What shapes US population health? Food, physical activity, or income?

Exploring the impact of food environment, physical activity, medical insurance and demographics on the prevalence of diabetes and obesity in US counties – Xinyu Zhang

The Problem

- Consistently high prevalence of diabetes and obesity in certain areas of US
- It is unclear how food environment, physical activity and medical care affects population health of US counties
- Food industry needs to identify the areas where improving food choice is the most urgent need
- An intervention strategy to reduce prevalence of diabetes and prevalence of obesity needs to be tailored and targeted at the US counties where diabetes and obesity are highly prevalent

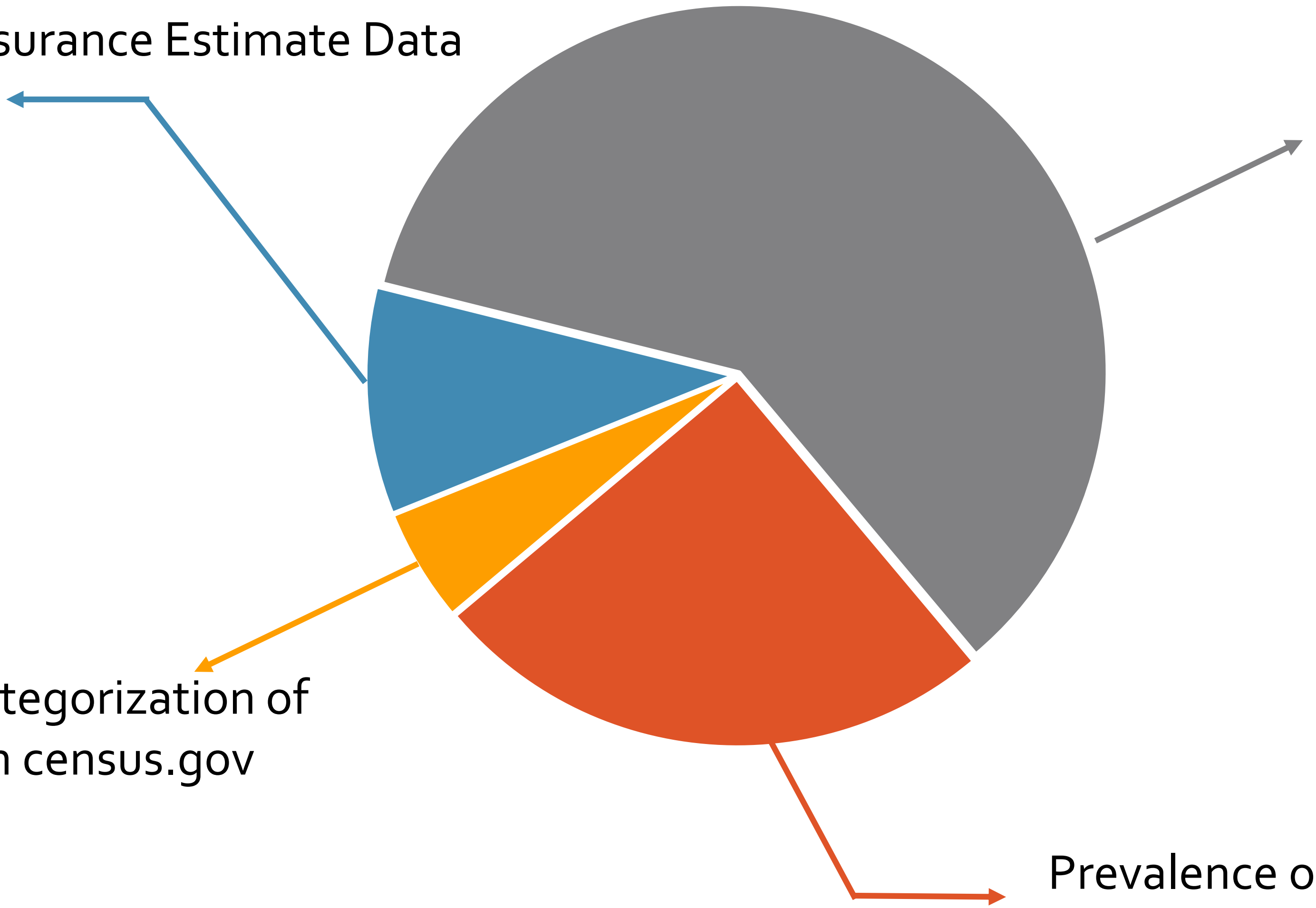
Questions Proposed to Explore

- Distribution of food environment, physical activity and medical insurance among US counties?

- Can we categorize US counties into distinct groups? What about those with high diabetes prevalence and high obesity prevalence?

- What determines the prevalence of diabetes and obesity?

Data Collection



Small Health Insurance Estimate Data
on census.gov

Food environment data
On data.gov

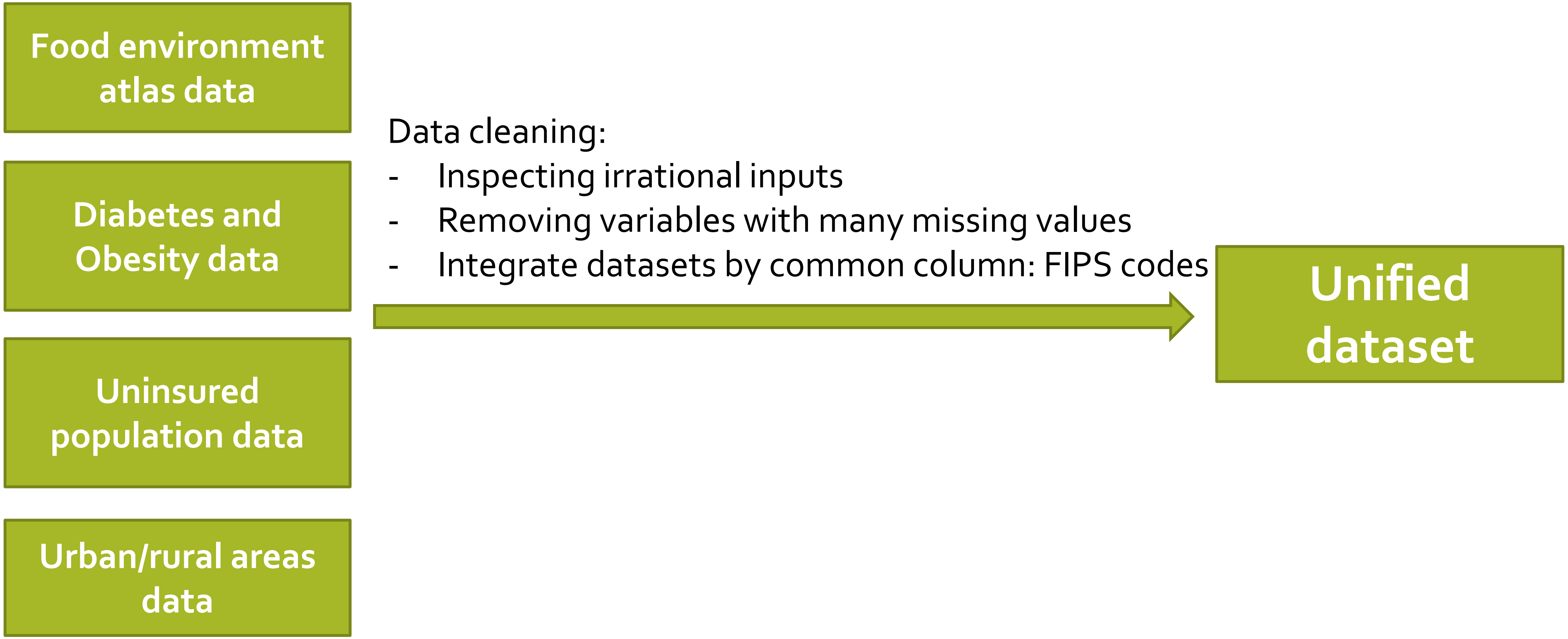
Urban-rural categorization of
US counties on census.gov

