Food Policy 99 (2021) 101985

Contents lists available at ScienceDirect

Food Policy

journal homepage: www.elsevier.com/locate/foodpol

Predicting access to healthful food retailers with machine learning☆

Modhurima Dey Amin [a], Syed Badruddoza [a], Jill J. McCluskey [b,*]

[a] The Department of Agricultural and Applied Economics at Texas Tech University, United States
[b] The School of Economic Sciences at Washington State University, United States

ARTICLE INFO

JELCodes:
I14
Q18
L81

Keywords:
Food deserts
Food swamps
Machine learning

ABSTRACT

Many U.S. households lack access to healthful food and rely on inexpensive, processed food with low nutritional value. Surveying access to healthful food is costly and finding the factors that affect access remains convoluted owing to the multidimensional nature of socioeconomic variables. We utilize machine learning with census tract data to predict the modified Retail Food Environment Index (mRFEI), which refers to the percentage of healthful food retailers in a tract and agnostically extract the features of no access—corresponding to a "food desert" and low access—corresponding to a "food swamp." Our model detects food deserts and food swamps with a prediction accuracy of 72% out of the sample. We find that food deserts and food swamps are intrinsically different and require separate policy attention. Food deserts are lightly populated rural tracts with low ethnic diversity, whereas swamps are predominantly small, densely populated, urban tracts, with more non-white residents who lack vehicle access. Overall access to healthful food retailers is mainly explained by population density, presence of black population, property value, and income. We also show that our model can be used to obtain sensible predictions of access to healthful food retailers for any U.S. census tract.

1. Introduction

About 117 million Americans suffer from chronic, preventable diseases attributable to the lack of healthful diet and physical activity (U.S. Department of Health and Human Services and U.S. Department of Agriculture (USDA), 2015). As evidence mounts regarding the impacts of the lack of balanced and healthful diet on health outcomes, such as obesity, cancer, diabetes and cardiovascular diseases (e.g., Liu et al., 2000; Riboli and Norat, 2003), policymakers are interested in interventions that encourage healthful eating patterns. Taste, cultural affinity, variety, and affordability of healthful food alternatives affect healthful food choices (McEntee and Agyeman, 2010; Ver Ploeg and Wilde, 2018). Many studies find that consumers know the difference between healthful and unhealthful foods and would prefer healthful foods if the prices were the same (Drewnowski and Barratt-Fornell, 2004; Breyer and Voss-Andreae, 2013; Patel et al., 2017).

A major concern is whether the retail food environment provides a broad assortment of affordable foods that underpin a healthful diet (Walker et al., 2011, Walker et al., 2012; Thomsen et al., 2015). Studies have shown that the built environment characterized by the socioeconomic factors, including presence, quantity, and accessibility of healthful food retailers contributes to healthful food consumption (e.g., Laraia et al., 2004; Rose and Richards, 2004; Moore et al., 2008; Keenan and Rosendorf, 2011; Fitzpatrick et al., 2015; 2019). However, the criteria used to define "access to healthful food" as well as the variables that predict access to healthful food vary widely (Walker et al., 2010, Alviola et al., 2013; Thomsen et al., 2015; Hager et al., 2016).

The current study applies machine learning to identify predictors of access to healthful food. The response variable, the modified Retail Food Environment Index (mRFEI), refers to the percentage of healthful food retailers in a U.S. Census tract. We utilize the Center for Disease Protection and Control (CDC) definition of mRFEI, which is a relative measure, to extract the features of areas with (1) no access to healthful food, and (2) low access to healthful food, rather than the USDA's absolute, multidimensional definition of food deserts e.g., distance, income, urbanization, and vehicle access (USDA, 2019). These two concepts correspond with the concepts of food deserts and food swamps, respectively (CDC, 2011a, 2011b; Mason et al., 2013; Luan et al., 2015; Berkowitz et al., 2018).

We exploit the mRFEI to categorize a census tract as a food desert if no healthful food retailer exists and food swamp if healthful retailers are disproportionally outnumbered by less healthful ones. We use U.S.

☆ The authors thank the session participants at the 2020 American Econ. Assoc. Annual Meeting and two anonymous reviewers for their helpful comments.
* Corresponding author.
E-mail addresses: modhurima.amin@ttu.edu (M.D. Amin), s.badruddoza@ttu.edu (S. Badruddoza), jjmccluskey@wsu.edu (J.J. McCluskey).

https://doi.org/10.1016/j.foodpol.2020.101985
Received 6 November 2019; Received in revised form 30 September 2020; Accepted 4 October 2020
Available online 16 October 2020

census tract information and search over a large set of variables and hyperparameters. This data-driven approach identifies demographic variables such as race, population density, urbanization, and vehicle availability predicting the access with high accuracy. Our model detects food deserts and swamps about 72% of the time in the independent test data. Given the unpredictable nature of cross-sectional and socioeconomic data, the variables offer a potent way to measure access to healthful foods for any U.S location. Results also suggest that food deserts and food swamps are intrinsically different, and they require separate policy attention.

1.1. Factors affecting access to healthful food

While there is no doubt about the importance of a healthful diet, current understanding of the access to healthful food is limited. For example, it is unclear whether the lack of access is a cause or consequence of an unhealthful diet. In socioeconomically deprived areas, healthful food choices are limited and therefore prices are higher (MacDonald and Nelson, 1991; Crockett et al., 1992; Sooman et al., 1993; Horowitz et al., 2004; Baker et al., 2006). Nutritious foods are often offered at premium prices, which low-income households cannot afford (Ver Ploeg et al., 2009). This creates a disincentive for healthful food retailers to locate in low-income neighborhoods. On the other hand, the lack of healthful food retailers in the area limits the food choice for people living there.

Some studies find that introduction of a supermarket in a food desert improves the economic well-being and health of the residents (Feather, 2003; Olsho et al., 2016; Richardson et al., 2017), while some others do not (Allcott et al., 2019; Sharpe et al., 2020). In fact, Allcott et al. (2019) observe that offering low-income households access to supermarkets had minor effects on diet and about 90% of the nutritional inequality was driven by differences in demand. The authors propose a subsidy to induce low-income households to purchase healthful groceries. Thus, increasing the demand for healthful food in deprived areas may be at least as important as the availability of healthful retailers or increasing their supply.

