*Springboard Data Science Capstone Project 1*

Correlations between poverty, food environment and diabetes in the United States

Shuangshuang Liu

January, 2020

Table of Contents

[1. Introduction: 1](#_Toc31747990)

[2. Methods 4](#_Toc31747991)

[2.1 The Dataset 4](#_Toc31747992)

[2.2 Data wrangling 4](#_Toc31747993)

[2.3 Statistical Analysis and Modeling 6](#_Toc31747994)

[2.3.1 Mean comparisons evaluating the effects of urbanity and poverty 6](#_Toc31747995)

[2.3.2 Predicting diabetes with food environment features (supervised learning) 6](#_Toc31747996)

[2.3.3 Finding patterns of food-economic environment among counties (unsupervised learning) 7](#_Toc31747997)

[3. Results 9](#_Toc31747998)

[3.1 Effects of urbanity and poverty on diabetes rate 9](#_Toc31747999)

[3.2 Important features predictive of diabetes rate 9](#_Toc31748000)

[3.3 The landscape of food environment across the United States 16](#_Toc31748001)

[4. Discussion 19](#_Toc31748002)

[4.1 Important features predictive of diabetes rate 19](#_Toc31748003)

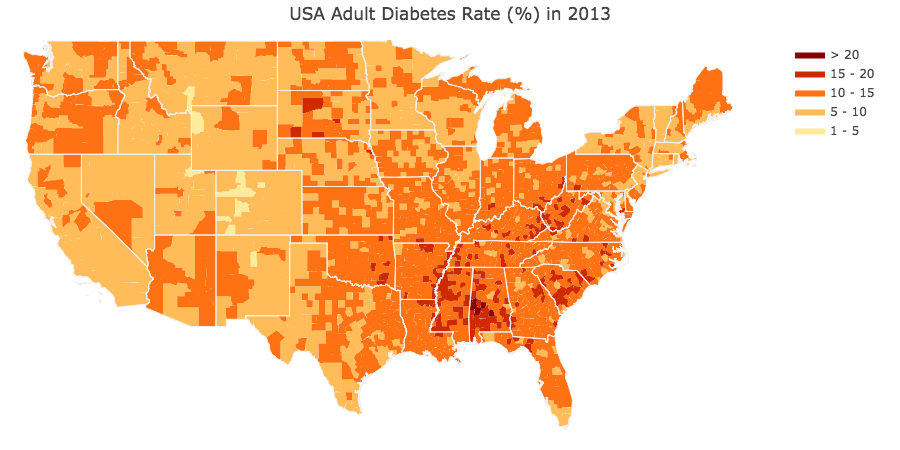
[4.2 The landscape of food environment across the United States 21](#_Toc31748004)

[References 22](#_Toc31748005)

# 1. Introduction:

Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar), which leads over time to serious damage to the heart, blood vessels, eyes, kidneys, and nerves (definition by World Health Organization, WHO). It is considered a major cause of blindness, kidney failure heart attacks, stroke and lower limb amputation, and is estimated to be the seventh leading cause of death in 2016. Type 2 diabetes, which is usually found in adults, is the most common type and its prevalence has increased significantly during the past three decades.

According to data from WHO, by 2014 there are 422 million people with diabetes, up from 108 million in 1980. In the United States, more than 30 million people (all ages, 2015) have diabetes and 1 in 4 of them do not know they have it. Moreover, in the last 20 years the number of number of adults diagnosed with diabetes has more than doubled as the American population has aged and become more overweight or obese (Diabetes quick facts, Centers for Disease Control and Prevention). An overview of U.S. adult diabetes rate is shown below (Fig. 1).



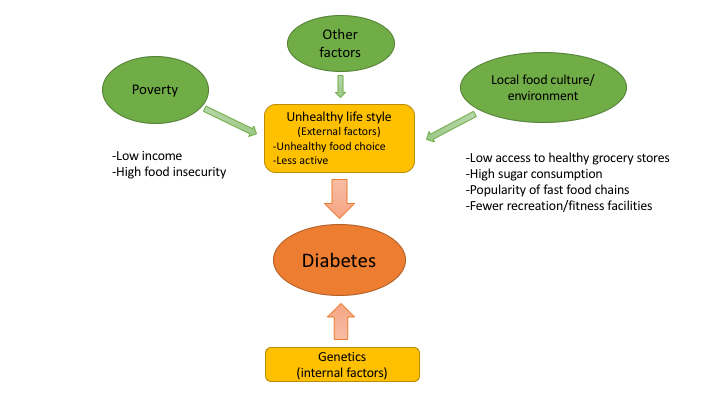
*Figure 1. Adult diabetes rate (%) in United States counties in 2013. Data source: Food Environment Atlas (https://www.ers.usda.gov/data-products/food-environment-atlas/)*

Measures to prevent or delay the onset of type 2 diabetes include healthy diet, regular physical activity, maintaining a normal body weight and avoiding tobacco use (International Diabetes Federation). Interestingly, the prevalence of diabetes has increased more rapidly in middle- and low- income countries (Diabetes fact sheets, WHO). This suggests that financial well-being may be correlated with the level of prevalence of diabetes. One explanation is that people in a poverty-persistent community may have less access to healthy foods and fitness facilities

such as football fields and gyms, or have less incentives to maintain a healthy lifestyle due to financial hardship.

Recent research on U.S. counties showed that low poverty/low-minority population counties had the greatest access to farmer’s markets and grocery stores, and that access to full services restaurants were significantly associated with lower prevalence of diabetes (Haynes-Maslow and Leone, 2017). However, the analysis in Haynes-Maslow and Leone (2017) has a few cavities, for example they categorized poverty level into low and high poverty groups which costed information loss; only used a multiple linear regression model on a handful of features and did not control for outliers. Other research had also found that high poverty neighborhoods have fewer supermarkets (Bower, et al., 2013), or a lower availability of healthy foods due to differential placement of types of stores and offerings of healthy foods within similar stores (Franco, et al., 2008).

Despite of previous research, we still lack a clear picture of the role of economic and food-environment indicators on diabetes prevalence. For example, which indicator is a better predictor of diabetes rate of a county, the poverty level or the availability of healthy foods? Are there any patterns in the food environment across U.S. counties?

In this analysis, we hypothesized that poverty level and local food culture/environment, as measured by factors such as access to grocery stores, among other possible factors, may contribute to unhealthy life style and thus be associated with diabetes prevalence. The hypothesis can be summarized in the following diagram (Fig. 2).

*Figure 2. Diagram illustrating the potential factors that may link to diabetes prevalence.*

Specifically, we asked:

1) Are there significant differences in adult diabetes rate between metro and non-metro counties (i.e., urbanity, which is related to access to different types of stores), as well as poverty-persistent and non-poverty counties?

2) What are the most important features that may contribute to diabetes rate? Among those features which ones are more significantly correlated with diabetes rate, indicators of poverty level or food environment?

3) Are there any patterns of food environment across U.S. counties?

We will utilize the "Food Environment Atlas" database (<https://www.ers.usda.gov/data-products/food-environment-atlas/>) and apply a suite of machine learning algorithms (supervised and unsupervised learning) to address the above questions.

# 2. Methods

## 2.1 The Dataset

The Food Environment Atlas dataset (Economic Research Service, U.S. Department of Agriculture) is used in this analysis. The dataset contains 275 features of the food environment factors in three broad categories for 3143 U.S. counties:

*1) Food Choices*—Indicators of the community access to and acquisition of healthy, affordable food, such as: access and proximity to a grocery store; number of food stores and restaurants; expenditures on fast foods; food and nutrition assistance program participation; food prices; food taxes; and availability of local foods.

*2) Health and Well-Being*—Indicators of the community success in maintaining healthy diets, such as: food insecurity; diabetes and obesity rates; and physical activity levels.

*3) Community Characteristics*—Indicators of community characteristics that might influence the food environment, such as: demographic composition; income and poverty; population loss; metro-nonmetro status; natural amenities; and recreation and fitness centers.

## 2.2 Data wrangling

*1) Feature selection and engineering*

We first excluded features measuring percent of change between years, eight categorical features related to SNAP program (food aid), four features classifying metro/non-metro and poverty-persistent counties. This yielded 138 features for clustering analyses (unsupervised learning), relating to our second question (i.e., identify patterns of food environment among counties).

