**Spatial Modeling of PM2.5 Concentration Before and During COVID-19 in the Southeastern U.S.**

Tyler Schappe

## Introduction

The COVID-19 pandemic currently underway has altered day-to-day activities for nearly every person in the U.S in 2020. Public health policies to combat the spread of the virus, as well as increased market uncertainty, has disrupted economic activity significantly relative to pre-pandemic times. The resulting economic slowdown is accompanied by a reduction in production and transportation of goods, less in-person retail, and millions of jobs lost. While these are obviously negative outcomes, one side effect is likely a reduction in air pollutants that result from these activities. Additionally, many workers who have not lost their jobs are able to work from home, further reducing emissions from transportation.

Particulate matter (PM) is a class of pollutants that includes both solid particles and liquid droplets found in the air. Many such particles are the result of chemical reactions that occur in factories, power plants, and internal combustion engines. PM2.5 is a subclass of these pollutants defined as particles that are 2.5 micrometers or smaller in size, which are the main cause of artificial haze and pose the greatest health risk for humans and other animals. Since these types of particles are a common by-product of the industrial and transportation sectors of the economy, I am interested in understanding how the COVID-19 pandemic, via its economic impacts, has affected PM2.5 quantities in the atmosphere. Specifically, I want to model the difference in PM2.5 levels, averaged over all of the days between April 1 and June 30, between 2020 and 2019 in the southeastern states of FL, GA, NC, SC, and VA.

I hypothesize that demographic, economic, and political factors affect PM2.5 pollution. First, regions with greater number of inhabitants are likely to contain more sources of PM2.5 and thus have the greatest potential for the largest magnitude of reduction of PM2.5 levels between 2020 and 2019. Therefore, I expect population density to be negatively associated with the difference in PM2.5 levels between 2020 and 2019 (greater density gives a larger negative difference). Second, it is generally the case that most ‘blue collar’ or ‘front-line’ jobs cannot easily be performed remotely and thus communities with predominantly these types of jobs have likely not altered their transportation patterns as much during the pandemic. Moreover, these types of jobs are associated with heavy industry economic activity in the region, which produce PM2.5 pollution. Therefore, I expect that a socioeconomic vulnerability index (SVI) [8], used as a proxy for proportion of jobs that can’t be worked remotely or are industrial in nature, will be positively associated with the difference in PM2.5 levels between 2020 and 2019 (greater SVI gives a less negative difference). Third, the partisan political environment can affect both state and local health policies, as well as behavior of individuals regarding self-quarantine and transportation patterns. Based on anecdotal evidence and COVID-19 infection data, right-leaning regions have tended to have less strict public health responses, including allowing retail stores and restaurants to remain open, having less strict or no local quarantine orders, and to re-open schools more widely. Because most of these activities involve transportation via automobile, these factors could result in a smaller reduction in PM2.5 levels compared to left-leaning regions that adopted more strict measures. Thus, I expect the percentage of votes that President Trump received in 2016 [5], a measure of the political response to the pandemic, to be positively associated with the difference in PM2.5 (greater percentage of votes gives a less negative difference). Finally, states reacted differently regarding policies and number of infections, so I expect the differences in PM2.5 to vary by state.

*Data description*

The PM2.5 data were downloaded in raw form from the US Environmental Protection Agency server for 2019 and 2020 for the states of Florida, Georgia, North Carolina, South Carolina, and Virginia [9]. Observations that occurred between April 1 and June 30 were extracted and only observations from measurement sites that were present in both years of data and that had at least ten observations were retained, resulting in 120 sites being included; negative values were converted to 0. For each year, the average PM2.5 value was calculated for each measurement location across all dates, then the average values from 2019 were subtracted from the average values from 2020 to calculate the final response variable used in the model (Fig. 1, left).

Based on the discussion above, the predictor variables were population density, socioeconomic vulnerability index (SVI), and percent of vote for Trump in 2016. Population density data based on the 2010 census and calculated on a grid of 30 arc-second increments across the U.S. were obtained from the Center for International Earth Science Information Network [2,3]. For each PM2.5 measurement location, the mean population density in a 200 x 200 grid square with sides measuring 6000 arc-seconds, or about 155 km, centered on the measurement location was calculated; this distance was somewhat arbitrary but reflected my belief that air pollution disperses across a fairly large distance. Second, each measurement site was assigned the SVI value for the county in which it is located. A similar procedure was performed for the by-county percentage of votes for Trump from the 2016 presidential election.

