
SPATIAL MODELING OF CARDIOVASCULAR DISEASE

INCIDENCE POSITIVELY ASSOCIATED WITH PM2.5

A PREPRINT

Johan Booc

Department of Statistics
Texas A&M University

jbooc24@tamu.edu

Christina Kim

Department of Statistics
Texas A&M University

christinaykim3@tamu.edu

Shombit Roy

Department of Statistics
Texas A&M University

shombit123@tamu.edu

April 24, 2024

ABSTRACT

Cardiovascular disease (CVD) is the leading cause of death in the United States. By using a spatial modeling technique (geographically weighted regression) and even after adjusting for median household income and unemployment rates (finding similar correlations studied in the past), we observed that the concentration of PM 2.5 is centered in the Stroke Belt region. Policymakers and

5 health practitioners can use these results to identify targeted interventions to curb the increasing rates
 6 of CVD, aiming to halt one of the world's deadliest diseases.

7 **Keywords** Fine Particulate Matter (PM2_5) • Cardiovascular Disease (CVD) • Cardiovascular Mortality (CVM)

8 1 Introduction

9 As cardiovascular disease is the leading cause of death in the US, numerous past studies have been done on this disease,
 10 specifically on the older population the 65+ year age group. Our study aims to further investigate the impact, focusing
 11 on the 18-44-year-old age population, since generally, there has been less interest in the effects of CVD on them.

12 Our approach involves analyzing this relationship to produce risk prediction estimates for our specified age group.
 13 There are several factors involved in influencing the incidence rates of CVD from genetics, lifestyle, diet, and smok-
 14 ing/alcohol habits. We look into the covariates - PM 2.5 concentration, median household income, and unemployment
 15 rates - to analyze the spatial variation for our specified age population, thus decreasing future CVD incidence rates.
 16 Specifically, we focus on PM 2.5 concentration as the world continues to depend on fossil fuels/natural gas and cli-
 17 mate change is an ongoing concern. Earlier studies such as (Warsito et al. 2018) and (Liu et al. 2020)highlight the
 18 modern-day threats from air pollution using a Robust GWR model and Bayesian-temporal model, respectively. From
 19 studying our covariates, we can examine the effects of certain CVD risk factors for the year 2015.

20 The geographically weighted regression model uses local variables and weights to produce statistical visualizations
 21 of the US's regional variation between our response variable (CVD deaths) and its covariates. Other methods, such
 22 as the traditional linear regression model and clustering, have been used in past CVD studies. However, we found
 23 the GWR model allows us to see each covariate's local significance and magnitude. The spatial distribution aspect
 24 of the GWR model highlights regions that are concentrated with CVD rates, whether incidence percentage rates are
 25 increasing or decreasing, and any significant correlations. This makes the GWR model efficient in analyzing the
 26 dynamic relationship of CVD rates and the factors that contribute to it.

27 Using a geographically weighted regression approach, the overall outcome of our study aids policymakers and health
 28 practitioners in implementing the necessary interventions for targeted regions that are being influenced the most by
 29 our covariates. For public health experts, it's ideal to tailor aid to each specific region of the US first to eventually
 30 reach a national decrease in CVD incidence rates.

31 2 Related Works

32 There's a growing popularity of using spatial models in the epidemiological domain to analyze the distribution and
 33 factors of a disease. The techniques often used differ between studies but all have the same goal: to reduce the disease
 34 incidence and mortality rates. Statistical regression models are popular methods for CVD studies. The underlying
 35 causes for CVD are dynamic and can be seen as an interconnected web. (Zelko et al. 2023) examines the relationship

36 between CVD and covariates similar to our study (air pollution, social determinants, and county-level data). From
37 using a GWR model, their results found that counties in the South had the highest exposure to PM 2.5 concentrations
38 whereas counties in the Northeast had the lowest. A strong correlation was also found between household income,
39 race, and healthcare access. Overall, there are several causes to consider with CVD, and where the covariates are
40 concentrated is another important consideration.

41 Compared to traditional regression models, the geographically weighted model (GWR) explores spatial data by con-
42 sidering varied coefficients for a certain spatial unit (Gebreab and Diez Roux 2012). The GWR model contains the
43 parameter, neighborhood (also known as bandwidth) and builds on the weighted least squares method to estimate the
44 regression coefficients. We want to see values closer to the point of interest since they carry more weight, thus having
45 a greater influence.

