IMPROVING POLLUTION SOURCE RESOLUTION FOR REAL TIME LOW COST SENSORS USING WIDELY AVAILABLE DATA RESOURCES

A PROOF OF CONCEPT

Presented: Air Sensors International Conference, Oakland, CA USA, Sep 12, 2018,

Edited 9/26/2018 for use as a circulated doc, rather than a presentation

L. Drew Hill (presenting), Ramboll, Environment & Health, San Francisco, USA

Ajay Pillarisetti, Division of Environmental Health Sciences, University of California, Berkeley, USA

Kirk R Smith, Division of Environmental Health Sciences, University of California, Berkeley, USA

Shari Libicki, Ramboll, Environment & Health, San Francisco, USA

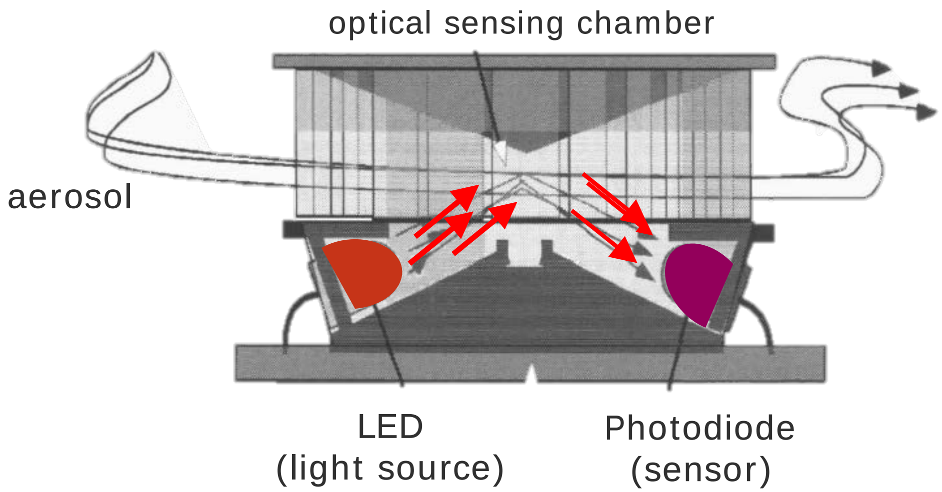
RAMBOLL IN BRIEF

- **Independent engineering and design consultancy** and provider of management consultancy
- Founded 1945 in Denmark
- 14,000 experts
- Close to 300 offices in 35 countries
- Particularly strong presence in the Nordics, the UK, North America, Continental Europe, Middle East and Asia Pacific
- Owned by Rambøll Fonden

Services across the markets:

- Buildings
- Transport
- Planning & Urban Design
- Water
- **Environment & Health**
- Energy
- Management Consulting

WHAT AFFECTS THE RELATIONSHIP BETWEEN SENSOR READINGS AND ACTUAL CONCENTRATIONS? (PM2.5, OPTICAL)



adapted from Litton et al 2004

Inter-device hardware inconsistencies

Environmental factors, cross-sensitivity

- Temperature, relative humidity

Aerosol properties

- Distributions of size and shape
- Aerosol refractive index
- Particle density

RAMBOLL

Berkeley School of Public Health

PROOF OF CONCEPT – METHODS

Machine Learning (ML)

- Very good at uncovering, assessing hidden and complex relationships
- Until very recently, the domain of mathematicians and computer scientists
- Computing advances, open source programming have made ML and Ensemble methods accessible to (more of) the general public
- One of the most important aspects of ML: picking the right variables
- ML is now the domain of subject matter experts (like us!) who have the ability to anticipate good predictors, related proxy variables

INFLUENCES OF LOCAL AEROSOL PROPERTIES, SENSOR OUTPUT

Specific makeup of local point, area sources

Traffic

- Time of day: Fraction of total ambient aerosols coming from mobile vs. point sources
- Ratio of diesel to non-diesel
- Ratio of clunkers to … not clunkers

Environmental phenomena, like wild fires

- Intermittent source
- Produce aerosols of size, shape, refractive index different from those of common urban sources

Meteorology

- Wind direction, speed
  - Regional and local transport
  - Determines upstream sources, dilution
- Precipitation, fog
- Air pressure

RAMBØLL

Berkeley School of Public Health

WHAT INFLUENCES THESE FACTORS?

Highly complex interrelationships

Specific makeup of local point, area sources

Traffic

- Time of day: Fraction of total ambient aerosols coming from mobile vs. point sources
- Ratio of diesel to non-diesel
- Ratio of clunkers to … not clunkers

