Simulating Air Pollution Dynamics in the United States With Kalman Filters and Machine Learning

Grady King and Ignacio Segovia-Dominguez, PhD
School of Mathematical and Data Sciences
West Virginia University, Morgantown, West Virginia 26506

Introduction

- **Air pollution**, PM2.5 (<2.5 micrometers)
- Associated with **respiratory** & cardiovascular diseases -> ↑ hospitalizations & mortality
- EPA's Community Multiscale Air Quality (CMAQ) models PM2.5 daily on 12km grid
- Uses emissions and meteorology data (ground sensors, satellite), utility, biogenic & agricultural emissions, mobile emissions
- However, requires enormous compute and data resources
- Goal: **predict PM2.5** on a grid with **only ground sensors** (AQS: PM2.5, Meteostat: wind direction & speed, barometric pressure)

Methodology

- Kalman filters can predict the next time step given uncertain information about a dynamic system

1) Prediction

$$\hat{\mathbf{x}}_k = \mathbf{F}_k\hat{\mathbf{x}}_{k-1} + \mathbf{B}_k\vec{\mathbf{u}_k}$$
$$\mathbf{P}_k = \mathbf{F_k}\mathbf{P}_{k-1}\mathbf{F}_k^T + \mathbf{Q}_k$$

Estimate at time k
Previous estimate
External influences
Uncertainty

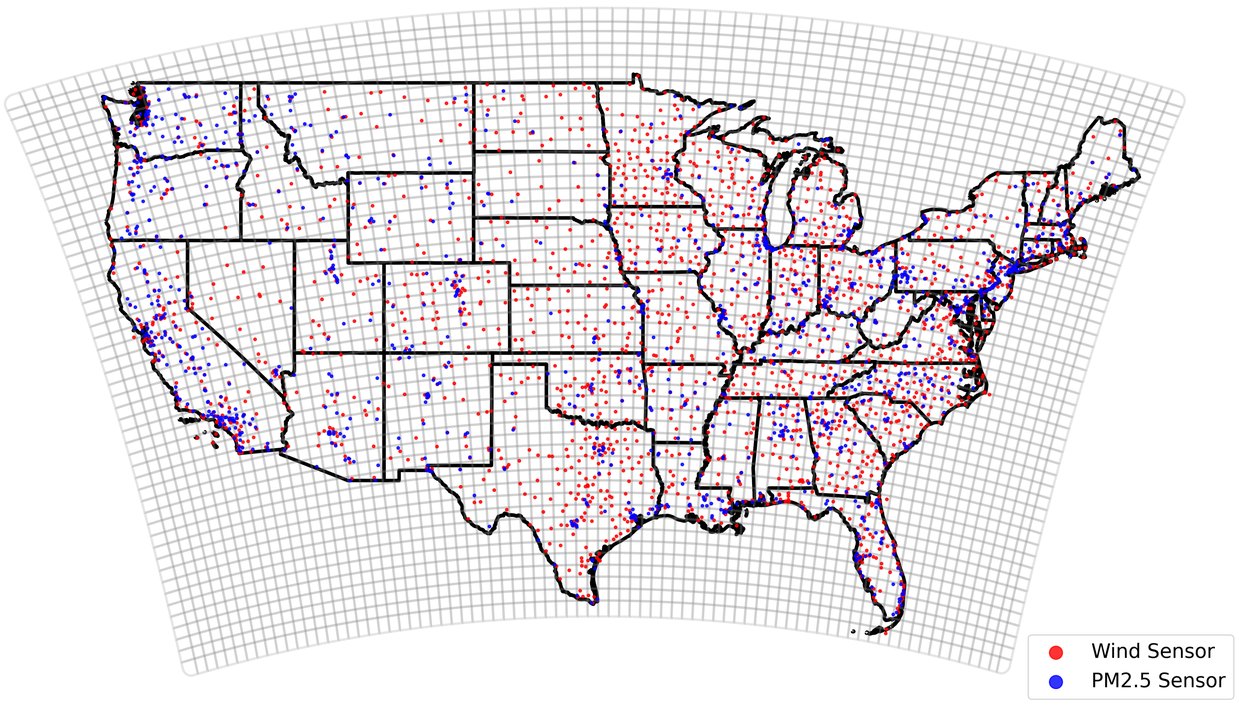
2) Update

$$\hat{\mathbf{x}}'_k = \hat{\mathbf{x}}_k + \mathbf{K}'(\vec{\mathbf{z}_k} - \mathbf{H}_k\hat{\mathbf{x}}_k)$$
$$\mathbf{P}'_k = \mathbf{P}_k - \mathbf{K}'\mathbf{H}_k\mathbf{P}_k$$
$$\mathbf{K}' = \mathbf{P}_k\mathbf{H}_k^T(\mathbf{H}_k\mathbf{P}_k\mathbf{H}_k^T + \mathbf{R}_k)^{-1}$$