Prevalence of diabetes, obesity
and physical inactivity
on CDC website

About Collected Data

- Data are stored in CVS files and excel files
- 21 datasets and to clean and integrate
- Over 300 variables to inspect (missing values and irrational inputs)
- Most statistics are measured at multiple times between year 2004-2014

Data preprocessing



Food environment atlas data

Diabetes and Obesity data

Uninsured population data

Urban/rural areas data

Data cleaning:

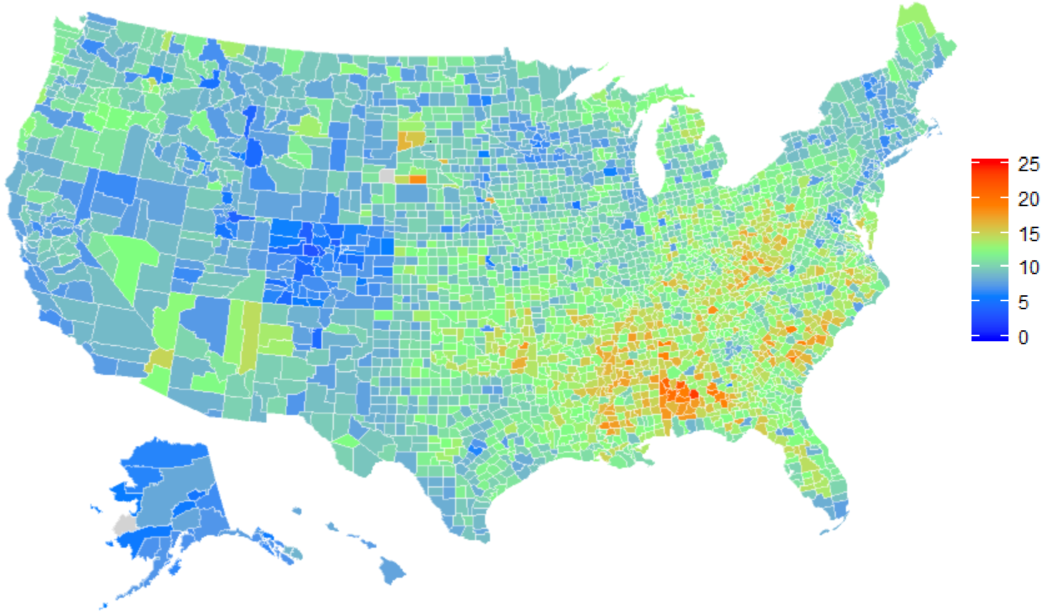
- Inspecting irrational inputs
- Removing variables with many missing values
- Integrate datasets by common column: FIPS codes

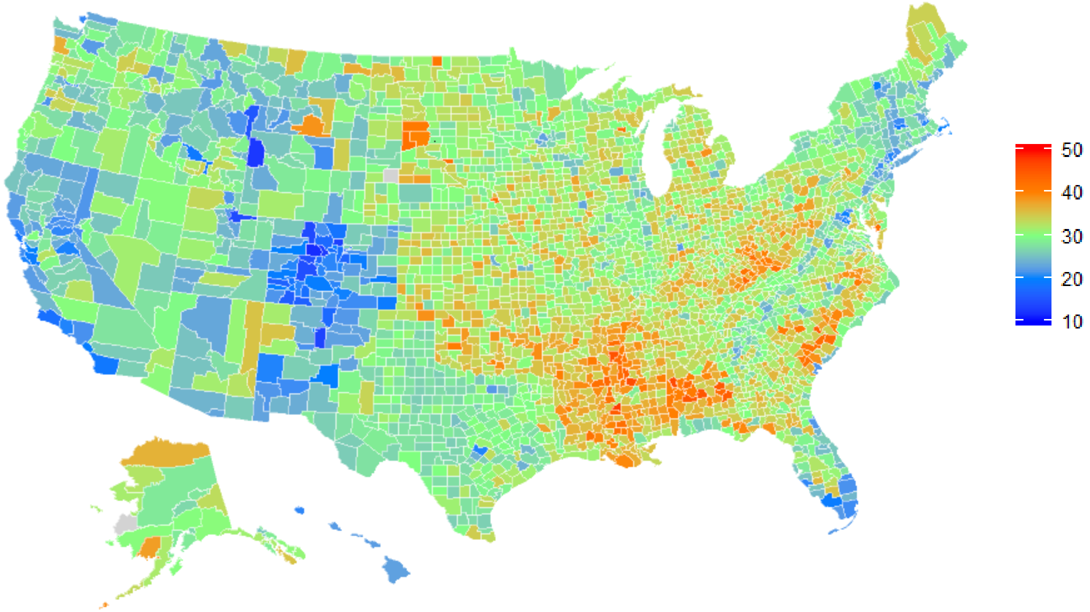
Unified dataset

Analysis scenarios

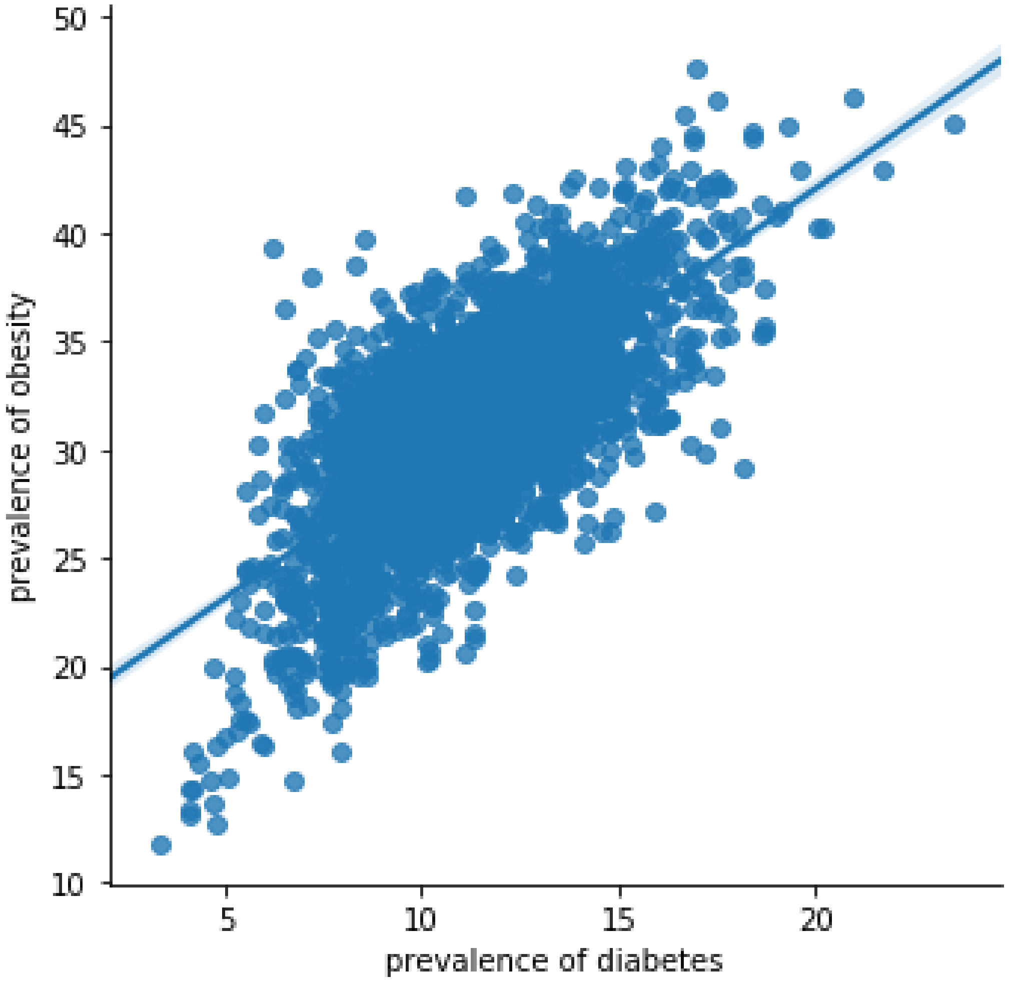
- Exploratory Data Analysis (EDA)
- Hierarchical Clustering Analysis
- Regression analysis with regularization