Extant studies that consider the association between socioeconomic features and the level of access find income inequality and racial segregation important for predicting urban deserts and vehicle access for rural areas (e.g., Alwitt and Donley, 1997; Dutko et al., 2012). Many studies generate mixed results due to local idiosyncracies in their samples (e.g., Mooney, 1990; Travers et al., 1997; Chung and Myers, 1999; Block and Kouba, 2006). One line of research suggests that residents of socioeconomically disadvantaged neighborhoods, especially non-white, may lack access to healthful food retailers (Alwitt and Donley, 1997; Chung and Myers, 1999; Morland et al., 2002; Zenk et al., 2006; Baker et al., 2006; Block and Kouba, 2006; Moore and Roux, 2006; Powell et al., 2007; Laska et al., 2010). In particular, some studies find low-income neighborhoods with higher percentages of black population have fewer supermarkets than others (Berg and Murdoch, 2008; Powell et al., 2007; Block et al., 2008; Larson et al., 2009). Other studies do not find an association or obtain the opposite result (Alwitt and Donley, 1997; Moore and Roux, 2006; Opfer, 2010; Sharkey and Horel, 2008).

Heterogeneous findings may occur due to methodological factors (e.g., Carroll and Samek, 2018). Access to healthful retailers is typically measured by distance to the store or the density of the stores in an area. If residents, who lack access to nearby healthful foods, require traveling to supermarkets outside their neighborhoods, the cost may be greater than the marginal utility of healthful food, given the financial and physical constraints to mobility (LeDoux and Vojnovic, 2013). The majority of studies approach the problem of access with categorizations of areas and examine administrative and demographic features of predefined geographic units. Such features include the presence of food stores, worksites, schools, and location identifiers, such as zip code, county, or state (Apparicio et al., 2007; McKinnon et al., 2009; Sage et al., 2013).

Table 1
Definitions of the predicted variables.

Variable	Description	Mean (SD)	Obs.
mRFEI	The modified Retail Food Environment Index (%) (Median = 9.09)	11.30 (12.00)	50,212
Desert	1 if tract has mRFEI = 0, 0 otherwise	0.23 (0.42)	11,730
Swamp	1 if 0 < mRFEI≤9.09 0 otherwise	0.24 (0.43)	12,266
Healthful	1 if mRFEI > 9.09 0 otherwise	0.53 (0.50)	26,216

Source: mRFEI is calculuated by Eq. (1) (CDC 2011b). Desert, swamp, and healthful dummies are created according to the CDC's definition.
Note: Standard deviations (SD) are in the parentheses.

The features can be standardized relative to population and the distance from the city centroid (e.g., Larsen and Gilliland, 2008; Ball et al., 2009; Chen and Clark, 2013). Ball et al. (2009) point out that coarse administrative measures of accessibility, e.g., census tract or zip code level observations, instead of actual distance from residents' home

Table 2
List of predictors with summary statistics.

Variable	Description	Mean (SD)
HH Income	Median household income in the past 12 months (thousand 2010 inflation-adjusted dollars)	54.75 (26.99)
Poverty rate	Tract poverty rate (measured with heterogeneous family-level threshold)	17.13 (12.73)
HH with SNAP	Tract housing units receiving Supplemental Nutrition Assistance Program (SNAP) benefits	219.17 (190.17)
Inequality	Gini Index for the tract (increases in inequality)	0.41 (0.06)
Unemployment	Unemployment rate among population 16 years and over	8.66 (5.61)
Below high school	% of population did not graduate high school	18.58 (14.52)
College no degree	% of population who went to college but did not complete	20.30 (6.18)
Some college	% of population with some college or associate's degree	38.62 (17.05)
Bachelor's or more	% of population with Bachelor's degree or higher	26.63 (18.31)
Property value	Median value of housing unit in thousand dollars	242.37 (192.99)
Public transport	% population using public transportation (excluding taxicab)	2.31 (4.86)
No vehicle	Number of housing units without a vehicle	160.37 (246.28)
Land area	Land area of the tract in square miles	45.83 (622.92)
Population density	Population density per square mile of land area (in thousand)	5.68 (11.78)
Black	African-American population (%)	13.50 (22.55)
Hispanic	Hispanic or Latino population (%)	10.95 (12.15)
Asian	Asian population (%)	3.83 (7.52)
Native	American Indian and Alaska Native population (%)	0.75 (3.37)
Pacific islander	Native Hawaiian and other Pacific Islander population (%)	0.13 (0.80)
Rural population	% of population living in rural part of the tract (measured by the distance from population-weighted centroid of a census tract)	17.47 (33.23)
Observations	Census tracts (2010)	50,212

Source: U.S. Census Bureau (2019a); (2019b;).
Note: Standard deviations (SD) are in the parentheses. Predictors are not standardized in the descriptive statistics but standardized in the predictive and feature extraction analysis to reduce the scale effect.

Table 3
Variable means and their differences by access to healthful food retailers.

Variable	Desert	Swamp	Healthful	Mean difference *t*-test with unequal variances		
				Desert -Healthful	Swamp -Healthful	Swamp -Desert
mRFEI	0.00	6.07	18.81	−18.81	−12.75	6.07
	(0.00)	(1.79)	(12.17)	[−250.39]	[−165.88]	[375.41]
HH Income	54.07	50.67	56.95	−2.89	−6.28	−3.40
	(26.83)	(25.60)	(27.46)	[−9.61]	[−21.92]	[−10.04]
Poverty rate	16.63	20.58	15.74	0.89	4.84	3.95
	(12.79)	(14.11)	(11.68)	[6.40]	[33.03]	[22.74]
HH with SNAP	192.16	265.49	209.59	−17.44	55.90	73.33
	(162.76)	(221.41)	(181.82)	[−9.29]	[24.38]	[29.32]
Inequality	0.41	0.42	0.41	<0.01	0.01	0.02
	(0.06)	(0.07)	(0.06)	[4.18]	[19.16]	[19.85]
Unemployment	8.75	9.72	8.13	0.62	1.59	0.97
	(5.95)	(6.25)	(5.04)	[9.89]	[24.69]	[12.27]
Below high school	19.37	19.09	17.99	1.38	1.10	−0.28
	(14.70)	(15.23)	(14.07)	[8.55]	[6.75]	[−1.44]
College no degree	20.83	19.62	20.38	0.45	−0.77	−1.21
	(6.00)	(6.44)	(6.10)	[6.64]	[−11.08]	[−15.10]
Some college	37.90	38.29	39.10	−1.20	−0.81	0.39
	(17.23)	(17.59)	(16.69)	[−6.33]	[−4.28]	[1.73]
Bachelor or up	23.73	26.94	27.79	−4.05	−0.85	3.20
	(16.71)	(18.92)	(18.55)	[−21.08]	[−4.12]	[13.92]
Property value	194.45	263.48	253.93	−59.49	9.55	69.04
	(160.18)	(195.89)	(201.33)	[−30.79]	[4.42]	[29.94]
Public transport	1.23	4.07	1.97	−0.74	2.11	2.84
	(2.54)	(6.45)	(4.55)	[−20.17]	[32.55]	[45.29]
No vehicle	101.26	254.14	142.95	−41.68	111.20	152.88
	(113.42)	(353.85)	(214.14)	[−24.71]	[32.16]	[45.47]
Land area	104.18	5.59	38.54	65.65	–32.95	−98.60
	(1,191.92)	(69.86)	(320.89)	[5.87]	[−15.84]	[−8.94]
Population density	2.62	10.32	4.87	−2.25	5.44	7.69
	(4.03)	(17.11)	(10.34)	[−30.48]	[32.56]	[48.42]
Black	14.02	18.87	10.76	3.26	8.11	4.85
	(24.23)	(25.82)	(19.45)	[12.85]	[30.94]	[15.02]
Hispanic	7.67	14.81	10.62	−2.95	4.19	7.14
	(9.97)	(13.31)	(11.98)	[−24.96]	[29.69]	[47.15]
Asian	2.32	4.69	4.10	−1.78	0.59	2.37
	(5.28)	(7.79)	(8.13)	[−25.41]	[6.84]	[27.70]
Native	0.99	0.61	0.71	0.28	−0.11	−0.39
	(5.21)	(1.30)	(2.96)	[5.45]	[−4.86]	[−7.79]
Pacific islander	0.10	0.13	0.15	−0.05	−0.02	0.04
	(0.71)	(0.86)	(0.80)	[−6.16]	[−1.66]	[3.49]
Rural population	34.60	2.76	16.68	17.92	−13.92	−31.84
	(42.42)	(12.07)	(31.75)	[40.90]	[−62.05]	[−78.30]
Observations (tracts)	11,730	12,266	26,216	Total=	50,212	