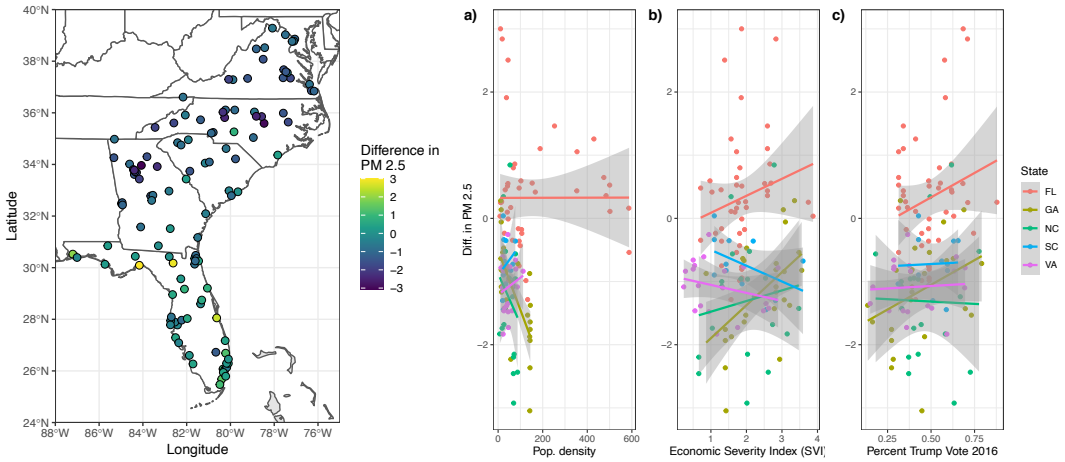


**Figure 1.** *Left:* Measured difference in mean PM 2.5 for the 120 sites from April 1 to June 30 between 2020 and 2019. *Right:* Scatter plots of the three predictor variables plotted against the response variable with separate least squares regression fits by state, along with corresponding confidence intervals, overlayed.

## Methods

I first examined the raw data visually by plotting each predictor variable against PM2.5 and overlaying a simple linear regression fit. As discussed above, I suspected that there would be difference among states, so I accounted for this when making the plots (Fig. 1, right). It appeared that only the association between PM2.5 and population density varied strongly by state, so I only considered that interaction in subsequent analyses. Directional variograms from residuals from an initial least squares model generated at 30-degree increments showed that at 60 degrees, the range may be greater or the partial sill smaller (Fig. 2), so I considered including anisotropy parameters in the model selection below. There was some evidence of differences in the covariance structure among states (Fig. 3), suggesting non-stationarity, but it was not overwhelming and it would be difficult to model, so subsequent models assumed stationarity.

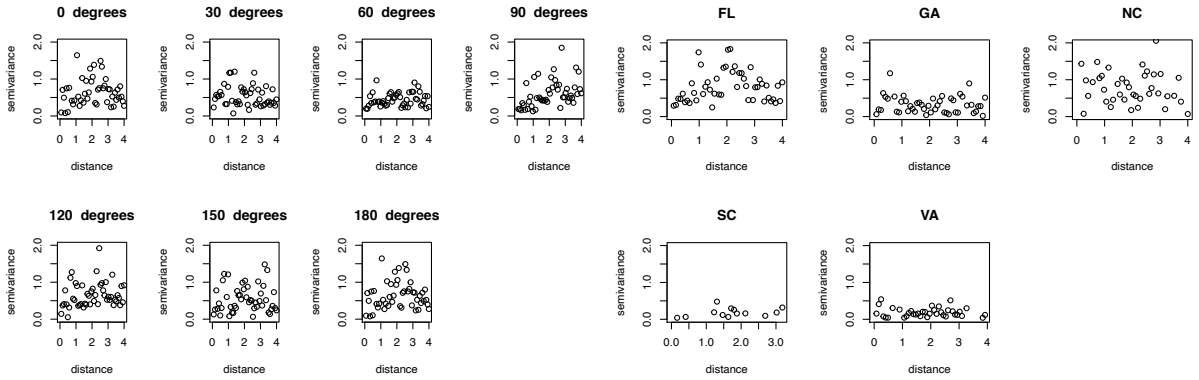


**Figure 2.** *Left:* Directional variograms in 30-degree increments created from residuals of a least squares fit using chosen predictor variables. *Right:* Variograms created for each state from residuals of a least squares fit.