EDA: Great variation in prevalence of diabetes and obesity among US counties and within some states

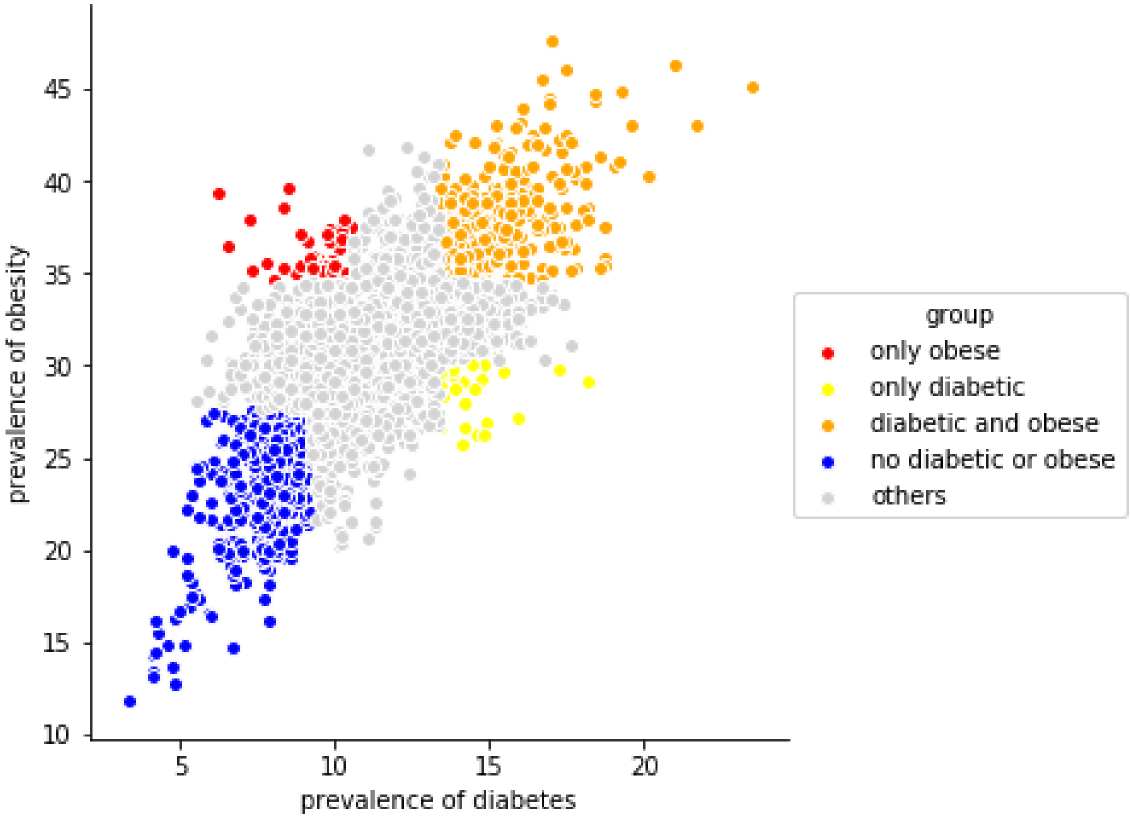

prevalence of diabetes

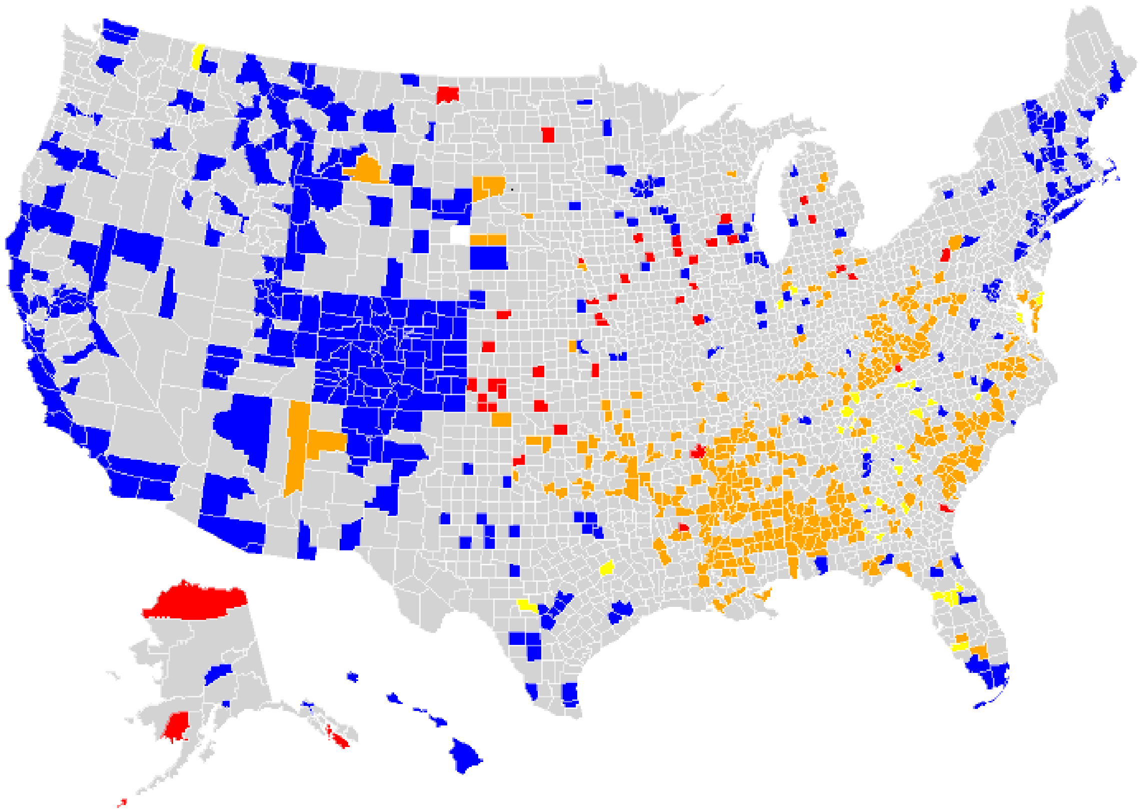

prevalence of obesity

EDA: Counties with high prevalence of diabetes also tends to have high prevalence of obesity