Note: The last three columns show the difference in variable means. Standard deviations appear in parentheses, t-statistic for the test on the equality of means are in brackets. Detail variable descriptions are in Table 2. Predictors are not standardized in the descriptive statistics but standardized in the predictive and feature extraction analysis to reduce the scale effect.

to retailers or not considering food range and prices within the stores, may underestimate the association between accessibility and deprivation. Researchers consider the areas distant from the centroid underserved (Chen and Yang, 2014).

However, the distance-based approach is multidimensional. In general, the larger the number of dimensions, the more difficult the application. For example, Dutko et al. (2012) use a three-dimensional measure for food desert: low income, low access, one-kilometer radius; while Rhone et al. (2017) use a combination of distance and income by census-tract type to generate a measure of access to healthful food retailers. The distance is measured either by Euclidian distance or road network distance (Smoyer-Tomic et al., 2006; Larsen and Gilliland, 2008; Wang et al., 2016). The standard definition of food desert includes distance to the nearest grocery store, percent of residents in poverty, type of the location (rural versus urban), and access to a vehicle (USDA, 2019; Goodman et al., 2020). Nevertheless, several studies note that low-income consumers shop outside food deserts when they can (e.g., Ver Ploeg, 2010; Allcott et al., 2019).

Therefore, the problem may not be the lack of access in the area, rather, the problem may be that, less healthful food is cheaper and more convenient to consume (Rose et al., 2009; Hager et al., 2016). Fewer grocery stores in the neighborhoods may make local fast food or carryout foods relatively more affordable. These neighborhoods with an excess of less healthful foods in relation to healthful foods are called "food swamps," and require a density-based approach. A density measure takes store competition into account, which is difficult to accommodate in a distance measure (e.g., Apparicio et al., 2007; Sparks et al., 2009). Household surveys overcome this problem, but surveys overlook the effects of mobility on procuring food, do not generate universal conclusions, and are costly to implement (McKinnon et al., 2009; Kestens et al., 2010). Designing a nationally representative sample of healthful food access is quite challenging.

The current study adds to the literature in two ways. First, the use of the modified Retail Food Environment Index (mRFEI) as a measure of healthful food retailers makes the analysis holistic and interpretable. The mRFEI places the overall retail environment on a continuous spectrum and offers a nationwide measure of food deserts and food swamps so that locations are easily comparable. Second, our data-driven

Table 4
Model performance statistics (test data).

Model	Statistic	RF	XGB	LASSO
Target = Desert, Benchmark = Healthful	Accuracy	0.714	0.715	0.705
	95% CI	(0.705, 0.722)	(0.706, 0.723)	(0.697, 0.714)
(Swamps not included in this sample, N = 11,383)	Sensitivity	0.916	0.916	0.924
	Specificity	0.262	0.265	0.218
	Kappa	0.210	0.213	0.170
Target = Swamp, Benchmark = Healthful	Accuracy	0.728	0.727	0.704
	95% CI	(0.720, 0.736)	(0.719, 0.735)	(0.696, 0.712)
(Deserts not included in this sample, N = 11,543)	Sensitivity	0.924	0.925	0.927
	Specificity	0.309	0.303	0.228
	Kappa	0.272	0.266	0.185
Target = Desert, Benchmark = Swamp	Accuracy	0.738	0.739	0.712
	95% CI	(0.728, 0.748)	(0.729, 0.749)	(0.702, 0.723)
(Healthful not included in this sample, N = 7,198)	Sensitivity	0.827	0.824	0.888
	Specificity	0.645	0.650	0.529
	Kappa	0.473	0.475	0.420
mRFEI (full test sample N = 15,061)	NRMSE	97.0	97.1	98.0

Note: N = respective observations, CI = confidence intervals, Sensitivity = True Positive Rate, Specificity = True Negative Rate, Kappa is the rate of agreement between actual and predicted classes beyond what is expected by chance. NRMSE = Normalized Root-Mean-Squared Error. All accuracy estimates are significant at p-values < 0.01. The total sample is (11,730 Desert + 12,266 Swamp + 26,216 Healthful) 50,212 tracts, all of which are used in the prediction of mRFEI. In each case, the sample is randomly split into training (70%) and test (30%) data. Training implies choosing the best hyperparameters from a grid of hyperparameters with 10-fold cross-validation. Test refers to the independent sample held out during training.

nonparametric approach does not impose prior assumptions on the outcome. The model evaluates many predictors to isolate the ones that are most influential in predicting the presence of healthful food retailers. We use data from almost all U.S. census tracts. Therefore, our study imposes minimum assumptions on sampling and modeling, which are two recurring challenges faced by previous studies. We also show that our model can be used to obtain a sensible prediction of access to healthful retailers for any U.S. census tract.