*Model comparisons*

I fit models with exponential and Matern covariance functions, with and without anisotropy and nugget parameters using maximum likelihood methods with the 'likfit()' function in geoR [7]. I then compared them using the Aikake Information Criterion and found that a model with an exponential covariance function with anisotropy parameters and without a nugget term was best supported (Table 1). However, I chose to include a nugget parameter in the final model because it was the second-best supported model, some of the variograms indicated a non-zero nugget, and it is difficult to believe that there is no non-spatial error. In this model, the anisotropy angle parameter was estimated to be 77 degrees and the anisotropy ratio was estimated as 22.37.

| Covariance Function | Nugget Term | Anisotropy | AIC | Model DF |
|---|---|---|---|---|
| Exponential | Included | Isotropic | 293.32 | 15 |
| Exponential | Excluded | Isotropic | 291.32 | 14 |
| Matern | Included | Isotropic | 294.26 | 16 |
| Matern | Excluded | Isotropic | 292.39 | 15 |
| **Exponential** | **Included** | **Anisotropic** | **288.31** | **17** |
| Exponential | Excluded | Anisotropic | 287.70 | 16 |
| Matern | Included | Anisotropic | 289.33 | 18 |
| Matern | Excluded | Anisotropic | 288.38 | 17 |

**Table 1.** Results of model selection procedure to determine optimal covariance function, inclusion of the nugget term, and inclusion of anisotropy terms. The final model is shown in bold.

I fit the final model chosen above using Bayesian estimation methods using the function 'krige.bayes()' in the geoR package [7]. Regarding priors, a multivariate normal distribution with mean vector of 0 and covariance matrix with variances of 10 and covariances of 0 was used for the vector of beta parameters, an inverse chi-squared distribution with 2 degrees of freedom and mean set to point estimate from MLE model was used for sigma-squared, a discrete squared reciprocal prior with 51 discrete support values between 0 and twice the maximum distance between data locations was used for phi (the default behavior), and a discrete uniform distribution with support values between 0 and 10 by 0.2 was used for the relative nugget parameter, which is the ratio of tau-squared to sigma-squared. I included in the model a grid of 3,867 equally spaced points 0.139 degrees (15 km) apart throughout the spatial domain, along with corresponding covariates, for Bayesian Kriging prediction; included among these prediction locations were the sampling sites themselves. Since this function does not allow estimation of the anisotropy parameters, they were set to be fixed at values estimated from the final MLE model above.

The parameterization for this model excluded baseline terms and thus the coefficients associated with the 'State' predictor represent the intercept and slope values for each state, not a difference from a baseline state. All predictors were centered and scaled before estimation. A total of 10,000 samples were made from the joint posterior distribution for all 15 model parameters and all 3,987 Kriging prediction values. The final model is specified as follows:

$$PM2.5_i = \beta_0 FL_i + \beta_1 GA_i + \beta_2 NC_i + \beta_3 SC_i + \beta_4 VA_i + \beta_5 pop.dens_i + \beta_6 SVI_i + \beta_7 Trump_i + \beta_8 FL_i * pop.dens_i + \beta_9 GA_i * pop.dens_i + \beta_{10} NC_i * pop.dens_i + \beta_{11} SC_i * pop.dens_i + \beta_{12} VA_i * pop.dens_i + Z_i + \varepsilon_i$$

$$\varepsilon_i \sim Normal(0, \tau^2) \qquad Cov(PM2.5_i, PM2.5_j) = \sigma^2 e^{-(d_{ij}^T B d_{ij})^{1/2}/\phi}$$

$$B = AA^T, A = \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{bmatrix} \begin{bmatrix} 1/r & 0 \\ 0 & 1 \end{bmatrix}$$

Here, α is the isotropy angle as measured from 0º (north), fixed at 77º for the Bayesian model, and *r* is the anisotropy ratio, fixed at 22.4 [12].

*Model checking*

The final Bayesian model was first checked for sampling convergence using trace plots, which all showed satisfactory mixing (Fig. 3).