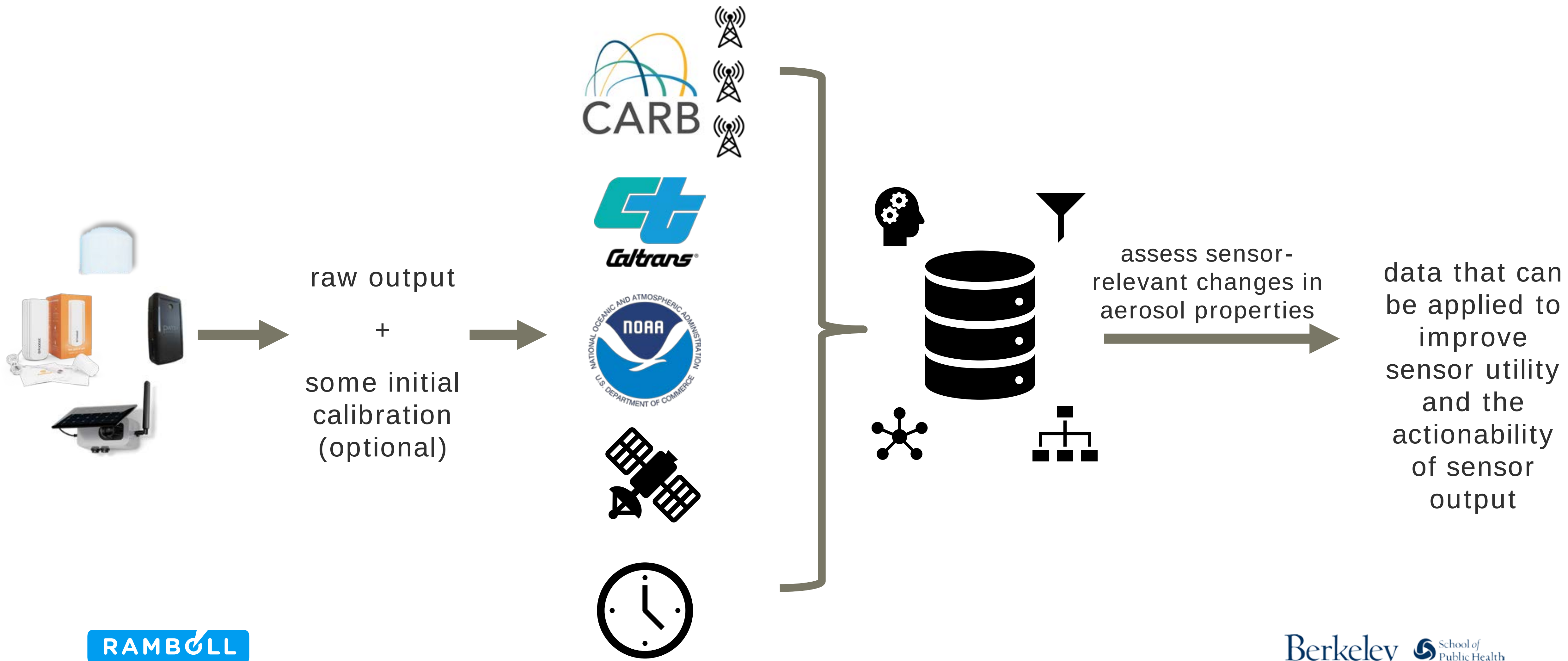
Environmental phenomena, e.g. forest fires

- Intermittent source
- Produce aerosols of size, shape, refractive index different from those of traffic, industrial sources

Meteorology (regional and local transport)

- Wind direction, speed
  - Determines upstream sources, dilution
- Precipitation, fog

USE PUBLIC DATA SOURCES, ADVANCED STATISTICS TO ASSESS AND EXPLOIT CHANGES IN THESE FACTORS RELEVANT TO SENSOR RESPONSE

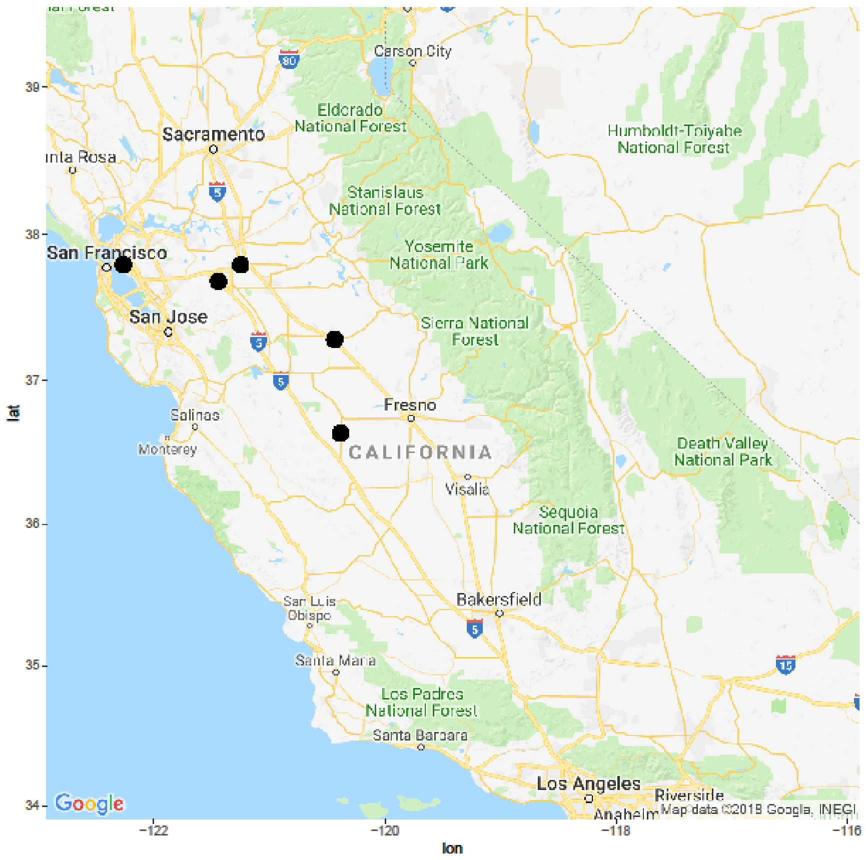


RAMBOLL

Berkeley School of Public Health

PROOF OF CONCEPT – METHODS

- Plantower sensor data (5 min.) from 5 Clarity Node devices throughout N. California, provided by Clarity
  - Concentration estimates of PM10, PM2.5, PM1.0; temperature; relative humidity
- Collocated with regulatory-grade monitors February – August 2018



