Spatial Modeling of Cardiovascular Disease Associated with Increasing Airborne Particulate Matter

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 16, 2024

Abstract

Cardiovascular disease (CVD) is the leading cause of death in the United States. By using a spatial modeling technique (geographically weighted regression), we found the concentration of PM 2.5 is centered in the Strokebelt region. Policymakers and health practitioners can use these results to

identify targeted interventions to curb the increasing rates of CVD, aiming to halt one of the world's deadliest diseases.

***K*eywords** Fine Particulate Matter (PM2_5) • Cardiovascular Disease (CVD) • Cardiovascular Mortality (CVM)

1 Introduction

- Broad strokes: Explain how cardiovascular disease is the leading cause of death in the US, our paper filing the gap by putting the focus on the 18-44-year-old age group
- Specifics: Explicitly state how the CVD incidence rate varies spatially in the US, for our covariates. Our approach involves analyzing this relationship to produce risk prediction estimates for our specified age group.
- Central thesis: Despite past studies focusing on the Stroke Belt region of the United States, we aim to put our focus on the nationwide impact of PM 2.5 concentration, median household income, and unemployment rates on cardiovascular disease deaths, necessitating regional interventions.
- Step back: The overall outcome of our study involves aiding policymakers and health practitioners develop the necessary interventions in the targeted regions that are most affected by CVD.Related Works

Paragraph 1) Review of regression approaches in cardiovascular disease mortality 2) Review of geographically weighted regression models 3) GWR with CVD 4) Risk prediction Model 5) Direct comparison with clustering approaches

2 Related Works

Paragraph 1. (Review of regression approaches in cardiovascular disease mortality) »» There's been a growing popularity for the use of spatial models in the epidemiological domain to analyze the factors and spatial distribution of a certain disease. The spatial models and techniques used differ between studies however the end goal remains the same, reduce the disease mortality rates for the population. Statistical regression models have been a popular choice, specifically for CVD-related studies. The factors that compromise the underlying causes for CVD can be seen as a dynamic and interconnected web. Zelko et. al examines the relationship between CVD and covariates similar to our study, such as air pollution, social determinants, and county-level data. By using a GWR model, they found a correlation between household income, race, and healthcare access. Their results also showed that counties in the South had the highest PM 2.5 concentrations.

Paragraph 2. (Review of geographically weighted regression models) »» The geographically weighted regression model (GWR) stands out from the traditional regression models to explore spatial data by doing so on the local scale as well as taking into account varied coefficients for a certain spatial unit (Gebread et. al). The GWR model uses the weighted least square method to estimate the regression coefficients. In general, we are interested in seeing values closer to the point of interest than values farther from it, as they carry more weight and have a greater influence. The GWR model includes the parameter, neighborhood (also known as the bandwidth).

Paragraph 3. (Direct comparison with OLS approaches) »» As seen in other studies, the Ordinary Least Squares model is a popular method for public health studies. Compared to the GWR model, the OLS model generates the regression coefficients on a global scale. Because of this, the OLS method may not be the ideal choice. as it is prone to hidden spatial variability. However, our paper takes into account the fact that socioeconomic factors are not constant on a time and space basis with cardiovascular disease by using a GWR model, as it's a much more complex relationship. That's not to say the OLS model has no use in epidemiological research studies. One of the main goals of public health is to better the population as a whole. The OLS model is efficient in capturing the average differences between covariates. To get to the root cause of CVD mortality rates, we need to start with the GRW model, to see overall improvement in the OLS model later on.

Paragraph 4. (GWR with CVD) »» The GWR model shows whether or not a linear relationship exists between cardiovascular disease deaths and the covariates, making it suitable to see where the areas of focus should be placed in terms of improvement as well as an increase presence of healthcare practitioners. The GWR model is better for analyzing disparities at the local-scale, see which socioeconomic factor affects different regions, and then implement a policy for that specific region. Past studies tend to soley focus on the trends of socioeconomic covariates and its spatial patterning. Our paper expands on past studies, by including the concentration of air pollutants as one of our covariates. By doing so, we are able to fully explore the dynamic relationship between geographical units and CVD incidence/mortality rates.

Paragraph 5. (Preventative risk measures + impact) »» Risk assessment and risk estimates helps to understand the key factors associated with CVD. The concentration of specific races vary by counties so by studying the underlying socioeconomic covariates of CVD mortality on the county-level, researchers and health practitioners are thus able to curate the necessary risk preventative measures and allocate resources to the areas that need it the most. Furthermore, the dose-response relationship of air pollutant concentrations in a region and its effects can be analyzed. By putting the focus on smaller units, CVD-mortality rates will start to see a decline, nationally and globally.