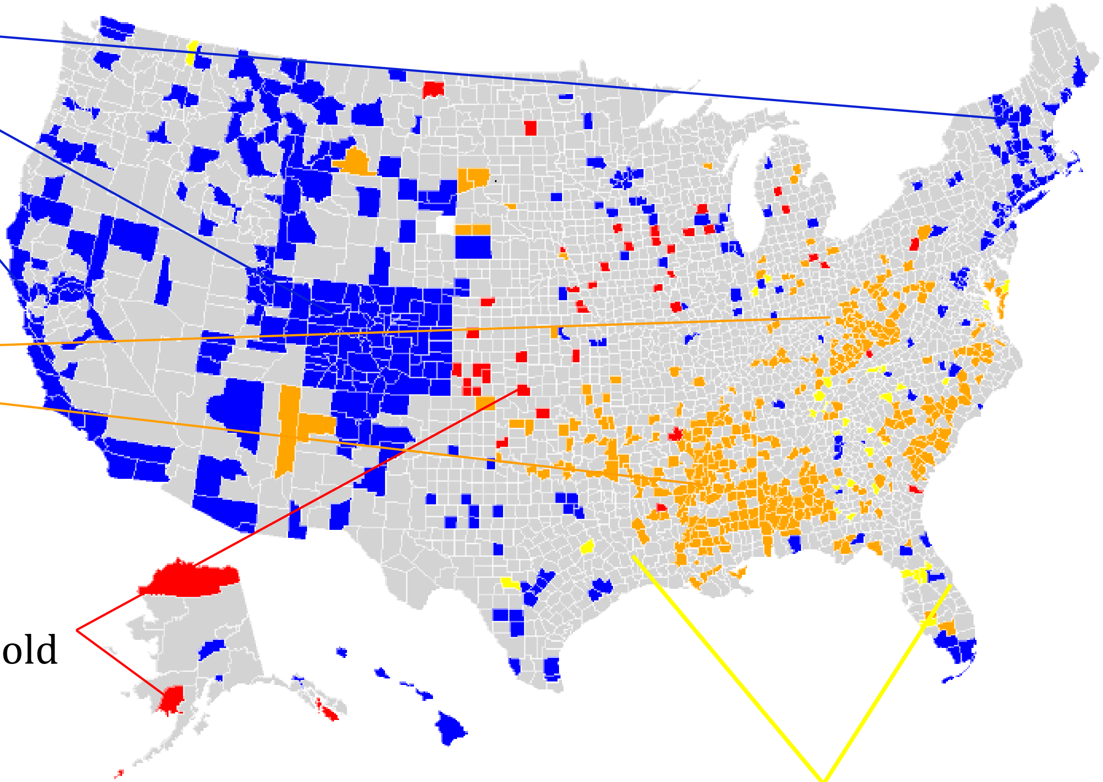


EDA: Counties with different combinations of diabetes prevalence and obesity prevalence





EDA: Counties with different combinations of diabetes prevalence and obesity prevalence differ greatly in physical activity and population age composition.



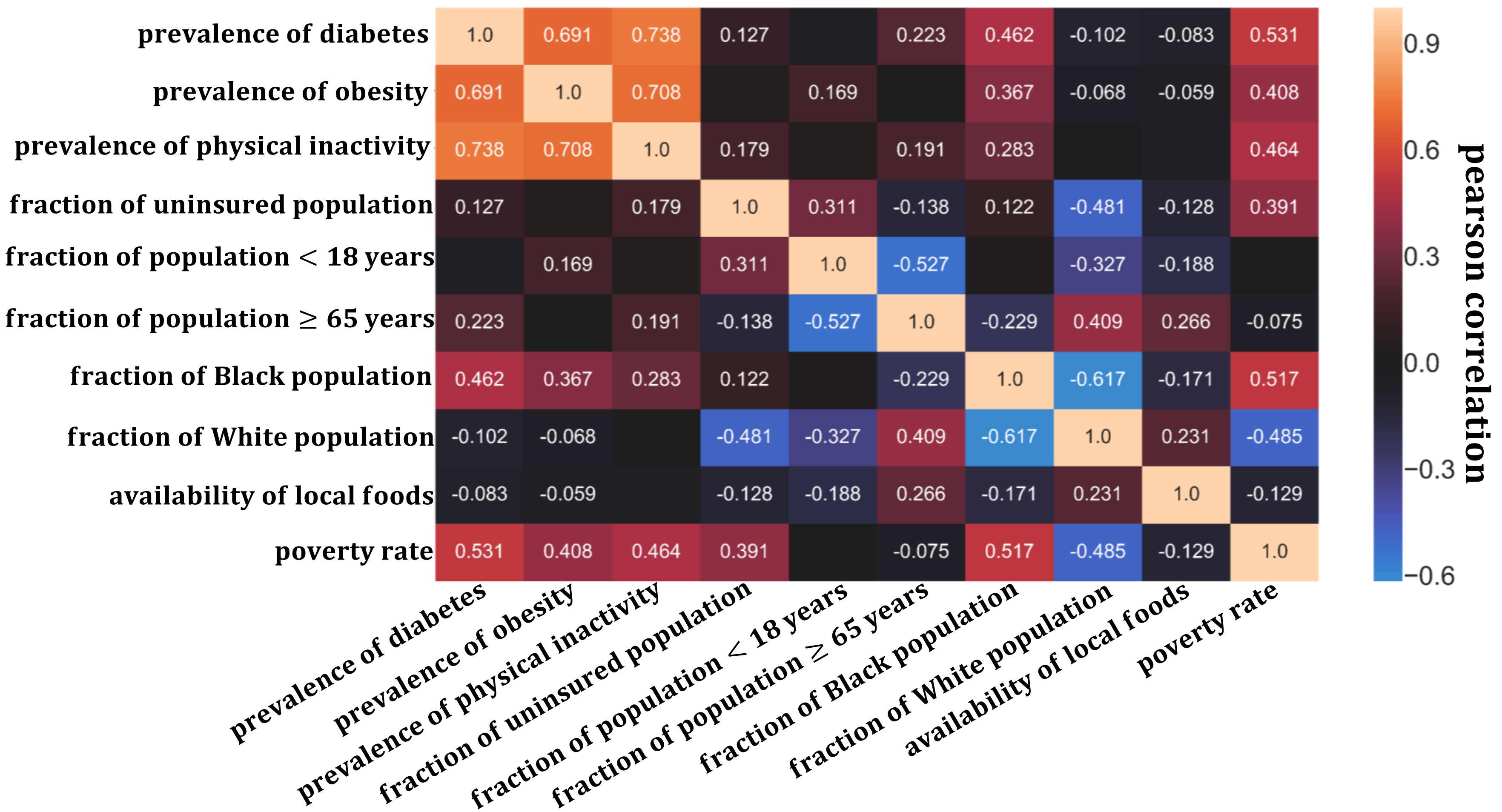
Low diabetes and obesity prevalence
Highly physically active

