ISyE 6416 – Spring 2021
Project Proposal

Team Member Names

Halle Kutsche
Hsin-Yi Lin
Yilun 'Elon' Zha

Project Title

Predicting census-tract level disease prevalence rates by physical and social determinants of health (SDOH)

Problem Statement

Health starts in our homes, schools, workplaces, neighborhoods, and communities. Healthy People 2020 highlights the importance of addressing **the social determinants of health** by including "Create social and physical environments that promote good health for all" as one of the four overarching goals for the decade.

Therefore, this study aims to build a statistical model to explore the potential nexus between health outcomes and environmental (physical, social, ...) determinants of health. The socio-ecological model implies that human behaviors and outcomes are determined by a series of macro-level to micro-level interconnected factors: from the policy environment, economic prosperity, to the adjacent physical environment, social environment, and to individual behavioral choices.

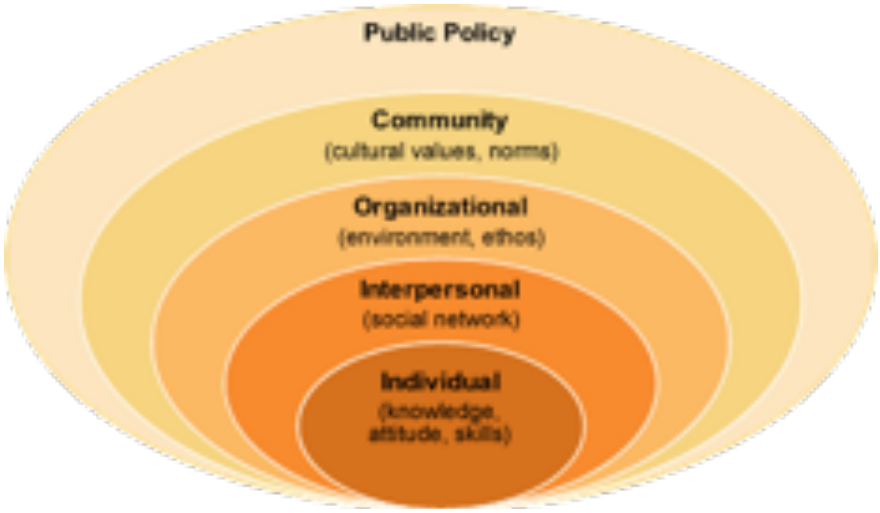

Fig.1 Diagram of an ecological model for human health

In particular, the research questions we are trying to answer in this study include: (1). To what extent can we predict the population health outcomes based on the attributes of the physical and social environments to which they are exposed? (2). Which health outcome is easier / harder to predict using the current attributes? (3). What is the relative contribution of each attribute in predicting health outcomes?

Data Source

There are three main sources of data included in this study: (1). Socio-economic status data from Census Gov; (2). Disease prevalence rate, preventive behavior, and risk factors data from CDC; (3). Restaurant Business data from Yelp API.

The variable to be predicted is **disease prevalence rate** (percentage), which is the ratio of the number of existing patients to the total population in a specified geographic unit. PLACES, a collaboration between CDC, the Robert Wood Johnson Foundation, and the CDC Foundation, provides population-level analysis to all counties, places (incorporated and census-designated places), census tracts, and ZIP Code Tabulation Areas (ZCTAs) across the United States. (CDC, 2020)

Explanatory variables can be categorized into three categories:
(a). **Socio-economic status**, which includes but is not limited to median household income, homeownership, education attainment, ethnicity, etc. A feature selection process will be performed to reduce dimensionality and avoid multicollinearity.

(b). **Physical environment attributes**. In particular, we are specifically interested in the association between urban foodscape and nutrient-related diseases. Hence, YELP restaurant business data and USDA SNAP retailers data will be integrated and pre-processed as a representation of the communities' food access.

(c). **Prevention and unhealthy behaviors**, which can also be obtained from CDC PLACES database. Prevention includes dentist visit, mammography, health insurance, etc, and unhealthy behaviors include drinking, smoking, and the list goes on.

Note: links of each dataset:
Census Gov: https://data.census.gov/cedsci/
CDC PLACES: https://www.cdc.gov/places/index.html
YELP API: https://www.yelp.com/developers/documentation/v3/business_search

Methodology

We will use a series of linear and non-linear models to fit the data. To be specific, linear methods include Multiple Linear Regression and its derivatives such as Ridge and Lasso. One specific technical consideration is a spatial lag term in order to cancel out the spatial spillover effect, which is often observed in the geographic distribution of many socio-economic variables. Non-

linear models including ensemble learning models will also be used to search for a better fit: ANN, Random Forest, etc.

When it comes to validation methods, we will use k-fold cross-validation to split the dataset into training and test sets. For feature selection, we will compare some key metrics of different learning models such as R2, AIC, BIC, Cp. Parameter paths (Ridge and Lasso) will also be plotted to facilitate the feature selection.

Evaluation and Final Results

In this project, we expected to deliver the following:

A model that can predict the health outcome based on the socio-economic status, physical environment attributes, and prevention and unhealthy behaviors.

Evaluate which health outcome is easier to predict based on our explanatory variables.

Identify the relationship between the socio-economic status, physical environment attributes, and prevention and unhealthy behaviors, and the health outcome. Like which attributes have a positive relationship with certain diseases.

The accuracy of the prediction model.

Bibliography

[1] Healthy People 2020 (2010): An Opportunity to Address the Societal Determinants of Health in the United States. July 26, 2010. Available from: http://www.healthypeople.gov/2010/hp2020/advisory/SocietalDeterminantsHealth.htm
[2] CDC (2020, December 08) PLACES: Local data for better health. . Retrieved April 01, 2021, from https://www.cdc.gov/places/index.html