46 The Ordinary Least Squares (OLS) model is a popular choice of method for public health studies because it generates
47 the regression coefficients on a global scale and captures the average differences between covariates. However, this
48 means spatial variability between units can be easily hidden for the OLS model. Our paper acknowledges that PM 2.5
49 concentration and socioeconomic factors are not spatially constant with cardiovascular disease. The GWR model is
50 more suitable for our study because the goal of public health is to improve the population as a whole. To get to the
51 root cause of CVD mortality rates and see future improvement, the OLS model should be treated as the ‘null’ model,
52 with the GWR model used to test and verify that it is a statistically significant better fit.

53 The GWR model spatially displays a relationship between CVD deaths and their covariates and analyzes disparities at
54 the local scale. Errors are also minimized between the actual model and any estimates. This makes the GWR model
55 suitable for seeing the socioeconomic factors that affect different regions and narrows down our focus to the areas in
56 need of improvement, which helps health practitioners implement policies for that region. Past studies (Zelko et al.
57 2023),(Terry et al. 2023), and(Singh et al. 2019) tend to focus on the trends of socioeconomic covariates and their spa-
58 tial patterning. However, our study extends past studies by including PM 2.5 concentrations as one of our covariates.
59 From this, we can fully understand the dynamic relationship between counties and CVD incidence/mortality rates.

60 Risk assessment and risk estimates uncover the key factors associated with CVD. Because the concentration of specific
61 races varies by county, we studied the dose-response relationship of PM 2.5 concentrations and the socioeconomic co-
62 variates of CVD mortality at the county level. This helps researchers and health practitioners to develop the necessary
63 risk-preventative measures and allocate resources to the areas that need them the most. By putting the focus on the
64 county level (rather than on individual states/nationally), resources can be allocated accordingly to the regions that
65 need the most assistance. This can lead to a reduction in CVD mortality rates, both nationally and globally.

66 **3 Methods**

67 **3.1 Data Collection**

68 A data set of Medicare services and claims from the Centers for Disease Control and Prevention (CDC) website was
 69 loaded to analyze Medicare claims data, specifically Cardiovascular death rates across different counties and States
 70 in the US. Racial and geographic data were retrieved from the Census and TIGER Bureau. The median income was
 71 extracted from the CENSUS API. The air quality index was extracted from the NASA PM 2.5 Concentration data.
 72 Finally, the unemployment rate was extracted from the CDC website. These data sources were collected and prepared
 73 for analysis to understand the relationship between these factors and Cardiovascular death rates. The code used to
 74 extract and filter the data is available in our GitHub repository.

75 **3.2 Data Preprocessing**

76 Several steps were taken to ensure the reliability and accuracy of the results when preparing the data for statistical
 77 analysis. The data from various sources was integrated into one file by year/county/latitude and longitude/geometries.
 78 This integration was achieved through coding in RStudio Version 4.3.2, specifically focusing on data from 2015,
 79 which allowed for a consistent time frame across all data sets. During the cleaning and processing phase, features with
 80 empty geometries were removed from the shapefile to ensure the removal of NA values in the dataset. Missing data
 81 entries were also omitted, which led to Nantucket County being removed as it had incomplete data and was removed
 82 from our analysis. Also, in this dataset, we are not analyzing Alaska and Hawaii due to their geographical isolation
 83 from the contiguous United States. Centroids of the multi-polygon geometries were calculated to provide a single
 84 point representing the location of these complex shapes. This step was essential for conducting precise location-based
 85 assessments. Lastly, the data frame was transformed into a SpatialPolygonsDataFrame, making it suitable for use in
 86 the Geographic Weighted Regression (GWR) model.

87 **3.3 Statistical Analysis**

88 A Geographically Weighted Regression (GWR) model was used in the statistical analysis, which was created using
 89 the GWmodel library. These packages were integral in setting up the model framework and executing the analysis,
 90 providing tools for spatial data manipulation, regression modeling, and data visualization. The optimal bandwidth
 91 for the GWR model was estimated using cross-validation with hyperparameters such as Gaussian kernel and fixed
 92 bandwidth. The Gaussian kernel was used because it assigns weights to the data points based on their distance from
 93 where the model is being calculated, with closer points receiving higher weights. This weighting falls off smoothly
 94 and symmetrically. The adaptiveness is FALSE because the goal is systematically comparing coefficients across
 95 different geographic regions. So, a fixed bandwidth can help ensure that each region is analyzed under the same
 96 spatial constraints. This function modeled the response variable, Cardiovascular death rate per 100,000 residents:

$$y = \beta_0 * \%White + \beta_1 * \%Black + \beta_2 * \%Hispanic + \beta_3 * \%Asian + \beta_4 * PM2.5 + \beta_5 * MedianIncome + \beta_6 * \%Unemployed$$