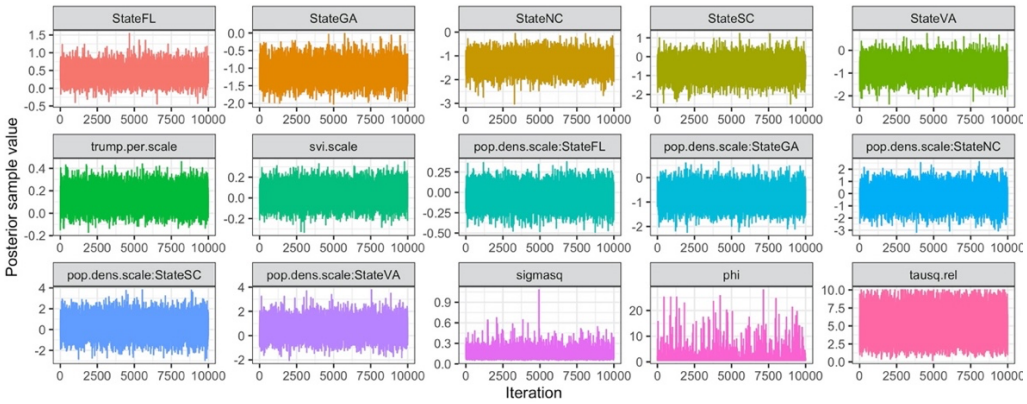


**Figure 3.** Trace plots for the samples of the marginal posterior distributions of each parameter in the model.

The final model assumes anisotropy (angle of 77 and ratio of 22.4) and stationarity as mentioned above which are justified by the patterns found in the variograms of the original least squares fit. Although the assumption of stationarity may be violated to some extent, it doesn't appear to be grossly violated. The final model also assumes normality of the nugget terms, linearity in the beta predictors, a constant nugget variance, and independence of nugget terms. Figure 4 (left) shows that the assumption of normality of the nugget is acceptably violated, linearity of the predictors is somewhat violated by the outliers with large negative residual values (center), and constant variance (center) and independence of the nugget terms (right) hold to a reasonable extent.

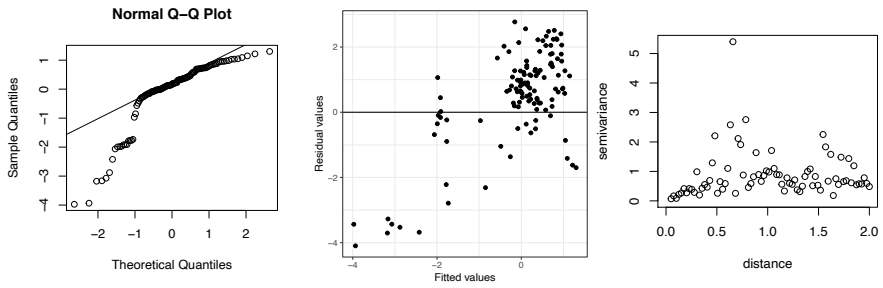


**Figure 4.** *Left*: Fitted values plotted against residual values from the final model. *Center*: Quantile-quantile plot with residuals from final model plotted against theoretical quantiles from a Gaussian distribution. *Right:* Variogram constructed with residuals from final Bayesian model.

## Results

Highest density intervals from the marginal posterior distributions for each model parameter are summarized in Table 1 below [4].

| | | Int. FL | Int. GA | Int. NC | Int. SC | Int. VA | Percent Trump 2016 | SVI | Pop. Density: FL | Pop. Density: GA | Pop. Density: NC | Pop. Density: SC | Pop. Density: VA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **95%** | **Lower HDI** | 0.02 | -1.61 | -1.93 | -1.53 | -1.50 | -0.03 | -0.14 | -0.28 | **-1.53** | -1.69 | -1.41 | -0.76 |
| | **Upper HDI** | 0.82 | -0.60 | -0.63 | 0.28 | -0.05 | 0.30 | 0.20 | 0.18 | **-0.06** | 1.22 | 2.00 | 1.96 |
| **80%** | **Lower HDI** | 0.16 | -1.45 | -1.70 | -1.20 | -1.23 | **0.02** | -0.09 | -0.20 | **-1.28** | -1.15 | -0.85 | -0.30 |
| | **Upper HDI** | 0.67 | -0.81 | -0.87 | -0.04 | -0.30 | **0.24** | 0.14 | 0.11 | **-0.34** | 0.73 | 1.33 | 1.49 |

**Table 1.** 95% and 80% highest density intervals (HDI) for all beta regression coefficients in the model calculated as the mean of their respective posterior distributions. Intercept parameters are shown in gray since their interpretation is not of primary interest. Parameters for which a given HDI does not include 0 are shown in bold.

