COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

Acknowledgments: NIEHS R01 ES028805, R01 ES030616, P30 ES009089, and F31 ES030263

Principal Component Pursuit for Source Apportionment from Block Missing Data: An Application to NYC $PM_{2.5}$ Data

Lawrence Chillrud[1]*, Jingkai Yan[2], John Wright[2], Jeff Goldsmith[3], Marianthi-Anna Kioumourtzoglou[1]

Departments of [1]Environmental Health Sciences, [2]Electrical Engineering, [3]Biostatistics, Columbia University, New York, USA

*Correspondence: lgc2139@columbia.edu

Background

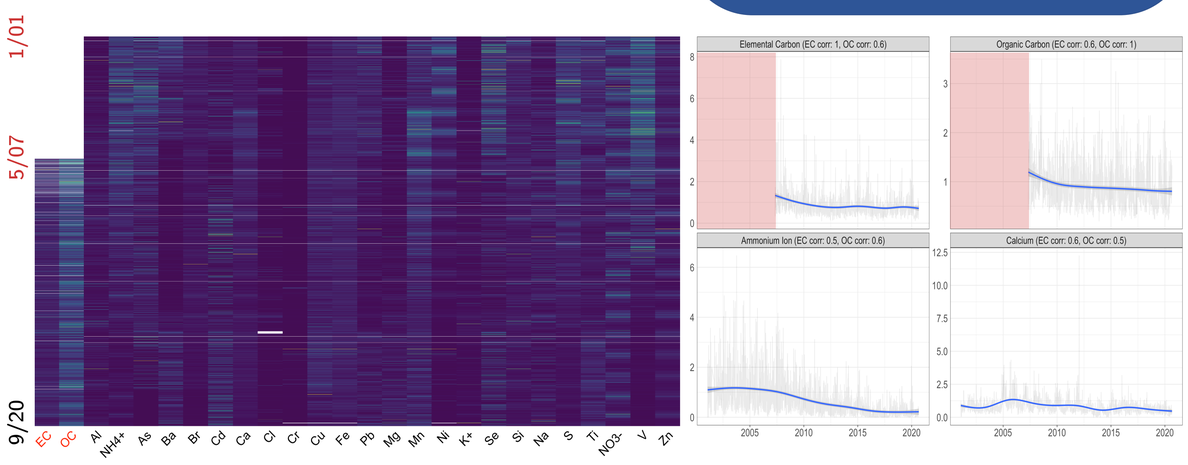
Linking identified patterns (e.g. sources) associated with adverse health outcomes could yield:

- Efficient policy/public health regulations
- Targeted interventions

Existing methods limited by:

- Subjective choice of k patterns
- Outliers may affect solution
- No standard for handling structured (block) missingness

Study aim: identify air pollution sources in NYC (2001-2020) & hindcast missing data for two $PM_{2.5}$ species (EC, OC)



EPA AQS PM2.5 data: NYC, 2001 – 2020

- 26 $PM_{2.5}$ chemical species measured across 2,378 days
- EC, OC missing all measurements from '01 – '07 (2.6% of overall mixture)

Methods

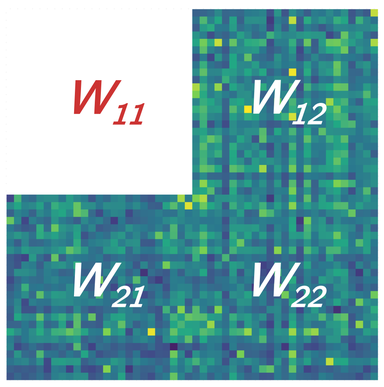
Principal Component Pursuit (PCP)

- Convex optimization algorithm from computer vision
- Dimension reduction by decomposing mixture into:

Original: environmental mixture = *Low-rank (L): consistent exposure patterns* + *Sparse (S): outlying exposure events*

Objective Fn: RRMC : $\min_{L,S} \mathcal{I}_{(L)\leq r} + \eta \|S\|_0 + \|L + S - D\|_F^2$

PCP & block missingness: how to reconstruct W_{11}?



Mixture w/missing block W_{11}

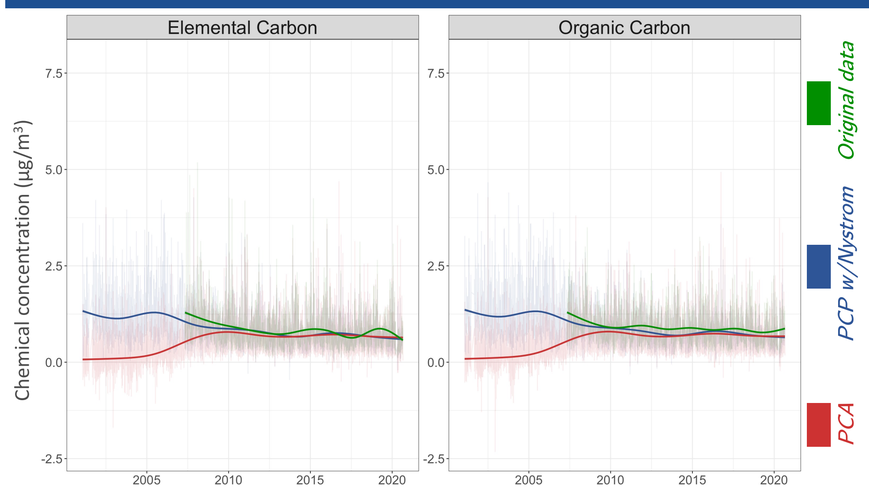
PCP's Nystrom extension:

