Predicting Breast Cancer Mortality on Greenness and Polluting Site Exposures

Statistical Learning Group Proposal

Haddon Sandefur, Scott Sun, Trevor Johnson

# Introduction:

In recent years, the philosophy surrounding cancer has evolved to emphasize preventative measures over symptomatic treatment and the pursuit of remission. Many preventative factors contributing to the genesis of cancers have been identified; smoking tobacco or exposure to asbestos for example, and it is now common knowledge to avoid these carcinogens. Given that exposure to these factors is often avoidable, it is imperative to pursue research pertaining to the discovery of more such carcinogenic factors.

Some studies have demonstrated linkages between the amount of vegetation or ‘greenness’, and mortality relating to cancer (James, Hart et. al, 2016); however, there is still an overall limited amount research pertaining to associations between greenness and cancer development and mortality (O’Callaghan-Gordo et. al, 2018). With this in mind, this study will seek to use discovery techniques to examine the association of greenness and breast cancer mortality. Breast Cancer is the second most prevalent cancer in women in the United States (National Institutes of Health, 2019); therefore, it is of the utmost importance to aggressively combat this disease through the discovery of factors related to its development and mortality.

# Data Description:

This study will utilize breast cancer mortality data from the Global Health Data Exchange (GHDx) data base alongside greenspace data and pollution site data collected from the University of Michigan’s National Neighborhood Data Archives (NaNDA). The entirety of the data has been acquired as of the drafting of this proposal, and has been concatenated by county FIPS codes, as well as census tracts to aid in the ease of analysis. The resulting dataset that will be employed contains 72,384 records for analysis with 148 fields including those for identifying information such as state, county, and year. This study will aim to utilize this data in order to identify factors from the NaNDA data that may be associated with breast cancer mortality as reported by the GHDx data base.

# Analysis Plan:

The analysis will consist of regression techniques for prediction purposes, utilizing breast cancer mortality counts by county as an outcome of interest. Factors presented in the greenspace data from the NaNDA data will be used as predictor variables. Polluting sites from the NaNDA data will be included as 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.

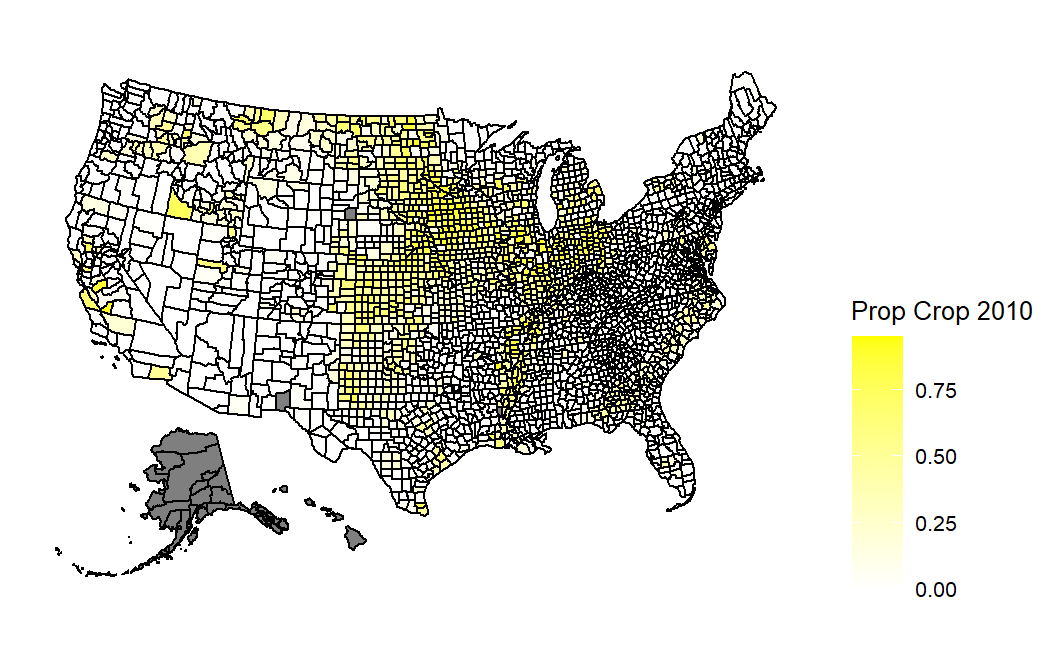
# Motivation from a Geographic Perspective:

Using the geographic identifiers alongside other information within the dataset, cartographic plots can be created to demonstrate similar densities amongst variables by county. These visualizations demonstrate the motivation behind the intention to use a clustering technique. On the following page maps plotting breast cancer mortality, county proportion of agricultural land use, and county proportion of wooded wetlands are presented. The elliptical shapes drawn on each map highlight areas in which there may be a relationship between these variables.

