A study of food environments and their effects on obesity rate in the US

JAVED AMAN & WILLIE CHANG

BACKGROUND

How can we use food environment and other demographic statistics to predict obesity rate?


- Obesity & other health factors are determined largely by diet.
- Food environments describe the availability of, access to, affordability at, and consumer spending at food outlets in a region.
- A common term for discussing food environments is **food desert**, defined by the US government as "rural, minority, or low-income areas" ... "that lack access to affordable fruits, vegetables ..., and other foods that make up a full and healthy diet".
- We also look at demographic statistics when predicting obesity.

RELATED WORK

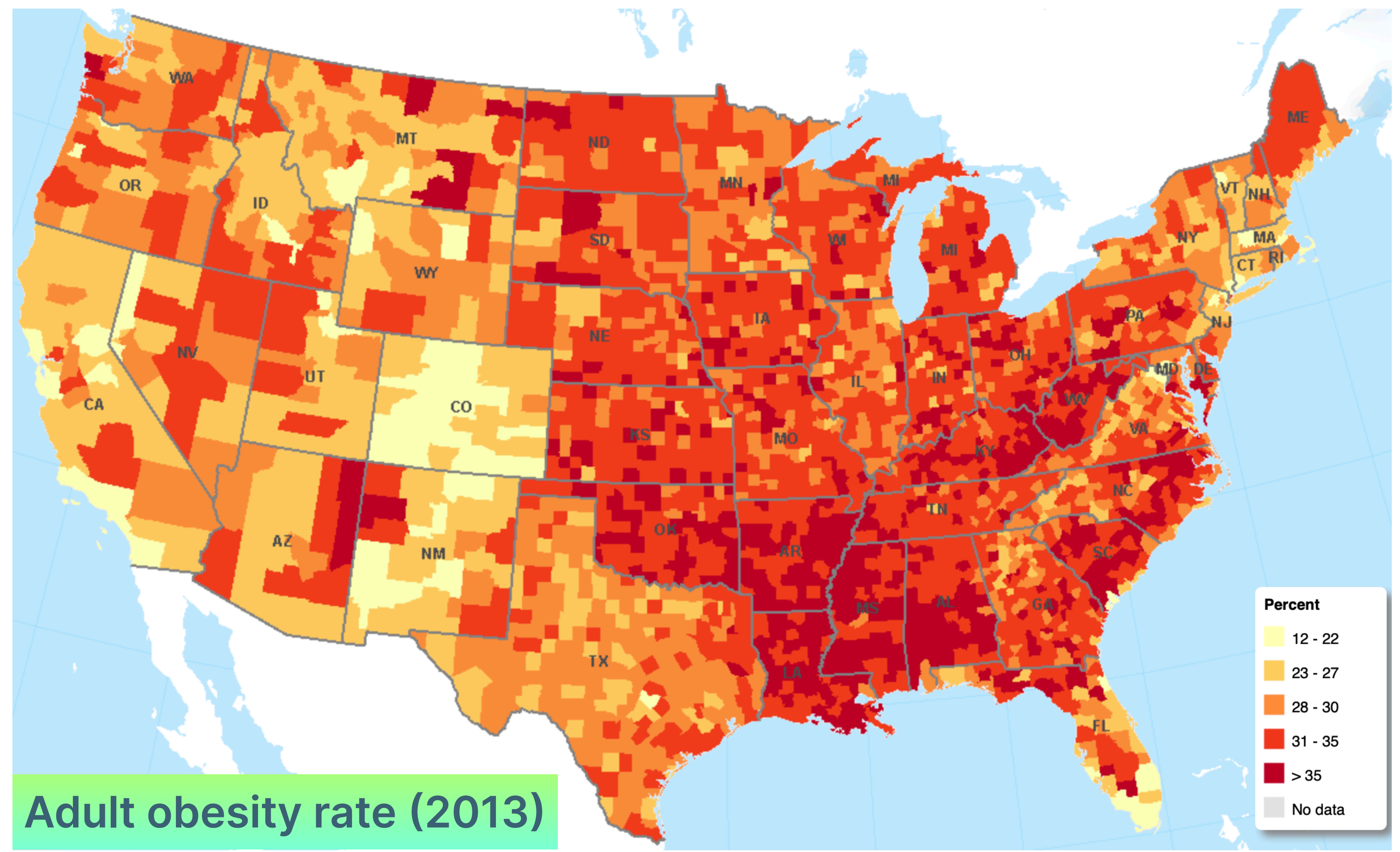
- Multivariate linear regression has been used to show relationship between household distances to food stores and consumption of fruits and vegetables. Does not directly relate environment to health outcomes [1].
- Geographic weighted regression demonstrates some relationship food environment and socioeconomic status with obesity rates. Uses older dataset and does not use more complex classifiers such as logistic regression and k-nearest neighbors [2].
- Food environments can exhibit clustering, but the relationship between clusters and health outcomes is not clear [3].
- Identifying high-risk counties has allowed for assisting public health officials in designing programs to reduce obesity rate [4].