To find potentially important features to address the first question (i.e., can food environment and poverty predict diabetes rate?), we first regressed state averages of each of the 138 features against the state average of adult diabetes rate in 2013 (target variable), which resulted in 66 candidate features that had significant correlations (p<0.05, unadjusted for multiple testing). Among the 67 features we selected 28 that are directly related to our hypothesis for regression analyses (supervised learning), as shown in Fig.3:



*Figure 3. Simple linear regression correlation coefficients between 28 Selected features against adult diabetes rates in 2013 based on state average values (p< 0.05, unadjusted for multiple testing). Note years of the features may vary between 2010 and 2016 (we assume diabetes rates in 2016, data of which not available, are highly correlated with that in 2013). When data of a feature for multiple years are available, only the most recent year is considered.*

To address the second question (i.e., identify patterns of food environment among counties), we selected a slightly different suite of features. When multiple years of data available for an indicator, only the most recent year was selected. Similar to previous selection process, features measuring percent of change between years and 12 categorical features were removed.

*2) Dealing with missing values*

We used sklearn.impute.IterativeImputer to impute missing values for county-level data. This imputer models each feature with missing values as a function of other features, and uses that estimate for imputation. The default estimator uses Bayesian Ridge regression and we set the minimum imputed value to be non-negative.

*3) Dealing with outliers*

We used scipy.stats.mstats.winsorize to treat outliers for the county-level data. The lowest and highest value are set to the 5th and 95th percentile, respectively.

## 2.3 Statistical Analysis and Modeling

### 2.3.1 Mean comparisons evaluating the effects of urbanity and poverty

We conducted Student’s *t*-test to compare means diabetes rates between metro (1167) and non-metro (1976) counties, and between poverty-persistent (353) and non-poverty (2790) counties. A Welch’s *t*-test was performed for the latter comparison, due to relatively large difference in data size between the groups which may lead to unequal variances.

### 2.3.2 Predicting diabetes with food environment features (supervised learning)

A suite of supervised learning models was selected to study the relationship between diabetes rate and the 28 features related food environment. The models were applied in an order of increased complexity, starting with multiple linear regression, then decision trees, random forest and lastly XGBoost.

Data was split into training set (60%) and testing set (40%) before the analyses.

#### 1) Multiple Linear Regression

Multiple linear regression models the linear relationship between the explanatory variables (features) and response variable. It is in essence an extension of ordinary least-squares regression that involves more than one explanatory variable. It assumes homoscedastic and normally distributed regression residuals, and that the explanatory variables are not highly correlated (absence of multicollinearity). As implemented in sklearn, the model minimized the residual sum of squares between the observed and predicted target variable.

This approach allows us to establish a baseline model to assess how well the features can explain diabetes rates at the simplest level of modeling. Here the 28 selected features (Fig. ?) (explanatory variables) were regressed on adulted diabetes rate (response variable). We examined the correlation matrix among the features to ensure there are no highly correlated features, and the residuals were checked after fitting the model.

#### 2) Decision Trees

Decision Trees are a non-parametric supervised learning method which can be used for both classification and regression problems. A tree is built by splitting the source set into subsets, the process of which repeats on each derived subset until the subset at a node has all the same values of the target variable, or when splitting no longer adds value to the predictions. Different algorithms use different metrics the determine the ‘best’ split for each (sub)set, such as Gini impurity (Decision tree learning, Wikipedia). A regression tree is used when the target (predicted) variable is continuous. For the DecisionTreeRegressor in sklearn, the default function to measure the quality of a split is ‘mse’ (mean squared error), which minimizes the L2 loss using the mean of each terminal node.

The advantages of Decision Trees is its interpretability. Nonetheless, the model has some disadvantages such as high variance (i.e., small changes in data can lead to different trees), tends to overfit and can only do axes-aligned splits. To avoid overfitting, hyperparameter tuning and cross-validation is required. In our case, we will focus on tuning the maximum tree depth, minimum samples required per split, minimum samples required per leaf node, and maximum features considered for the best split. The class GridSearchCV implemented in sklearn will be used for this task.

#### 3) Random Forest

Random forest builds upon the idea of bagging (bootstrap aggregating). It fits a number of decision trees, each based on a bootstrapped (drawn with replacement) sub-sample of the dataset, and then averages the trees to improve predictive accuracy and control over-fitting. In addition, the node splits are calculated from random feature subsets, which introduces another level of randomness into the model. In comparison, in a Decision Tree all features are examined and the best one is selected for each split.

Similarly as Decision Trees, GridSearchCV will be applied for hyperparameter tuning and cross validation. We will tune the same set of hyperparameters plus the number of trees.

#### 4) XGBoost

XGBoost (eXtreme Gradient Boosting) is an optimized distributed gradient boosting library which implements machine learning algorithms under the Gradient Boosting framework. It is highly efficient, flexible and portable (XGBoost Github webpage). Distinctive features of XGBoost include clever penalization of trees, a proportional shrinking of leaf nodes, Newton Boosting, and extra randomization of parameters (XGBoost, Wikipedia).

We will also use GridSearchCV to tune hyperparameters of the XGBoost model, including max\_depth, min\_child\_weight, gamma, subsample, colsample\_bytree, reg\_alpha and reg\_lambda.

### 2.3.3 Finding patterns of food-economic environment among counties (unsupervised learning)

#### 1) Principal component analysis (PCA)

PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables called principal components (Principal component analysis, Wikipedia). PCA is an effective tool for reducing the dimensionality of a dataset. The resulted principal components can be further used in K-means clustering to explore potential patterns within the dataset.

In our case, we will conduct PCA on four groups of variables: 1) the 15 features selected based on the linear regression above, which contains both economic and food-environment indicators; 2) 45 variables which belong to the ‘Local’ subcategory of the original dataset, including indicators on availability of local foods such as number of farmer’s market and vegetable farms; 3) 6 variables which belong to the ‘Restaurants’ subcategory, containing indicators on availability of dining choices and people’s average expenditures on those venues; 3) 12 variables which belong to the ‘Stores’ subcategory, with indicators on the availability of groceries including those participating in food assistance programs (i.e., SNAP). The subcategories ‘Local’, ‘Restaurants’ and ‘Stores’ were selected because they are most representative of local food environment. Data were scaled using StandardScaler implemented in sklearn prior to PCA analyses.

#### 2) K-means clustering

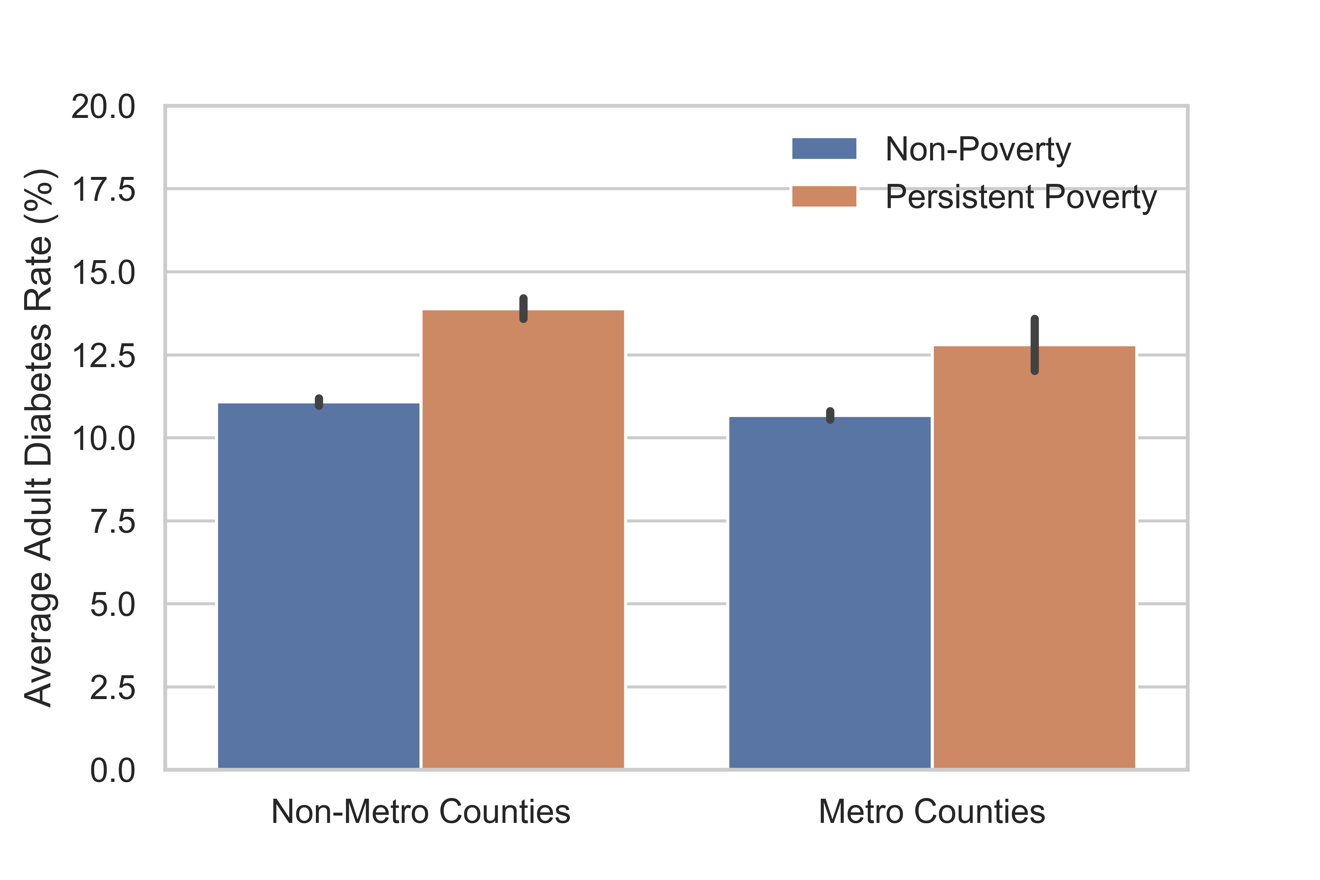
K-means clustering partitions *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (k-means clustering, Wikipedia). The algorithm minimizes within-cluster variances, which are calculated as squared Euclidean distances. It is guaranteed to converge, but it is sensitive to outliers and the results depend on initialization (i.e., not guaranteed to converge to the global optimum). The number of clusters (*k*) needs to be chosen for initiation, and can be determined via plotting within groups sum of squares against number of clusters (the ‘Knee’ or ‘Elbow’ plot).

