Greenspace, Breast Cancer, and Prediction

Trevor Johnson, Haddon Sandefur, Scott Sun
BIOSTAT 707

https://github.com/TrevorWJohnson/BIOSTAT707-Project

Duke University Honor Code Acknowledgement:
Trevor Johnson, Haddon Sandefur, and Scott Sun

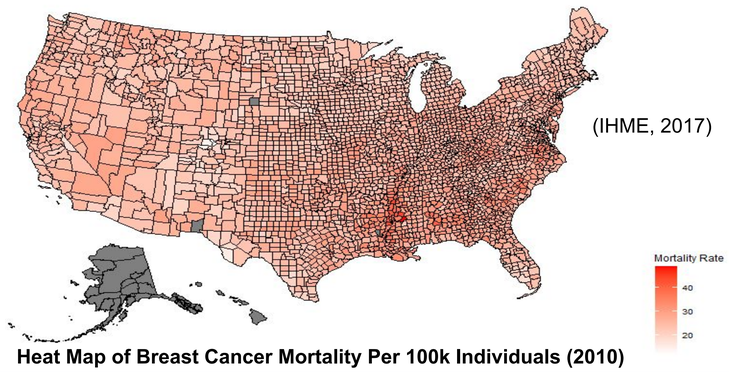
Project Background

Breast Cancer is the second most prevalent cancer in women in the United States (National Institutes of Health, 2019). Overtime, clinical researchers have placed a greater emphasis on cancer prevention, searching for factors that can dampen both incidence and mortality of cancers. Some studies have demonstrated linkages between the amount of environmental greenspace and mortality relating to cancer (James, Hart et. al, 2016). However, there is a general lack of research pertaining to this subject (O'Callaghan-Gordo et. al, 2018). We seek to contribute to this matter by seeing how well we can predict breast cancer from environmental greenspace, as well as which types of greenspace contribute most to this prediction. This study uses the University of Michigan's NaNDA greenspace dataset (Clarke et. al, 2019).

Objectives

A) Use multiple models to predict breast cancer mortality from greenspace density variables.

B) Select the best models and interpret their contributing variables.



(IHME, 2017)

Heat Map of Breast Cancer Mortality Per 100k Individuals (2010)

Models and Methods

Classification

- Distinguishing types of greenspace can be difficult. Therefore, latent classes were fitted to types of greenspace to reveal patterns.

Prediction

- Four models are utlized for the sake of comparison: KNN, LASSO Regression, Random Forest, and XGBoost.
- Each model underwent a cross-validation or bootstrapped process to develop confidence intervals for the RMSE (Root Means Squared Error).

Find Classes Within Data

↓

Fit Predictive Models

↓

Evaluate Performance

↓

Interpret Results

Results

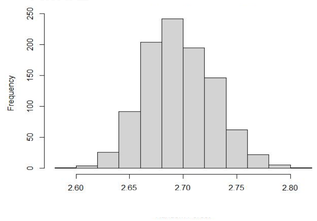
Poor Predictive Performance:

Of the four methods, Random Forest and XGBoost provided the most accurate results. However, all of the models did not predict breast cancer mortality rates well. As such, there is no indication that the greenspace variables collected from the NaNDA dataset are holistically predictive of breast cancer mortality.

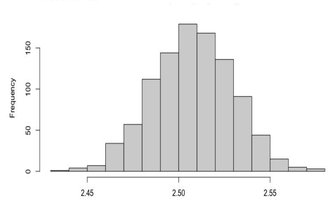
Table of model performances:

	Testing RMSE	95% Confidence Interval
LASSO Regression	2.82	(2.78, 2.85)
KNN	2.74	(2.74, 2.75)
Random Forest	2.70	(2.64, 2.76)
XGBoost	2.71	(2.65, 2.77)

Distribution of Random Forest RMSE

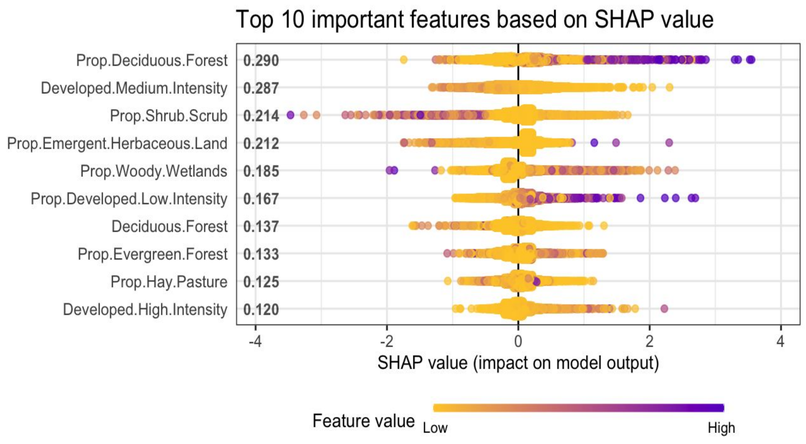


Distribution of XGBoost RMSE



Certain Types of Greenspace are More Important than Others:

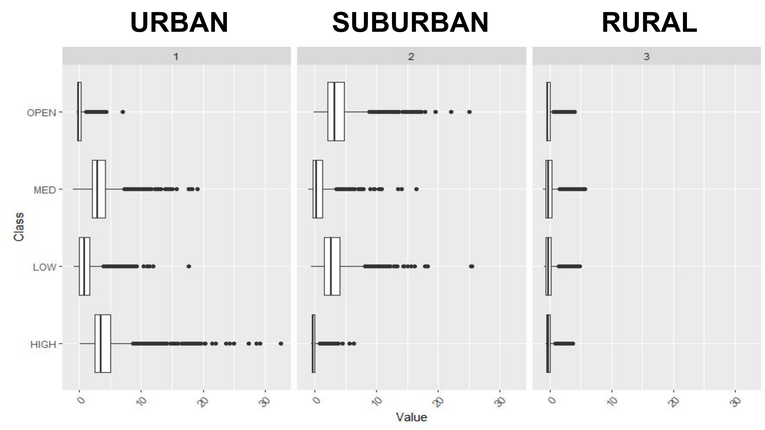
Each model provided different 'most important' important variables (variables contributing most to accurate predictions). However, human structural development, deciduous forests, shrub/scrub laden land, and herbaceous land were common important variables to the XGBoost, Random Forest, and LASSO regression models. However, XGBoost was able to convey whether or not variables contributed to higher or lower mortality rates.



Waterfall plot of SHAP values. Most important variables are listed in descending importance. Right-skewed variables are ones which generally predict higher mortality rates. Left-skewed variables predict lower mortality rates.

Latent Classes did NOT Improve Predictive Capability:

Latent Classes were fit on forest/vegetation-type and structural development-type variables and were selected via Entropy and AIC metrics. However, these classes did not meaningfully enhance predictive performances in our models.



Boxplots of a fitted three latent class model on structural development-type greenspace data with titled interpretations.

Conclusions and Future Work

Takeaways:

- Greenspace did **not** predict breast cancer mortality rates well on a county-level basis via Random Forest, XGBoost, LASSO, and KNN.
- Types of greenspace that predicted mortality comparatively well are likely artifacts of corresponding population desnities.

Future Work:

- More examination of the relationship between influential greenspace variables and breast cancer mortality is needed. Recording of new greenspace variables could also impact future prediction efforts.

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. 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.

4. 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.

5. 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.