Bioinformatics Summer School Report

Name: Mohammed Al-Hubail & Ali Alsalman
Internship Organization: Duke University
Supervisor's Name: Prof. Mercedes Bravo
Project title: Exploring the Impact of Environmental and Climate Exposures and Social Determinants of Health (SDoH) on Multiple Sclerosis (MS)

1. Introduction

Multiple Sclerosis (MS) is a disease in which the immune system attacks the myelin sheath that protects neurons throughout the body, which can slow communication between the brain and the body. Symptoms include partial or complete vision loss, numbness or weakness in limbs, spasms, fatigue, depression, and walking difficulties.
There have been some studies suggest that MS symptoms are exacerbated with an increase in body temperature or ambient temperature. [3]

Objective: Identify associations between SDoH, environmental, and climate stresses with prevalence of Multiple Sclerosis (MS) The study will provide, for the first time, a picture of the association between SDoH, environmental and climate variables, and MS prevalence for Duke patients residing in North Carolina

Primary Outcome: Prevalence of MS by Census Tract Study Population: Adults at Duke Health who have been diagnosed with MS since January 1, 2016 (ICD 10: G35) Negative Associations: Spatial healthcare access, education attainment, median household, food security Positive Associations: Percent of households below the poverty line, unemployment rate, humidity, air pollution, temperature, social vulnerability index, percent uninsured, percent Hispanic, percent non-white, average age, average BMI

2. Materials and Methodology

Phase 1: Data Collection and Cleaning

- Collect data on Social Determinants of Health and Climate Exposure.
- Clean and preprocess the data to ensure quality and consistency.

Phase 2: Data Integration and Analysis

- Merge datasets and create visual correlation plots.
- Identify and remove highly correlated variables to prevent multicollinearity.

Phase 3: Model Development

- Develop and run Poisson Regression, Quasi-Poisson, and CARleroux Spatial Models to analyze the data.

Here's an extended version of your text with more detail:

Data Sources and Methodology: The analysis leveraged data from multiple authoritative sources, including the United States Department of Agriculture (USDA), the American Community Survey (ACS) conducted by the U.S. Census Bureau, Environmental Protection Agency (EPA) Frequently Asked Questions on Data (FAQSD), Duke Health Enterprise Research Database (HERD), County and Rural Development Model (CRDM), ReferenceUSA, and the Daymet R Package for climate data. These data sources provided a comprehensive set of variables relevant to the study, encompassing socioeconomic, demographic, environmental, and climatic factors.

Geographic Scope: The study was confined to specific counties in North Carolina, namely Caswell, Person, Granville, Vance, Warren, Alamance, Durham, Orange, Franklin, Nash, Wake, Wilson, Johnston, Harnett, Lee, and Chatham. These counties were selected based on their relevance to the study's objectives, focusing on areas with varying degrees of socioeconomic challenges.

Analytical Approach: The data analysis was conducted using R, with two primary statistical models being developed:

- **Poisson Regression Model:** This model was used to analyze count data, typically incidents or occurrences within census tracts, to identify key predictors among the variables collected. Poisson regression is particularly suitable for modeling the frequency of events when the variance is equal to the mean, common in count data scenarios.

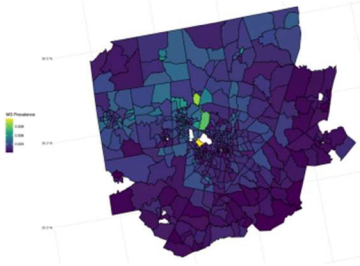
- **CARleroux Spatial Model:** The Conditional Autoregressive (CAR) model with Leroux prior was employed to account for spatial correlation and non-independence between census tracts. This model is critical in spatial epidemiology and public health research, as it adjusts for spatial autocorrelation, ensuring that the influence of neighboring areas is properly accounted for in the analysis. The CARleroux model was specifically chosen for its robustness in handling spatially structured random effects.

- **Variable Selection and Exclusions:** Several socioeconomic variables, including median income, educational attainment, unemployment rate, and the percentage of the population below the poverty line, were excluded from the final model due to their high correlation with the Social Vulnerability Index (SVI). The SVI itself is a composite measure that encompasses these variables, so their exclusion was necessary to prevent multicollinearity and to maintain the integrity of the model.