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$$

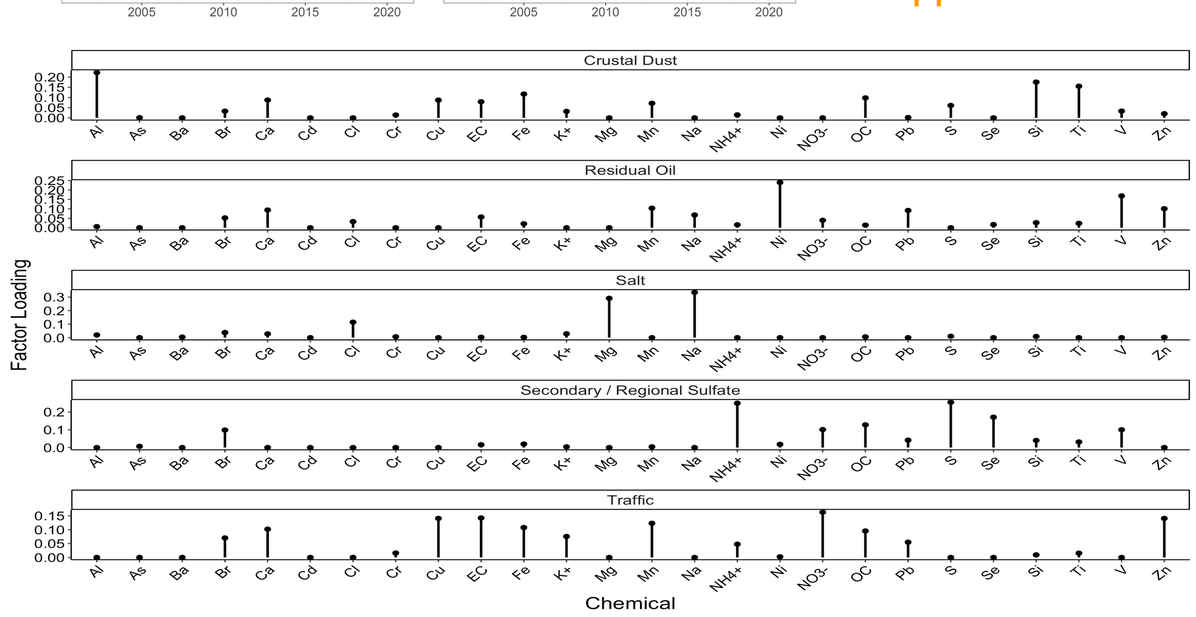
$$\widehat{W} = \begin{bmatrix} W_{12}[\mathcal{P}_r(W_{22})]^\dagger W_{21} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$$

Main idea: the missing block is reconstructed from observed data

Results

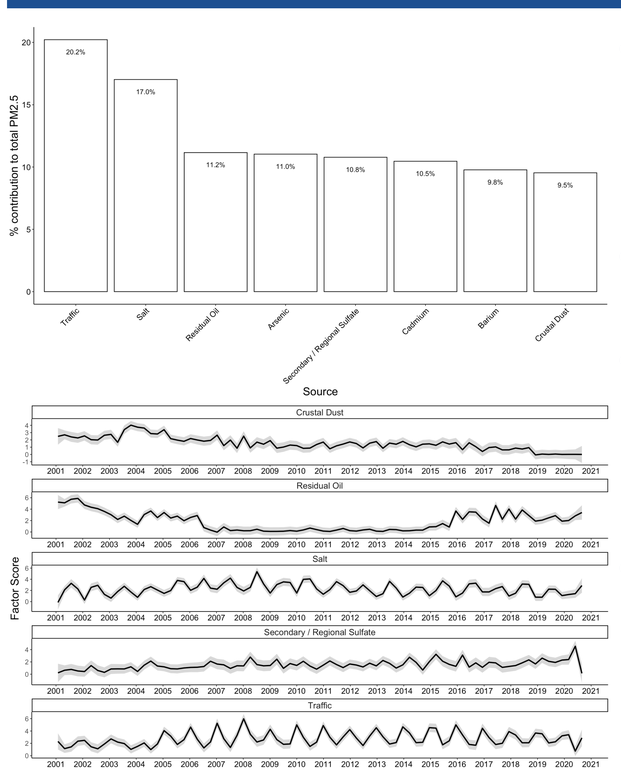


- PCP+Nystrom outperforms PCA in hindcasting missing EC, OC
- This paves way for more powerful source apportionment



- PCP identified 5 sources of $PM_{2.5}$ pollution: crustal dust, residual oil, salt, secondary sulfate, & traffic, as well as 3 single-constituent components (not shown): As, Ba, & Cd.

Conclusions



- Traffic contributed most to total $PM_{2.5}$ concentrations (20.2%) across study period
- Traffic peaked weekdays
- Traffic peaked in winters, including during block missing period
- Spikes observed in K^+ concentrations around each fourth of July (including during missing period)

Final takeaways:

- Nystrom extension improves recovery of missing block
  - Exact recovery in no-noise conditions
  - As noise increases, harder to recover missing block
  - Main assumption: Missing block characterized by same patterns governing observed blocks
- PCP w/Nystrom offers reliable pattern recognition allowing researchers to leverage more data
- Sources can be used in subsequent health models