Analyzing US County Level Food Deserts' Demographics, Food Access, and Diet Related Disease Prevalence

Morgan L. Ford fordm2@rpi.edu

[1]Rensselaer Polytechnic Institute, Troy, NY, United States,

Abstract

23.5 million Americans live in food deserts[1]. Many Americans also live with diet related health conditions. This poster will examine a potential relationship between access to grocery stores and access to healthy foods at grocery stores, the differences between urban and non-urban food deserts, the relationship between population demographics and food deserts, and the relationship between food deserts and the prevalence or mortality rates of diet-related health conditions. I hypothesis that **food deserts are less likely to have a variety of healthy foods in their limited food sources** compared to areas that are not food deserts. I also hypothesize that **food desert communities are more likely to be at risk for diet related health issues** than than areas that are not food deserts.

Introduction

According to the USDA, a food desert is defined as an area that has either:

- A poverty rate >= 20%
- A median family income <= 80% of the median family income (urban areas)
- A median family income <= 80% of statewide median family income (nonurban areas)

and at least 500 people or 33% live further than:

- Urban: 1 miles from nearest large grocery store
- Rural: 10 miles from nearest large grocery store

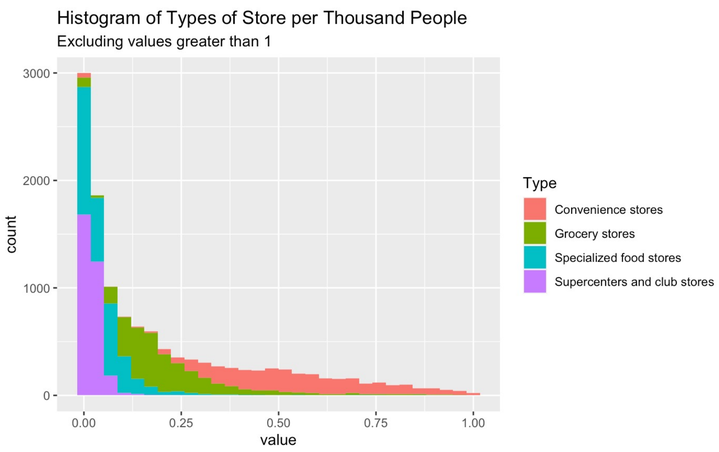
Diet related health conditions include:

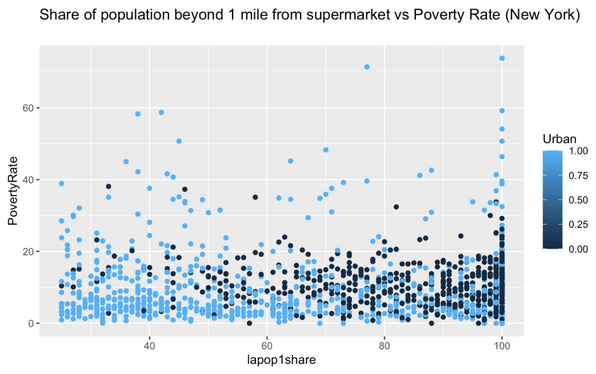
- Type 2 Diabetes
- Heart Disease
- Obesity
- Breast, Colon, and Uterine Cancers

The Data

1. **Food Environment Atlas**
   - **County level data about types of accessible food**
2. **Food Access Research Atlas**
   - **Census tract level data about access to food**
3. **US Health Map**
   - **Variety of data about health conditions**

EDA





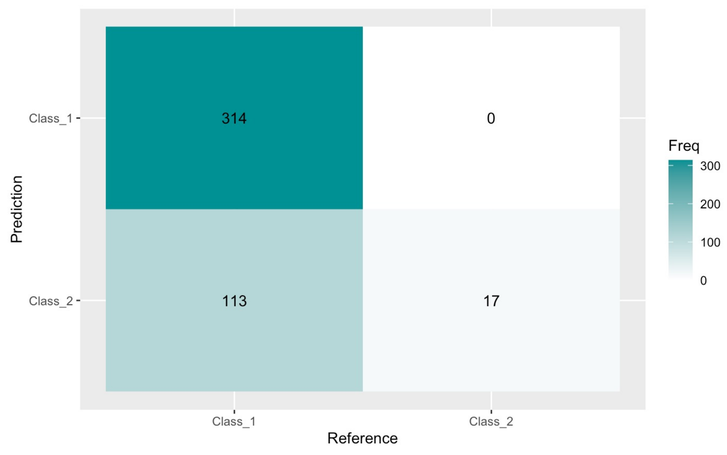
Models

Demographics SVM

A support vector machine model was model to classify if a county that was a food desert or not, using the variable that flagged if a tract that flagged at low access for 1 mile for urban areas and 10 miles for non-urban areas, and the following demographic variables about population income and ethnicity.