High diabetes and obseity prevalence
physically inactive

High obseity prevalence
More people < 18 years old

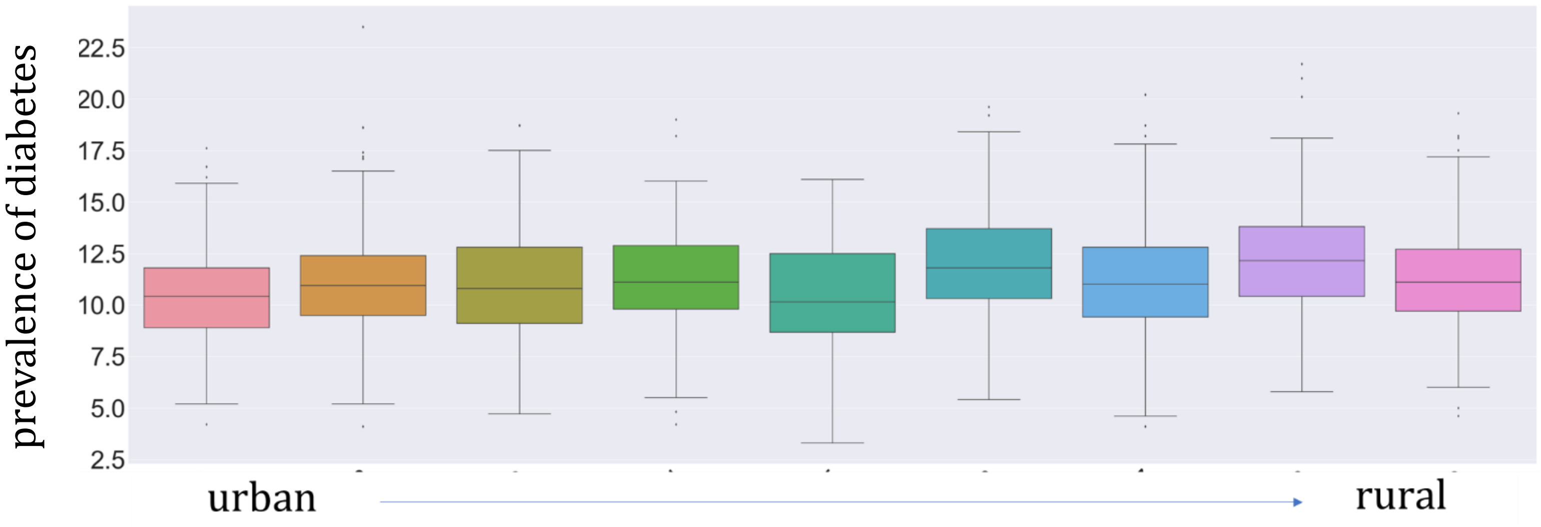
High diabetes prevalence
More people ≥ 65 years old

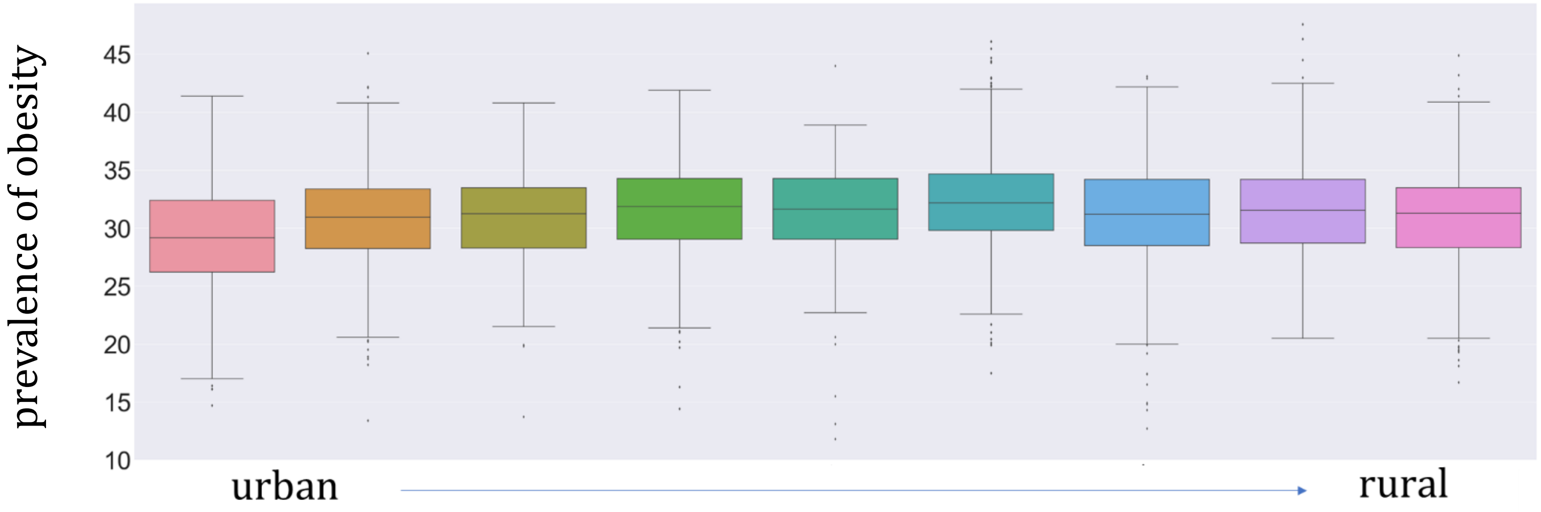
EDA: prevalence of diabetes and prevalence of obesity are significantly correlated with multiple variables



	prevalence of diabetes	prevalence of obesity	prevalence of physical inactivity	fraction of uninsured population	fraction of population < 18 years	fraction of population ≥ 65 years	fraction of Black population	fraction of White population	availability of local foods	poverty rate
prevalence of diabetes	1.0	0.691	0.738	0.127		0.223	0.462	-0.102	-0.083	0.531
prevalence of obesity	0.691	1.0	0.708		0.169		0.367	-0.068	-0.059	0.408
prevalence of physical inactivity	0.738	0.708	1.0	0.179		0.191	0.283			0.464
fraction of uninsured population	0.127		0.179	1.0	0.311	-0.138	0.122	-0.481	-0.128	0.391
fraction of population < 18 years		0.169		0.311	1.0	-0.527		-0.327	-0.188	
fraction of population ≥ 65 years	0.223		0.191	-0.138	-0.527	1.0	-0.229	0.409	0.266	-0.075
fraction of Black population	0.462	0.367	0.283	0.122		-0.229	1.0	-0.617	-0.171	0.517
fraction of White population	-0.102	-0.068		-0.481	-0.327	0.409	-0.617	1.0	0.231	-0.485
availability of local foods	-0.083	-0.059		-0.128	-0.188	0.266	-0.171	0.231	1.0	-0.129
poverty rate	0.531	0.408	0.464	0.391		-0.075	0.517	-0.485	-0.129	1.0

EDA: No significant difference in distribution of prevalence of diabetes or obesity among urban and rural areas (examined using permutation tests)