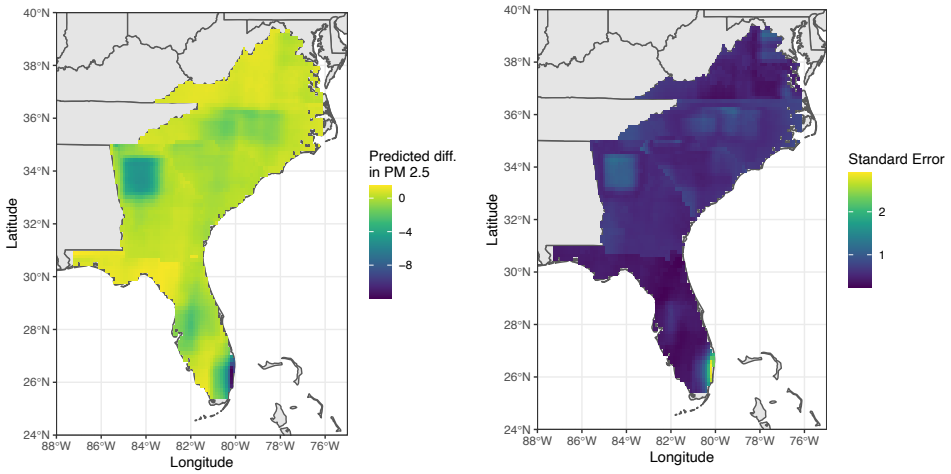


**Figure 5.** *Left:* Bayesian Kriging predictions of the difference in PM 2.5 between 2020 and 2019. Predictions were generated by finding the mean of the posterior distribution of the prediction. *Right:* Standard error of Bayesian Kriging predictions of the difference in mean PM 2.5 between 2020 and 2019 calculated as the standard deviation of the posterior distribution of the predicted values.

At the 95% credible level, the only coefficient that is significantly different from 0 is for the relationship between population density and PM2.5 in the state of Georgia, and it is estimated to be negative. From this, we can infer that more densely populated areas in GA are associated with a larger decrease in PM2.5

during the pandemic, but other states did not have this trend. This can be seen in both Fig. 5 (left) and Fig. 6 (left) as the large area around suburban Atlanta where a large decrease in PM2.5 is predicted. At the 80% credible level, the coefficient corresponding to percentage of votes cast for Donald Trump in 2016 is significantly greater than 0. This suggests that on average across all five states, a larger percentage of votes for Trump is associated with a smaller decrease in PM2.5, or in some cases an increase. Since the coefficients in a multiple linear regression model are interpreted as partial effects, the association with percentage of votes for Trump accounts for all other predictors in the model, including population density, and thus is likely not confounded with a theorized rural vs. urban split of voters for Trump.

*Spatial prediction*

Bayesian Kriging predictions show interesting patterns in all states, in a addition to the pattern regarding Atlanta mentioned above. In Florida, predictions vary greatly by location, with a larger decrease in PM2.5 predicted for urban areas and vice versa. There is a large area in central North Carolina for which the model predicts that PM2.5 has decreased significantly, which corresponds generally to areas around Raleigh-Durham, Winston-Salem-Greensboro, and Charlotte (Fig. 6 left); these are areas with relatively high population densities and also with many left-leaning counties (Fig. 6 right). In South Carolina, the largest area with a predicted decrease in PM2.5 appears to surround Columbia, which is a major metropolitan area and left leaning (Fig. 6 left), but does not include other surrounding counties with lower population density that also did not vote majority Trump.