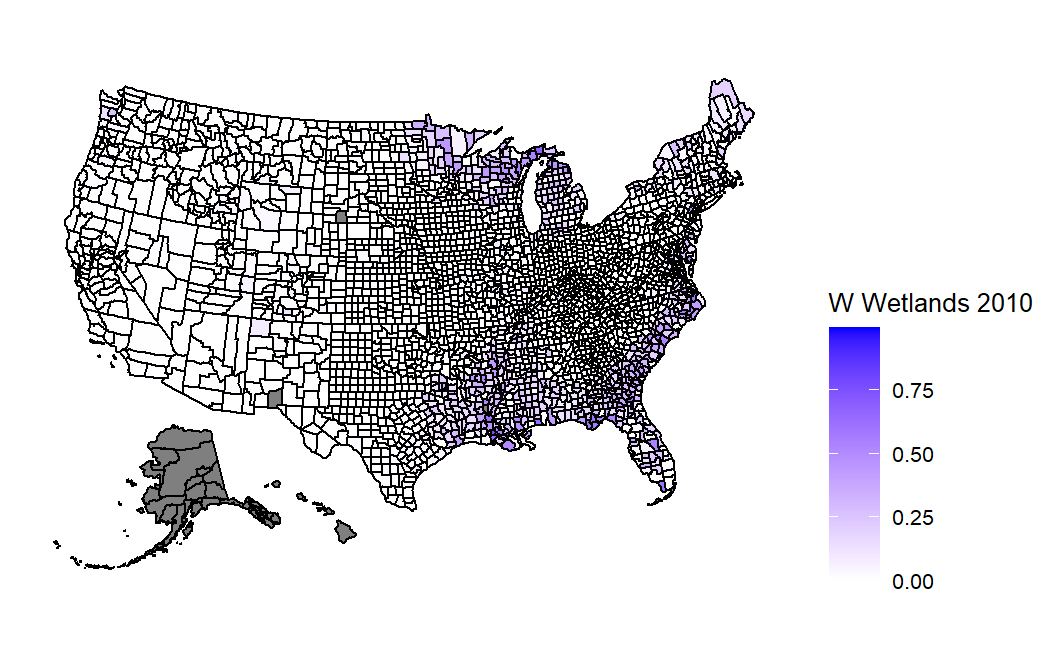
## Breast Cancer Mortality by County (2010):

# 

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



## Proportion of Wooded Wetlands by County (2010):



# Principal Component Analysis (PCA):

# PCA was conducted on the NaNDA data, which was used for the predictor variables in all of the potential learning models. The scree plot below visualizes the variance, and the cumulative variance, that was explained by each single principal component. Based on the plot, a perfect “elbow” point was not observed; therefore, the predictor variables could not be represented using only a few principal components.

Chart, line chart

Description automatically generated

# K-Means Clustering

K-Means Clustering was performed on the first and second principal components. Three different values for the number of clusters were utilized and the results were visualized in the plots below. The plots below appeared to indicate the presence of 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. National Institutes of Health (NIH). National Cancer Institute. SEER Cancer Statistics Review (CSR). 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.
4. Institute for Health Metrics and Evaluation (IHME). United States Cancer Mortality Rates by County 1980-2014. Seattle, United States of America: Institute for Health Metrics and Evaluation (IHME), 2017. https://ghdx.healthdata.org/record/ihme-data/united-states-cancer-mortality-rates-county-1980-2014. Accessed on 10/10/2022.
5. Centers for Disease Control and Prevention (CDC). Agency for Toxic Substances and Disease Registry (ATSDR). Geospatial Research, Analysis, and Services Program. CDC/ATSDR Social Vulnerability Index 2010 Database US. https://www.atsdr.cdc.gov/placeandhealth/svi/data\_documentation\_download.html. Accessed on 10/10/2022.
6. Finlay, Jessica, Melendez, Robert, Esposito, Michael, Khan, Anam, Li, Mao, Gomez-Lopez, Iris, … Chenoweth, Megan. National Neighborhood Data Archive (NaNDA): Polluting Sites by Census Tract, United States, 2000-2018. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2022-01-20. https://doi.org/10.3886/E159961V1.
7. Clarke, Philippa, and Melendez, Robert. National Neighborhood Data Archive (NaNDA): Land Cover by Census Tract, United States, 2001-2016 . Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-09-11. https://doi.org/10.3886/E110663V1. Accessed on 10/10/2022.