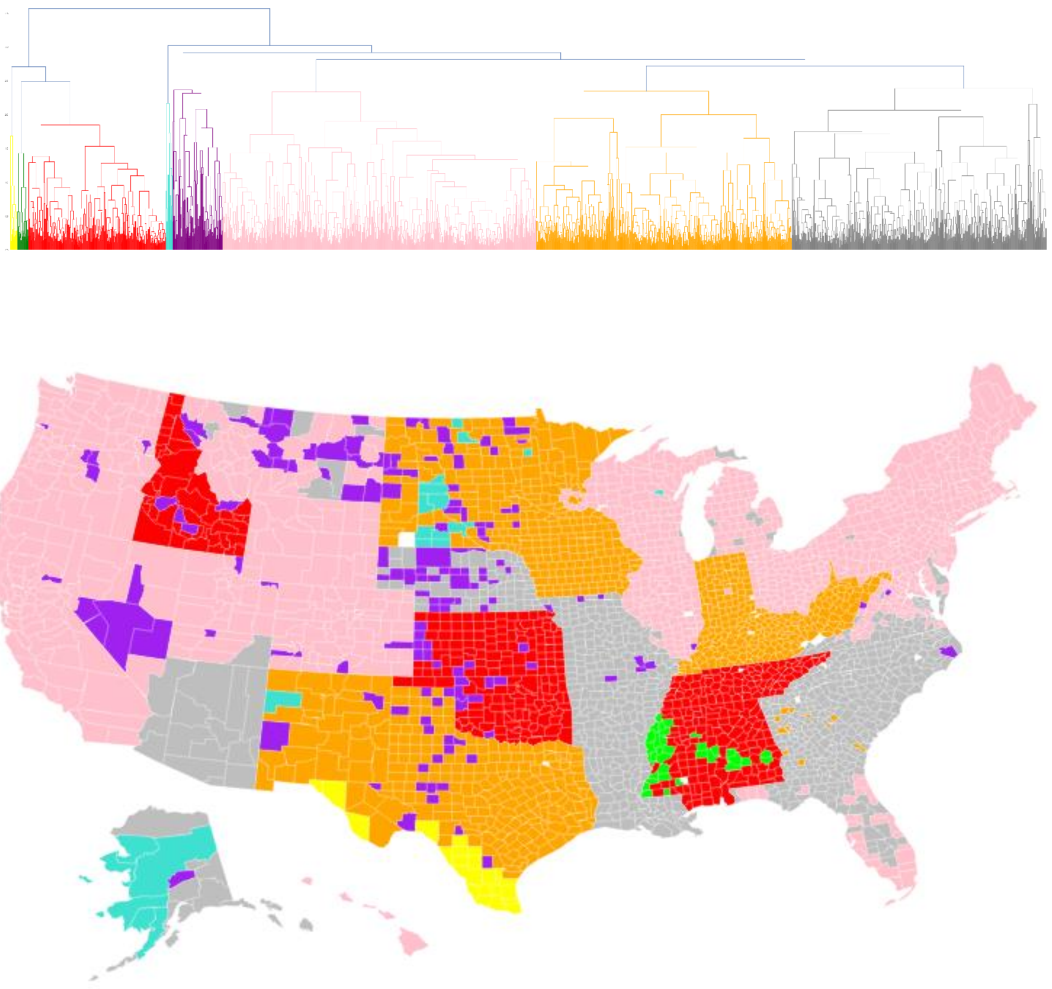




Hierarchical Clustering Analysis: Can we identify distinctive groups of US counties?

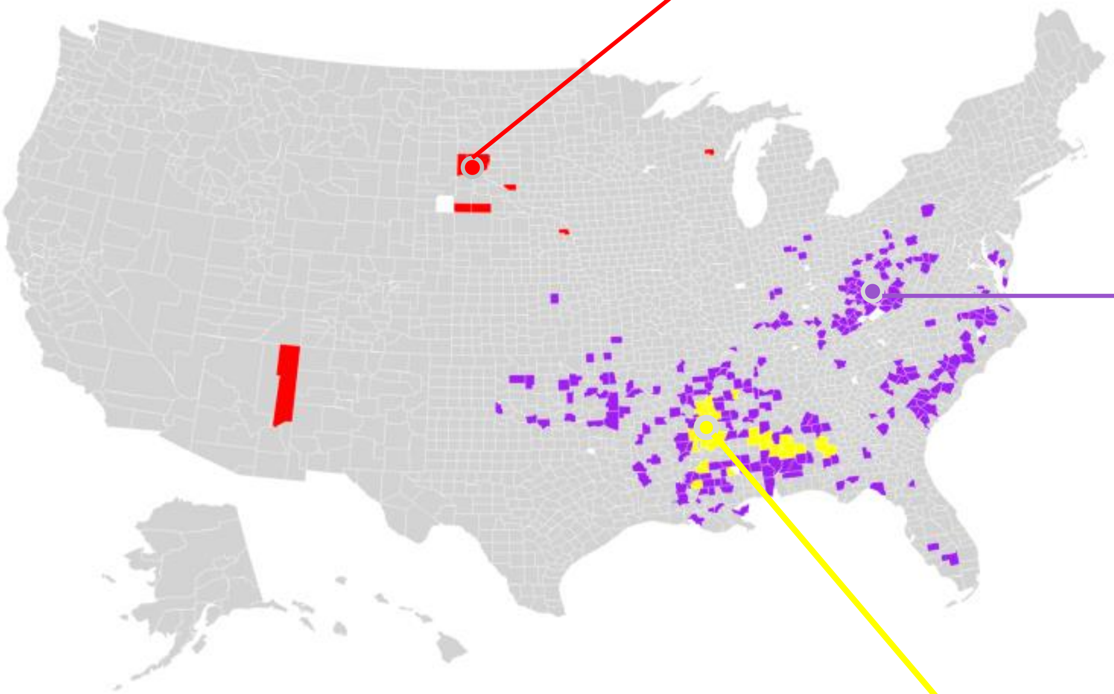
- Normalizing data: each column is rescaled between 0 and 1
- Principal Component Analysis (PCA) reduces the dimensionality and control for multicollinearity
- Using complete linkage method to find clusters with clear distinctions

Hierarchical clustering identifies eight clusters of US counties



- 🟨 Hispanic dominated, young, poor, low prevalence of diabetes and obesity, physically active
- 🟩 Black dominated, poor, high prevalence of diabetes and obesity, highly physically inactive
- 🟥 White dominated, moderately old, moderately high prevalence of diabetes and obesity, physically inactive
- 🟦 American Indian and Alaska Native dominated, young, poor, high prevalence of obesity, moderate physical activity
- 🟪 White dominated, old, low prevalence of diabetes and obesity, moderately physically active, low poverty rate
- ▨ White dominated, low prevalence of diabetes and obesity, highly physically active
- 🟧 White dominated, moderately old, moderate prevalence of diabetes and obesity, moderately physically active
- ⬜ White dominated with Black as second dominant race, moderately high prevalence of diabetes and obesity, physically inactive

Three distinct groups of the counties with highest prevalence of diabetes and highest prevalence of obesity



American Indian and Alaska Native dominated
moderate physical activity,
high poverty rate