RAMBOLL

Berkeley School of Public Health

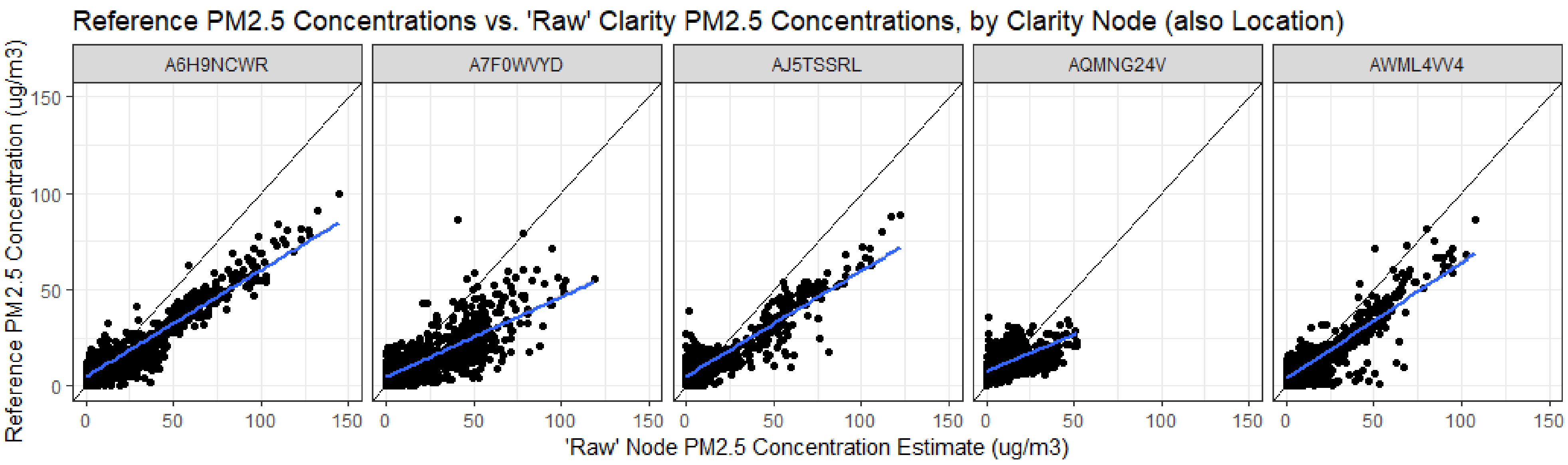
PROOF OF CONCEPT – METHODS

Mean$_{Ref}$: 9.0 ug/m3

σ$_{Ref}$: 8.1 ug/m3

Mean$_{CN_raw}$: 7.6 ug/m3

σ$_{CN_raw}$: 13.0 ug/m3



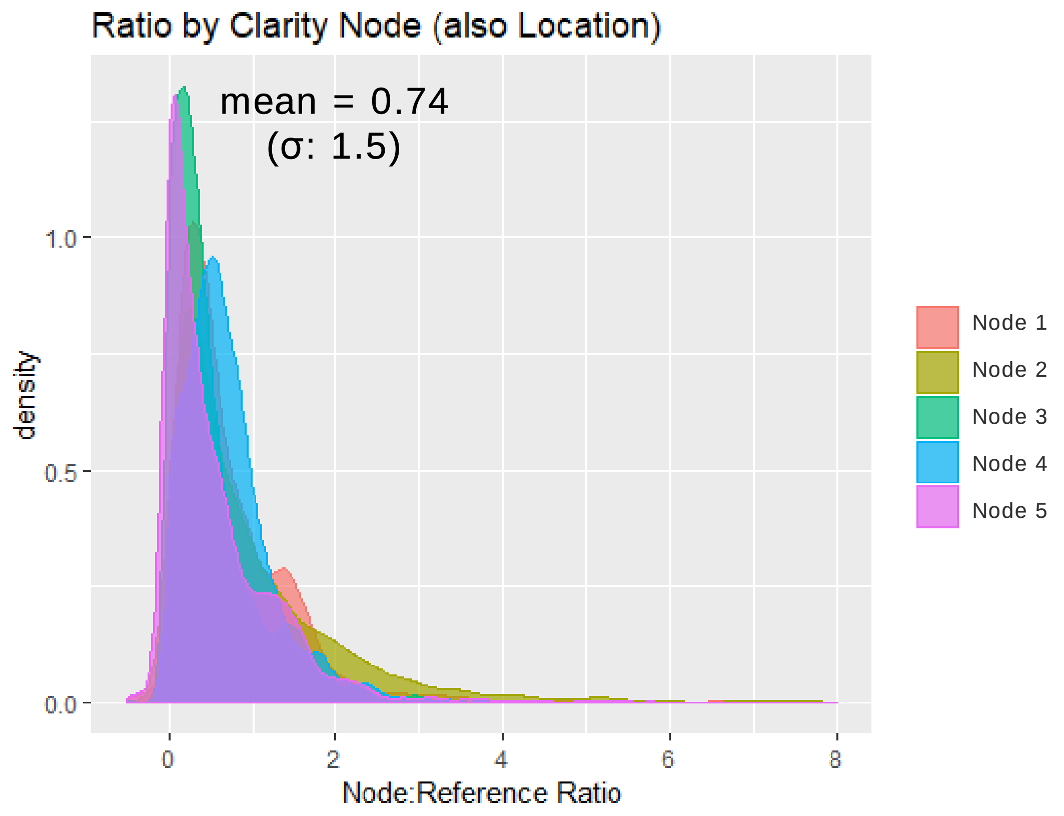
Reference = 5.0 +0.52('Raw' Sensor Estimate)

