5441–5458, 2020, https://doi.org/10.5194/amt-13-5441-2020
2. Barkjohn, K, K (2020), Development and Application of a United States wide correction for $PM_{2.5}$ data collected with the Purple Air sensor, https://doi.org/10.5194/amt-2020-413
3. Feenstra, B.(2019), Performance evaluation of twelve low-cost $PM_{2.5}$ sensors at an ambient air monitoring site, Atmospheric Environment, https://doi.org/10.1016/j.atmosenv.2019.116946
4. Datta, A. (2020),Statistical field calibration of a low-cost $PM_{2.5}$ monitoring network in Baltimore, Atmospheric Environment, https://doi.org/10.1016/j.atmosenv.2020.117761
5. Liu, H, Y. (2019), Performance Assessment of a Low-Cost $PM_{2.5}$ Sensorfor a near Four-Month Period in Oslo, Norway, http://dx.doi.org/10.3390/atmos10020041
6. Kelleher, S (2018), A low-cost particulate matter ($PM_{2.5}$ monitor for wildland fire smoke, Atmos. Meas. Tech., 11, 1087–1097, 2018, https://doi.org/10.5194/amt-11-1087-2018
7. Jayaratne, R (2018), The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog, https://doi.org/10.5194/amt-11-4883-2018
8. Hopke P. K,(2018), Hourly land-use regression models based on low-cost PM monitor data, Environmental Research 167(2018) 7-14, https://doi.org/10.1016/j.envres.2018.06.052
9. He, M., Kuerbanjiang, N., and Dhaniyala, S.: Performance characteristics of the low-cost Plantower PMS optical sensor, Aerosol Science and Technology, 54, 232–241, 2020, https://doi.org/10.1080/02786826.2019.1696015
10. Jelsema, C. M., & Peddada, S. D. (2016). CLME: An R Package for Linear Mixed Effects Models under Inequality Constraints. *Journal of Statistical Software*, *75*(1), 1–32. https://doi.org/10.18637/jss.v075.i01