**Ecological Factors Associated with Self-Reported Mental Health Status**

Andrew S Cistola, Alyssa Berger

1. Project Description

Ecological factors including socioeconomic status and health resource availability have been shown to relevant predictors of many health outcomes. However, the scope of research identifying these factors is relatively limited and many possible effects and interactions have not been identified.

1.1 Specific Aim

Identify ecological variables associated level self-reported mental health status using open source computational resources and public access datasets.

1.2 Research Design:

1. Collect zip code level self-reported health status from the 2020 release of the CDC and RWJF PLACES dataset (formerly 500 Cities project) for all populated U.S. zip codes. This represents the health outcome or dependent variable.

2. Collect approximately 400 zip code level socio-economic variables in the detailed profile tables from the 2020 release of the US Census American Community Survey. These represent possible predictor or independent variables.

3. Collect approximately 2000 county level health resource variables from the 2020 release of the HRSA Area Health resource File. These represent possible independent variables.

4. Connect poor mental health population rates with socio-economic percentage estimates by Zip code. Remove observations with missing outcomes and impute missing data for predictor variables using median values. Standard scale all variables.

5. Conduct Principal Component Analysis among predictors to identify the largest eigenvectors for each independent variable. Select variables with above average eigenvectors.

6. Create a Random Forest model predicting the outcome using all predictors to identify the gini impurity value for each independent variable. Select variables with above average gini impurity.

7. Use Recursive Feature Elimination with predictors present in both groups to identify the smallest set of predictors with the highest AIC value using cross-validation.

8. Use Moran’s Global I test to confirm spatial autocorrelation for the final zip code variables.

9. Connect zip codes to counties using the most recent HUD crosswalk file where zip codes are assigned to the county where the largest portion of their population is located.

10. Use local bayes smoothing to calculate a second zip code outcome. Repeat steps 5, 6, 7 and seven using county health resource variables. Retain the final county variables.

11. Calculate geographic weighted regression to determine weighted coefficients for the final zip code variables from step 8.

12. Average these weights by county and calculate the z-score of each zip code variable by county. Create labels for county based on the highest absolute value of the z-score of each zip code variable.

12. Create a multinomial support vector machines model identifying coefficients for all county variables. Select the top county variable for each zip code variable as the final zip interaction variables.

13. Evaluate the final zip code variables, the final county variables, and interaction variables for relevance to domain knowledge and current research. Withdraw variables only if no known theory can possibly account for their inclusion. Identify a final variable set.

14. Create a final multiple linear regression model using the final variable set and interaction terms. Utilize original raw data with all missing observations removed. Run statistical tests to determine OLS assumptions/violations and determine an appropriate regression modeling approach. Use this model to identify parameter estimates.

15. Create a multi-layered perceptron to predict a binary version of the final outcome (top quartile) and calculate the c-statistic from a ROC. Compare the AUC values for 1) all possible zip code and county variables 2) a random selection of variables equal in quantity to the final variable set 3) the final variable set. Use the results to identify whether the final set adds any predictive capability.