PROOF OF CONCEPT – PM2.5 DATA SUMMARY

(uncalibrated) Clarity Output : Reference, by unit

- Variation within units over time
- Variation between units

Overall, the ratio observed is not steady over the assessment period (σ: 1.5)



PROOF OF CONCEPT – METHODS

Concurrent data collected from publicly accessible sources:

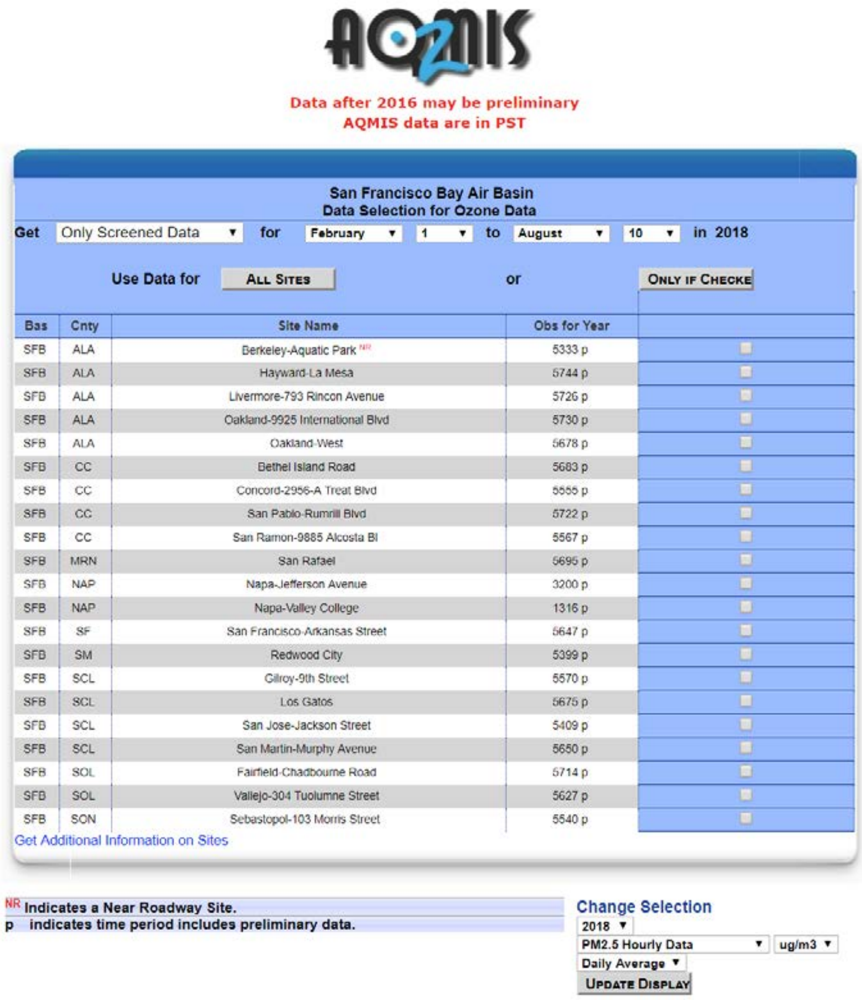
- Meteorology (3 closest NOAA ISD-listed stations to each location)

```r
# ISD for local met station (HOURLY resolution)
# Find monitors near a station -- takes about 60 seconds
#### Note ISD time appears to be in UTC.
library(rnoaa)

dt_isd_stations <- data.table('ref_name'= NA,'usaf'= NA,'wban'= NA,'icao'= NA
,'distance'= NA,'latitude'= NA,'longitude'= NA,'elev_m'= NA)

for (i in c(1:nrow(coordinates)) ){
    stations= as.data.table(isd_stations_search(lat = coordinates$latitude[i],
                     lon = coordinates$longitude[i],
                     radius= 75)) #km
    # take only those with data reaching into 2018
    stations_recent = stations[substr(end,0,4) > 2017]
    # add ref_name
    stations_recent[,ref_name := coordinates$id[i]]
    # take top 3 and insert into dataframe
    dt_isd_stations <- rbind(dt_isd_stations[!is.na(ref_name)], stations_recent[
,.(ref_name,usaf,wban,icao,distance,latitude,longitude,elev_m)][1:3])
}

## note, closest stations are same for Tracy and Manteca
save(dt_isd_stations, file = 'data\\isd_stations.Rda')
```

PROOF OF CONCEPT – METHODS

Concurrent data collected from publicly accessible sources:

- Meteorology (3 closest NOAA ISD-listed stations to each location)
- Hourly average PM2.5 concentrations from BAAQMD, SJVAPCD sites (excluding those used in colocation)