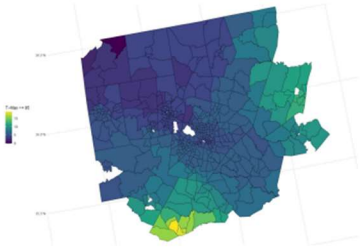
  Handling Missing Data and Outliers: Census tracts with missing data were excluded from the analysis. These exclusions were due to non-response or a lack of population, ensuring that only complete and reliable data were used. Additionally, the analysis accounted for outliers, retaining census tracts that were identified as socioeconomically challenged to provide a more focused analysis of vulnerable populations.

- **Results and Discussion**

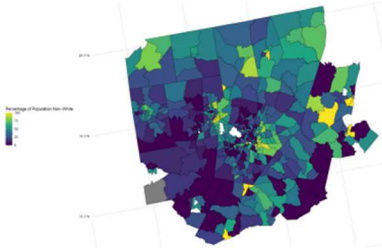
MS Prevalence by 2010-2019 Census Tracts



Number of Days in Summer 2015 with Daily High of 95 or Above by 2010-2019 Census Tracts



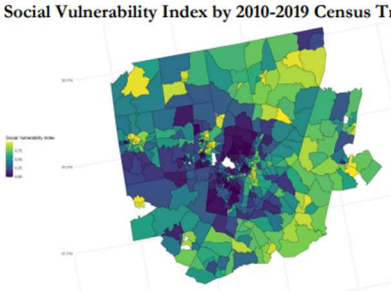
Percent Non-White by 2010-2019 Census Tracts



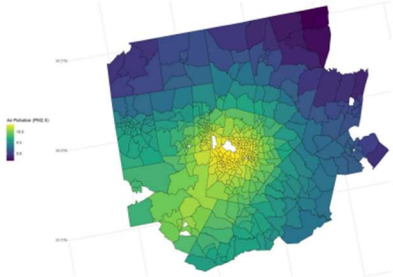
Percent Uninsured by 2010-2019 Census Tracts



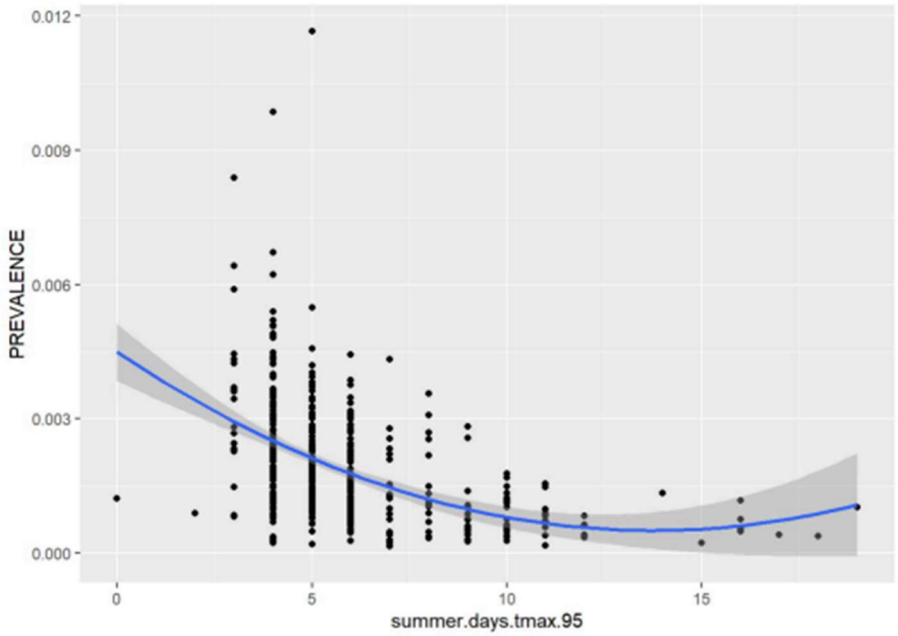
Social Vulnerability Index by 2010-2019 Census Tracts



Air Pollution (PM2.5) by 2010-2019 Census Tracts



Quadratic Regression of Scatterplot of the Number of Days in Summer 2015 with Daily High of 95 or Above



CARleroux Spatial Model Coefficients and Confidence Intervals for Significant Variables