1.2. Measuring access to healthful food

We utilize the mRFEI as a measure of healthful food retailers within census tracts. The CDC originally featured the index in Children's Food Environment State Indicator Report (CDC, 2011a). The mRFEI measures the percentage of healthful food retailers within census tracts, and is calculated using the following formula,

$$mRFEI = 100 \times \frac{number\ of\ healthful\ retailers\ in\ tract}{number\ of\ healthful\ retailers\ in\ tract + number\ of\ unhealthful\ retailers\ in\ tract} \quad (1)$$

where the classification of retailers follows the North American Industry Classification Codes (NAICS), which depends on typical food offerings in specific types of retail stores. Healthful food retailers include supermarkets and other grocery (except convenience) stores (NAICS 445110), warehouse clubs (NAICS 452910), and fruit and vegetable markets (NAICS 445230) within census tracts or half a mile from the tract boundary. Less healthful food retailers refer to fast food restaurants (NAICS 722211), small grocery stores, and convenience stores (NAICS 445120) within census tracts or half a mile from the tract boundary. Small groceries or convenience stores are characterized by stores that primarily engage in retailing a limited line of goods that generally includes milk, bread, soda, and snacks (NAICS, 2007; CDC, 2009, 2011b; Grimm et al., 2013).

The definition above does not represent consumers' choice or the availability of healthful food within the store, which may be a more appropriate measure (Gustafson et al., 2012; Stern, et al., 2016). However, a large number of studies in the literature consider larger retailers to be healthful because they offer a wider variety of foods at a lower price, hence consumers are more likely to have access to affordable healthful alternatives (e.g., Ver Ploeg et al., 2009; Volpe et al., 2013; Courtemanche et al., 2019).

An mRFEI score of zero indicates that there are no healthful food retailers within the census tract, and a low score indicates that census tracts contain relatively many convenience and fast-food restaurants compared to the number of healthful food retailers. The CDC (2011b) specifies that a score of zero generally corresponds with the concept of food deserts, and nonzero but low scores correspond with the concept of food swamp. Although this classification differs from the USDA's multidimensional definition based on income, distance, location, and vehicle access (USDA, 2009), it is commonly used in the literature undertaking density-based approaches (e.g., Gustafson et al., 2012; Luan et al., 2015; Berkowitz et al., 2018; Allcott et al., 2019; Testa and Jackson, 2019; Goodman et al., 2020; Yang et al., 2020).

The CDC's classification suits our objective because we want to extract the demographic features of areas with (1) no access and (2) low access to healthful food retailers. In addition, since there is no multidimensional definition of food swamps, the USDA's multidimensional definition for food deserts is not useful for differentiating the factors that predict food deserts versus food swamps. As discussed earlier, if an unhealthful diet is an outcome of differences in demand rather than supply, using the CDC's definition will generate the demographic features that underscore the differences between the two tracts.

Food swamps deserve separate recognition because the consumers who reside in food swamps are less likely to choose healthful foods, *ceteris paribus*, compared to those who reside in non-desert, non-swamp tracts just because a relatively smaller selection of healthful choices is available (e.g., Hager et al., 2016). It is important to understand the distinction between food deserts and food swamps. Food desert tracts contain limited-option food retailers such as convenience stores and fast-food restaurants that mainly offer packaged, processed, and energy-dense foods (Drewnowski and Specter, 2004; Maguire et al., 2015). However, *food swamps* are tracts where both healthful and less healthful retailers are available, but the availability of foods high in sugar, salt, or fat dominate the availability of healthful alternatives such as fruits, vegetables, whole grains, dairy or dairy alternatives, and other foods that make up the full range of a healthful diet (Rose et al., 2009; CDC, 2011b). In fact, Cooksey-Stowers et al. (2017) find evidence that food swamps are a distinct and separate phenomenon to food deserts as they predict U.S. adult obesity rates better than food deserts. The authors define food swamps as a ratio of healthful to unhealthful retailers. Previous studies have defined food swamps as areas with healthful food outlets from over 0% to 10% of total retailers—as households are more likely to purchase healthier foods if the healthful outlets are over 10% (e.g., Mason et al., 2013; Luan et al., 2015).

Thus, tracts can be categorized into three types using mRFEI, resulting in three mutually exclusive and exhaustive groups:

$$mRFEI = \begin{cases} 0 \Rightarrow \text{The tract is a food desert} \\ \in (0, \text{median(mRFEI)}] \Rightarrow \text{The tract is a food swamp} \\ \in (\text{median(mRFEI)}, 100] \Rightarrow \text{The tract has good access to healthful food.} \end{cases} \quad (2)$$

The median of the mRFEI is 9.09, which is slightly below the mean mRFEI of 11.30. We chose median as a cut-off point instead of mean to reduce the influence of extreme observations, and control the class imbalance problem. In the median U.S. tract, about 9% of the food retailers are healthful, whereas food deserts have none, and food swamps have between zero and 9% healthful retailers. The above definition is consistent with the literature (e.g., CDC, 2011b; Mason et al., 2013; Luan et al., 2015) and generates 11,730 food deserts, 12,266 food swamps, and 26,216 tracts that have good access to healthful food (see Table 1). Our objective is to predict food deserts, food swamps, and the mRFEI using socioeconomic variables at the census tract level. The selection of predictors is accomplished with machine learning models, and hence they will be presented after the methodology section.

We specify four models based on the nature of the response variables: (1) target = desert, benchmark = healthful, swamp tracts are not included in the sample, (2) target = swamp, benchmark = healthful, desert tracts are not included in the sample, (3) target = desert, benchmark = swamp, healthful tracts are not included in the sample; and (4) mRFEI. The first two models help extract features of deserts and swamps, and the third one unravels the contrasts between deserts and swamps. Finally, the continuous response variable mRFEI is used to predict a continuous level of access to healthful retailers in a tract. We do not include a multiclass model (all three categories together as response) because it would complicate the interpretation of characteristics. For example, high relative importance of a demographic variable in predicting all three responses together is not informative enough to understand what type of tract is associated with that variable. Moreover, the continuous response mRFEI is more informative than a multiclass response because it has more variation. In short, the rationale for conducting separate paired evaluations is to understand their similarities and differences with minimum noise.