97 The variable “y” represents the CVM per 100,000 residents in each county. The GWR was performed using the
 98 identified optimal bandwidth, incorporating the same predictors as those used for bandwidth selection. This ensures the
 99 model’s findings are reliable and applicable to the predictors. Lastly, the p-values were found to determine significance
 100 by using the t-values obtained from the local regression coefficients from the GWR model. Then, the cumulative
 101 distribution function of the T-distribution is used to calculate the p-values. It was done as a two-tail test as we tried to
 102 find significant differences, and degrees of freedom were found from the GWR diagnostic output.

103 **3.4 Mapping**

104 We generated geographic plots to visualize the spatial distribution of the dependent variable (deaths from CVD per
 105 100,000 people) across different counties in the US, highlighting the significance of the variables, and plotted the
 106 significance of each variable within different regions to assess the overall effect on the CVD death rate.

107 This approach allows for examining how various socioeconomic, environmental, and demographic factors influence
 108 health outcomes across different regions in the United States. We showed this in simulated data where we used a
 109 piecewise function to the coordinates, which divides the geographical space into four quadrants and assigns different
 110 coefficients to them to model for spatial heterogeneity. Also, a spatial autocorrelation test (Moran’s I) is performed
 111 on the residuals of the GWR model. From the table in our appendix and the graph in our appendix, we have a
 112 high standard deviation and low p-value, strongly suggesting that the residuals of the GWR model are not randomly
 113 distributed but show significant spatial autocorrelation. The graph shows that the majority of the residuals are close
 114 to zero, which means the predicted values are pretty close to the observed values, indicating that the GWR model has
 115 accounted for spatial variation effectively and that it is a good fit.

116 This methodological outline ensures that each step of the data handling, analysis, and visualization process is docu-
 117 mented, providing transparency and reproducibility of the research findings.

118 **4 Results**

119 **4.1 Regression Summary**

120 The GWR regression showed two models: global regression and GWR regression output. They show the relationships
 121 between socioeconomic, demographic, and environmental variables and death rates. The first is a global regression
 122 model that does not consider spatial correlation despite revealing that all predictors are statistically significant. Hence,
 123 while the model does suggest that our variables are indeed important, the global model may overlook local variations
 124 that are crucial in understanding the true nature of the data. This can be seen in Figure 2, which shows the residual

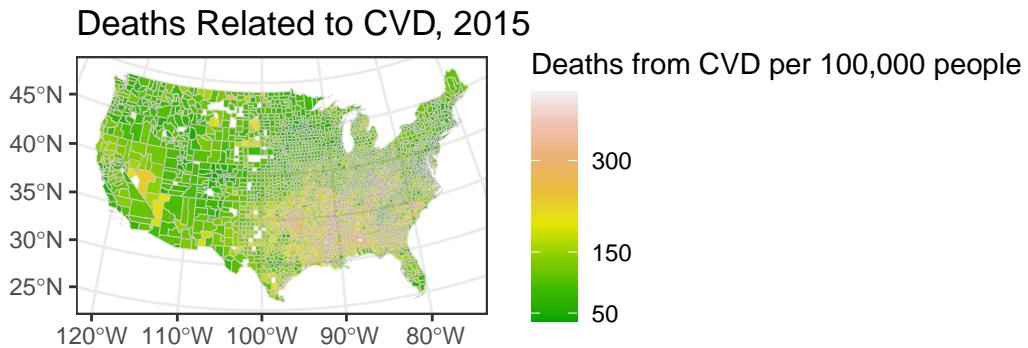


Figure 1

125 of the global model, and there appears to be a spatial pattern to the residuals, with specific areas showing clusters of
 126 higher residuals and the other regions showing clusters of lower residuals. This clustering of residuals suggests that
 127 the global regression model may not be capturing all the spatial variation in the data. This implies that the relationship
 128 between the independent and dependent variables might differ across different locations.

