U.S. Food Environment Clustering and Analysis

STEPHANIE CUNNINGHAM, STUDENT – MASTER'S OF APPLIED DATA SCIENCE, BAY PATH UNIVERISTY

Examination of Food Environment Features and Outcomes with Machine Learning Techniques

INTRO

"Food environment" – culmination of socioeconomic and physical factors which shape the food habits of a population. A theme of food environment research is the idea of a "food desert," generally defined as an area that is both low-income and low-access to food markets and thought to result in poor nutrition and health problems such as diabetes. Recent research calls this into question, with claims that choices drive poor health outcomes, rather than lack of nutritional access.

Goal – determine whether clustering areas by common food environment can reveal how underlying factors and health outcomes (here, diabetes rate) vary beyond food desert classification.

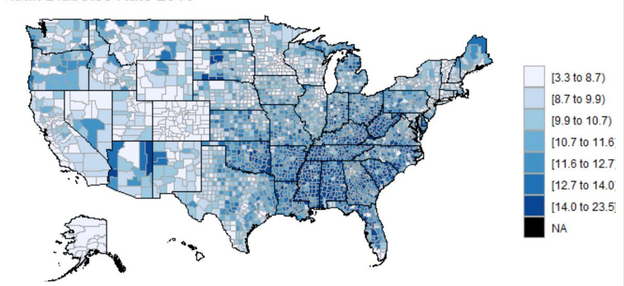
DATA

2020 Food Environmental Atlas (USDA)
2020 Food Access Research Atlas (FARA) (USDA)
2018 Labor Participation and Educational Attainment Data (US Census Bureau)

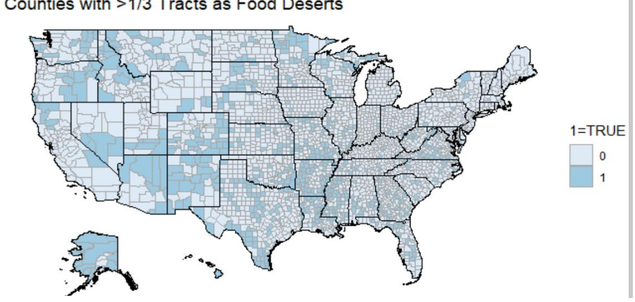
290 total features on all 3143 counties

Outcome "food desert status" labeled for counties with >1/3 tracts defined as food deserts based on FARA.

Adult Diabetes Rate 2013



Counties with >1/3 Tracts as Food Deserts

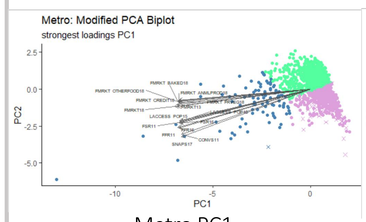


CONCLUSIONS

While food environment, food desert status, and diabetes rate are inter-related, certain factors associated with food desert status are more correlated with poor health outcomes than others. While poverty and access are important to food environment, this is not a simple causal relationship and further clustering work could reveal more personalized solutions for populations in need.

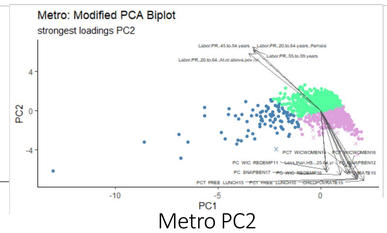
RESULTS

PCA & Clustering (with strongest loadings by PC)



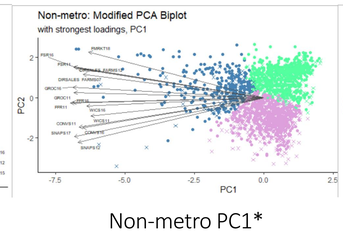
Metro PC1

Variance drivers: Population & resources



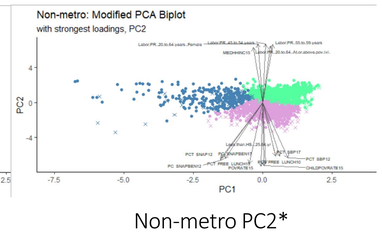
Metro PC2

Employment, assistance & income



Non-metro PC1*

Population & resources



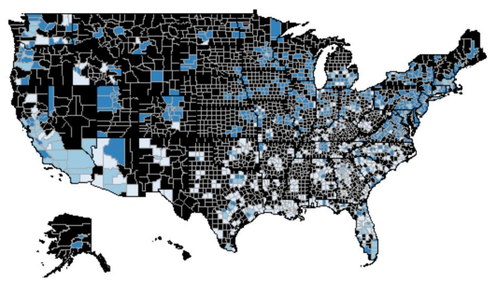
Non-metro PC2*

Employment, edu., assistance, income

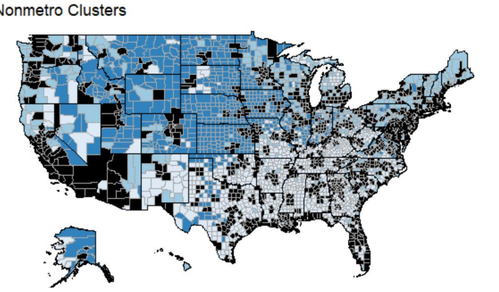
*Non-metro had 4 significant PCs. PC3 variance driver mainly store access, and PC4 showed cultural implications (e.g. Native American populations, etc.)