2. Methodology

This section introduces machine learning (ML) models used for

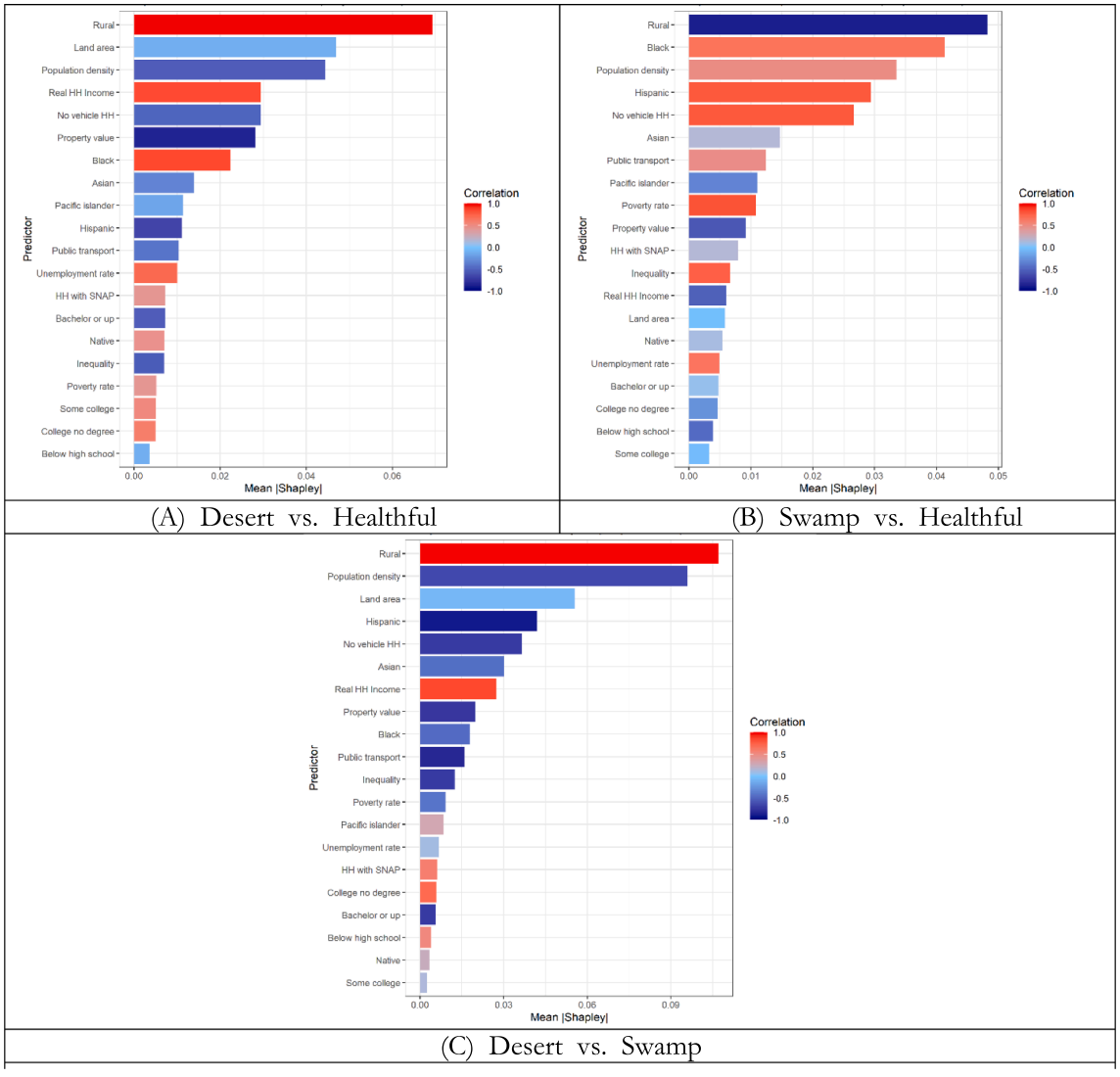

Fig. 1. Top predictors of food deserts and swamps. The correlation between the predictor and SHAP values indicate the direction of association (red for positive and blue for negative), whereas the mean SHAP values show the predictor's marginal contribution in prediction. Predictor descriptions are provided in Table 2. We have three models that use binary response: (A) Target = desert, benchmark = healthful, swamp tracts are not included in the sample, (B) target = swamp, benchmark = healthful, desert tracts are not included in the sample, and (C) target = desert, benchmark = swamp, healthful tracts are not included in the sample.

feature extraction and prediction. ML has advantages over traditional econometric models in certain situations. We argue it is useful in our context. A large number of factors contribute to the retail food environment that are likely to be endogenous. A parametric approach inherits the researcher's prior expectations, thus imposes a structure on data and may lead to functional form misspecification. Such an approach does not accommodate high-dimensional data and hence may suffer from omitted variable bias. Further, it may not be used for weeding out correlated predictors due to computational complexity (Breiman, 2001). In contrast, ML models do not presume a formal structure and directly learn the associations between variables from the data. Although some ML algorithms use hyperparameters, such as the number of trees and learning rates, these hyperparameters are optimally chosen over a broad grid, hence they are also driven by the data.

ML often creates uninterpretable results and has limited use for causal inference (Athey et al., 2019), but it fits our purpose of predicting the retail food environment, particularly because the data are large and predictors are convoluted. Given the success of ML in prediction assignments for computing and business, we expect the method to generate practical directions for the U.S. food policy. To the best of our knowledge, researchers have not explored ML models for food access predictions. The study applies random forests (RF)—a tree-based ML model—for predicting access to healthful food retailers, and complements it with eXtreme Gradient Boost (XGB) and Least Absolute Shrinkage and Selection Operator (LASSO) for robustness. We selected these ML models over others because they scale with the volume of information without damaging statistical efficiency, are more interpretable than Neural Networks, versatile than support vector machines, and typically demonstrate good predictive accuracy as they are robust to outliers due to repeated sampling.

2.1. Random forests (RF)

The RF algorithm constructs a multitude of decision trees at training time with bootstrapping and randomly chosen predictors, and then aggregates them for final prediction. Tree models without randomization may produce weak predictions if the trees are correlated (Hastie et al., 2009, p.587). RF grows each tree to a resampled part of the data, which makes the trees different and de-correlates them. RF offers better comprehension than neural-network-type algorithms for its tree-like structure, hence it is gaining use in the social sciences (e.g. Davis and Heller, 2017).