129 On the other hand, the geographically weighted regression (GWR) model incorporates spatial variation, which is a
 130 critical factor given the data context. We can see that the AIC and BIC values are lower for the GWR regression than
 131 the Global regression, which makes it the most preferred model given the data. In summary, by accommodating the
 132 spatial component present in the data, the GWR model provides a more realistic interpretation of how various factors
 133 influence death rates across different regions.

134 4.2 Local Significance Plots

135 In Figure 1, the plot shows an exploratory data analysis (EDA) plot, specifically a choropleth map displaying the
 136 number of deaths from cardiovascular disease (CVD) per 100,000 people across the contiguous United States for the
 137 year 2015. The particulate concentrations include all the covariates mentioned in the methods section. The regions
 138 along the higher latitudes (nearing 45 N) and towards the eastern section (approaching 80 W) display darker shades,
 139 suggesting higher CVD death rates in these areas, specifically the midwest and southwest regions.

140 Figure 3a represents the local significance and magnitude of the '% White' demographic parameter on cardiovascular
 141 disease outcomes. We plotted for the local significance and magnitude because it can help identify areas where the
 142 predictor variable has a stronger or weaker influence on the outcome, leading to targeted insights that would not be

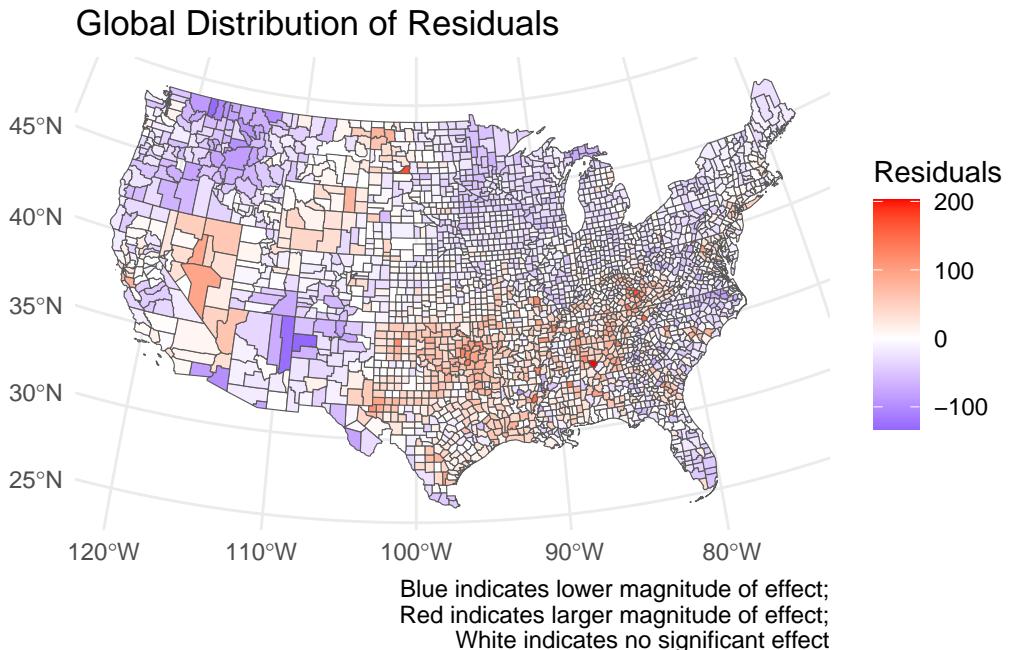


Figure 2

143 possible with a global model. From the plot, the prominent dark purple areas in the central part of the country, extending
 144 towards the southeastern regions, indicate a significant negative correlation between the percentage of the white
 145 population and CVD outcomes in these areas. Like the Northwest region, however, there are few areas, particularly in
 146 the southwest near New Mexico and Arizona, where there are positively correlated to CVD death rate, with the white
 147 regions of counties suggesting no significance and this is white region is similar for all the figures.

148 Figure 3b represents the local significance and magnitude of the '% African' demographic parameter on cardiovascular
 149 disease outcomes. From the plot, areas located approximately in the northern central region indicate a significant
 150 positive correlation to CVD rate, and areas, notably in the central to southeastern areas of the map, suggest a significant
 151 negative correlation.