Cluster Mapping and Outcome Comparison

Metro Clusters



Nonmetro Clusters



Cluster main findings:
Many factors underlie FD status, which is loosely associated with DB rate. Clustering can help connect with geographical and cultural influences.

Cluster feature summaries:

METRO

1- Smaller populations, fewer resources, higher poverty rates

2-Higher populations, more resources, higher poverty rates but medium labor PRs

3-Lower poverty rates, high labor PRs

NON-METRO

1- Higher poverty rates, low edu. attainment, low labor PRs

2- Higher populations, more resources

3- Lower poverty rates, smaller populations, high labor PRs, high educational attainment

Cluster (Metro)	Median Diabetes Rate	% Food Deserts
1	12.3%	20%
2	9.3%	<1%
3	9.8%	4%

Outcomes of interest are significantly different between clusters based on Chi-Square and ANOVA.
Post hoc tests show more detail:

Cramer's V for cluster association with FD status: 0.2746 (weak)
Dunn test for cluster vs. diabetes rate multiple comparison:

	Comparison	Z	P.unadj	P.adj
1	1 - 2	11.736925	8.242892e-32	1.648578e-31
2	1 - 3	17.906065	1.057519e-71	3.172556e-71
3	2 - 3	-2.012332	4.418492e-02	4.418492e-02

Cluster (Non-metro)	Median Diabetes Rate	% Food Deserts
1	13.1%	41%
2	10.6%	19%
3	10.2%	24%

Outcomes of interest are significantly different between clusters based on Chi-Square and ANOVA.
Post hoc tests show more detail:

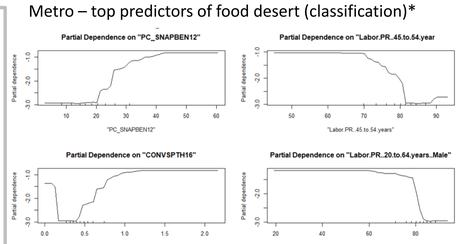
Cramer's V for cluster association with FD status: 0.2091 (weak)
Dunn test for cluster vs. diabetes rate multiple comparison:

	Comparison	Z	P.unadj	P.adj
1	1 - 2	15.382619	2.141079e-53	4.282158e-53
2	1 - 3	23.256222	1.230582e-119	3.691745e-119
3	2 - 3	3.497849	4.690260e-04	4.690260e-04

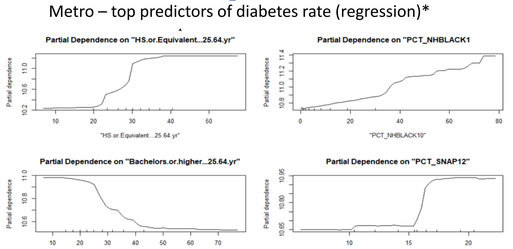
Random Forest – Variable Importance & Partial Dependence

Variable permutation importance graphs (not shown) main findings:
Demographics (age, race, educational attainment) more important to health outcome than to food desert status; labor participation more important to food desert status than to health outcome; SNAP benefit participation more important to both outcomes (positively related).

Metro – top predictors of food desert (classification)*



Metro – top predictors of diabetes rate (regression)*



*Non-metro not shown. Similar variable importance and relationships except:
Demographics more important to both outcomes in non-metros than in metros
Convenience stores per capita a major predictor in metro only

METHODS

Bayesian PCA

Selected for dimensionality reduction due to its proficiency at handling missing value imputation. Counties were separated by metro status to reduce influence of population over PCA results.

K-Means Clustering

3 food environment clusters were identified for both Metro and Non-metro counties. Strongest loadings were plotted on each principal component to impart meaning to clusters.

Statistical Analysis

Significance (Chi-Sq, ANOVA) and post hoc tests (Cramer's V and Dunn's, respectively) – (how) do clusters associate with food desert status and with health outcomes (diabetes rate)?

Random Forest

This supervised learning portion does not involve the clusters, only attributes and outcomes. Classification of food desert status and regression on diabetes rates. Used to compare predictive importance of variables (excepting those which directly define food desert e.g. poverty rate)

FURTHER RESEARCH

Additional clustering/stratifying work to identify more specific needs – variance in this research still driven largely by population size and then by income levels.

2017 policy change for SNAP-authorized stores to encourage healthier offerings – data not yet reflective. In both 2012 and 2017 data, partial dependence showing authorized SNAP stores per capita is a strong positive predictor of diabetes rate, and patterns follow closely with number of convenience stores. Research on policy effects is needed.