Following Biau and Scornet (2016), we assume there exists a vector of observed predictors $\mathrm{x} \in \chi \subset \mathbb{R}^{\mathrm{p}}$ that can be used to predict an observed response $y \in \mathbb{R}$ by estimating a function $m(\mathrm{x}) = \mathbb{E}[Y|\mathrm{x} = \mathrm{X}]$ using a training sample $\mathscr{D}_n = ((\mathrm{x}_1, y_1), \cdots, (\mathrm{x}_n, y_n))$. If the response is continuous, the function takes the form of a regression model. Under supervised classification,

$$m(\mathrm{x}) = \begin{cases} 1 & \text{if } \mathbb{P}[y = 1|\mathrm{x} = \mathrm{X}] > \mathbb{P}[y = 0|\mathrm{x} = \mathrm{X}] \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The RF algorithm grows a collection of $t = 1, \cdots, T$ randomized trees such that each tree $m_n(\mathrm{x}; \Theta_t, \mathscr{D}_n)$ contains the independent random parameter Θ_t that is used to resample the training set before growing the tree, and to select the successive directions for splitting. Using observations $i = 1, \cdots, n$ the t^{th} tree estimate is defined by,

$$m_n(\mathrm{x}; \Theta_t, \mathscr{D}_n) = \sum_{i \in \mathscr{D}_n^*(\Theta_t)} \frac{\mathbb{I}(\mathrm{x}_i \in A_n(\mathrm{x}; \Theta_t, \mathscr{D}_n))Y_i}{N_n(\mathrm{x}; \Theta_t, \mathscr{D}_n)} \tag{4}$$

where $\mathscr{D}_n^*(\Theta_t)$ is the sub-sample, $A_n(\mathrm{x}; \Theta_t, \mathscr{D}_n)$ contains predictors x, $\mathbb{I}$ is an indicator function, and $N_n(\mathrm{x}; \Theta_t, \mathscr{D}_n)$ is the length of set $A_n(\mathrm{x}; \Theta_t, \mathscr{D}_n)$. The aggregation of T trees by average gives the RF estimate,

$$m_T(\mathrm{x}; \Theta_1, \cdots, \Theta_T, \mathscr{D}_n) = \frac{1}{T}\sum_{t=1}^{T} m_n(\mathrm{x}; \Theta_t, \mathscr{D}_n) \tag{5}$$

The aggregation is conducted via a majority vote in the classification context,

$$m_T(\mathrm{x}) = \begin{cases} 1 & \text{if } \frac{1}{\mathrm{T}}\sum_{t=1}^{T} m_n(\mathrm{x}; \Theta_t, \mathscr{D}_n) > 0.5 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

A typical RF algorithm has the following steps:

(a) Randomly sample from the training dataset with replacement.
(b) Fully grow a decision tree for each sample drawn in (step 1), i.e., no further splits are possible.
(c) At each node of the tree, select the best split among a randomly selected q of p predictors such that $q \subset p$, where the number of predictors in the subset q can be defined by the user.
(d) Repeat until T trees are grown.

Once the trees are grown, RF reports the overall pattern of the trees either by averaging (for regression) or by majority voting (for classification).

2.2. eXtreme Gradient Boost (XGB)

The RF steps discussed above involve bootstrap sampling, the random selection of some predictors in the split of each node, growing full depth decision tree, repeating the process until the desired number of trees are created, and calculating error on the samples which were not selected during bootstrap sampling. The bootstrap aggregation of trees is called bagging, and the error calculation part is known as out-of-bag (OOB) error estimation. Conversely, a gradient booster (GB) reduces the error and improves the prediction of RF by growing additional trees with the residuals as an outcome. For each subsample, the GB grows an extra tree to the prediction model, which shrinks the errors of estimation. A GB algorithm iterates the process of growing trees, re-estimating the predictor weights many times on the updated residuals, such that poorly predicted observations get increasing weight in each repetition (hence the term, boosting). The adjustment of weights is conducted via a gradient descent algorithm, which is using the partial derivatives of the loss function to find the direction for updating the model parameters. Analogous to RF, the final estimate is then a vote or an average across the collection of individual estimates (Varian, 2014; Athey and Imbens, 2019).

The eXtreme Gradient Boosting (XGB) is an advanced form of GB that utilizes a rigorous penalty structure (e.g., L1/L2 regularization) and second partial derivatives to minimize the loss function efficiently (Chen and Guestrin, 2016). The regularization term depends on the tree leaves and the weights on the predictors, and helps prune complex and large trees to avoid overfitting. Thus, XGB requires more parameters to be tuned, but its training is typically faster than RF or GB and can be parallelized or distributed across clusters.

2.3. Least absolute shrinkage and selection Operator (LASSO)

The LASSO regularizes the regression coefficients by adding a penalty term to the loss function. The LASSO estimates for a model $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$ are,

$$\widehat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \gamma \sum_{j=1}^{p} |\beta_j| \right\} \tag{7.1}$$