$$K(x, x') = exp(-\frac{||x - x'||^2}{2(0.203)^2})$$

It performed with **74.55%** accuracy and the confusion matrix is below.



Access Linear Regression

A linear regression model was created to examine feature importance and to examine the difference between urban and non-urban areas. The urban linear model is likely overfit due to the lack of data. The non-urban linear model performs with residual standard error of **0.3614** and multiple R-squared of **0.04345**. This model may also be overfit, but does perform with relatively low error.

Coefficient (per thousand, 2016)	Urban Estimate	Non-Urban Estimate
Grocery Stores	-15.8089	0.13203
Super-centers and Club Stores	-45.6754	3.03135
Convenience Stores	6.9483	0.10744
Specialty Stores	NA	0.27167
SNAP Authorized Stores	NA	-0.17213
WIC Authorized Stores	NA	0.13954
Fast Food Restaurants	NA	-0.02529
Food Service Restaurants	NA	-0.02453

Table 1. The coefficents for the linear regression model for the number of food deserts in a county compared between urban and non-urban areas.

Access Random Forest

A random forest classifier was made with 50 trees and 3 variables tried at each split to see if it was possible to correctly classify a county as having 33% or more food desert tracts based on the food environment. It performed with **97.3%** accuracy.

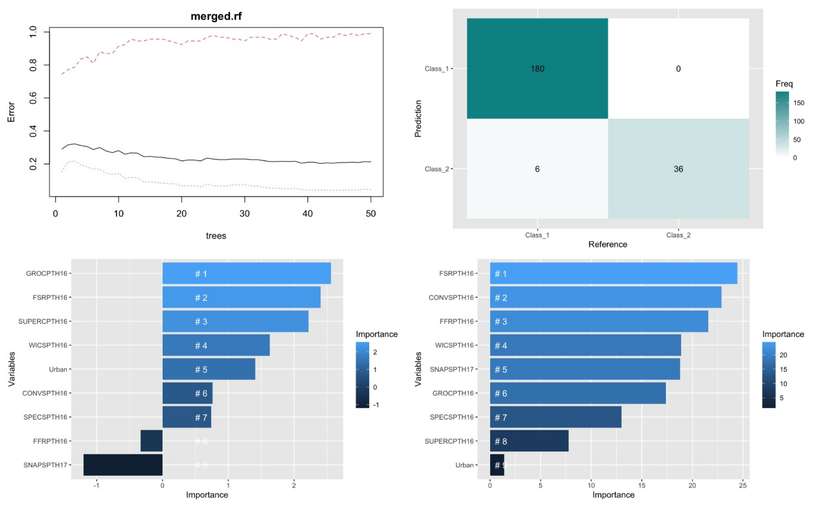
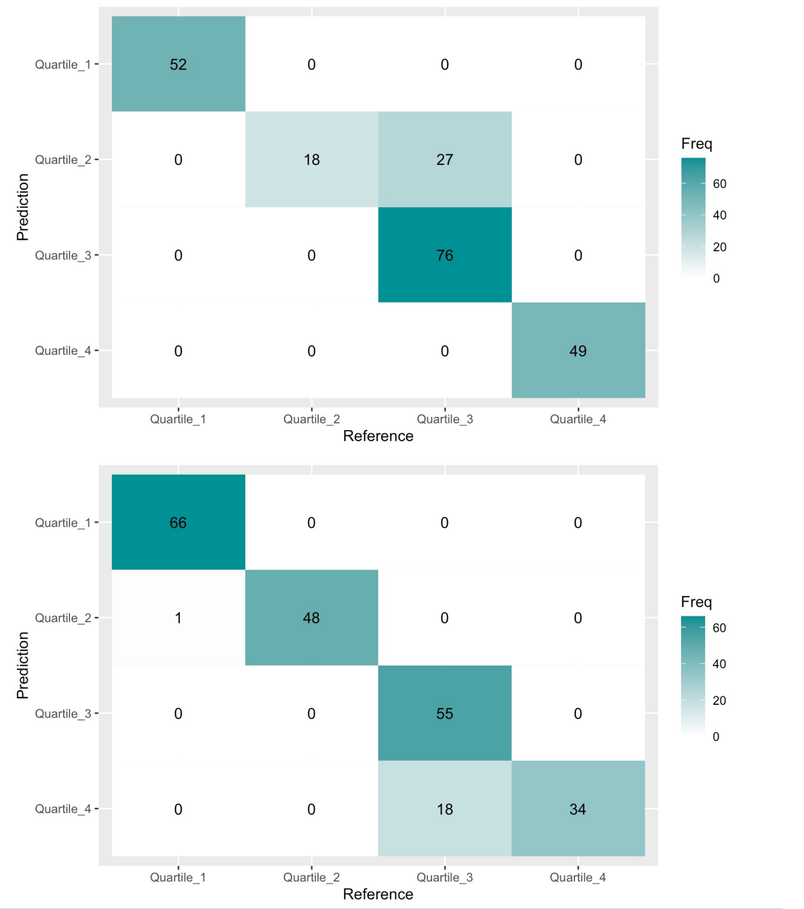