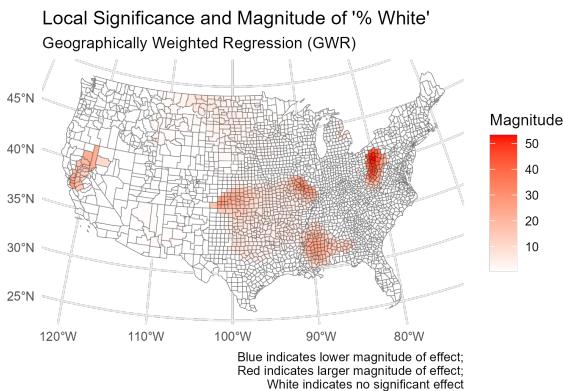
152 Figure 3c represents the local significance and magnitude of the '% Hispanic' demographic parameter on cardiovascular
 153 disease outcomes. From the plot, areas across the central and southeastern regions are shaded in purple, suggesting
 154 a significant negative correlation between the percentage of the Hispanic population and CVD outcomes. The isolated
 155 red patches in the north-central region indicate areas where an increased Hispanic population correlates with higher
 156 CVD outcomes.

157 Figure 3d represents the local significance and magnitude of the '% Asian' demographic parameter on cardiovascular
 158 disease outcomes. From the plot, a substantial portion of the map, particularly across the central to eastern regions, is
 159 colored in various shades of purple. This suggests that in these areas, an increased percentage of the Asian population
 160 correlates with lower CVD outcomes. However, there are some parts of the West where higher percentages of the
 161 Asian population are associated with an increase in CVD outcomes.

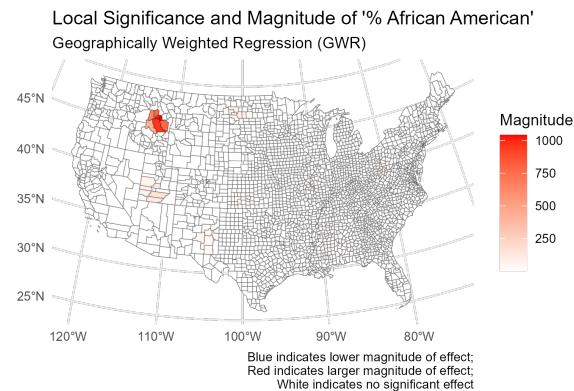
162 Figure 4 represents the local significance and magnitude of the “%p2.5 air quality” parameter on cardiovascular disease
 163 outcomes. From the plot, in the southern and northeastern regions, there increased levels of PM2.5 colored as red in
 164 the plot, meaning they are associated with higher rates of CVD. There are some regions in the West where it was
 165 negatively correlated with CVD.

166 ?@fig-8 represents the local significance and magnitude of the “median income” parameter on cardiovascular disease
 167 outcomes. From the plot, most of the region suggests a significant negative correlation between the median income
 168 parameter and CVD outcomes. However, some parts of Texas indicate a positive correlation to CVD outcomes.

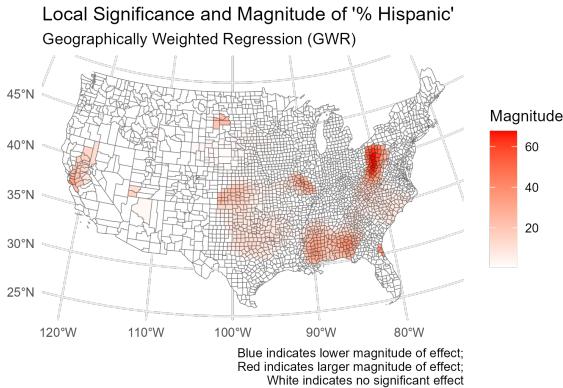
169 ?@fig-9 represents the local significance and magnitude of the “Unemployment” parameter on cardiovascular disease
 170 outcomes. From the plot, there is a positive correlation to CVD rates in most of the central region, and in the southwest
 171 region, there is a negative correlation.



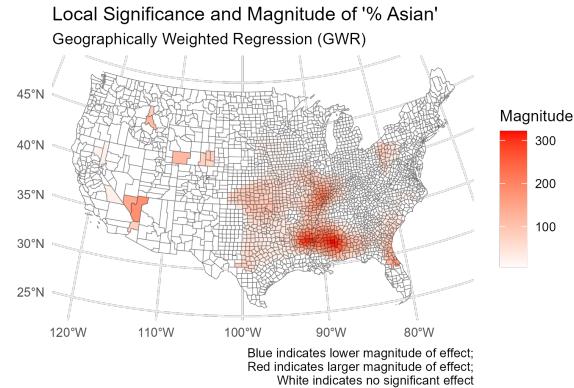
(a) Figure 3



(b) Figure 4



(c) Figure 5



(d) Figure 6

- 172 • Demographic Impact: The percentage of white and Hispanic populations shows significant regional variations
 173 in association with death rates. Higher proportions of white populations correlate with lower death rates in
 174 central areas, whereas higher percentages of Hispanic populations are linked to lower death rates in the West
 175 and Southwest. Conversely, higher percentages of African American populations are associated with higher
 176 death rates in certain Midwestern and Southeastern regions.

