Haddon Sandefur

Scott Sun

Trevor Johnson

Statistical Learning Group Proposal

Predicting Breast Cancer Mortality on Greenness and Polluting Site Exposures

# Introduction:

The philosophy surrounding cancer treatment has evolved to emphasize preventative measures. Multiple preventative factors contributing to the genesis of cancers are common knowledge, such as smoking or asbestos exposure. Given that exposure to these factors is often avoidable, research pertaining to the discovery of more such factors has been important. Currently, there’s limited research pertaining to exposure to the amount of vegetation or ‘greenness’, and its association with cancer development and mortality (O’Callaghan-Gordo et. al, 2018). Certain studies additionally demonstrate linkages between greenness and mortality, particularly relating to cancer (James, Hart et. al, 2016). With this in mind, we wish to use discovery techniques to study the association of greenness and breast cancer mortality. Breast Cancer is one of the most prevalent diseases in the United States, being the second most prevalent cancer in women nationally (NIH, 2019).

We will utilize county breast cancer mortality data from GHDx’s data base for our analysis. This data will be paired with greenspace data and pollution site data collected from the University of Michigan’s National Neighborhood Data Archives (NaNDA). Consolidation will occur by county FIPS codes, as well as census tracts. Our analysis will consist of regression techniques for prediction purposes, utilizing breast cancer mortality counts (by county) as an outcome of interest. We intend to utilize greenspace data as our predictor variables. Polluting sites will be included for the sake of being an environmental control. This analysis will be performed cross-sectionally for the year 2010, and thus cannot account for shifts in vegetation density over time.

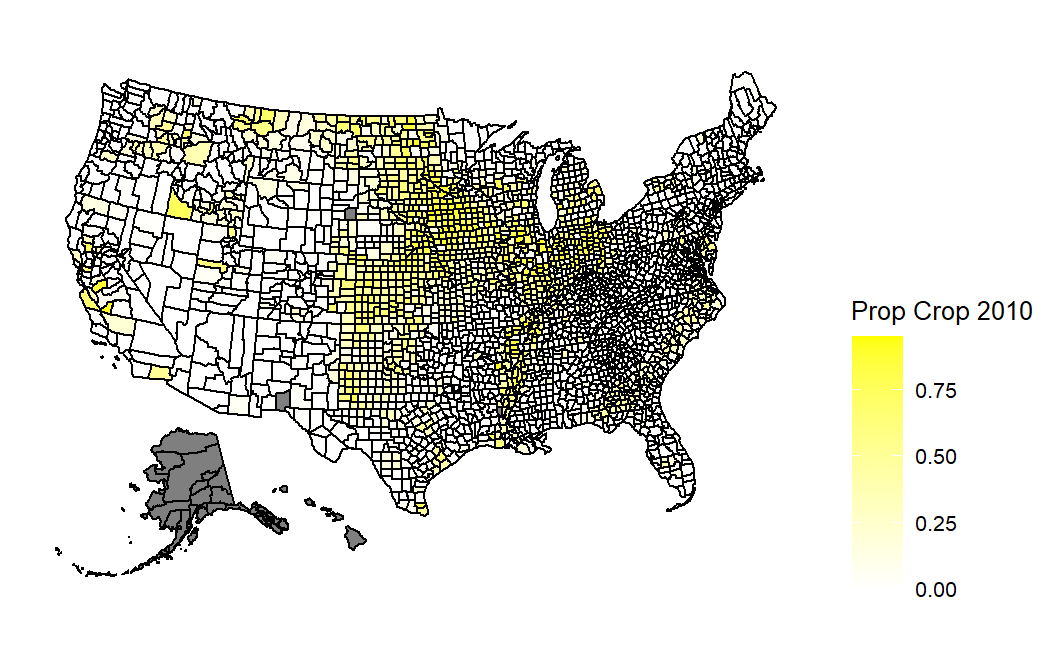
# Motivation from a Geographical Perspective:

Using the geographical information in our dataset, we can create cartographic plots to demonstrate similar densities amongst variables by county. These visualizations can demonstrate the motivation behind the intention to use a clustering technique. Below, you will find maps plotting breast cancer mortality, county proportion of agricultural land use, and county proportion of wooded wetlands. The elliptical shapes drawn on each map highlight areas in which there may be a relationship between these variables.