Fig. 6. The plots for the random forest model to classify a county having food deserts.

The most important factors for accuracy: the number of grocery stores, food service restaurants, and convenience stores.

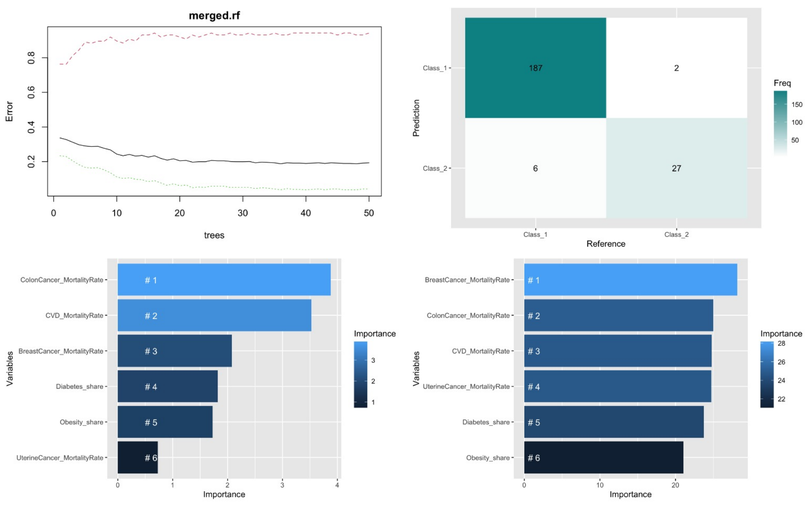
Quality of Life KNN

A knn model was created with **72.3%** accuracy to classify food deserts based on length of life. Another knn model was created with **94.15%** accuracy to classify food deserts based on quality of life. The respective confusion matrices are below.



Health Random Forest

A random forest classifier was made with 50 trees and 3 variables tried at each split to see if it was possible to correctly classify a county as having 33% or more food desert tracts based on the prevalence of obesity and diabetes as well as the mortality rates of cardiovascular disease, breast, colon, and uterine cancers. It performed with **94.1%** accuracy.



The most important factors for accuracy: colon cancer mortality rate and cardiovascular disease mortality rate.

Conclusion

By t-test alone, my hypothesis is disproven. However, from the models that I have generated, I find there to be a significant link between food deserts, food access, quality of life, and diet-related health conditions.

Note:

When creating urban and non-urban linear models, a county was counted as urban if the percent of urban census tracts in the county was higher than 50%.

When classifying entire counties as food deserts, a county would be classified as a food desert if 33% or more of the census tracts in the county were food deserts.

Sources:

https://www.ers.usda.gov/webdocs/publications/45014/30940_err140.pdf
http://www.ers.usda.gov/Publications/AP/AP036/
https://www.ers.usda.gov/data-products/food-environment-atlas/
https://www.ers.usda.gov/data-products/food-access-research-atlas/
https://www.cdc.gov/chronicdisease/resources/publications/factsheets/nutrition.htm