Bas	Cnty	Site Name	Obs for Year	
SFB	ALA	Berkeley-Aquatic Park NR	5333 p	
SFB	ALA	Hayward-La Mesa	5744 p	
SFB	ALA	Livermore-793 Rincon Avenue	5726 p	
SFB	ALA	Oakland-9925 International Blvd	5730 p	
SFB	ALA	Oakland-West	5678 p	
SFB	CC	Bethel Island Road	5683 p	
SFB	CC	Concord-2956-A Treat Blvd	5555 p	
SFB	CC	San Pablo-Rumrill Blvd	5722 p	
SFB	CC	San Ramon-9885 Alcosta Bl	5567 p	
SFB	MRN	San Rafael	5695 p	
SFB	NAP	Napa-Jefferson Avenue	3200 p	
SFB	NAP	Napa-Valley College	1316 p	
SFB	SF	San Francisco-Arkansas Street	5647 p	
SFB	SM	Redwood City	5399 p	
SFB	SCL	Gilroy-9th Street	5570 p	
SFB	SCL	Los Gatos	5675 p	
SFB	SCL	San Jose-Jackson Street	5409 p	
SFB	SCL	San Martin-Murphy Avenue	5650 p	
SFB	SOL	Fairfield-Chadbourne Road	5714 p	
SFB	SOL	Vallejo-304 Tuolumne Street	5627 p	
SFB	SON	Sebastopol-103 Morris Street	5540 p	

https://www.arb.ca.gov/aqmis2/aqdselect.php

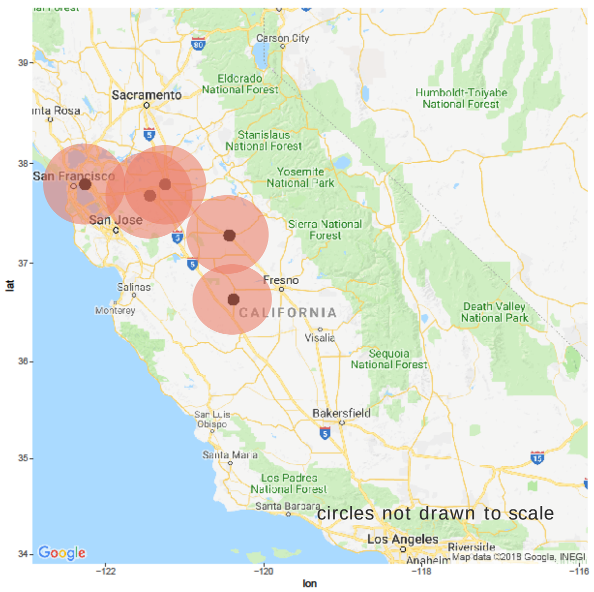
PROOF OF CONCEPT – METHODS

Wildfire Automated Biomass Burning Algorithm

http://www.ssd.noaa.gov/PS/FIRE/Layers/ABBA/abba.html

Concurrent data collected from publicly accessible sources:

- Meteorology (3 closest NOAA ISD-listed stations to each location)
- Hourly average PM2.5 concentrations from BAAQMD, SJVAPCD sites (excluding those used in colocation)
- Daily indicator of nearby wildfires (> mid-March)
  - ABBA, geosphere package (75 km radius)



RAMBOLL

Berkeley School of Public Health

PROOF OF CONCEPT – METHODS

Machine Learning (ML), Ensemble Methods

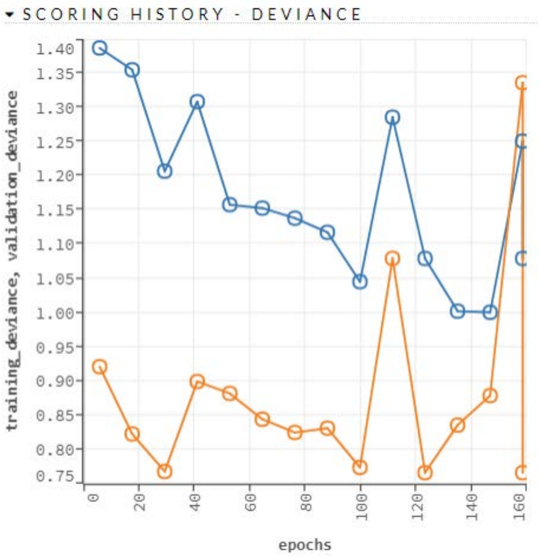
1. Deep Neural Net
   - Multi-layer, feed-forward perceptron
   - 18710 data points, 126 covariates (~ 2.4 million cells)
   - 90%/10% cross validation
2. A ensemble of
   - Random Forests
   - Support Vector Machines
   - GLM, GLM net
   - Ultimate sample size: 5586 data points, 66 covariates (~ 370,00 cells)
   - 10-fold cross validation

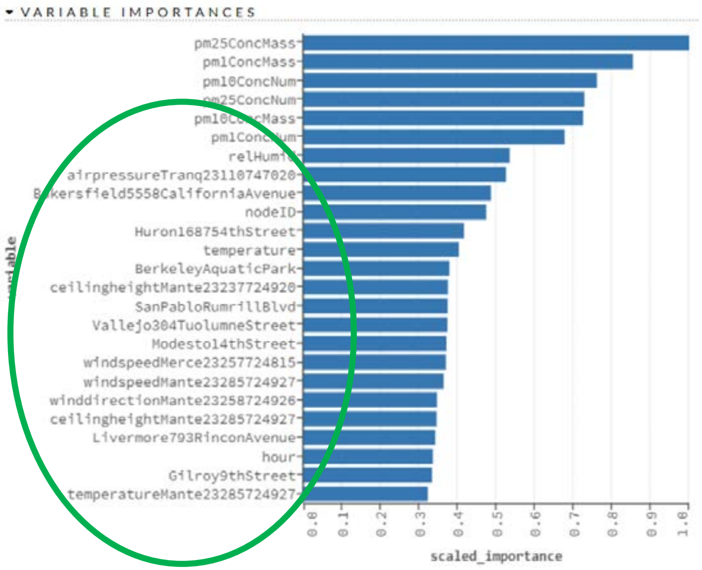
$$\varphi = \frac{\text{Raw Clarity PM2.5 Estimate (ug/m3)}}{\text{Reference PM2.5 Value (ug/m3)}}$$

