NO_2 Concentration Analysis based on Mathematical and Geospatial Approaches

Riley Chen [1], Mason Lu [2] , Matilda Slosser [3], Aneesh Srinivas [4]

[1] Emory University, [2] Murray State University, [3] Smith College, [4] University of California, Berkeley

Abstract

Motivation: NO_2 is one combustion byproduct associated with multiple adverse health outcomes

Data: Air Quality System (AQS) NO_2 monitoring networks over the contiguous United States of the Environmental Protection Agency (EPA) [1] from 2000–2016

Goals:

- predict average daily NO_2 concentration for contiguous US
- find potential correlations between NO_2 concentration and socioeconomic status

Model-Driven Approach

- Mathematical model of NO_2 average daily concentration
- Exponential decay and seasonal oscillation

$$y_{\text{model}}(t; \mathbf{p}) = p_1 + p_2 e^{-p_3(t-p_4)} + p_5 \cos(2\pi p_6(t+p_7))$$

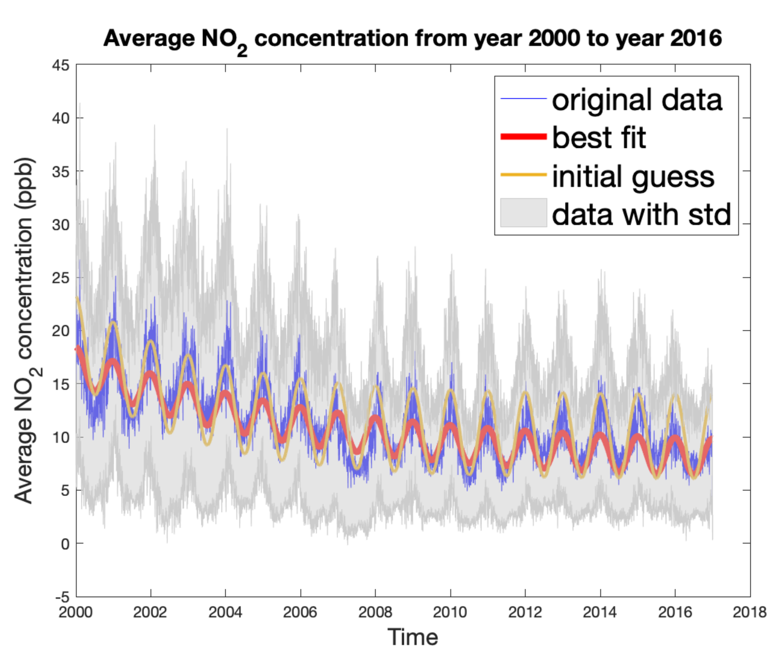
p_1: average NO_2 concentration in US before 2000	p_5: oscillation amplitude
p_2: scale factor	p_6: oscillation frequency
p_3: decay rate	p_7: shift
p_4: initial time	

Data Fitting Approach

Nonlinear least squares problem

$$\arg\min_{\mathbf{p}} \|\mathbf{W}(\mathbf{y}_{\text{model}}(\mathbf{p}) - \mathbf{y}_{\text{data}})\|_2^2$$

$\mathbf{W}$: diagonal weight matrix (standard deviation (std)$^{-1}$)	$\mathbf{y}_{\text{data}}$: original data
$\mathbf{y}_{\text{model}}$: data predicted from the model	



Optimization via Nelder-Mead method (MATLAB `fminsearch`)

Bayesian Approach

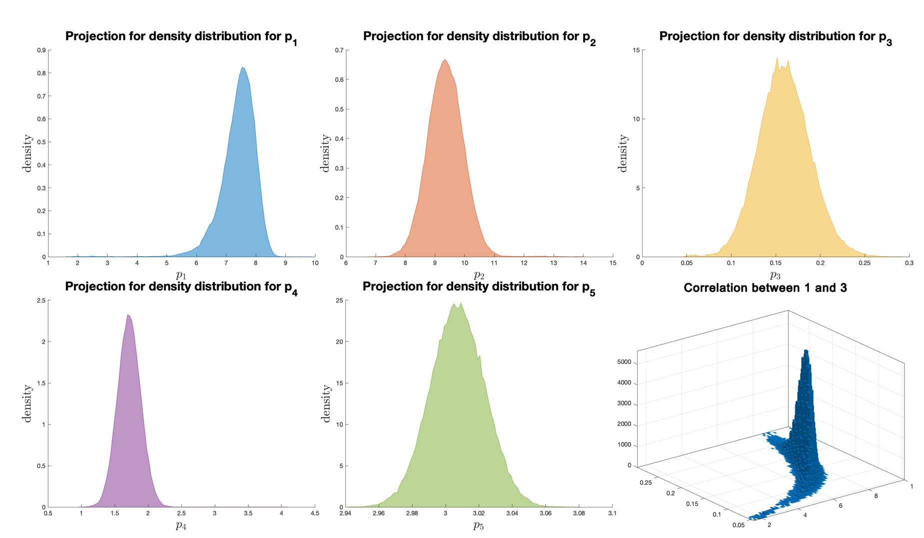
Use Bayes' Theorem [2]