We will use the Clustering package implemented in SciPy (Johns et al., 2001) to conduct K-means clustering analysis.

# 3. Results

## 3.1 Effects of urbanity and poverty on diabetes rate

Both urbanity and poverty had significant effects on adult diabetes rates among the counties (p<0.05 for both cases). On average, metro counties had lower diabetes rates than non-metro counties, and non-poverty counties had lower diabetes rates than persistent counties (Fig. 4).

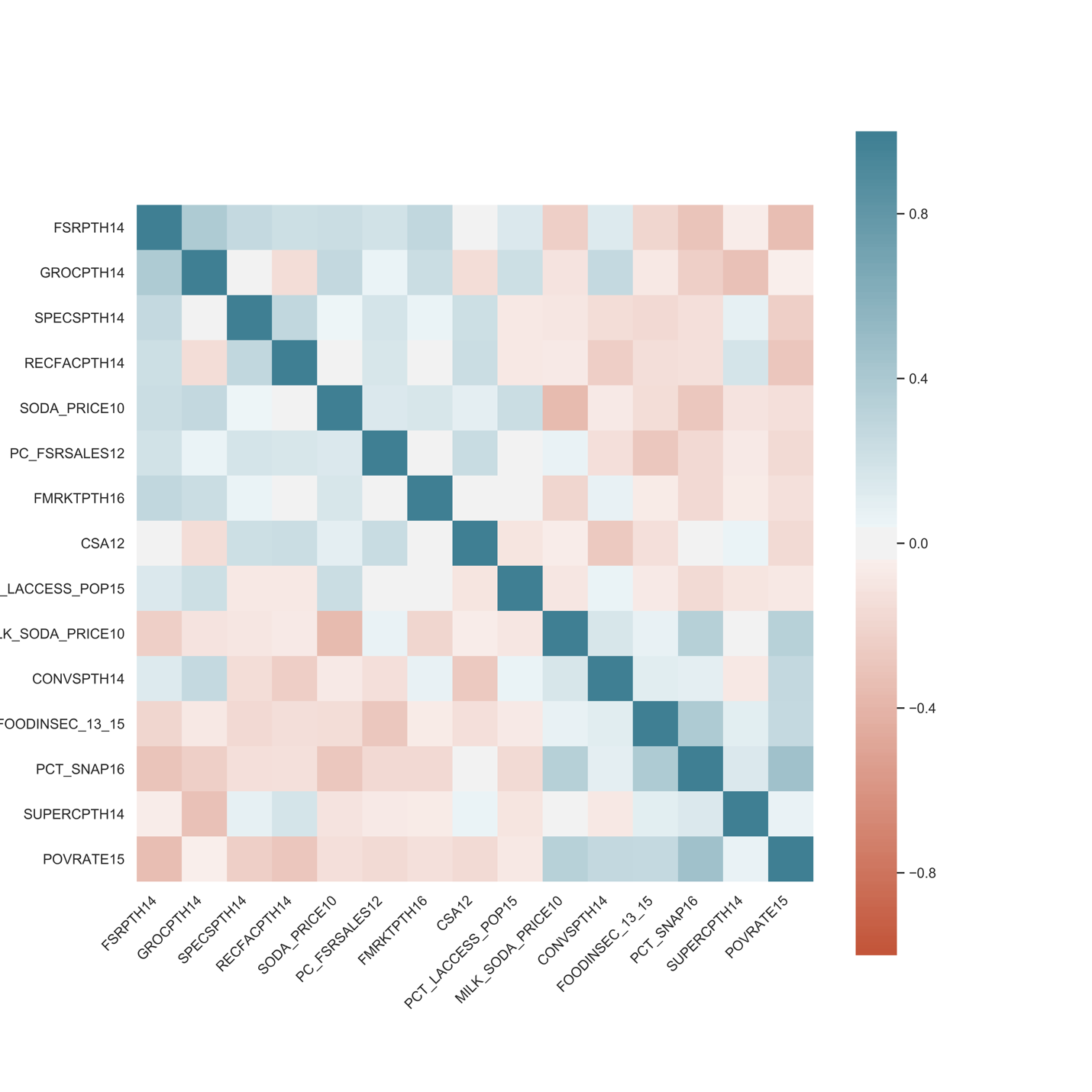


*Figure 4. Average adult diabetes rate in 2013 (%) among 3143 counties in the United States.*

## 3.2 Important features predictive of diabetes rate

#### 1) Multiple Linear Regression results

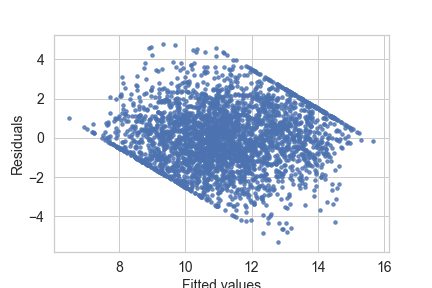
The correlation matrix suggests high correlations among a few variables, such as between medium household income and poverty rate, and between number of vegetable farms and CSA farms. These features (10 in total) were removed from initial regression. Afterwards, three features (agriculture tourism, fast food expenditures per capita and direct farm sales per capita) which had near-zero correlation coefficients were also removed. This resulted in 15 features in the final model (Fig. 5).



*Figure 5. Correlation matrix among the 15 selected features from the food environment atlas.*

Using the above 15 features to fit a linear regression model, we obtained an R2= 0.51 for the training dataset and an R2= 0.56 for the testing dataset. Model assumptions on the homoscedasticity and normality of residuals were confirmed via the fitted value – residuals plot and the QQ plot (Fig. 6), respectively.

**Summary:** linear regression model is slightly underfitting. A more complex model is desired for better explaining diabetes rates.



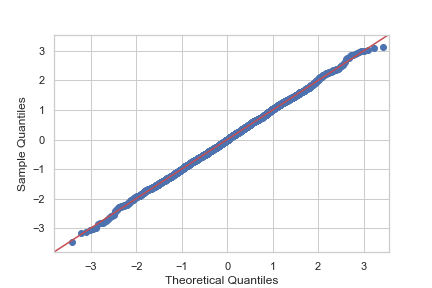


Figure 6. Residuals plot and QQ plot of the linear regression model.

#### 2) Decision Trees Results

After tuning hyperparameters and cross validation, our best decision tree obtained an R2= 0.58 for the training dataset and an R2= 0.55 for the testing dataset. The model accuracy is better than linear regression but is slightly overfitting.