Local Significance and Magnitude of PM2.5 Parameter Geographically Weighted Regression (GWR)

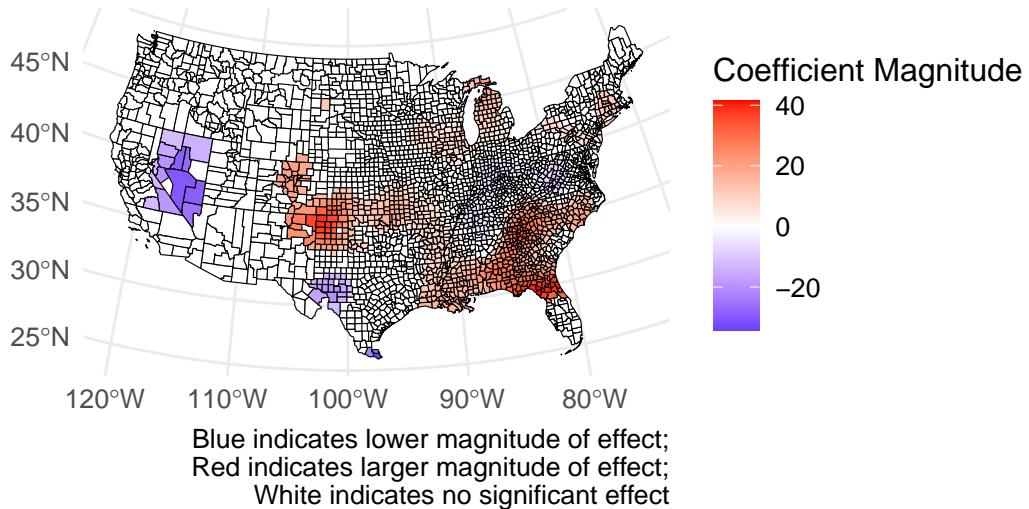


Figure 4

- 177 • Environmental Influence: Air quality, indicated by PM2.5 levels, demonstrates a significant positive relationship with death rates, particularly east of the Rockies, highlighting environmental health as a major concern.
- 178

Local Significance of Median Income Parameter Geographically Weighted Regression (GWR)