$$\pi_{\text{post}}(\mathbf{p} \mid \mathbf{y}_{\text{data}}) = \frac{\pi_{\text{like}}(\mathbf{y}_{\text{data}} \mid \mathbf{p})\pi_{\text{prior}}(\mathbf{p})}{\pi_{\text{marg}}(\mathbf{y}_{\text{data}})}$$

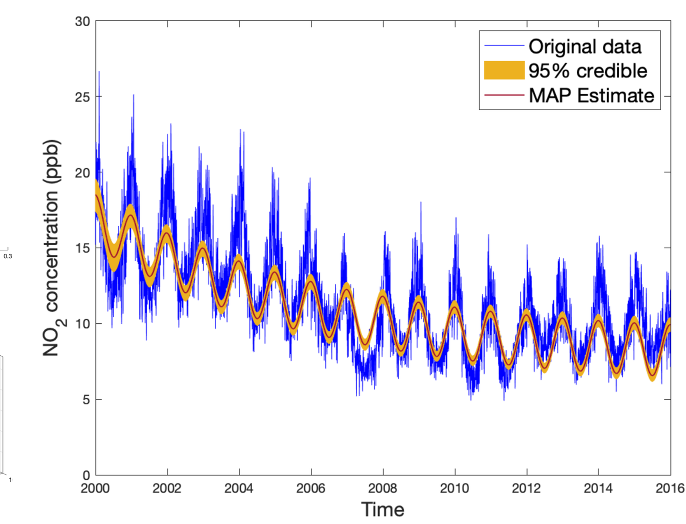
- Generate random samples from posterior distribution using Adaptive Metropolis (fixed $p_4 = 2000$ and $p_6 = 1$) with maximum a-posteriori estimate (MAP)

$$\mathbf{p}_{\text{MAP}} = \arg\max_{\mathbf{p}} \ \pi_{\text{post}}(\mathbf{p}|\mathbf{y}_{\text{data}})$$

Projections of Posterior Distribution:



Model Predictions:



Hybrid Model and Data-Driven Approach

Goal: Train a Long-Term Short-Term Memory Model (LSTM)[3] to predict the residual $\mathbf{r} = \mathbf{y}_{\text{model}} - \mathbf{y}_{\text{data}}$ of the "Model-Driven Approach"

Computational Approach:

- 60 time points used to predict the next time point
- Train on the first 5000 time points, test on the last 1000

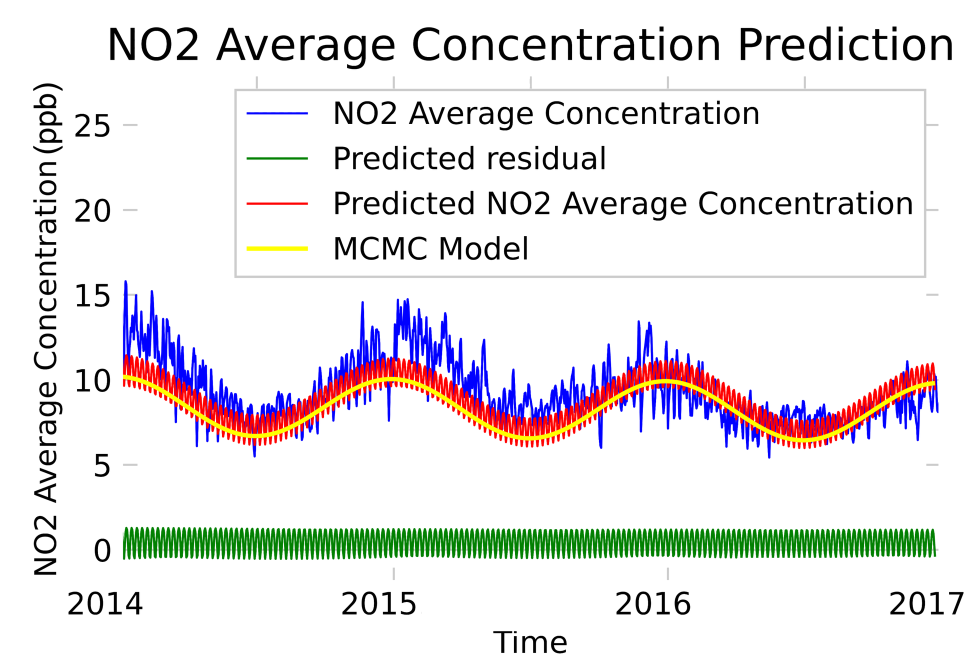
$$\min_{\boldsymbol{\theta}} \ \|\boldsymbol{\Phi}(\mathbf{r}; \boldsymbol{\theta}) - \mathbf{r}\|_2^2$$