According to the resulted best tree (Fig. 7), the most important feature (root node) is ‘restaurant expenditures per capita’ (PC\_FSRSALES12). Features for the 2nd and 3rd splits include:

2nd split:

- Price of low-fat milk/price of sodas (MILK\_SODA\_PRICE10)

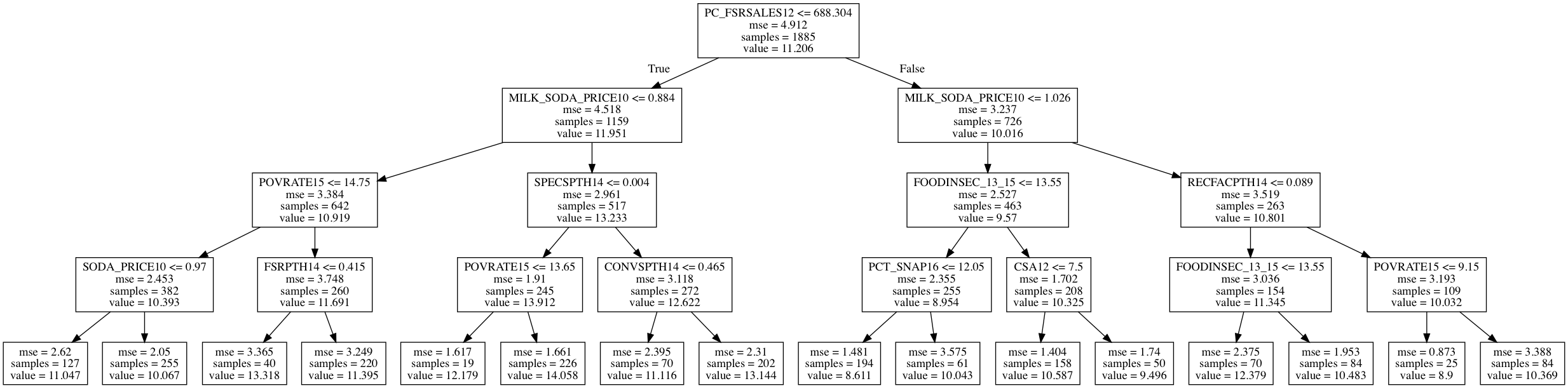
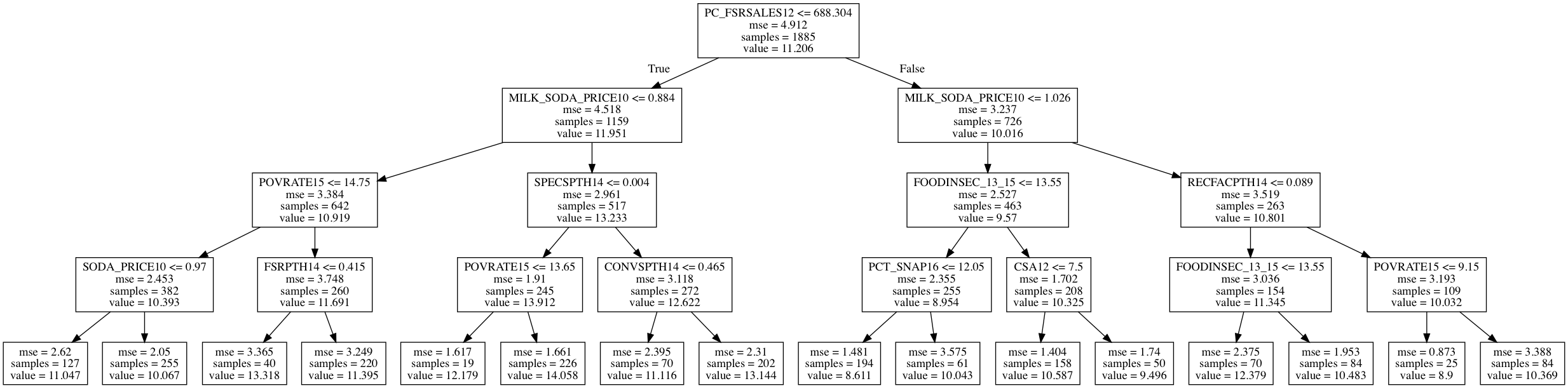
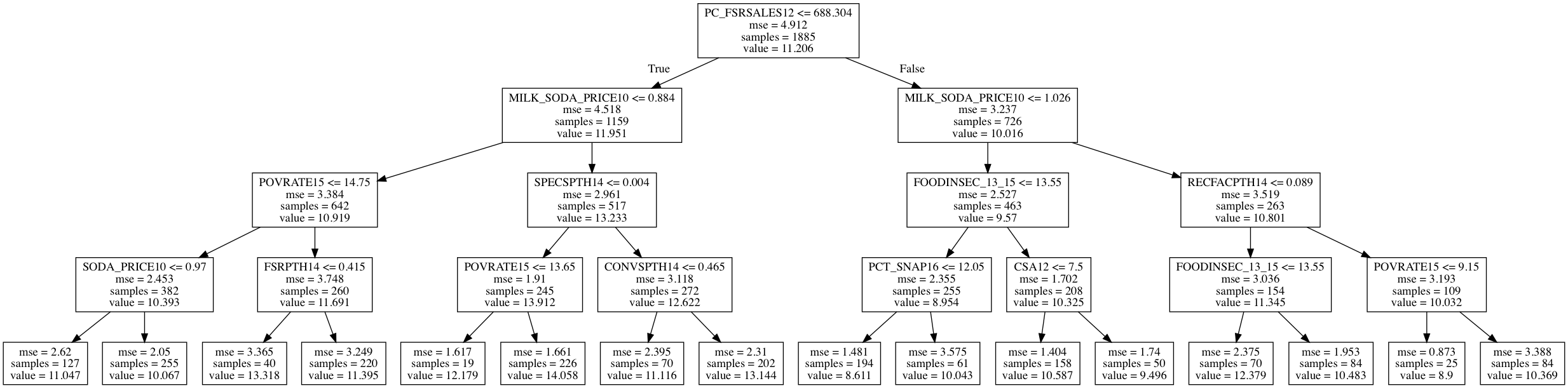
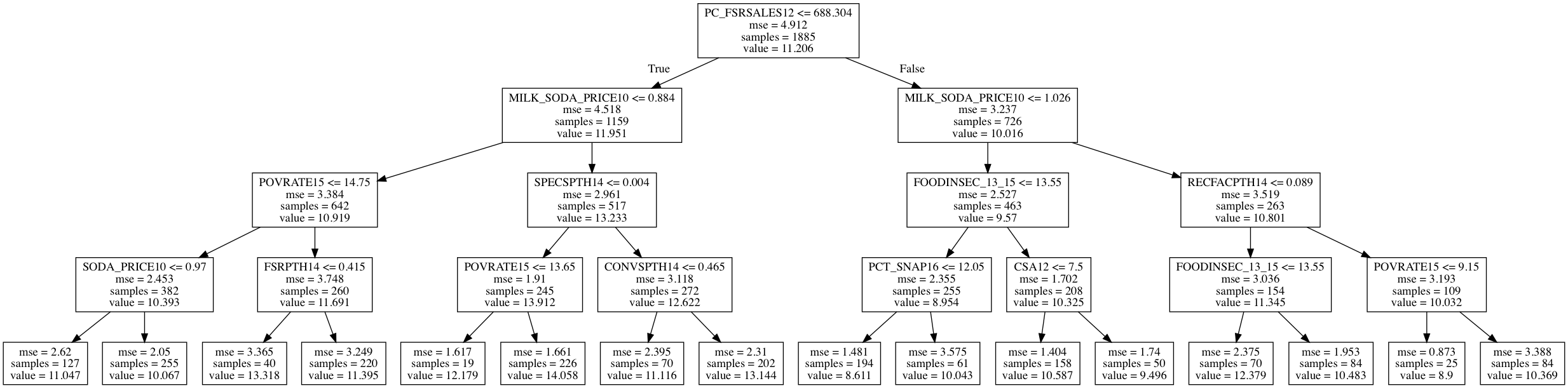
3rd split:

- Poverty rate (POVRATE15)

- Specialized food stores/1,000 pop (SPECSPTH)

- Food insecurity, three year average (FOODINSEC)

- Recreation & fitness facilities/1,000 pop (RECFACPTH14)