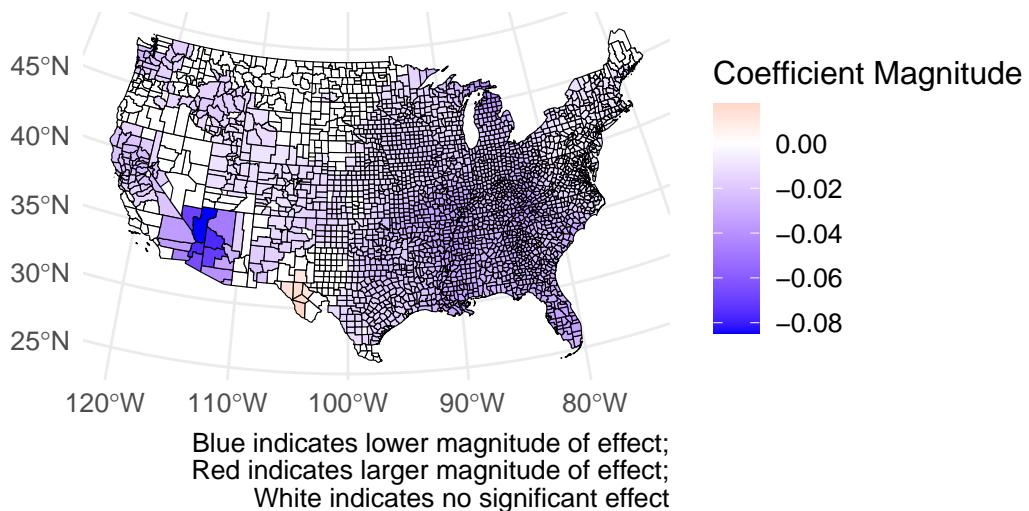


Figure 5

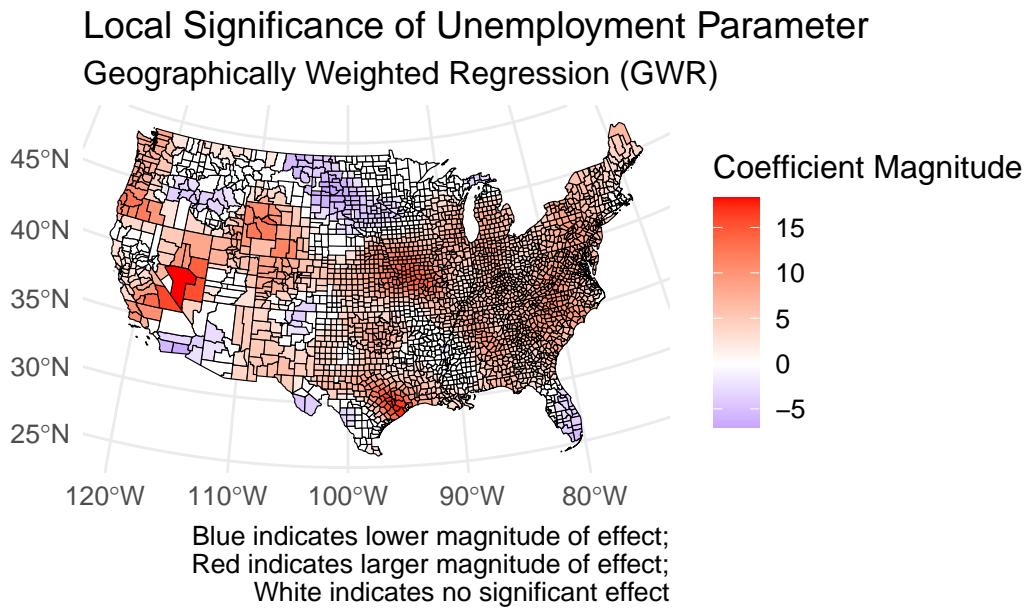


Figure 6

- 179 • Socio-economic Correlation: Median income levels across many regions show a consistent negative association
 180 with death rates, suggesting that higher income areas generally experience fewer deaths.

181 5 Discussion

182 The purpose of this study was to fill a gap in prior studies where the impacts of CVD were primarily studied within
 183 the Stroke Belt, rather than the country at large.

184 Our goal was to put our findings towards answering the following question: what are the socioeconomic and environmental factors affecting CVD rates in the United States?

186 These maps reveal that the relationships between race, socio-economic factors, environmental quality, and death rates
 187 are complex and highly localized. The significance and strength of these relationships vary considerably across different parts of the United States. In contrast, some of the socio-economic factors such as median income show
 189 widespread, consistent significance, implying that the significance of the relationship with CVD outcomes is nearly
 190 constant across various locations. This differs from the heterogeneous local significance that we observed in the
 191 majority of our other variables.

192 Another factor to consider from the plots is the percentage of each race that inhabits each county. We can observe
 193 that for African Americans, Hispanics, and Asians that the impacts are significant in localized areas of the country,
 194 which may reflect underlying health disparities in access to medical care. One limitation of race percentages is the
 195 misreporting of medical records affecting minorities (Tabb et al. 2020). However, this oversight lends further credence

Table 1
 % latex table generated in R 4.3.2 by xtable 1.8-4 package % Wed Apr 24 14:47:05 2024

	1	2	3	4	5	6	7
[ht]	183.22	196.12	7.24	2.56	1.15	2.44	2.62
	81.32	93.80	4.14	2.53	1.77	3.39	2.65
	302.16	313.78	16.89	4.06	1.27	2.42	2.09
	140.83	151.46	5.25	2.13	1.10	2.35	2.73
	131.77	128.39	8.33	2.67	1.07	2.34	2.32
	282.66	294.23	14.14	3.77	1.24	2.44	2.22

196 to the fact that intervention is needed in order to combat racial health disparities. We can look at the Variance Inflation
 197 Factor (VIF) to determine if multi-collinearity has an impact on our results:

198 Typically, when examining the VIF of a GWR model, we want the coefficients for each of the seven columns—which
 199 represent our seven independent variables—to be less than 15. We can see in Table 1 that this standard holds for most
 200 cases of the columns outside the % White column, which has extremely high values. This represents a limitation in
 201 our model, as it shows multicollinearity with respect to the white variable, however, this is to be expected given that
 202 the white population makes up a majority of the United States.