where $\boldsymbol{\Phi}$ is an LSTM network with network parameters $\boldsymbol{\theta}$

- 50 epochs for the training via 'Adam' optimizer

Observations:

- Hybrid approach captures oscillation trend
- Large deviations still exist
- Data-driven approaches require larger datasets



References/Acknowledgement

[1] Qian Di et al. "Assessing NO2 concentration and model uncertainty with high spatiotemporal resolution across the contiguous United States using ensemble model averaging". In: *Environmental science & technology* 54.3 (2019), pp. 1372–1384.

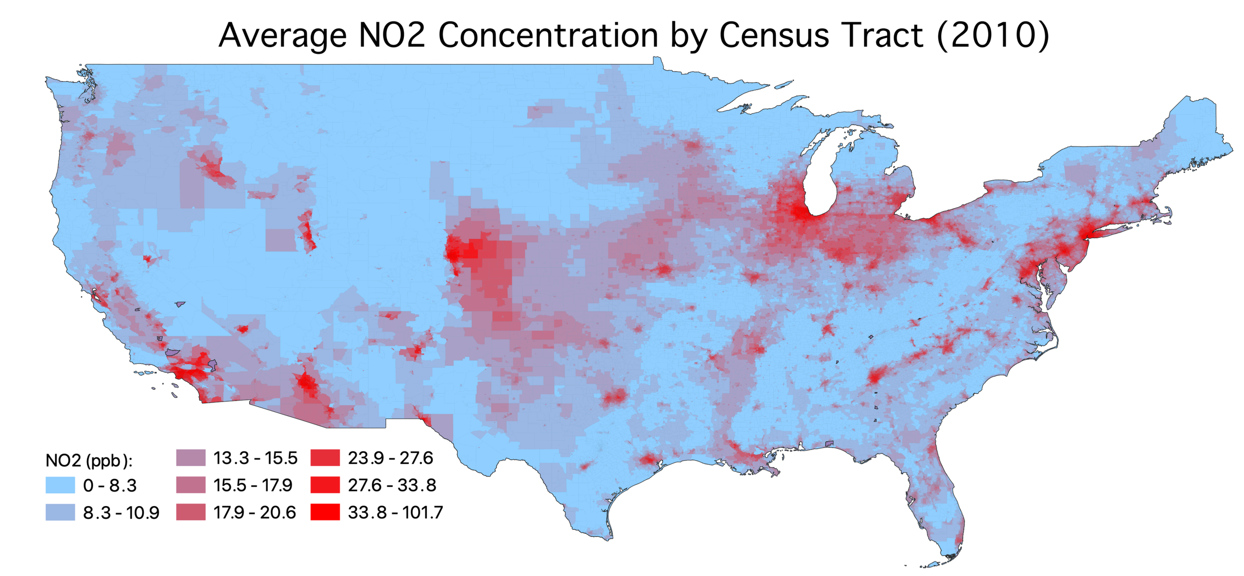
[2] Arianna Krinos and Aimee Maurais. "Parameter and Uncertainty Estimation for a Model of Atmospheric CO2 Observations". In: *SIAM Undergraduate Research Online* (2019).

[3] Palash Sharma. *Keras LSTM Layer Explained for Beginners with Example.* MLK - Machine Learning Knowledge. Feb. 1, 2021. URL: https://machinelearningknowledge.ai/keras-lstm-layer-explained-for-beginners-with-example/ (visited on 06/29/2023).