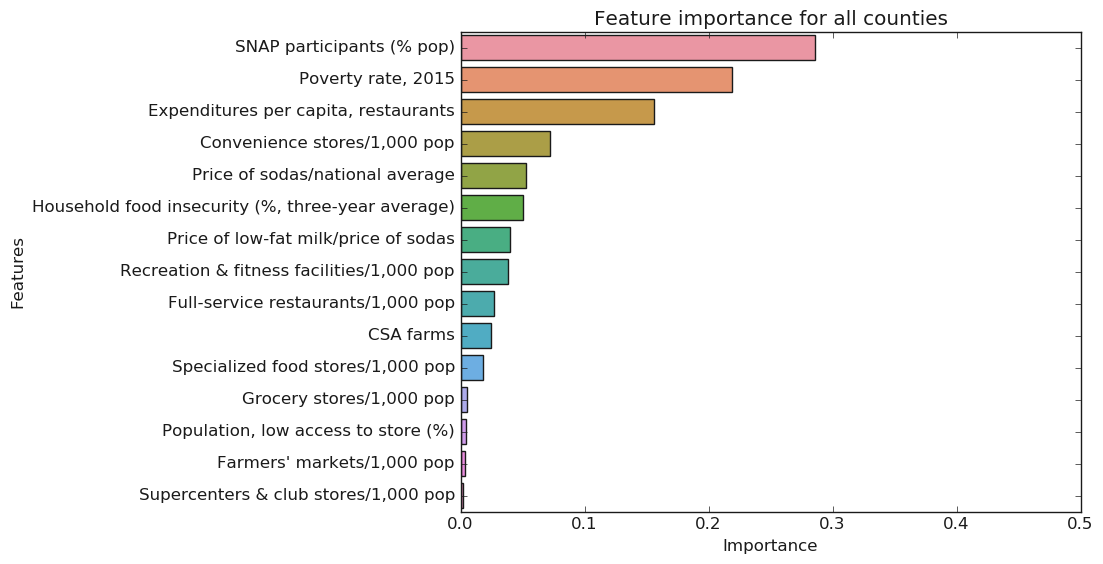
*Figure 7. Decision Tree based on best parameter estimates (split to left and right parts for easy viewing), using 15 food environment features as explanatory variables and diabetes rates as target variable.*

In general, higher restaurant expenditures and recreation and fitness facilities predict lower diabetes rate, whereas lower milk/soda price ratio, lower specialized food stores, higher poverty rate and food insecurity rate predict higher diabetes rate.

#### 3) Random Forest Results

After tuning hyperparameters and cross validation, our best random forest model obtained an R2= 0.68 for the training dataset and an R2= 0.63 for the testing dataset. The model accuracy improved comparing to linear regression a single decision tree. It may be slightly overfitting.

The Random Forest model is less interpretable than a decision tree, but it can rank the importance of features based on how much each feature contribute to reducing average variances (Fig.8). The top three most important features include percentage of SNAP (supplemental nutrition assistance program) participants, poverty rate and expenditures at restaurants per capita. (seems that economic factors more important than food factors



*Figure 8. Feature importance based on the Random Forest model for all counties.*

We also conducted Random forest on metro and non-metro counties separately. The best model for metro counties had an R2= 0.65 for the training dataset and an R2= 0.60 for the testing dataset. The rank of features importance is slightly different (Fig. 9). In particular, the number of convenience stores became the most important feature while poverty rate ranked 5th, compared to 2nd when considered all counties.

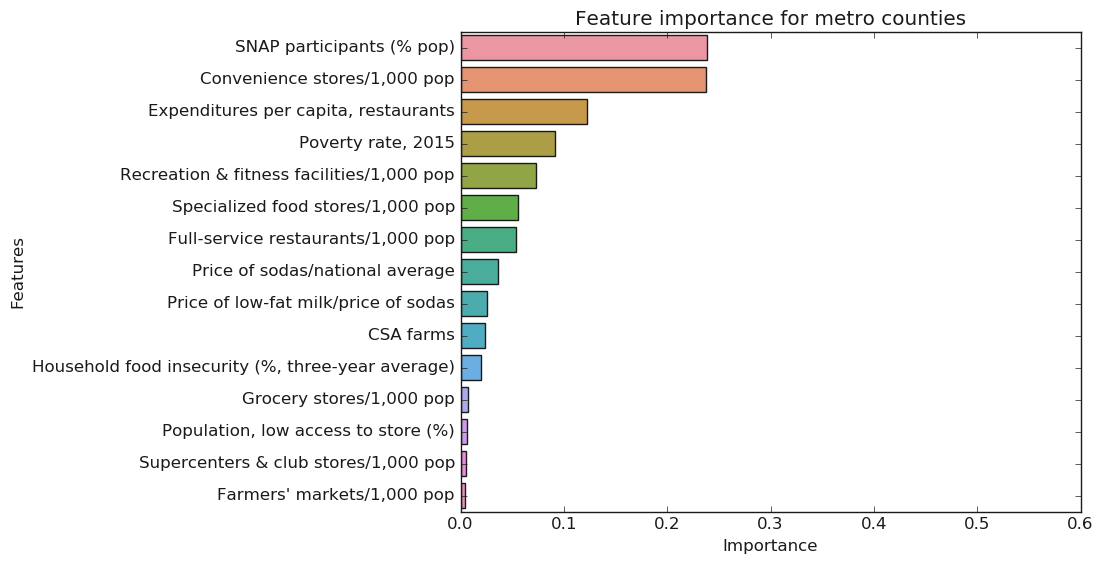


Figure 9. *Feature importance based on the Random Forest model for metro counties.*

On the other hand, the best model for non-metro counties had an R2= 0.53 for the training dataset and an R2= 0.50 for the testing dataset. The rank of features importance is shown below (Fig. 10). Price of low-fat milk/price of sodas became the 2nd most important feature, which was ranked 8th for metro counties. Expenditures at restaurants ranked the 6th, compared to the 3rd for metro counties.

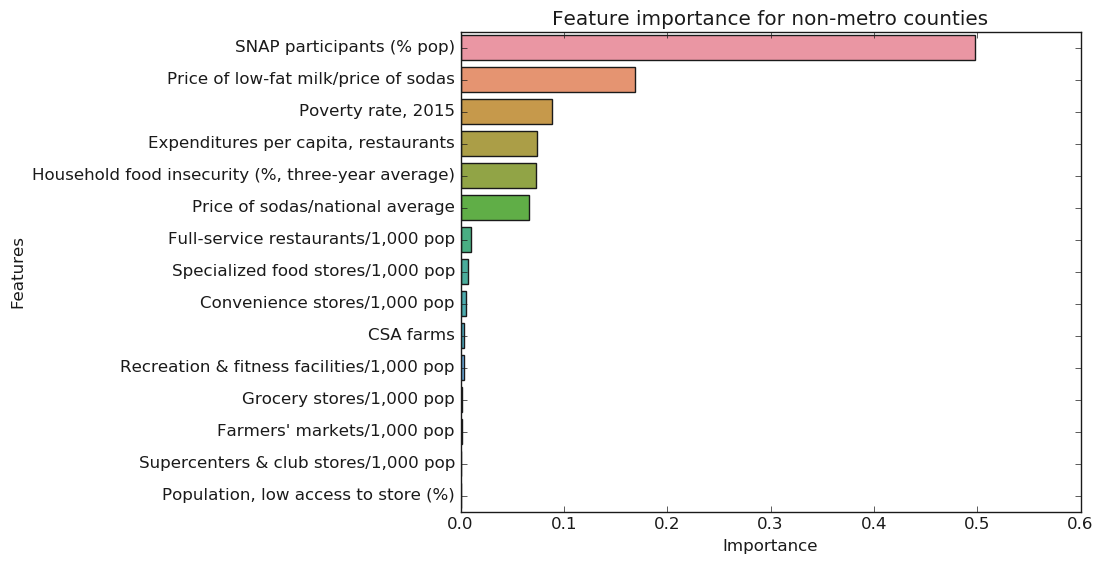


Figure 10. *Feature importance based on the Random Forest model for non-metro counties.*

Similarly, we analyzed non-persistent poverty counties versus persistent poverty counties (Fig 11-12).

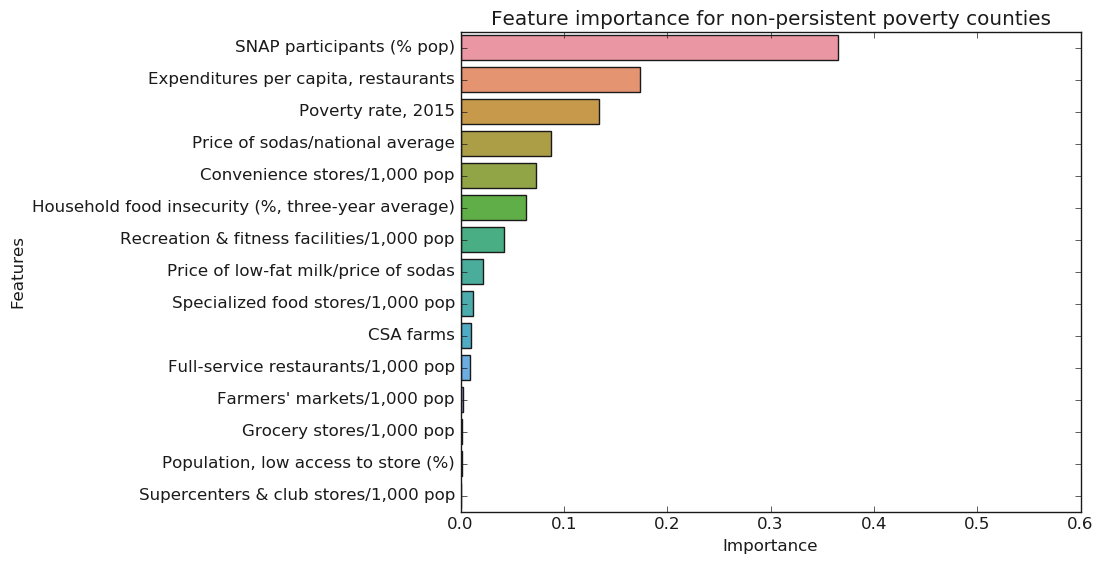


Figure 11. *Feature importance based on the Random Forest model for non-persistent poverty counties.*