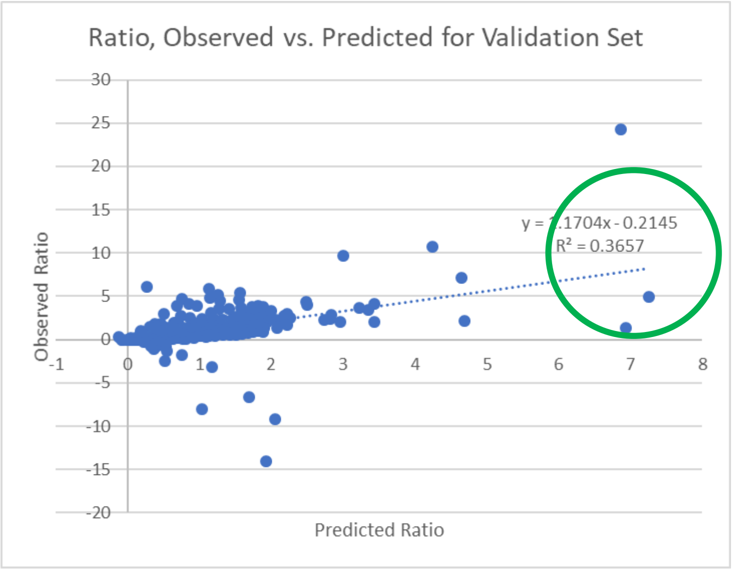
PROOF OF CONCEPT – RESULTS

- Deep Neural Network:
  - Moderate predictive power, well-fit, moderate error
  - Variable importance: nearby NOAA and regulatory monitor data show high importance

Mean φ observed	**Mean φ** predicted	**r^2** Obs. Vs. Pred	**β^2** Obs. Vs. Pred	**RMSE** validation	**RMSE** train
0.67 (σ: 1.1)	0.76	~ 0.35	1.17 *17% underestimation*	0.88	1.04







RAMBOLL

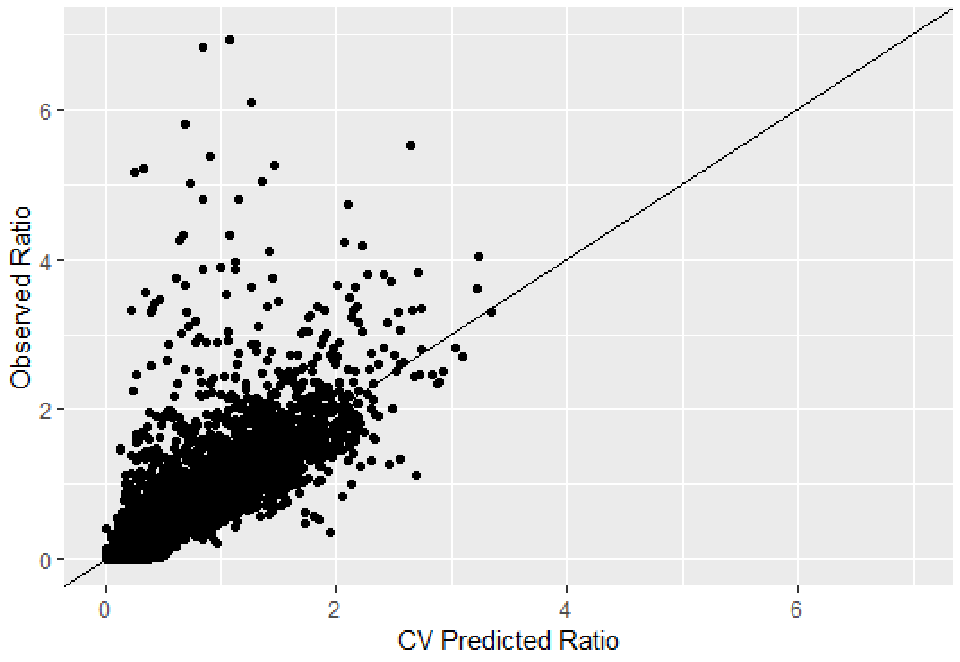
Berkeley School of Public Health

PROOF OF CONCEPT – RESULTS

- Ensemble (RF, SVM, GLM, GLM net):
  - Low bias, moderate error
    - Strongly predicted ratio as it changed
    - Thus, likely a strong predictor of changes in aerosol properties and potentially nearby source characteristics

Mean φ observed	Mean φ predicted	Ensemble Avg. RMSE	β Obs. Vs. Pred (obs < 7)	$adj\text{-}r^2$ Obs. Vs. Pred (obs < 7)
0.71 (σ: 1.4)	0.66 (σ: 0.51)	1.60	1.04 (SE: 0.01) *4% underestimation*	0.48