203 The concentration of PM2.5 and the consequential reduction in air quality has a significant impact localized in the
 204 central and southeastern regions of the United States, suggesting environmental health concerns that might require
 205 region-specific intervention.

206 We have shown that by using a GWR model to analyze the relationship between CVM and socioeconomic covariates,
 207 the factors that have the most significant impact on death rates vary by area of the country. This highlights the need to
 208 identify efficient methods of intervention in order to curb one of the deadliest groups of diseases on the planet (Marlow
 209 1994).

210 6 Appendix

211 The R code used for this study can be found in our public GitHub repository: <https://github.com/jbooc117/STAT489-Project.git>.

213 **References**

- 214 Gebreab, Samson Y., and Ana V. Diez Roux. 2012. "Exploring Racial Disparities in CHD Mortality Between Blacks
 215 and Whites Across the United States: A Geographically Weighted Regression Approach." *Health & Place* 18 (5):
 216 1006–14. <https://doi.org/10.1016/j.healthplace.2012.06.006>.
- 217 Liu, Yi, Jingjie Sun, Yannong Gou, Xiubin Sun, Dandan Zhang, and Fuzhong Xue. 2020. "Analysis of Short-
 218 Term Effects of Air Pollution on Cardiovascular Disease Using Bayesian Spatio-Temporal Models." *International
 219 Journal of Environmental Research and Public Health* 17 (3): 879. <https://doi.org/10.3390/ijerph17030879>.
- 220 Marlow, Hilary F. 1994. "The Pharmaceutical Industry Viewpoint." *Cardiology* 85 (1): 102–12. <https://doi.org/10.1159/000176769>.
- 221 Singh, Gitanjali M., Ninon Becquart, Melissa Cruz, Andrea Acevedo, Dariush Mozaffarian, and Elena N. Naumova.
 222 2019. "Spatiotemporal and Demographic Trends and Disparities in Cardiovascular Disease Among Older Adults
 223 in the United States Based on 181 Million Hospitalization Records." *Journal of the American Heart Association* 8
 224 (21): e012727. <https://doi.org/10.1161/JAHA.119.012727>.
- 225 Tabb, Loni Philip, Angel Ortiz, Suzanne Judd, Mary Cushman, and Leslie A. McClure. 2020. "Exploring the Spatial
 226 Patterning in Racial Differences in Cardiovascular Health Between Blacks and Whites Across the United States:
 227 The REGARDS Study." *Journal of the American Heart Association* 9 (9): e016556. <https://doi.org/10.1161/JAHA.120.016556>.
- 228 Terry, Katrina, Mohamed Makhlouf, Salah E. Altarabsheh, Vaishali Deo, Fanny Petermann-Rocha, Yakov Elgudin,
 229 Khurram Nasir, Sanjay Rajagopalan, Sadeer Al-Kindi, and Salil Deo. 2023. "Trends in Cardiovascular Disease
 230 Mortality by County-Level Social Vulnerability Index in the United States." *Journal of the American Heart Asso-
 231 ciation* 12 (20): e030290. <https://doi.org/10.1161/JAHA.123.030290>.
- 232 Warsito, Budi, Hasbi Yasin, Dwi Ispriyanti, and Abdul Hoyyi. 2018. "Robust Geographically Weighted Regression of
 233 Modeling the Air Polluter Standard Index (APSI)." *Journal of Physics: Conference Series* 1025 (May): 012096.
 234 <https://doi.org/10.1088/1742-6596/1025/1/012096>.
- 235 Zelko, Andrea, Pedro R. V. O. Salerno, Sadeer Al-Kindi, Fredrick Ho, Fanny Petermann Rocha, Khurram Nasir, Sanjay
 236 Rajagopalan, Salil Deo, and Naveed Sattar. 2023. "Geographically Weighted Modeling to Explore Social and
 237 Environmental Factors Affecting County-Level Cardiovascular Mortality in People With Diabetes in the United
 238 States: A Cross-Sectional Analysis." *The American Journal of Cardiology* 209 (December): 193–98. <https://doi.org/10.1016/j.amjcard.2023.09.084>.
- 239