NO2 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: NO2 is one combustion byproduct associated with multiple adverse health outcomes.

Data: Air Quality System (AQS) NO2 monitoring networks over the contiguous United States of the Environmental Protection Agency (EPA) [1] from 2000-2016

Goals:
- predict average daily NO2 concentration for contiguous US
- find potential correlations between NO2 concentration and socioeconomic status

Model-Driven Approach

- Mathematical model of NO2 average daily concentration
- Exponential decay and seasonal oscillation

\[
h_{\text{model}}(t) = p_1 + p_2 e^{-t} + p_3 \cos(2\pi t)
\]

Data Fitting Approach

Nonlinear least squares problem

\[
\min_{p} ||W(y_{\text{model}}(p) - y_{\text{data}})||^2
\]

Bayesian Approach

Use Bayes’ Theorem [2]

\[
\pi_{\text{post}}(p | y_{\text{data}}) \propto \pi_{\text{prior}}(p) \pi_{\text{likelihood}}(y_{\text{data}} | p)
\]

- Generate random samples from posterior distribution using Adaptive Metropolis (fixed \(p_4 = 2000\) and \(p_5 = 1\)) with maximum a-posteriori estimate (MAP)

\[
p_{\text{MAP}} = \arg \max_p \pi_{\text{post}}(p | y_{\text{data}})
\]

Model Predictions:

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 \(r = y_{\text{model}} - y_{\text{data}}\) 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

\[
\min_{\Theta} \left| \Phi(r, \Theta) - r \right|^2
\]

where \(\Phi\) is an LSTM network with network parameters \(\Theta\)

- 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


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.

Conclusions

- A model-driven approach with appropriately selected parameters can provide good predictions of average daily NO2 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 NO2 concentration, most noticeable in 2010.

2D NO2 Maps Analysis

- Averaged NO2 values over each census tract for years 2000, 2010, 2014, and 2016

Regression Table: Average NO2 explained by SVI

<table>
<thead>
<tr>
<th>Social Vulnerability Index (SVI)</th>
<th>NO2 Average Concentration</th>
<th>NO2 Average Concentration</th>
<th>Predicted residual</th>
<th>Predicted NO2 Average Concentration</th>
<th>MCMC Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>No SVI</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
</tr>
<tr>
<td>Low SVI</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
</tr>
<tr>
<td>Moderate SVI</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
</tr>
<tr>
<td>High SVI</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
</tr>
<tr>
<td>Very High SVI</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
<td>20.8</td>
</tr>
</tbody>
</table>

Optimization via Nelder-Mead method (MATLAB fminsearch)

SIAM Undergraduate Research Online (2019).