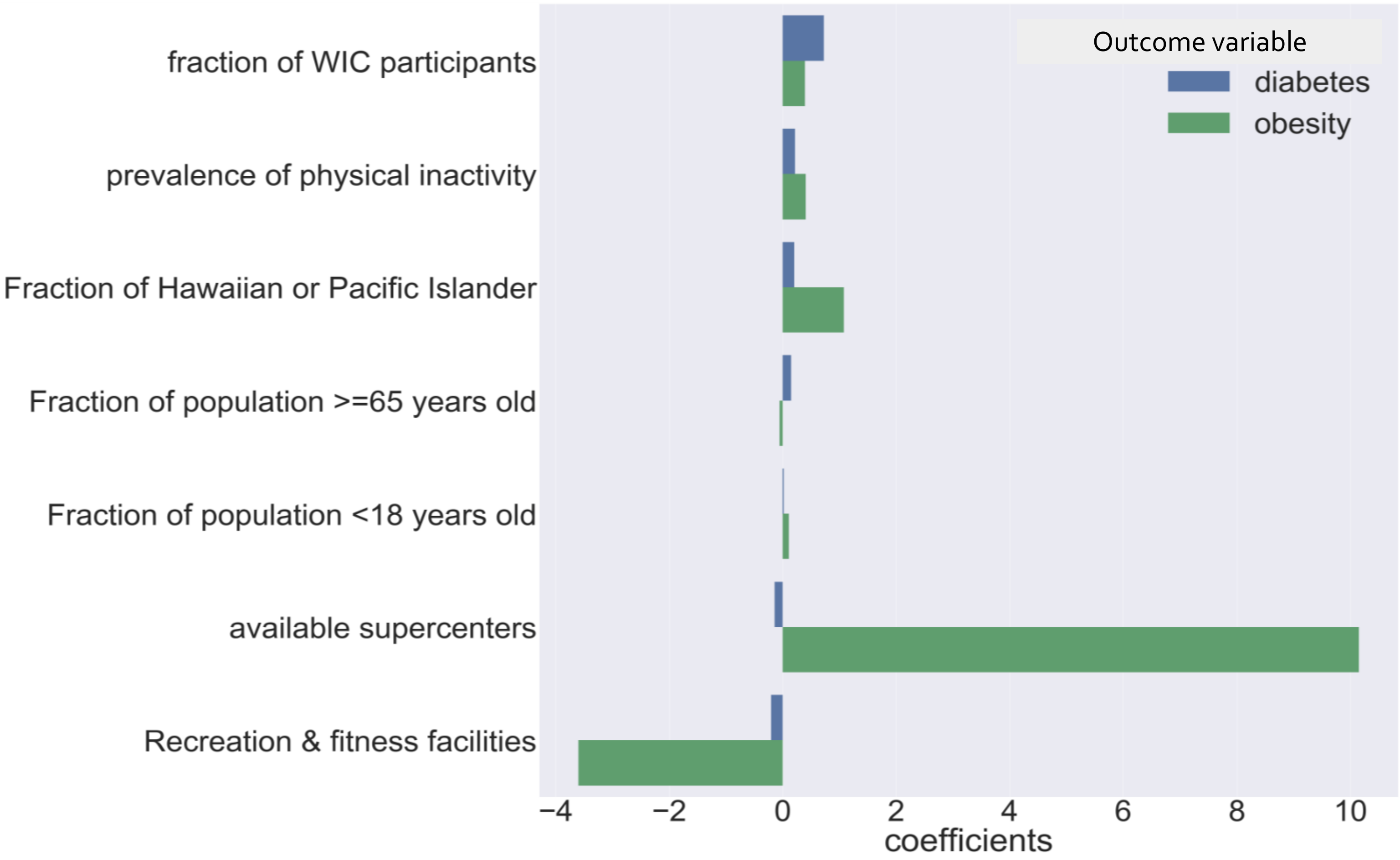
White dominated
Black as second dominant race
physically inactive

Black dominated
highly physically inactive
high poverty rate

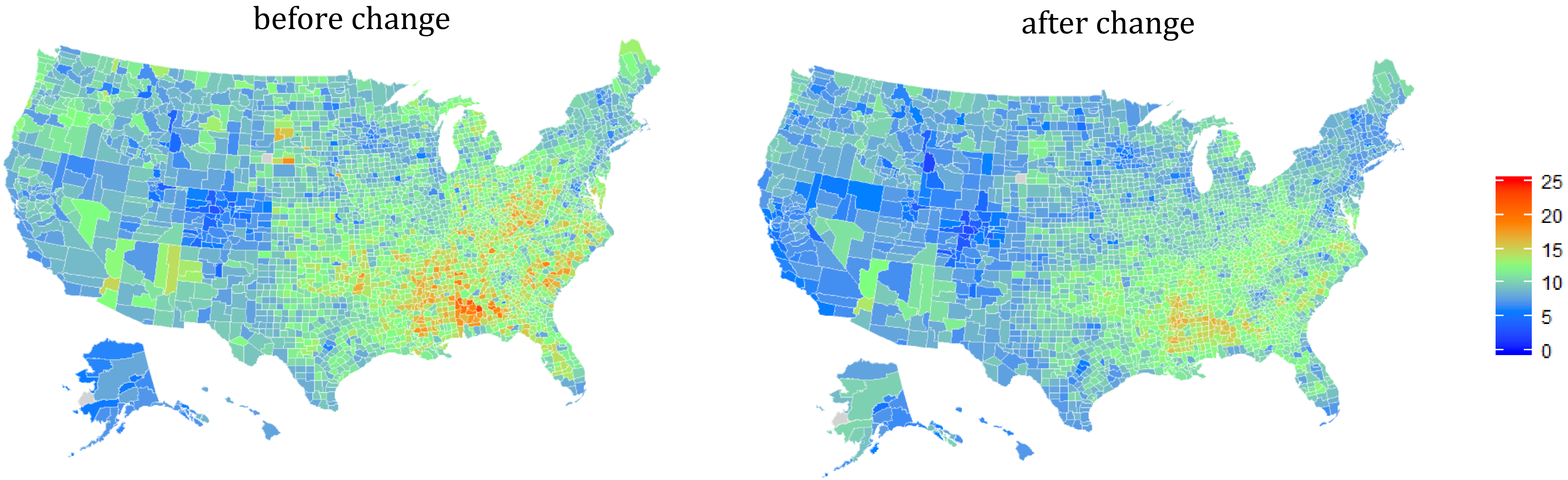
Regression analysis: What are important predictors for US counties prevalence of diabetes and obesity?