REFERENCES

1. Bodor, et al. Neighbourhood fruit and vegetable availability and consumption: the role of small food stores in an urban environment. Public Health Nutrition, 11(4):413420, 2008.
2. Sang-Hyun Chi, et al. Can Geo-graphically Weighted Regression improve our contextual understanding of obesity in the US? Findings from the USDA Food Atlas. Applied Geography, 44:134 – 142, 2013.
3. S. Bryn Austin, et al. Clustering of fast-food restaurants around schools: A novel application of spatial statistics to the study of food environments. American Journal of Public Health, 95(9):1575–1581, 2005. PMID: 16118369.
4. Centers for Disease Control and Prevention (CDC). Estimated county-level prevalence of diabetes and obesity - United States, 2007. Morbidity and mortality weekly report, 58(45):1259–1263, November 2009.
5. Economic Research Service (ERS), U.S. Department of Agriculture (USDA). Food Access Research Atlas, https://www.ers.usda.gov/data-products/food-access-research-atlas/
6. Economic Research Service (ERS), U.S. Department of Agriculture (USDA). Food Environment Atlas. https://www.ers.usda.gov/data-products/food-environment-atlas/

METHODS

Dataset
USDA's Food Environment Atlas (2018), a county-level dataset [6]



Adult obesity rate (2013)

Supervised methods for estimating adult obesity rate (2013)

Linear regression
Search for hyperparameters (e.g. L1/L2 norm, number of features, etc.) with the lowest mean sq. error on the dev set.

Popular classifiers
Classify counties with over 30% obesity rate using popular classifiers, including logistic regression and naive Bayes.

Unsupervised methods for characterizing data

Clustering
Illustrate county clusters with t-SNE and characterize clusters by coloring in principal component and feature values.

Gaussian mixture models (GMMs)
Demonstrate the efficacy of mixed membership (MM) models with the AIC metric.

CONCLUSIONS

- Obesity rate is overwhelmingly correlated with **socioeconomic covariates** such as income, education, and social benefit usage.
- % change in **fast food** sales is a superb indicator of obesity rate.
- **Access to food stores**, **presence of farmers' markets**, and **presence of fitness centers** are good indicators of obesity rate, although the presence of and access to said facilities may just be side-effect characteristics of low-income neighborhoods.
- Counties are difficult to stereotype; many clusters are needed.
- MM models may be more suitable for characterizing counties.

RESULTS

Features most correlated, both positively and negatively, with the outcome, adult obesity rate (2013)

FEA
1. Adult diabetes rate*
2. Restaurant expenditure PC
3. Median household income
4. SNAP benefits per capita (PC)
5. Child poverty rate

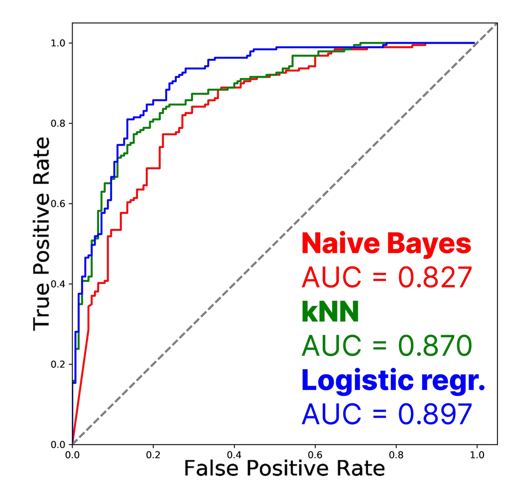
Appended
1. % adults w/ college degree
2. % w/ HS diploma or GED
3. Per capita (PC) income
4. Creative population
5. % adults w/ partial college

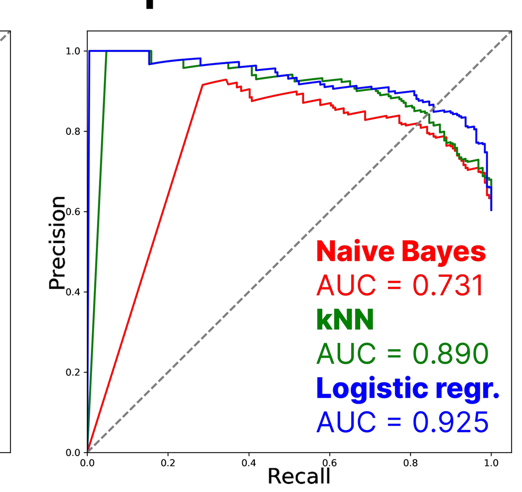
** Feature not used for classification or regression*

Logistic regression: MSE = 5.87 pct², 30% obesity cutoff F = .783

Obesity	15%	20%	25%	30%	35%	40%	45%	50%
Predicted	1	6	41	146	106	12	2	0
Actual	2	11	50	135	100	14	1	1

Classification: ROC and precision-recall curves





For $P_{threshold}(>30\%) = .5$
Naive Bayes F = 0.811
k-nearest neighbors F = 0.819
Logistic regression F = 0.872

t-SNE plots and GMM AIC, BIC plots (metro counties only)