Josh Villarreal
joshuavillarreal@college.harvard.edu

1 Summary

For my final project for a data science course I took in the spring of 2020, my team chose to work with data provided by a biostatistics professor at the Harvard T.H. Chan School of Public Health to investigate whether air pollution increases COVID-19 mortality rates. This turned into a more in-depth research project that continued over the entire summer, where I developed a Bayesian negative binomial regression model with air pollution as a predictor. Through this analysis, I was able to verify other experimental results that showed that even a $1\mu g/m^3$ increase in $PM_{2.5}$ (fine particulate matter; the "soot" produced from air pollution), all things being equal, tends to increase the mortality rate from COVID-19 by 10.5% (95% confidence interval 4.3%-17%). To complete this project, I used the `pymc3` package in Python.

2 Appendix

2.1 Model specifications

We'd like to model the number of deaths in county i, y_i. I suppose $y_i \sim \text{NBin}(\mu_i, \alpha)$ for mean μ_i and Gamma parameter α, where $\mu_i = c_i \exp \eta_i$, where $\eta_i = \beta_{0j} + \sum_k \beta_k x_{ki}$ for:

- c_i population offset in county i,
- random intercept $\beta_{0j} \sim \mathcal{N}(0, \sigma_0)$ for state j, and
- predictors x_{ki} include mean $PM_{2.5}$, median household income, and other socioeconomic and geographic predictors,

with α, σ_0, having uniform priors and β_k normal priors.

2.2 Plots

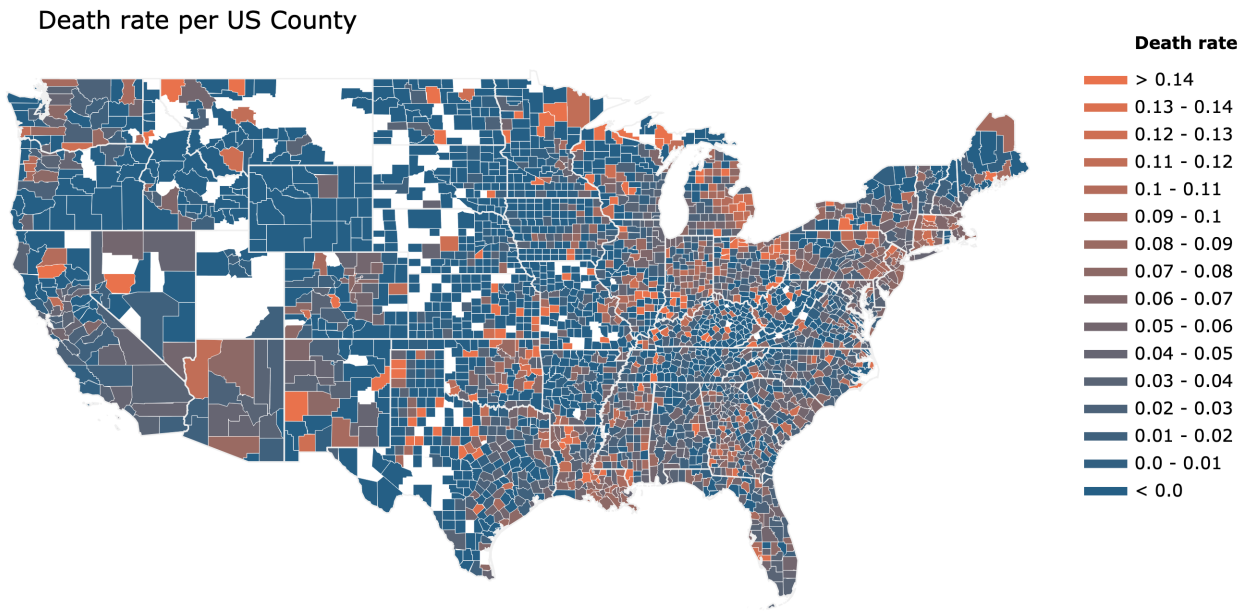

Figure 1: COVID-19 mortality rates in the USA by county (as of 7/1/2020).

Josh Villarreal
joshuavillarreal@college.harvard.edu

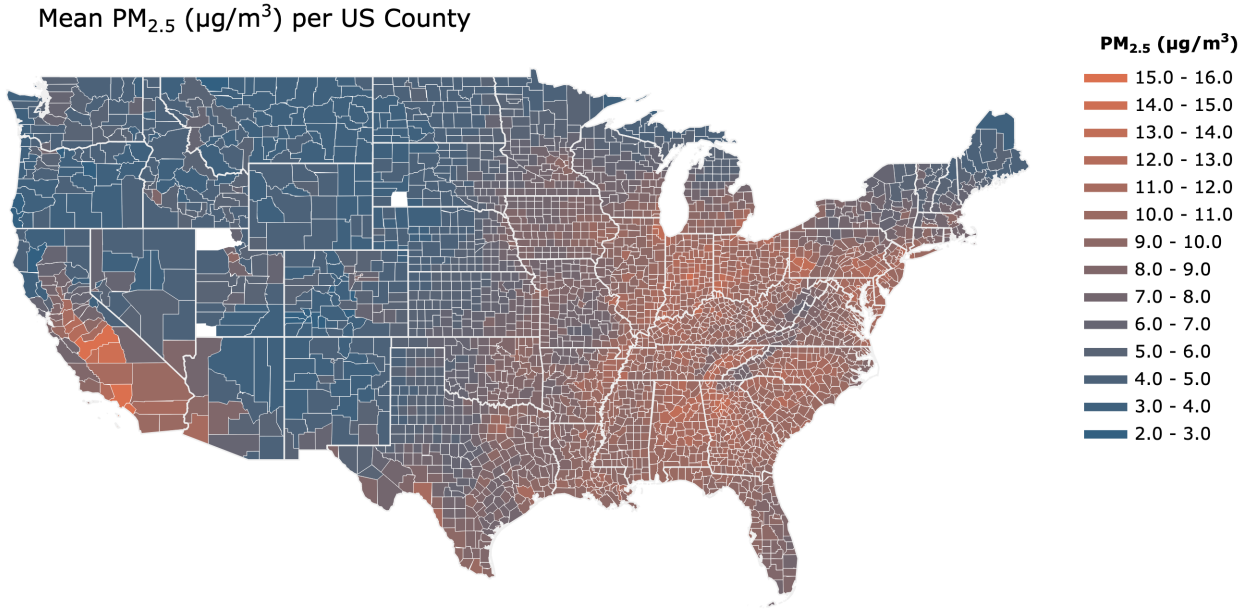

Figure 2: 1-year average amount of air pollution (measured in $\mu g/m^3$ of $PM_{2.5}$.) in the USA by county (as of 7/1/2020).