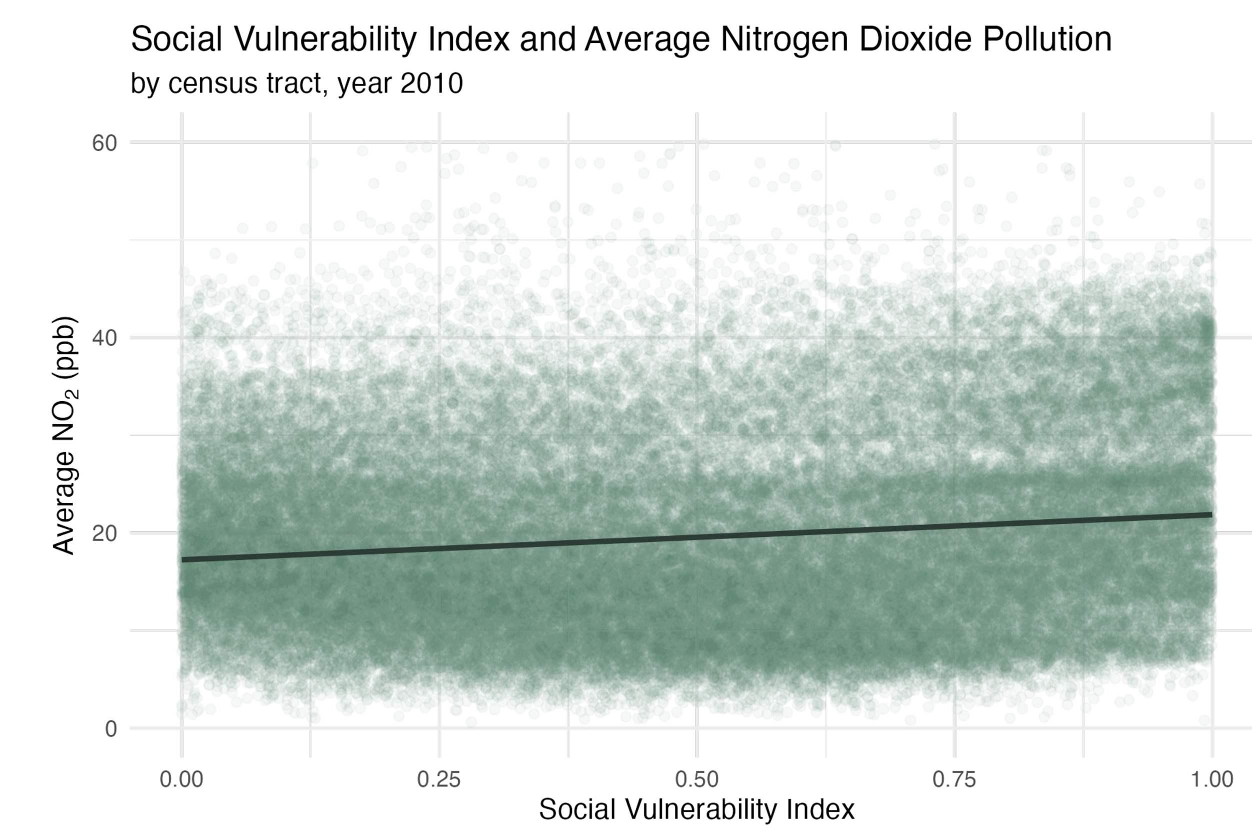
We would like to thank our amazing mentors, Dr. Matthias Chung and Dr. Julianne Chung for their guidance in this project. This work is supported in part by the US NSF award DMS-2051019.

2D NO_2 Maps Analysis

- Averaged NO_2 values over each census tract for years 2000, 2010, 2014, and 2016



- Combined geographical data of census tracts with their Social Vulnerability Indexes (SVI)



Regression Table: Average NO_2 explained by SVI

	estimate	std	p val	lower CI	upper CI
intercept	17.387	0.073	0	17.245	17.530
slope	4.507	0.126	0	4.260	4.754

Conclusions

- A model-driven approach with appropriately selected parameters can provide good predictions of average daily NO_2 concentrations. Including a weight matrix in the objective function resulted in a better data fit.
- Posterior MCMC samples suggest high levels of agreement and demonstrate little uncertainty in their predictions.
- The LSTM model was not ideal for our small data set. A future step is to analyze the frequency of oscillations in the residuals.
- Although weak for some years, we observe correlations between the SVI and NO_2 concentration, most noticeable in 2010.