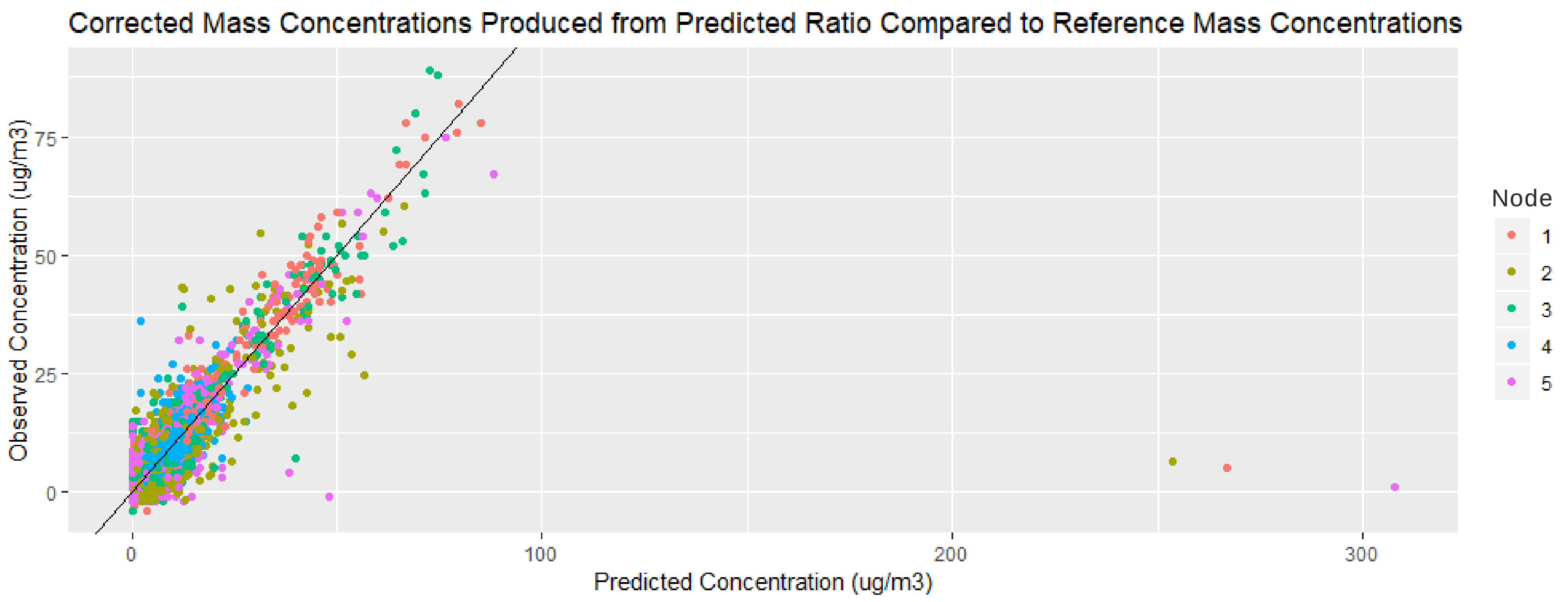
10-Fold CV-Predicted and Observed Clarity-Reference ratio



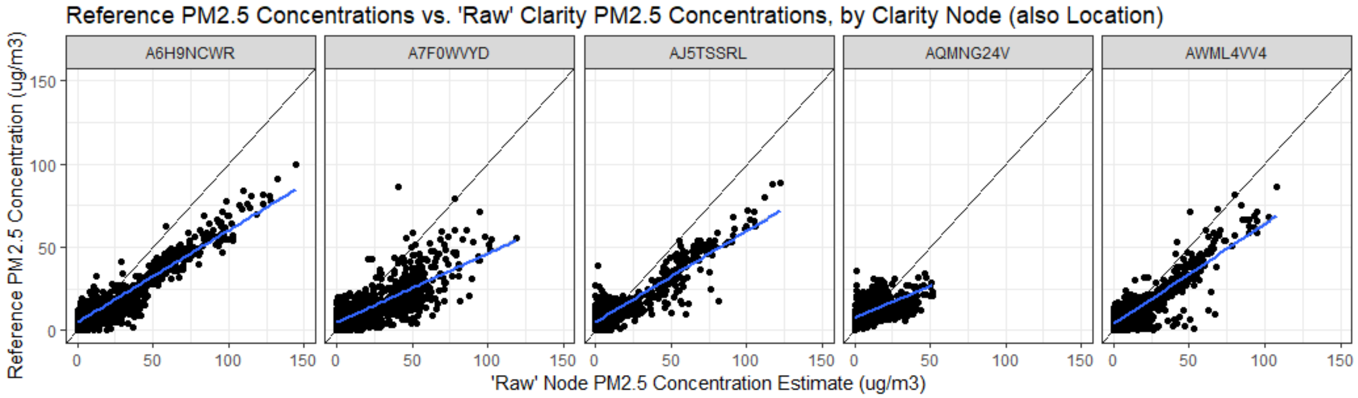
PROOF OF CONCEPT – RESULTS

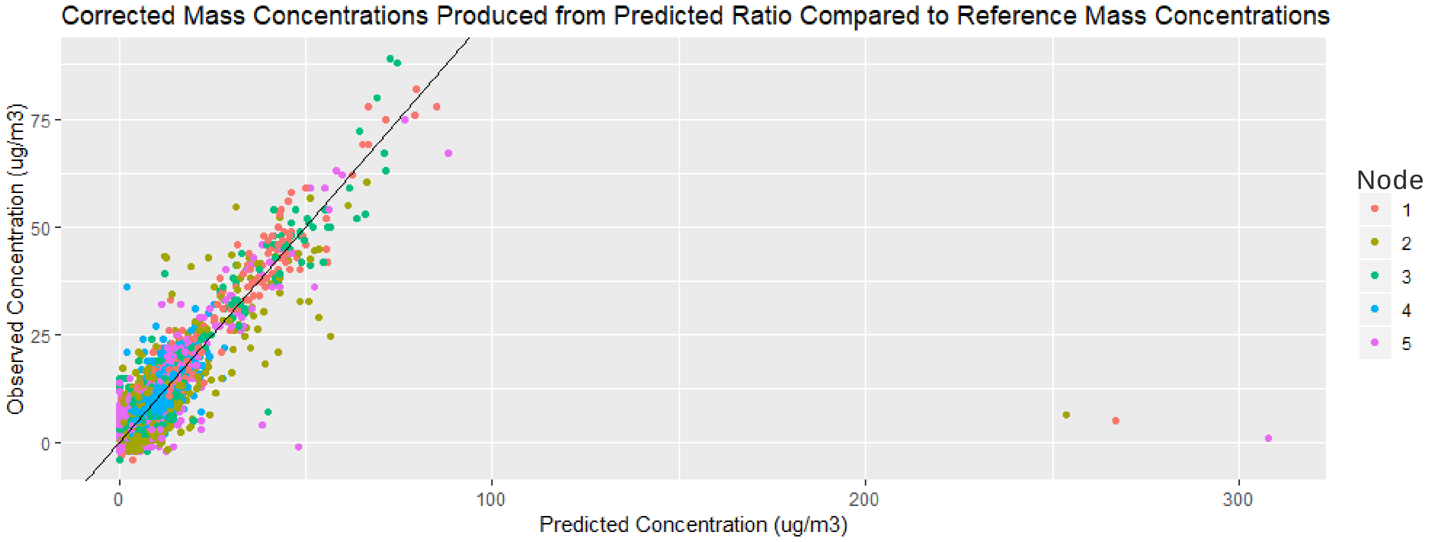
Ensemble (RF, GLM, GLM net, SVM):