- Missing value imputation
- Data normalization
- 25% as test data and 75% as training data
- Elastic net regression analysis
- Five-fold cross validation

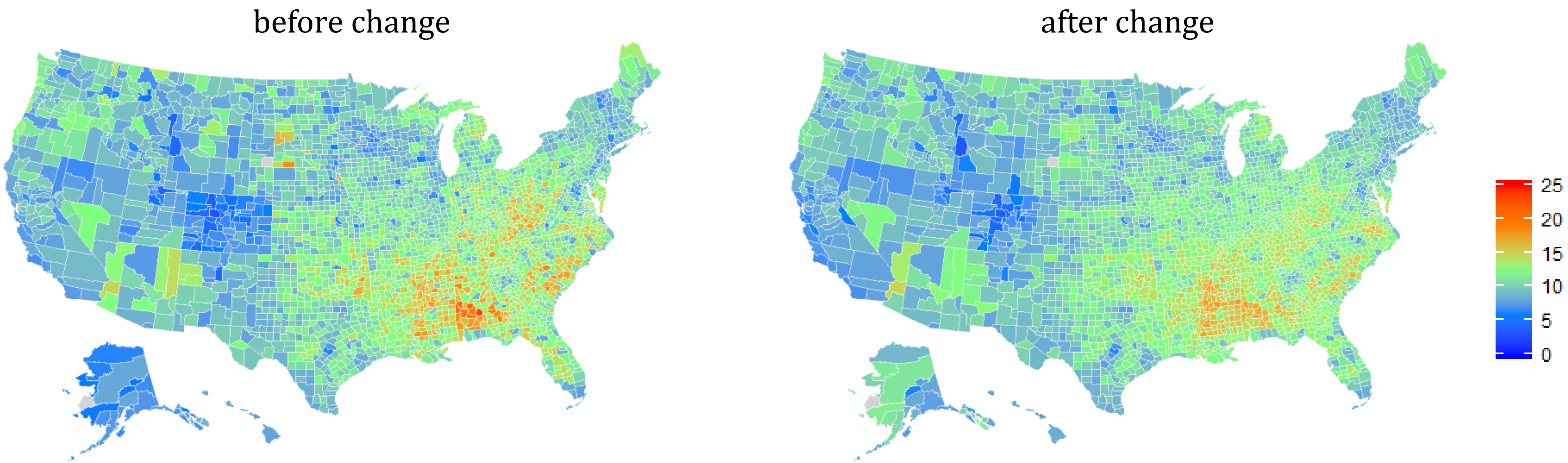
Some predictors have similar impact for prevalence of diabetes and prevalence of obesity while some others show different impacts on two outcome variables



Predicted Prevalence of Diabetes if fraction of population with physical inactivity decreases by 5%



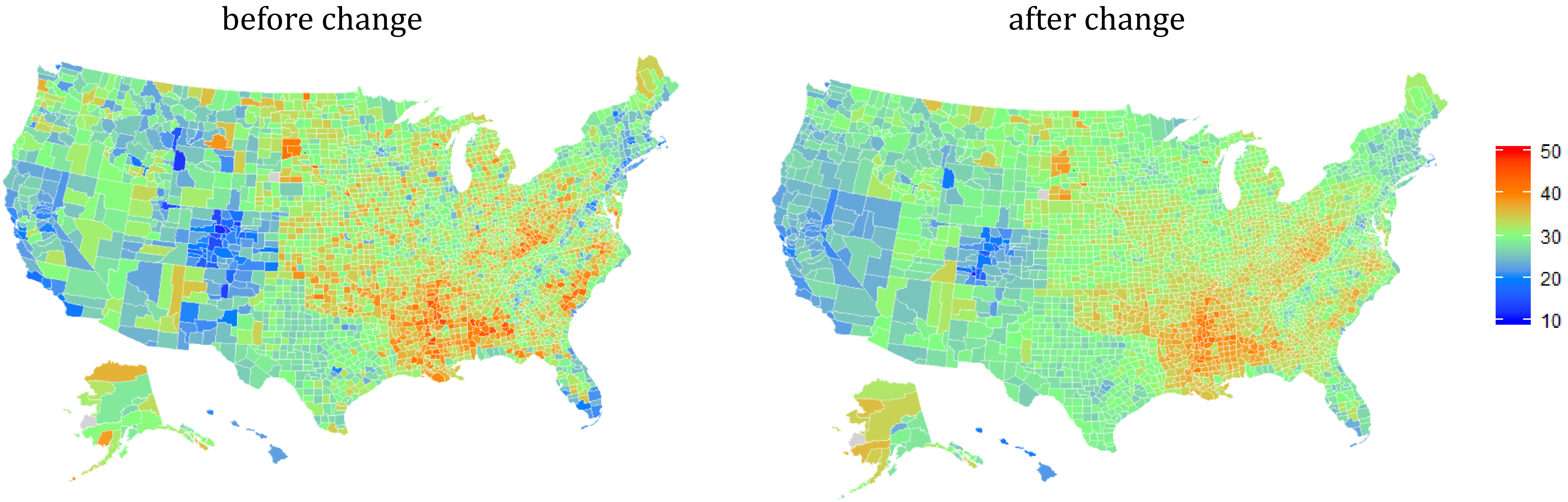
Predicted Prevalence of Diabetes if fraction of grocery stores increases by 5%



before change

after change

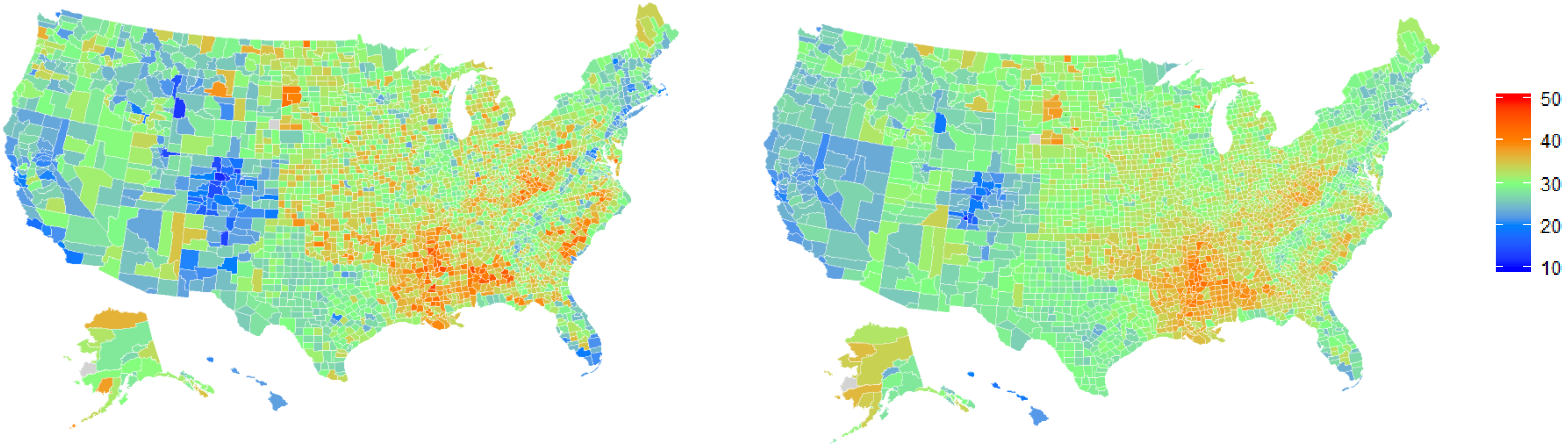
Predicted Prevalence of Obesity if recreation facilities increases by 5%



before change

after change

Predicted Prevalence of Obesity if available local foods increases by 5%



before change

after change

Messages

- Counties have high prevalence of diabetes and high prevalence of obesity are mostly in Southeast US. Improving physical activity and food choice likely reduce prevalence of diabetes and obesity in these areas.

- Improving recreation facilities and food choice can considerably reduce prevalence of obesity, especially counties in Southeast US, Midwest US and Northen Alaska.

- High level of physical activity is likely the reason of low prevalence of diabetes and obesity in counties along shoreline of ocean and lakes and along Rocky mountains. Food environment likely plays little role in population health in these areas.