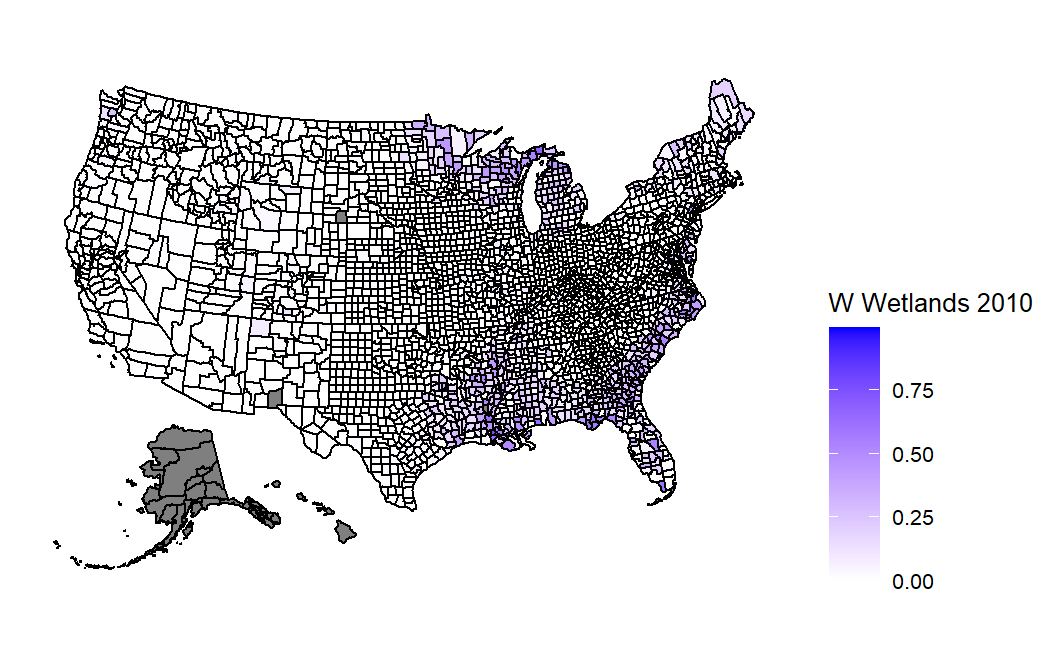
## Breast Cancer Mortality by County (2010)

# 

## Proportion of Crop Land Use by County (2010)



## Proportion of Wooded Wetlands by County (2010)



# Principal Component Analysis

# We conduct PCA on the NaNDA data, which are used as the predictor variables in all the potential learning models. The scree plot below visualizes how much variance is explained by each single principal component and the cumulative variance explained. Based on the plot, we do not observe a perfect “elbow” point. The predictor variables cannot be represented by only a few principal components.

Chart, line chart

Description automatically generated

# K-Means Clustering

We conduct K-Means Clustering on the first and second principal components. We try with three different values for the number of clusters and visualize the results. Based on the plot below, we observe that there seems only one cluster for the first two principal components.

Chart

Description automatically generated

# References

1. O'Callaghan-Gordo C, Kogevinas M, Cirach M, Castaño-Vinyals G, Aragonés N, Delfrade J, Fernández-Villa T, Amiano P, Dierssen-Sotos T, Tardon A, Capelo R, Peiró-Perez R, Moreno V, Roca-Barceló A, Perez-Gomez B, Vidan J, Molina AJ, Oribe M, Gràcia-Lavedan E, Espinosa A, Valentin A, Pollán M, Nieuwenhuijsen MJ. Residential proximity to green spaces and breast cancer risk: The multicase-control study in Spain (MCC-Spain). Int J Hyg Environ Health. 2018 Sep;221(8):1097-1106. doi: 10.1016/j.ijheh.2018.07.014. Epub 2018 Aug 1. PMID: 30076044.
2. Hari S. Iyer, Jaime E. Hart, Peter James, Elise G. Elliott, Nicole V. DeVille, Michelle D. Holmes, Immaculata De Vivo, Lorelei A. Mucci, Francine Laden, Timothy R. Rebbeck. Impact of neighborhood socioeconomic status, income segregation, and greenness on blood biomarkers of inflammation. Environment International, Volume 162, 2022. 107164, ISSN 0160-4120. <https://doi.org/10.1016/j.envint.2022.107164>.
3. Lifetime Risk (Percent) of Dying from Cancer by Site and Race/Ethnicity: Females, Total US, 2014-2016 (Table 1.19). https://seer.cancer.gov/csr/2016\_2018/results\_merged/topic\_lifetime\_risk.pdf. 2019. Accessed November 5, 2021.