- Ratios can be used to reliably produce estimates of true hourly average local PM2.5 mass concentrations
- Low bias across nodes, low/moderate error
- Ratio & Clarity output allowed reliable reconstruction of reference values
  - Better in some nodes than other



PROOF OF CONCEPT – METHODS





RAMBOLL

Berkeley School of Public Health

INSIGHTS, NEXT STEPS

Using publicly available data, a machine learning-enhanced statistical model can be trained to:

- strongly predict hourly changes in the relationship between sensor output and PM2.5 concentrations
  - Identify key changes in local pollution source contributions, important events
- account for location-based and inter-unit differences with good accuracy

Such a model leverages and highly relies upon local, sophisticated low-cost sensor output

- Clarity Node provides estimates of PM1 and PM10, allows model to consider changes in size distribution

Such a model can reliably produce estimates of true hourly average local PM2.5 concentrations

Future work should explore the ability of such a model to predict low-cost sensor calibration factors in near real-time (~ hourly)

Future models should explore local traffic data

REFERENCES

1. Institute for Health Metrics and Evaluation. 2018a. GBD Compare. Accessed: 9/9/2018. Available at: https://vizhub.healthdata.org/gbd-compare/
2. Institute for Health Metrics and Evaluation. 2018b. GBD Results Tool. Accessed: 9/9/2018. Available at: http://ghdx.healthdata.org/gbd-results-tool
3. Hill, LA. 2017. A Breath of Fresher Air: improving methods for PM2.5 exposure assessment from Mongolia to California. Dissertation, University of California, Berkeley. Available at: https://escholarship.org/uc/item/2bs6d62s
4. United States Environmental Protection Agency. 2016. Air Quality System data mart [monitor listing]. Accessed 3/14/2017. Available at: http://www.epa.gov/ttn/airs/aqsdatamart
5. Litton, C. D., Smith, K. R., Edwards, R., & Allen, T. (2004). Combined optical and ionization measurement techniques for inexpensive characterization of micrometer and submicrometer aerosols. *Aerosol Science and Technology*, *38*(11), 1054-1062.
6. Pillarisetti, A.; Allen, T.; Ruiz-Mercado, I.; Edwards, R.; Chowdhury, Z.; Garland, C.; Hill, L.D.; Johnson, M.; Litton, C.D.; Lam, N.L.; Pennise, D.; Smith, K.R. Small, Smart, Fast, and Cheap: Microchip-Based Sensors to Estimate Air Pollution Exposures in Rural Households. *Sensors* **2017**, *17*, 1879.
7. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
8. C. Agostinelli and U. Lund (2017). R package 'circular': Circular Statistics (version 0.4-93). URL https://r-forge.r-project.org/projects/circular/
9. Matt Dowle and Arun Srinivasan (2018). data.table: Extension of `data.frame`. R package version 1.11.4. https://CRAN.R-project.org/package=data.table
10. D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf
11. Robert J. Hijmans (2017). geosphere: Spherical Trigonometry. R package version 1.5-7. https://CRAN.R-project.org/package=geosphere
12. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
13. Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.
14. Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.
15. Eric Polley, Erin LeDell, Chris Kennedy and Mark van der Laan (2018). SuperLearner: Super Learner Prediction. R package version 2.0-24. https://CRAN.R-project.org/package=SuperLearner
16. Scott Chamberlain (2017). rnoaa: 'NOAA' Weather Data from R. R package version 0.7.0. https://CRAN.R-project.org/package=rnoaa
17. NOAA National Satellite Information Center: GOES Wildfire Automated Biomass Burning Algorithm. http://www.ssd.noaa.gov/PS/FIRE/Layers/ABBA/abba.html
18. AQMIS2. BAAQMD and SJVAPCD hourly PM2.5 data, Feb – Aug 2018. [Accessed August 2018]. https://www.arb.ca.gov/aqmis2/aqdselect.php
19. NOAA National Centers for Environmental Information (2001): Global Surface Hourly [ISD]. NOAA National Centers for Environmental Information. [Accessed August 2018] via R package “rnoaa”.

THANK YOU!

For providing the Node/FEM colocation datasets.

Collaborator **Shari Libicki** for good feedback on early drafts, and the organization for allowing me to utilize our resources to pursue this area of work.

Collaborators **Ajay Pillarisetti** and **Kirk Smith**, who've tolerated years of brainstorming and provided good comments.