3 Methods

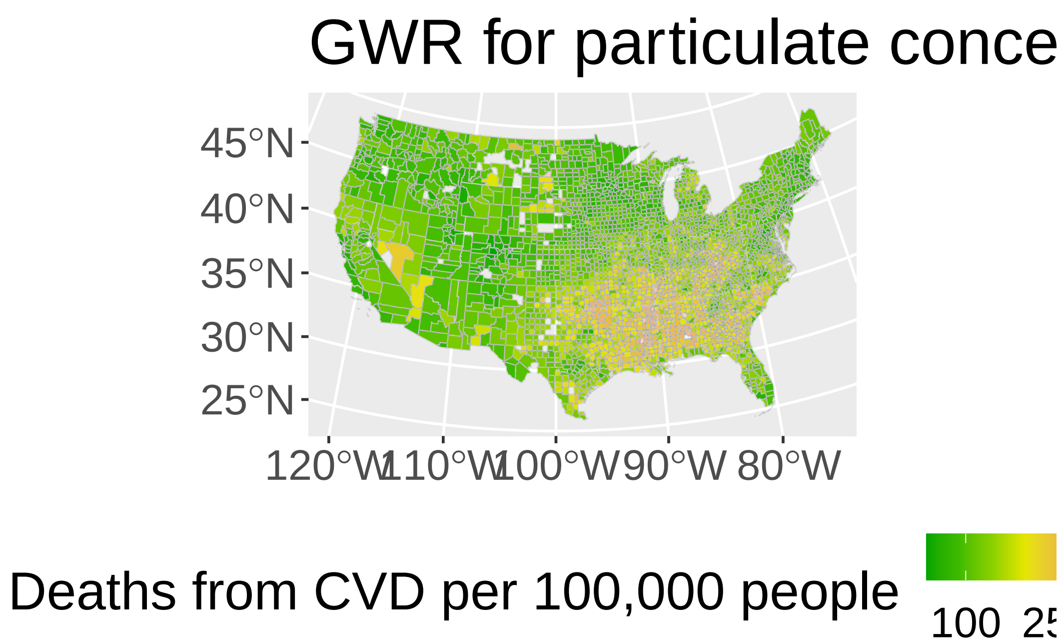
3.0.1 Data

- Combination of four data sets (will add citations later): CDC cardiovascular disease death rates per US counties, NASA particulate matter concentrations for the year 2015, US Census Data for 2015, and US Vaccination rate by county data
  - Vaccination rates were not considered in the study, rather, the dataset was used for its geometry listing, which will be important for our GWR section.

- We merged the four data sets together to analyze multiple variables against the CVM per 100,000 residents in each county. Variables included % of race present in county, PM2.5 concentrations, median income, and unemployment rates.

3.0.2 Geographically Weighted Regression Model

- Local Significance
  - Coefficients have t-values and SE values, converted t-values into p-values with formula from code, extracted these p-values for each covariate and then plotted significance from p-values. The results of this should show which covariates have higher effects per county
  - Per our (source), different regions of the United States should have different variables with greater significance relating to CVD rates.
    - For example, in the midwest, food insecurity was found to be the most significant factor, while in the West it did not play much of a role compared to income, PM2.5, and access to healthcare.