Variable Name	Correlation Coefficient	95% Credible Interval
Number of Days in Summer with Daily High of 95 or Above	-0.3989 0.0396 (Quadratic)	-0.5033, -0.2988 0.0002, 0.0766 (Quadratic)
Percent Non-White	0.1996	0.1185, 0.2818
Percent Uninsured	0.1243	0.0594, 0.1871
Social Vulnerability Index	-0.2642	-0.3505, -0.1793
Age	0.0929	0.0239, 0.1615

- **Skills and Knowledge Gained**

  I have developed strong teamwork skills through collaborative projects, where I successfully worked with diverse teams to achieve common goals. My ability to communicate effectively, coordinate tasks, and support team members has been crucial in driving project success. In addition to my teamwork abilities, I have honed my technical expertise in the R programming language, particularly in the field of spatial and analytical statistics within bioinformatics. My experience includes performing complex data analyses, developing statistical models, and interpreting results to extract meaningful insights from biological data. These skills have enabled me to contribute to advanced research projects, where precise data analysis and spatial modeling are critical for understanding biological phenomena.

- **Challenges and Solutions**
- Throughout the project, several challenges arose that required careful consideration and mitigation strategies. One significant issue was the presence of flaws in the data, which stemmed from a non-random sample of Multiple Sclerosis (MS) patients. The data was strictly limited to patients within the Duke Health System, introducing potential biases that could affect the generalizability of the study's findings. This limitation meant that the sample might not accurately represent the broader population of MS patients, which could influence the statistical inferences drawn from the analysis.
- Additionally, the temperature data posed its own challenges. The calculations were interpolated from satellite data to 1 km^2 grids, which, while providing a fine spatial resolution, also introduced potential errors due to the interpolation process. This method assumes a level of uniformity across each grid that may not fully capture microclimatic variations, potentially impacting the accuracy of temperature exposure assessments for individual patients.

- The aggregation of patient data from the Duke EHR CRDM (Electronic Health Record and Clinical Research Data Mart) was another area of concern. This aggregation was specific to MS patients, and while it provided a valuable dataset for analysis, the aggregation process might have led to a loss of granularity in the data, limiting the ability to conduct more nuanced analyses at the individual level.
- Finally, the interpolation of PM2.5 (particulate matter) data from monitoring stations to estimate exposure levels was a crucial yet challenging aspect of the project. The spatial distribution of monitoring stations is often uneven, leading to potential inaccuracies in the interpolated values, particularly in areas with fewer monitoring points. This could result in less precise estimations of air pollution exposure, which is a critical factor in studies examining environmental impacts on health outcomes.

6. Conclusion

Quadratic Associations: There was a statistically significant quadratic relationship between the number of days in the summer of 2015 with daily high temperatures of 95°F or above and the outcome variable. The Likelihood Ratio Test confirmed that the quadratic regression model provided a significantly better fit compared to a linear model.
Negative Associations: The Social Vulnerability Index was found to have a negative association with the outcome, indicating that higher social vulnerability is linked to a decrease in the measured effect.
Positive Associations: Several factors showed positive associations with the outcome, including the percentage of the population that is uninsured, the percentage of the population that is non-White, and the age of the population. These factors were identified as contributing to an increased effect in the analysis.

7. Acknowledgments

Thank you to KAUST Academy for sponsoring this opportunity and to Duke University Data+ and Climate+ Management for their invaluable support.

8. References (if applicable)

Noorimotlagh Z, Azizi M, Pan HF, Mami S, Mirzaee SA. Association between air pollution and Multiple Sclerosis: A systematic review. Environmental Research. 2021;196:110386. doi:10.1016/j.envres.2020.110386 Sisodiya SM. Hot brain: practical climate change advice for neurologists. Practical Neurology. Published online November 2023:pn-2023-003777. doi:10.1136/pn-2023- 003777 Thermosensitive People with Multiple Sclerosis: A Randomized Controlled Trial. Tambasco N, ed. Acta Neurologica Scandinavica. 2024;2024. doi:10.1155/2024/4882755 Dobson R, Rice DR, D'hooghe M, et al. Social determinants of health in multiple sclerosis. Nature Reviews Neurology. 2022;18(12):723-734. doi:10.1038/s41582-022- 00735-5