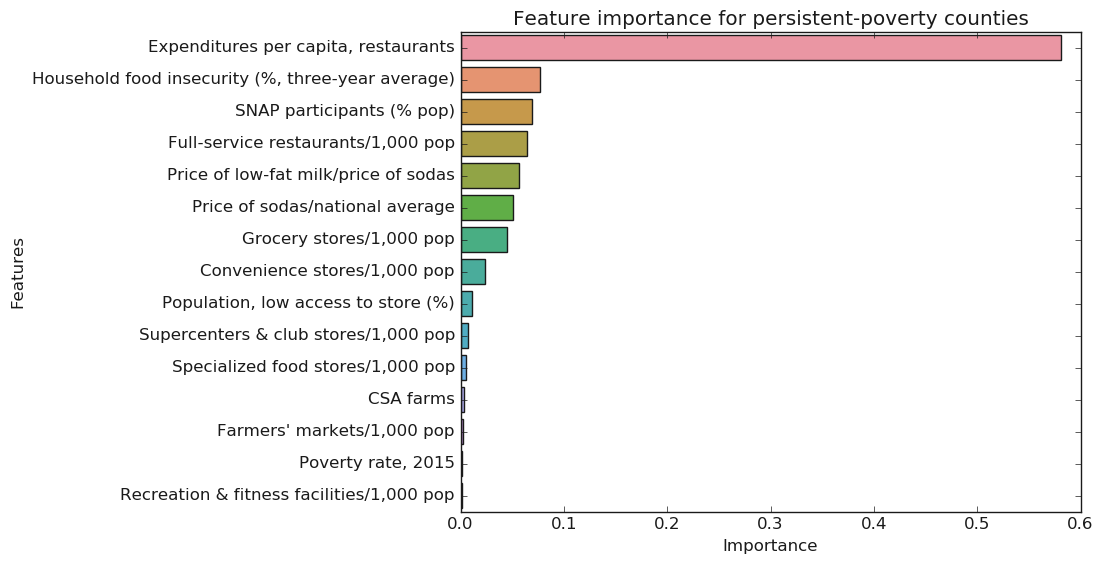


Figure 12. *Feature importance based on the Random Forest model for persistent poverty counties.*

#### 4) XGBoost results

After tuning hyperparameters and cross validation, our best random forest model obtained an R2= 0.83 for the training dataset and an R2= 0.69 for the testing dataset. The model accuracy improved comparing to random forest. However, it is a bit overfitting.

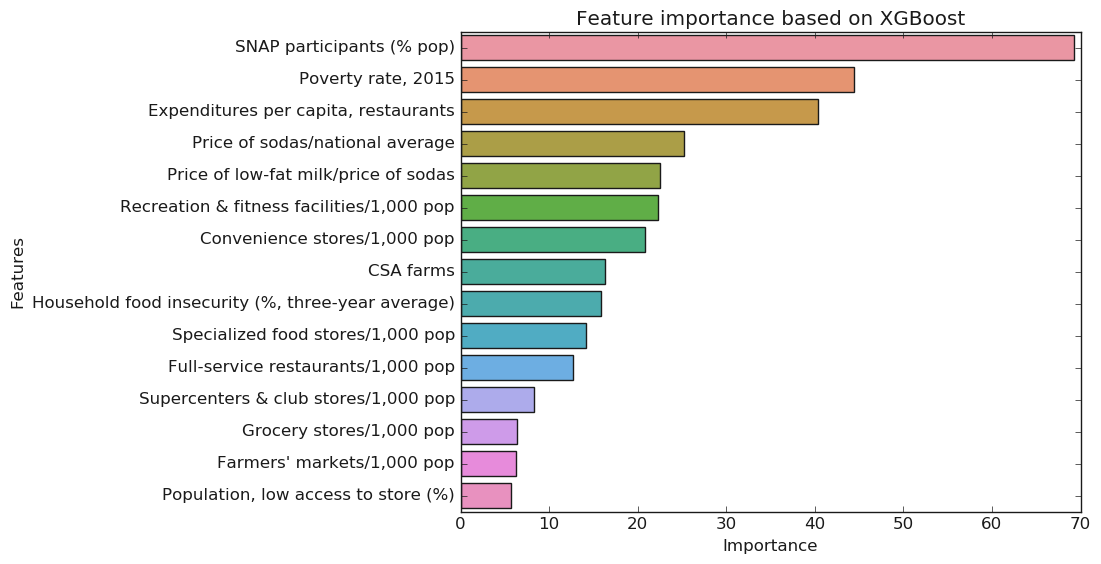
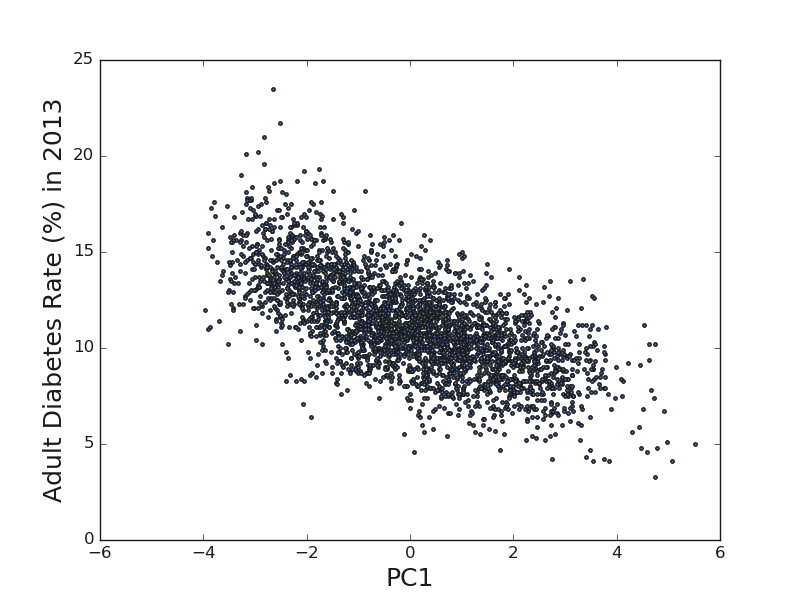


Figure 13. *Feature importance (‘gain’, i.e.,* the average gain across all splits the feature is used in*) based on the XGBoost model for all counties.*

## 3.3 The landscape of food environment across the United States

#### 1) PCA results

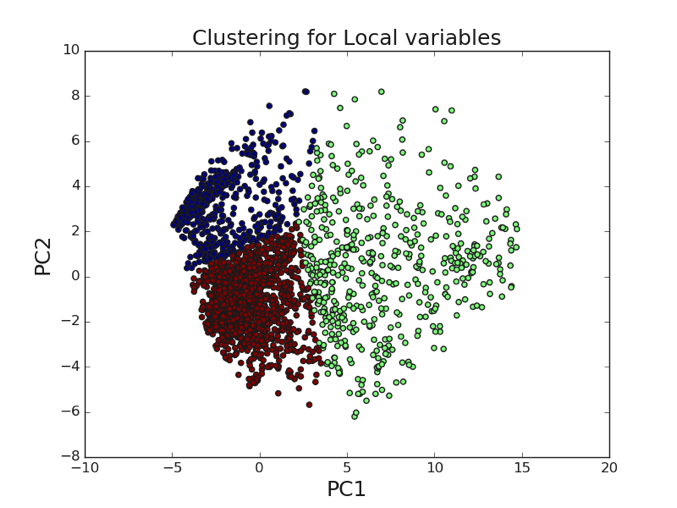
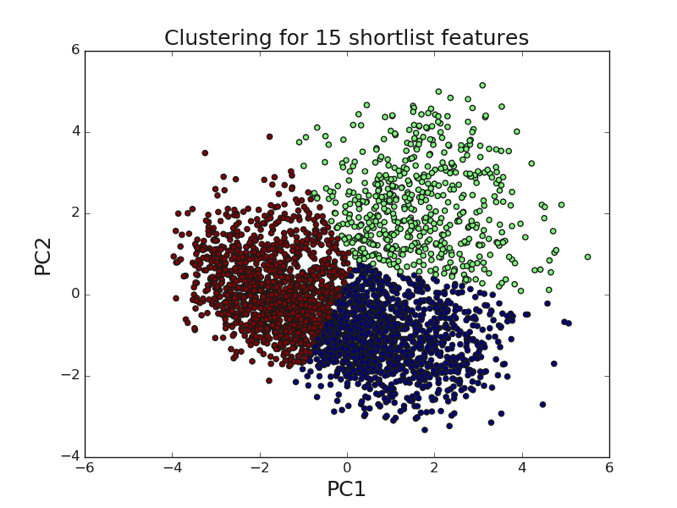
For the group of 15 shortlist features, the first four PCs explained 51% of variance. Higher PC1 values is correlated with lower diabetes rate (Fig. 11). For the ‘Local’ group, first three PCs explained 51% variance; for the ‘Restaurants’ and ‘Stores’ groups, the first two PCs explained 61% and 66% variances, respectively.