4 Results

```
**********************************************************************
*                       Package   GWmodel                            *
**********************************************************************
Program starts at: 2024-04-16 11:57:24.063415
Call:
```

```
gwr.basic(formula = Data_Vl ~ prc_wht + prc_frc + prc_hsp + perc_sn +
 p2_5_20 + estimat + U__2015, data = mergedf_spatial, bw = opt_bandwidth,
 kernel = "gaussian", adaptive = FALSE, parallel.method = "omp")

Dependent (y) variable:  Data_Vl
Independent variables:  prc_wht prc_frc prc_hsp perc_sn p2_5_20 estimat U__2015
Number of data points: 3057
***********************************************************************
*                    Results of Global Regression                     *
***********************************************************************

Call:
 lm(formula = formula, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-133.061  -22.813   -5.661   19.637  202.340

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.388e+02  1.024e+01  23.313  < 2e-16 ***
prc_wht     -8.158e+01  9.500e+00  -8.587  < 2e-16 ***
prc_frc      5.526e+01  9.988e+00   5.532 3.42e-08 ***
prc_hsp     -9.321e+01  1.023e+01  -9.107  < 2e-16 ***
perc_sn     -2.090e+02  3.425e+01  -6.101 1.18e-09 ***
p2_5_20      6.311e+00  4.453e-01  14.170  < 2e-16 ***
estimat     -2.120e-03  7.150e-05 -29.648  < 2e-16 ***
U__2015      3.736e+00  4.118e-01   9.072  < 2e-16 ***

---Significance stars
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 35.64 on 3049 degrees of freedom
Multiple R-squared: 0.6105
Adjusted R-squared: 0.6096
F-statistic: 682.6 on 7 and 3049 DF,  p-value: < 2.2e-16
***Extra Diagnostic information
```

```
Residual sum of squares: 3873809
Sigma(hat): 35.6093
AIC:  30534.31
AICc:  30534.37
BIC:  27603.76
***********************************************************************
*          Results of Geographically Weighted Regression              *
***********************************************************************

*********************Model calibration information*********************
Kernel function: gaussian
Fixed bandwidth: 128251.7
Regression points: the same locations as observations are used.
Distance metric: Euclidean distance metric is used.

****************Summary of GWR coefficient estimates:******************
                 Min.     1st Qu.      Median     3rd Qu.      Max.
Intercept -1.8237e+02  1.6625e+02  2.7812e+02  4.1353e+02 2293.3931
prc_wht   -1.7708e+03 -2.3314e+02 -1.1902e+02 -3.3864e+01  309.4663
prc_frc   -1.9197e+03 -2.0396e+02 -6.1895e+01  5.2242e+01 3229.3222
prc_hsp   -1.7720e+03 -2.8864e+02 -1.5801e+02 -6.4279e+01  504.0406
perc_sn   -1.9199e+03 -6.5045e+02 -3.4826e+02 -1.4864e+02 1635.4916
p2_5_20   -4.8476e+01 -9.9089e-01  3.7131e+00  9.5388e+00   41.5626
estimat   -8.0369e-03 -2.6333e-03 -1.8226e-03 -1.1243e-03    0.0017
U__2015   -7.0110e+00  2.6200e+00  5.3167e+00  7.4156e+00   18.3070
************************Diagnostic information*************************
Number of data points: 3057
Effective number of parameters (2trace(S) - trace(S'S)): 576.851
Effective degrees of freedom (n-2trace(S) + trace(S'S)): 2480.149
AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): 28146.12
AIC (GWR book, Fotheringham, et al. 2002,GWR p. 96, eq. 4.22): 27581.59
BIC (GWR book, Fotheringham, et al. 2002,GWR p. 61, eq. 2.34): 27506.99
Residual sum of squares: 1290938
R-square value:  0.8701873
Adjusted R-square value:  0.8399823
```

```
************************************************************************
Program stops at: 2024-04-16 11:57:25.225759
```

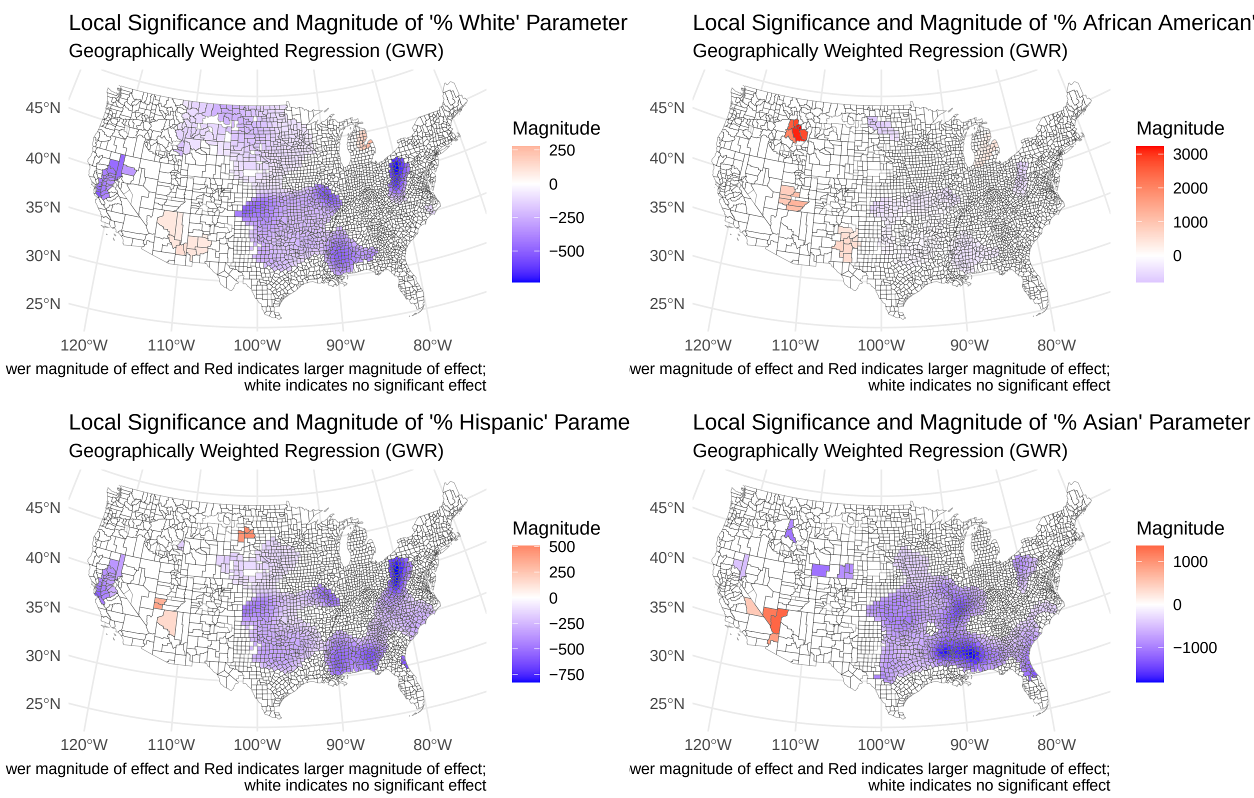
Table Summary interpretation: The analysis compares two distinct statistical models to explore the relationships between socio-economic, demographic, and environmental variables and death rates. The first is a global regression model which, despite revealing that all predictors are highly statistically significant, does not take into account spatial correlation. Hence, while the model does suggest that our variables are indeed important, the global model may overlook local variations that are crucial in understanding the true nature of the data.