$$\text{subject to} \sum_{\mathrm{j}=1}^{\mathrm{p}} |\beta_j| \leq \Omega$$

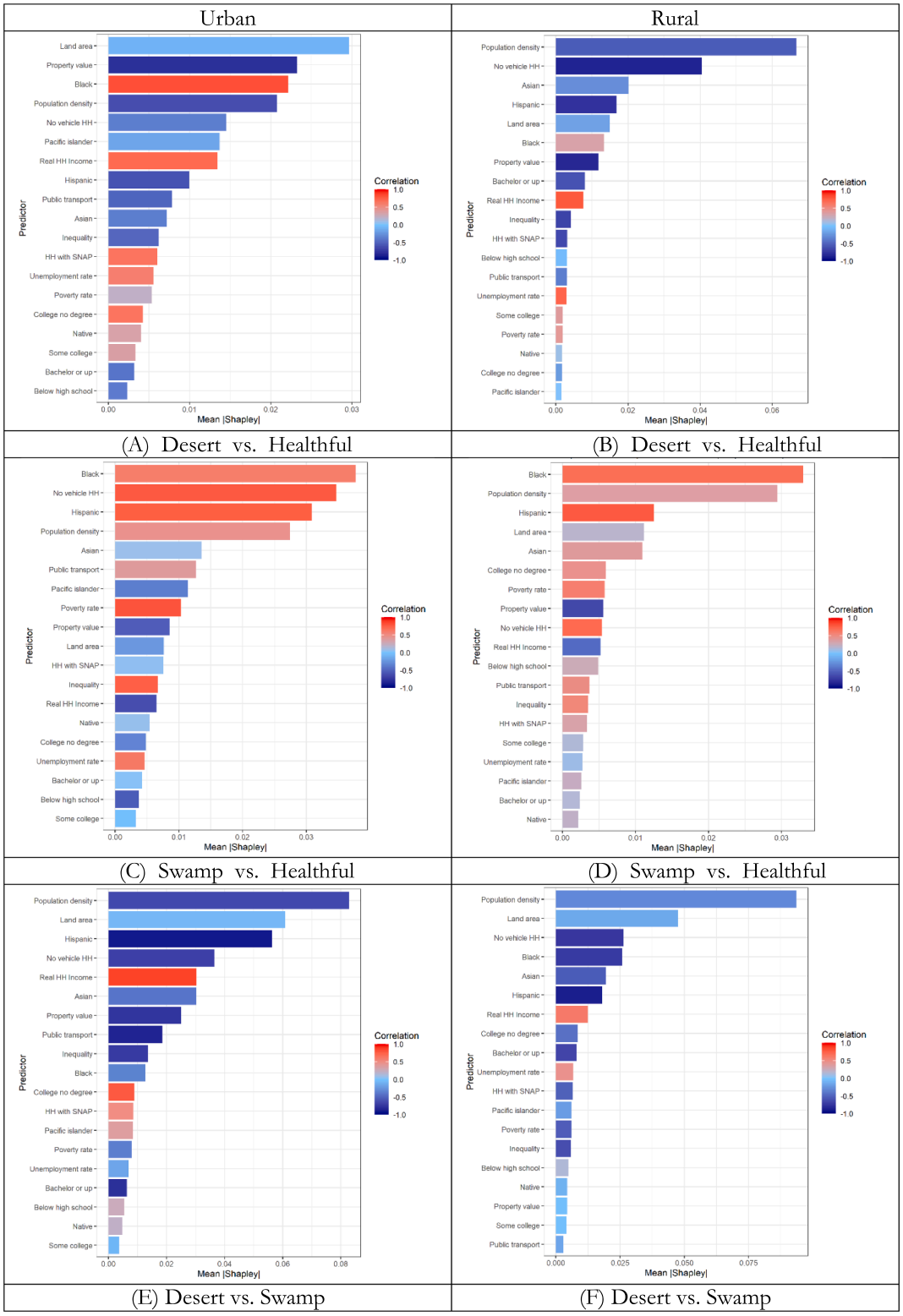

Fig. 2. Determinants of Access: Urban vs. Rural. Top predictors of food deserts and swamps for urban tracts on the left column and rural tracts on the right column. The correlation between the predictor and SHAP values indicate the direction of association (red for positive and blue for negative), whereas the mean SHAP values show the predictor's marginal contribution in prediction. Predictor descriptions are provided in Table 2. We have three models that use binary response: (A and B) Target = desert, benchmark = healthful, swamp tracts are not included in the sample, (C and D) target = swamp, benchmark = healthful, desert tracts are not included in the sample, and (E and F) target = desert, benchmark = swamp, healthful tracts are not included in the sample.

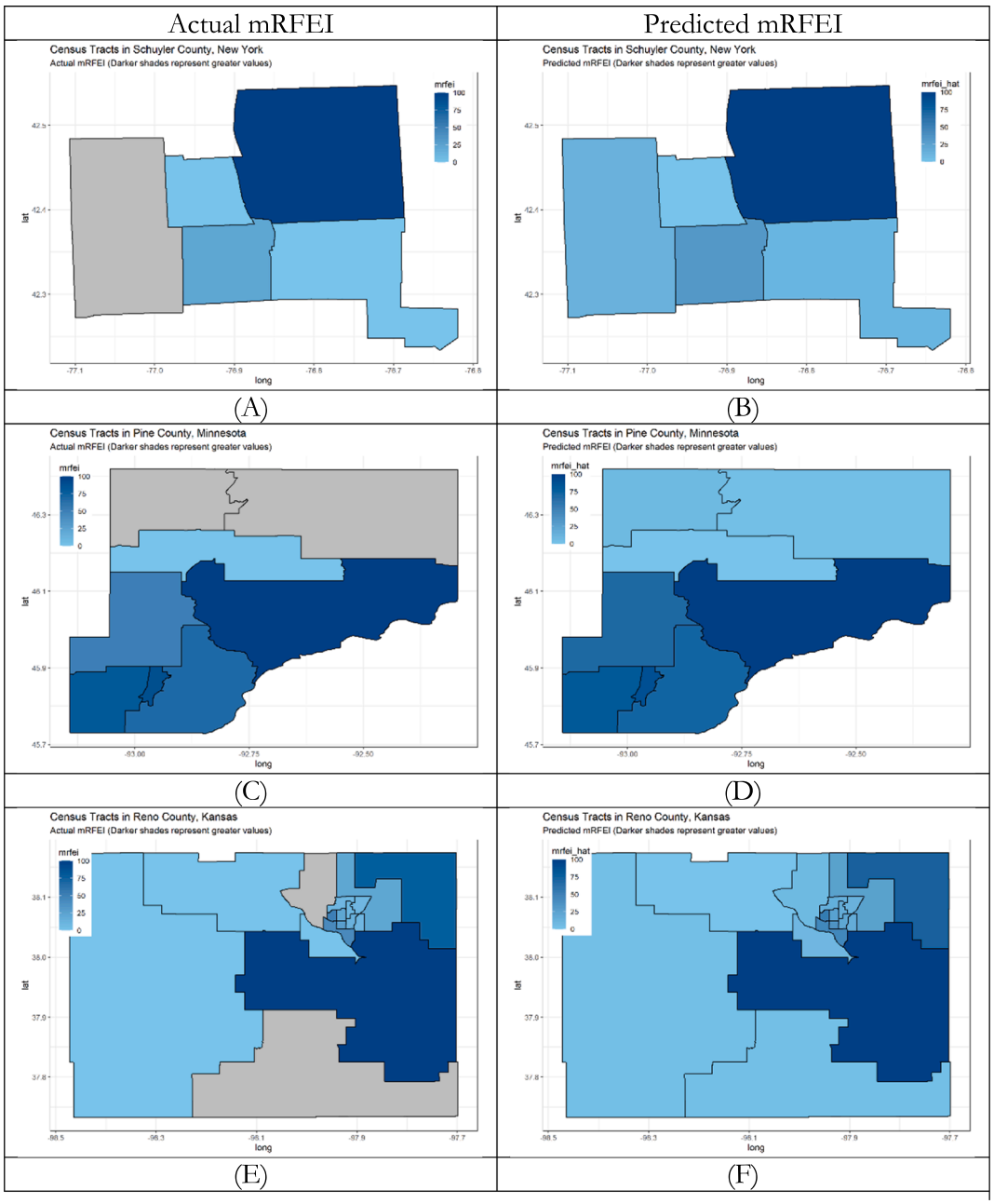

Fig. 3. Actual and predicted modified Retail Food Environment Index (mRFEI). Tracts shaded grey on the left column indicate missing mRFEI values in the data. All tracts are predicted by ML on the right column, including the tracts with missing mRFEI. Darker shades represent greater values of mRFEI.

where the βs are the parameters to be estimated, γ represents the penalty parameter, and N is the sample size. Choosing $\Omega > \sum_1^p \left|\widehat{\beta}_j\right|$ gives the ordinary least squares (OLS) estimates. Define a shrinkage factor $\lambda = \Omega / \sum_1^p \left|\widehat{\beta}_j\right|$ such that $\lambda = 1$ yields the least squares estimates, and the coefficients reduce to zero as $\lambda \to 0$ (Hastie et al., 2009). The shrinkage factor is optimally chosen through the ten-fold cross-validation. The key idea is to drop the xs that have βs close to zero, so it is important to standardize the xs beforehand for a fair comparison.