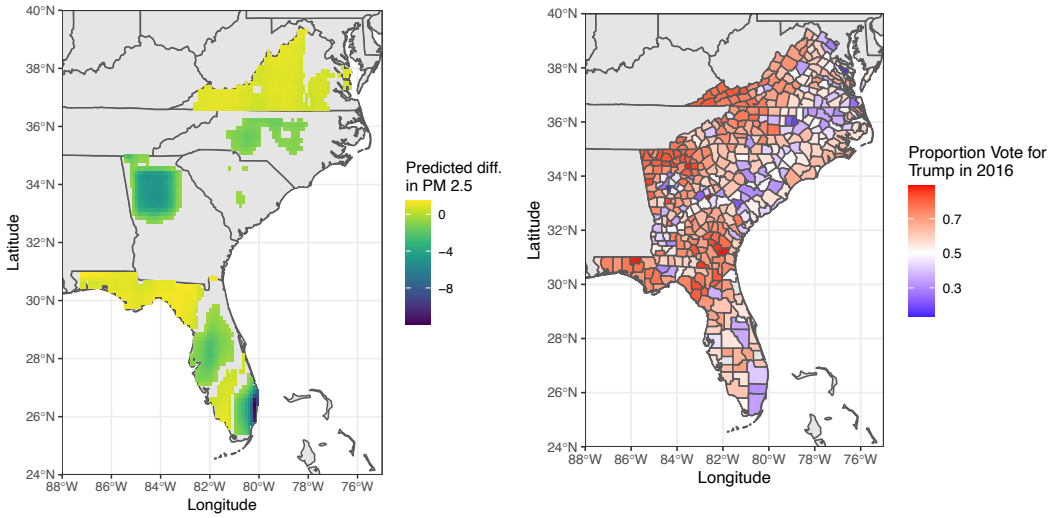


**Figure 6.** *Left:* Bayesian Kriging predictions of the difference in mean PM 2.5 between 2020 and 2019 for which the 95% highest density interval of the posterior distribution of the predicted values does not include 0. *Right:* Proportion of votes for Donald Trump in the 2016 presidential election by county.

*Conclusions*

COVID-19 has impacted many aspects of the everyday lives of people living in the U.S. In this study, we used a Bayesian spatial model to examine how population density, socioeconomic vulnerability, and political context are associated with particulate matter pollution, and to make spatial predictions for pollution. We found that population density was strongly associated with a decrease in PM2.5 in Georgia, percentage of votes for Trump in 2016 was marginally associated with an increase in PM2.5 levels across the entire study region, and socioeconomic vulnerability was not associated with PM2.5 levels. Furthermore, we found evidence that the largest reductions in PM2.5 levels during COVID-19 are predicted to have predominantly occurred in densely populated metropolitan areas that also voted for Clinton in 2016.

## Appendix

### References


1. Becker, Richard A., Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka and Alex Deckmyn. (2018). maps: Draw Geographical Maps. R package version 3.3.0. https://CRAN.R-project.org/package=maps

2. Center for International Earth Science Information Network - CIESIN - Columbia University, United Nations Food and Agriculture Programme - FAO, and Centro Internacional de Agricultura Tropical - CIAT. 2005. *Gridded Population of the World, Version 3 (GPWv3): Population Count Grid*. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://dx.doi.org/10.7927/H4639MPP. Accessed 18 Sep 2020.

3. Center for International Earth Science Information Network - CIESIN - Columbia University. 2013. *Environmental Treaties and Resource Indicators (ENTRI) Query Service*. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://sedac.ciesin.columbia.edu/entri. Accessed 18 Sep 2020.

4. Meredith, Mike and John Kruschke (2020). HDInterval: Highest (Posterior) Density Intervals. R package version 0.2.2. https://CRAN.R-project.org/package=HDInterval

5. MIT Election Data and Science Lab Github. 2018. *2016 Presidential Election Results by County*. https://github.com/MEDSL/2018-elections-unoffical/blob/master/election-context-2018.csv. Accessed 18 Sep 2020.

6. Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009

7. Ribeiro, Paulo J. Jr, Peter J. Diggle, Martin Schlather, Roger Bivand and Brian Ripley (2020). geoR: Analysis of Geostatistical Data. R package version 1.8-1. https://CRAN.R-project.org/package=geoR

8. US Centers for Disease Control and Prevention (CDC), Department of Health and Human Services. 2018. *Social Vulnerability Index*. https://healthdata.gov/dataset/social-vulnerability-index-2018-united-states-county. Accessed 18 Sep 2020.

9. US Environmental Protection Agency. Air Quality System Data Mart [internet database] available via https://www.epa.gov/airdata. Accessed Sep. 18, 2020.

10. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

11. Wickham, H. (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

12. Y. Shen, A.E. Gelfand. Exploring geometric anisotropy for point-referenced spatial data. 2019. Spatial Statistics. Volume 32, 100370. ISSN 2211-6753