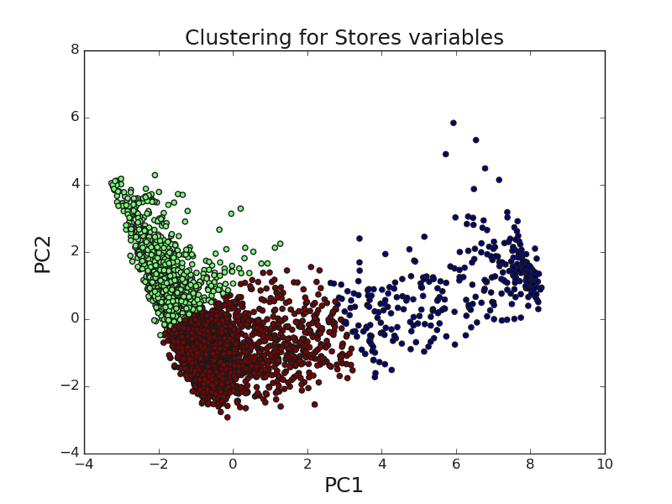
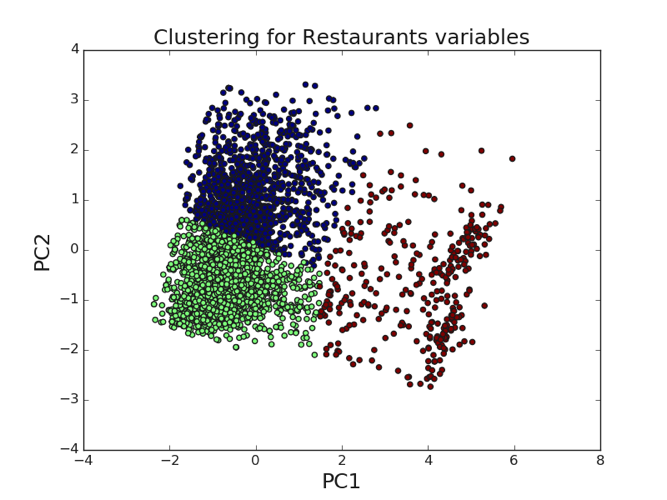


*Figure 14 Correlations between PC1 of 15 shortlist features and adult diabetes rate.*

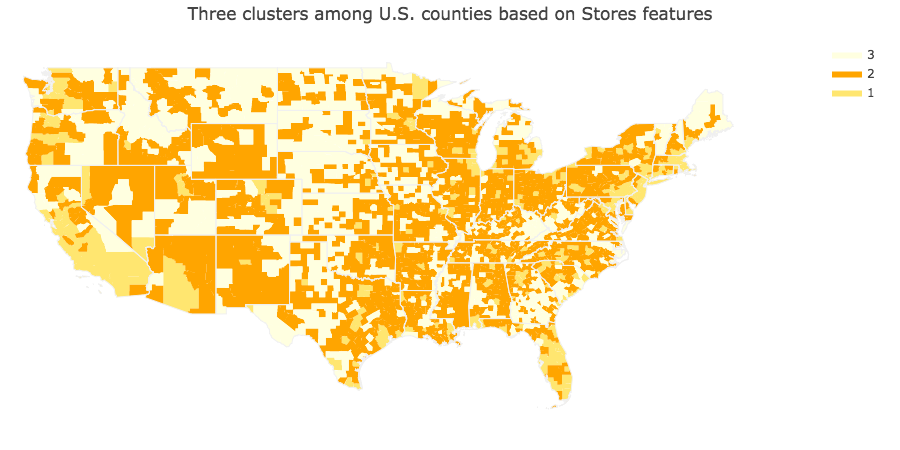
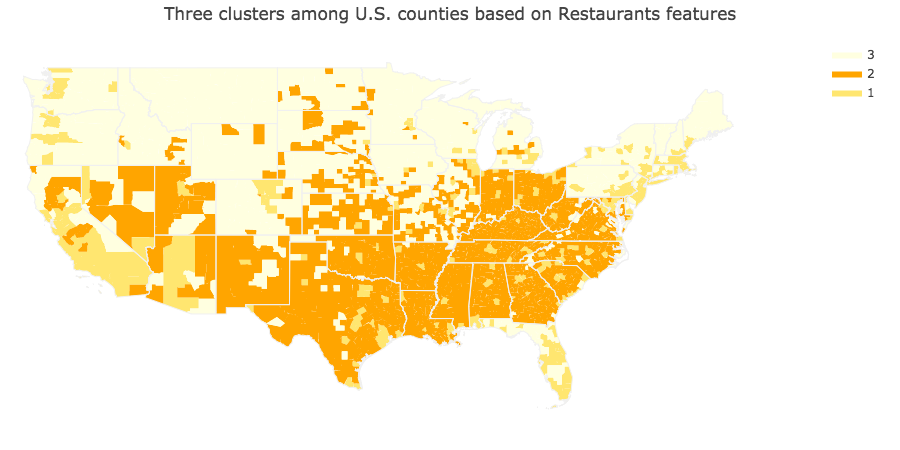
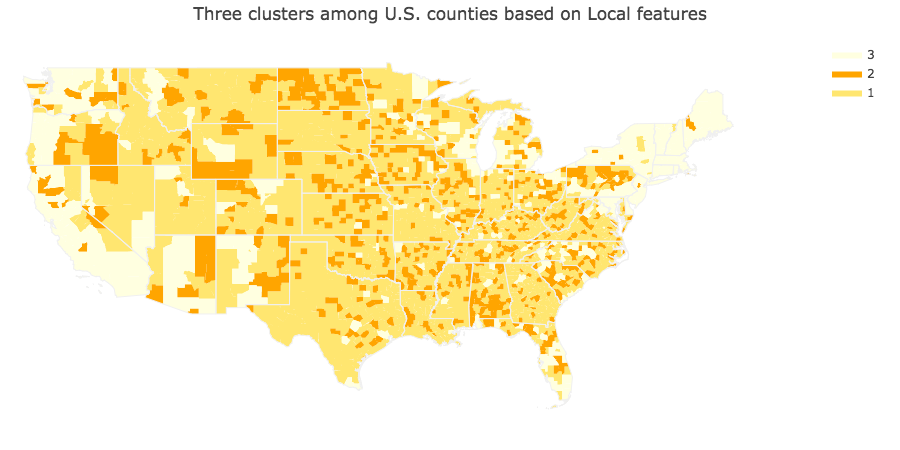
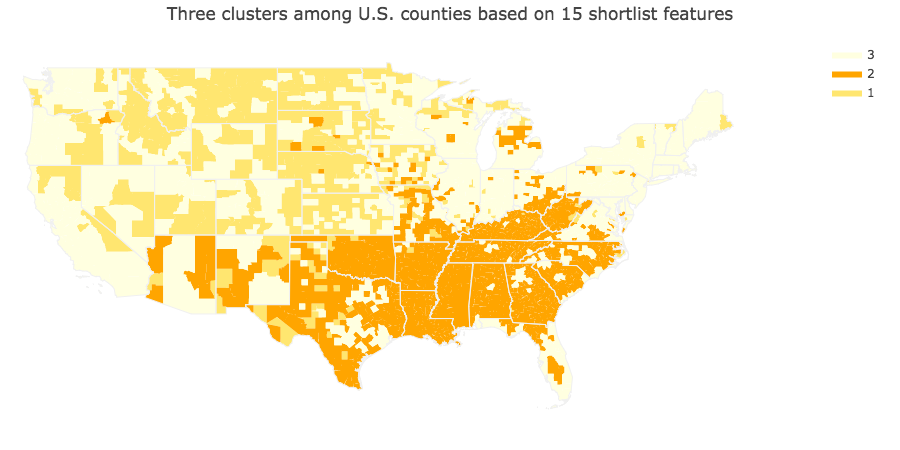
#### 2) K-means clustering based on first two PC components

Fig. 12 shows K-means clustering results with k=3, on data of the first two PC components from above PCA results.