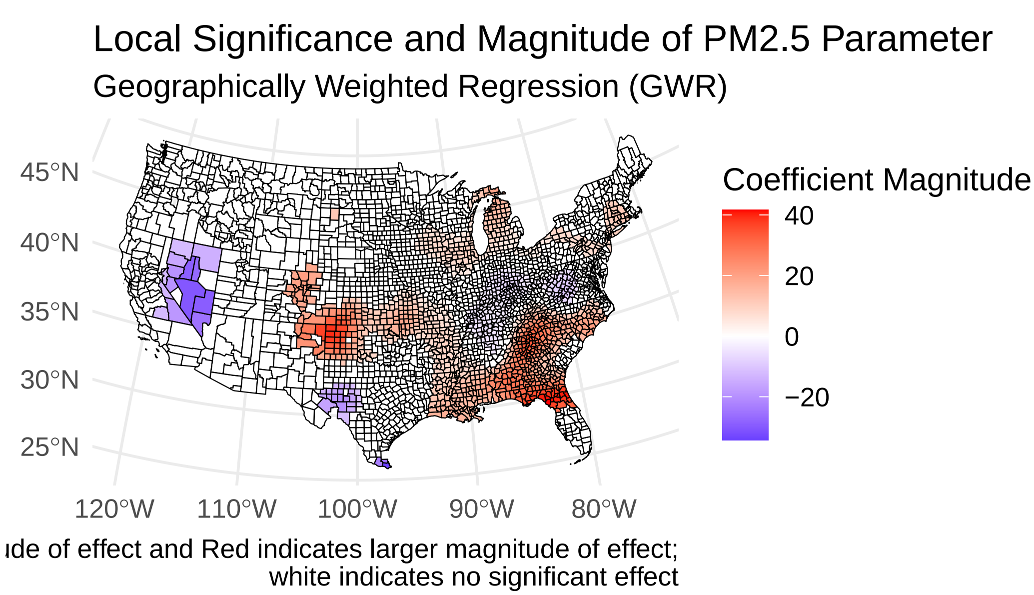
On the other hand, the geographically weighted regression (GWR) model incorporates spatial variation, which is a critical factor given the context of the data. We can see how well the GWR model worked by looking at the R-squared value, which is .870, a significant improvement over the global model. This R-squared value, along with the use of localized spatial statistics, confirms the non-uniform relationship between the predictors and the response variable across different geographical areas. In summary, the GWR model, by accommodating the spatial component present in the data, provides a more realistic interpretation of how various factors influence death rates across different regions.