Observed sensor reading
Combine reading with estimate

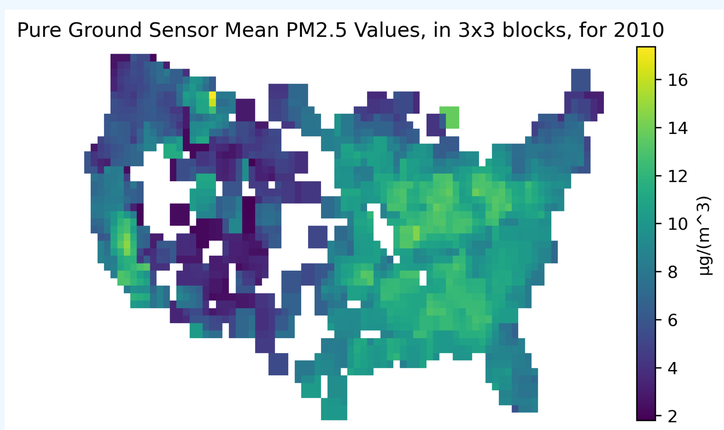
- Learn $\mathbf{F}_k$ and $\mathbf{B}_k$ using differential evolution (optimization technique)

US 2010-2021 Wind & PM2.5 Sensors



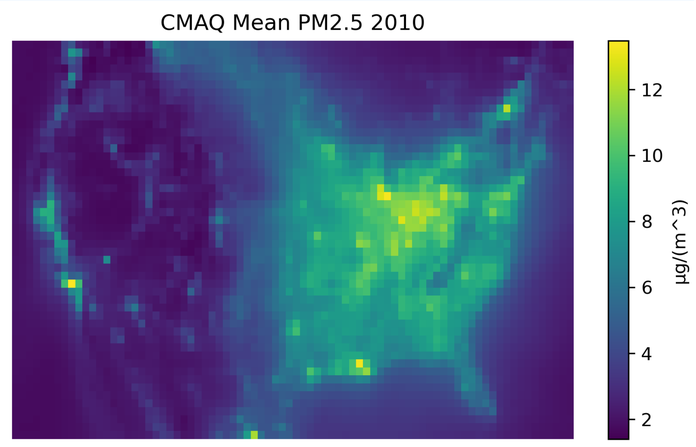
Preliminary Work

Ground Sensors vs CMAQ (2010)

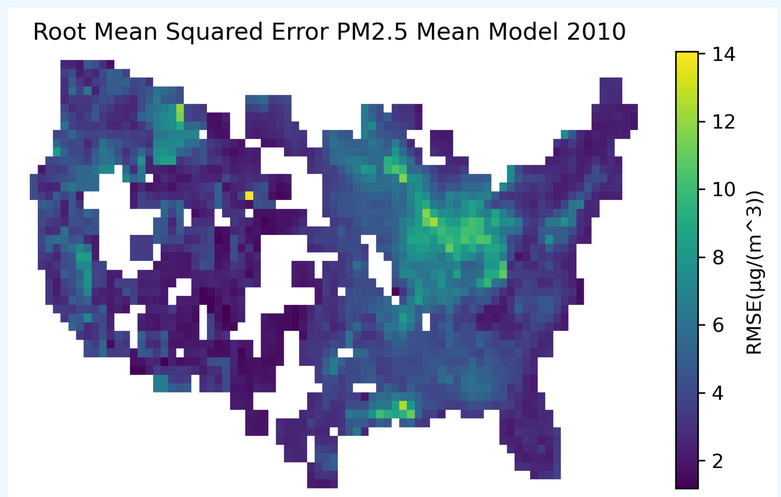


Missing values, difficult to model with

Similar trend, but mean model doesn't capture full picture



Mean model: $\mathbf{Y} = a\mathbf{X}$, a optimized w/ differential evolution



Train: Jan-Sep 2010
Test: Oct-Dec 2010

Train RMSE: **3.79**
Test RMSE: **4.54**

Future Analysis

- Linear Kalman Filter, optimized $\mathbf{F}_k$ and $\mathbf{B}_k$
- Unscented Kalman Filter, function from LSTM
  - Non-linear relationships learned from data

Code & Citations

NumPy
GeoPandas

WestVirginiaUniversity.
EBERLY COLLEGE OF ARTS AND SCIENCES