*Figure 15. K-means clustering results with k=3, on data of the first two PC components from PCA on the 15 shortlist features, and groups ‘Local’, ‘Restaurants’, and ‘Stores’.*



*Figure 16. Clustering results plotted on U.S. map for each of the four clustering analyses.*

# 4. Discussion

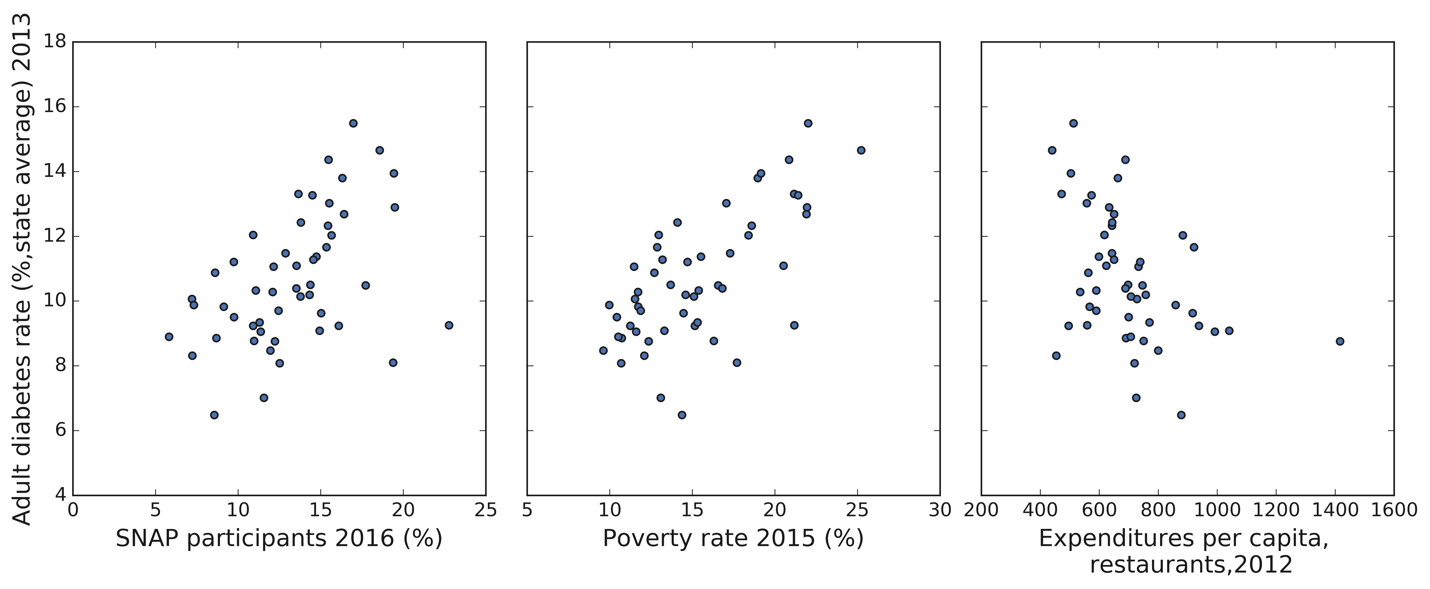
## 4.1 Important features predictive of diabetes rate

Among the supervised-learning models applied, random forest and XGBoost had the best accuracy (Table 1). XGBoost outperformed random forest but tend to be more overfit, despite of hyperparameter tuning. Therefore, we focused on random forest and further divided counties according to their urbanity (metro vs. non-metro counties) and poverty level (persistent poverty vs. non-persistent poverty counties).

*Table 1. Comparison of supervised-learning model performances*

|  |  |  |
| --- | --- | --- |
| **Model (All counties)** | **Training dataset R2** | **Testing dataset R2** |
| Linear regression | 0.51 | 0.56 |
| Decision tree | 0.58 | 0.55 |
| Random Forest | 0.68 | 0.63 |
| XGBoost | 0.83 | 0.69 |

Based on random forest model results, when considering all counties, features related to economic status generally played the most important role in predicting diabetes rate. The top ranked features included SNAP participants, poverty rate and expenditures at restaurants (Fig. 8, 17, Table 2). After dividing counties according to urbanity, SNAP participants remained the most important feature for both metro and non-metro groups. However, for metro counties the number of convenience stores/1000 population became the 2nd place (up from 4th), while for non-metro counties the price of low-fat milk/price of sodas became the 2nd place (up from 7th). These features are all positively correlated with diabetes rate (Fig. 3). Number of convenience stores in urban counties may indicate the popularity/availability of type of foods relating to diet choice, whereas a higher ratio between milk and sodas suggests more expensive milk price and/or cheaper soda price, thus may drive customers’ selection of drinks.



*Figure 17. Correlations between the top three features and adult diabetes rate (state averages).*

*Table 2. Rankings of top features by random forest model when considering all counties, urbanity or poverty level.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature name** | **All counties** | **Metro counties** | | **Non-Metro counties** | **Non-persistent poverty counties** | **Persistent poverty counties** |
| SNAP participants (% pop) | 1 | 1 | 1 | | 1 | 3 |
| Poverty rate | 2 | 4 | 3 | | 3 | 14 |
| Expenditures per capita, restaurants | 3 | 3 | 4 | | 2 | 1 |
| Convenience stores/1000 pop | 4 | 2 | 9 | | 5 | 8 |
| Price of low-fat milk/price of sodas | 7 | 9 | 2 | | 8 | 5 |
| Household food insecurity  (%, three-year average) | 6 | 11 | 5 | | 6 | 2 |

For non-persistent poverty counties, SNAP participants, Expenditures per capita at restaurants and poverty rate are also top features predictive of diabetes rate. In contrast, among persistent-poverty counties the expenditures at restaurants became the most predictive feature, followed by household food insecurity and SNAP participants. It is possible that when comparing counties of similar poverty level (i.e., non-persistent poverty or persistent poverty), habits/cultures related to dining-out spending may be associated with people’s fitness.

However, other features such as number and types of farms, number of different types of stores (except for convenience stores) and fitness facilities are playing a less significant role.

## 4.2 The landscape of food environment across the United States

Principal components from PCA on 15 our focal features revealed three major clusters of U.S. counties (Fig. 16). The counties located at central south and southeast forms one cluster, featuring higher poverty rate and diabetes rate; counties at west and east coast and the great lakes area forms a second cluster, which have lower poverty rate and diabetes rate. The rest of counties forms the third cluster.

To summarize, overall financial status is the major factor predictive of diabetes rates of counties. Counties at central south and southeast are high-risk areas for the prevalence of adult diabetes, likely strongly attributed to the lower economic status of these areas than other counties. Governments and health organizations should consider allocating more resources to these areas to improve the prevention and treatment of diabetes.

# References

Bower K.M., et al. 2014.The intersection of neighborhood racial segregation, poverty, and urbanicity and its impact on food store availability in the United States. *Preventive Medicine*, 58:33-39. ﻿doi: 10.1016/j.ypmed.2013.10.010

Decision tree learning, Wikipedia. <https://en.wikipedia.org/wiki/Decision_tree_learning>

Diabetes quick facts, Centers for Disease Control and Prevention. <https://www.cdc.gov/diabetes/basics/quick-facts.html>

Diabetes prevention. International Diabetes Federation. https://www.idf.org/aboutdiabetes/prevention.html

Economic Research Service (ERS), U.S. Department of Agriculture (USDA). Food Environment Atlas. <https://www.ers.usda.gov/data-products/food-environment-atlas/>

Franco, M et al. 2008. Neighborhood Characteristics and Availability of Healthy Foods in Baltimore. *American Journal of Preventive Medicine*, 35: 6, 561 – 567. doi: ﻿10.1016/j.amepre.2008.07.003

Haynes-Maslow, L., Leone, L.A. 2017. Examining the relationship between the food environment and adult diabetes prevalence by county economic and racial composition: an ecological study. BMC Public Health 17, 648. doi:10.1186/s12889-017-4658-0

Jones, E. et al., 2001. SciPy: Open source scientific tools for Python, Available at: <http://www.scipy.org/>

Pedregosa et al., 2011. Scikit-learn: Machine Learning in Python, JMLR 12:2825-2830.

World Health Organization, Diabetes fact sheets. <https://www.who.int/news-room/fact-sheets/detail/diabetes>

XGBoost Github webpage <https://github.com/dmlc/xgboost>

XGBoost, Wikipedia <https://en.wikipedia.org/wiki/XGBoost>