Significance plot:-

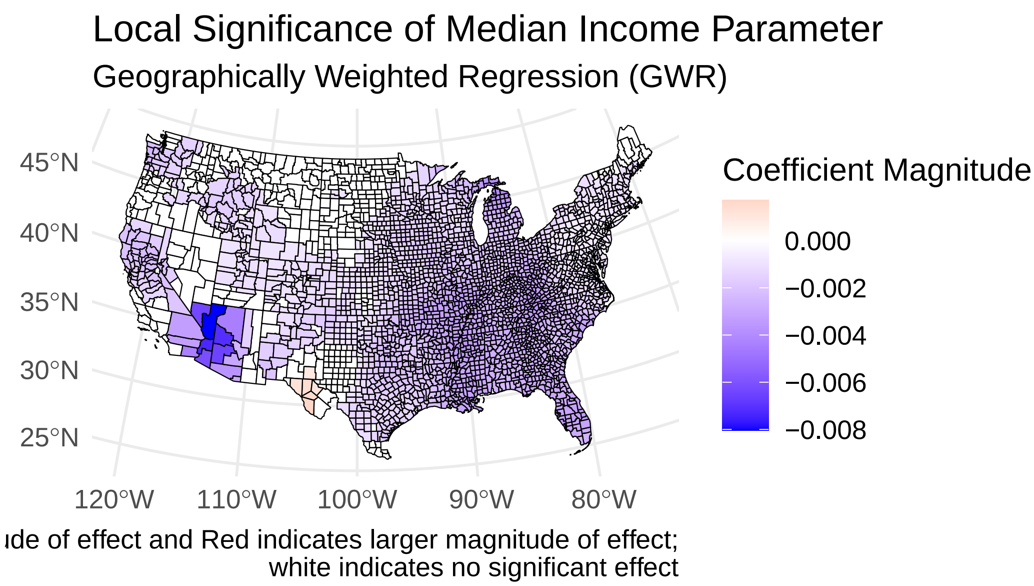
The plots provided depict the local significance and magnitude of different parameters (covariates) from a geographically weighted regression (GWR) model, focusing on their relationship with death rates across the United States. The areas in red indicate regions where the covariates are statistically significant, and the color's intensity represents the effect's magnitude.

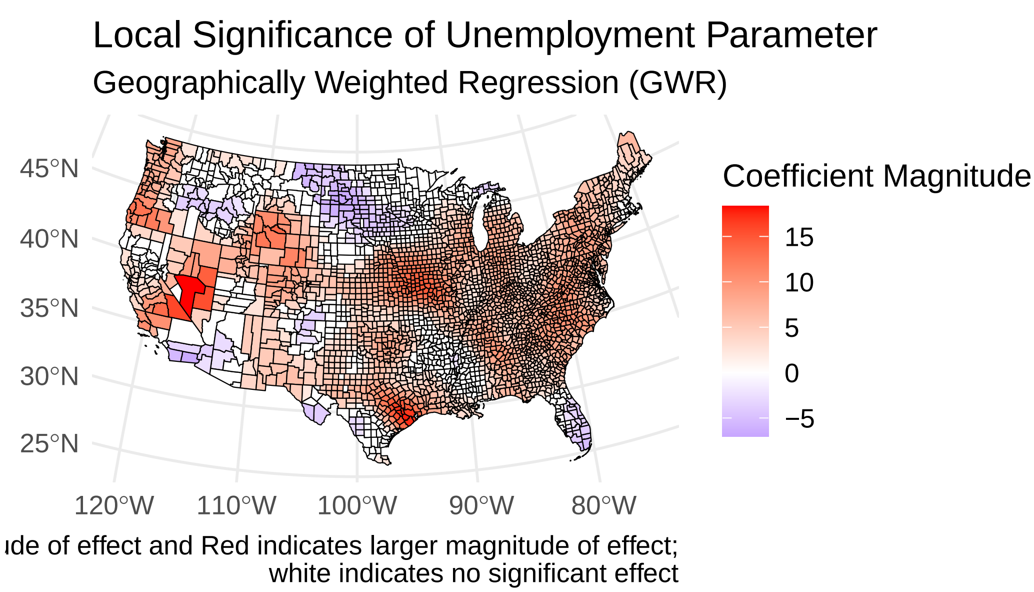
- Demographic Impact: The percentage of white and Hispanic populations shows significant regional variations in association with death rates. Higher proportions of white populations correlate with lower death rates in central areas, whereas higher percentages of Hispanic populations are linked to lower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





- 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.





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

5 Discussion

These maps reveal that the relationships between race, socio-economic factors, environmental quality, and death rates are complex and highly localized. The significance and strength of these relationships vary considerably across different parts of the United States. For example,

- Socio-economic factors like income show widespread significance, implying a consistent relationship with health outcomes across various locations.
- Race impacts are significant in certain areas, which may reflect underlying health disparities, access to care, or other social determinants of health.
- Air quality has a broad impact, suggesting environmental health concerns that might require region-specific policies.

References

n.d.

(n.d.)