An equivalent classification problem where the response is binary, e.g., y={Yes, No}, the logistic regression model implies $\log \frac{\Pr(y=\text{Yes}|x)}{\Pr(y=\text{No}|x)} = \beta_0 + \sum_j^p x_{ij}\beta_j$, the LASSO estimates are,

$$\widehat{\beta}_{LASSO} = \underset{\beta}{\text{argmax}} \left\{ \frac{1}{N} \sum_i^N [I(y_i = \text{Yes}) \log p(\text{x}_i) + I(y_i = \text{No}) \log(1 - p(\text{x}_i))] - \gamma \sum_{j=1}^p |\beta_j| \right\} \tag{7.2}$$

where, $p(\text{x}_i)$is the probability $\frac{1}{1+e^{-(\beta_0 + \sum x_{ij}\beta_j)}}$ (Friedman et al., 2010).

Setting $\gamma = 0$ yields the usual logit model. Note that we do not include logit model as the benchmark because LASSO specification reduces to logit for a classification problem and to ordinary least squares (OLS) for a regression problem when its penalty restriction is set equal to zero. Since LASSO conducts a grid search and chooses the optimal penalty factor, including logit or OLS would be be redundant.

The LASSO shrinkage causes the estimates of the non-zero coefficients to be biased and inconsistent because an estimate close to zero does not mean the respective predictor can be omitted from the model (Hastie et al., 2009). However, it serves our purpose of prediction and feature extraction.

2.4. Evaluation of model performance

For binary target variables, we evaluate the level of accuracy and its 95% confidence interval (CI), true positive rate (Sensitivity), true negative rate (Specificity), and Cohen's Kappa (Cohen, 1960). As defined earlier, let Desert = 1 if a tract is a desert and 0 otherwise. Then a 2×2 confusion matrix has elements $a_{row,column}$ with predicted conditions $\hat{y} = \{1, 0\}$ on rows and true conditions $y = \{1, 0\}$ on columns. The statistics are defined by,

$$Accuracy = \frac{a_{11} + a_{22}}{a_{11} + a_{12} + a_{21} + a_{22}} \tag{8}$$

$$Sensitivity \text{ or } True\ Positive\ Rate\ (\text{TPR}) = \frac{a_{11}}{a_{11} + a_{12}} \tag{9}$$

$$Specificity \text{ or } True\ Negative\ Rate\ (\text{TNR}) = \frac{a_{22}}{a_{21} + a_{22}} \tag{10}$$

In this example, accuracy refers to the portion of tracts where a food desert is predicted as a food desert and a healthful tract is predicted as healthful. We calculate the 95% confidence interval using accuracy's standard deviation generated through iterations. Sensitivity implies how many deserts are correctly identified as deserts, and specificity indicates how many healthful tracts are identified as healthful. Cohen's Kappa statistic measures the agreement for categorical variable relative to what would be expected by chance. That is,

$$\kappa = 1 - \frac{1 - \text{Accuracy}}{1 - \text{EP}} \tag{11}$$

$$EP = \frac{(a_{11} + a_{12})(a_{11} + a_{21}) + (a_{21} + a_{22})(a_{12} + a_{22})}{(a_{11} + a_{12} + a_{21} + a_{22})^2} \tag{12}$$

where, accuracy of prediction is the observed probability, and *EP* is the probability of random agreement or expected probability. Cohen's Kappa becomes zero if there is no agreement among the predicted and observed response other than what would be expected by chance.

For the continuous variable, *mRFEI*, we estimate the normalized root mean square error (NRMSE),

$$NRMSE(\%) = 100 \times \frac{\sqrt{\frac{\sum_i^N \left(y_i - \hat{y}_i\right)^2}{N}}}{sd(y_i)} \tag{13}$$

which offers a straightforward comparison across models. A NRMSE close to zero implies high predictive accuracy (Hossain et al., 2019).

2.5. Interpretation of features with Shapley values

Both RF and XGB calculate the variable importance, which is the increase in the mean squared error (MSE) of prediction when a variable is randomly permuted from the model. A high importance of a predictor indicates a substantial increase in the MSE of prediction due to the omission of a predictor. The importance factor reveals the influence of independent variables in predicting the dependent variable, hence offers a way to reduce the number of predictors by trimming the variables with low importance (Altmann et al. 2010). However, the importance factors generated by RF and XGB do not provide the direction of the association between the predictors and response.

We derive Shapley values (SHapley Additive exPlanations or SHAP) to make the results from ML more interpretable. The SHAP value, named after Shapley (1953), is a solution concept in cooperative game theory and is the weighted sum of the prediction gaps with and without the predictor. The weight is generated from all possible combinations of predictors with ordering. Formally,

$$\phi_j\left(\hat{y}_i\right) = \sum\nolimits_{s \subseteq \{x_j \notin p\}} \frac{|s|!(p - |s| - 1)!}{p!} \left[\hat{y}_i(s \cup x_j) - \hat{y}_i(s)\right] \tag{14}$$

where s is a set of predictors out of all possible sets of predictors excluding x_j, $|s|$ is the length of s, and $\hat{y}$ is the predicted response. Notice that ϕ_j is generated for each observation, making the Shapley value matrix Φ having the same dimension as the predictor matrix. The sum of all predictors' Shapley values is the unexplained part of the model for each observation, that is, $\sum_{j=1}^{p} \phi_j\left(\hat{y}_i\right) = y_i - \bar{\hat{y}}$, where $\bar{\hat{y}}$ is the overall mean of all predictions generated by the model.

Shapley values provide an interpretable association of predictors with the response. Intuitively, a positive (negative) SHAP value implies an increase (decrease) in the predicted response ($\hat{y}_i$) compared to the overall mean prediction ($\bar{\hat{y}}$) due to the inclusion of a certain predictor (x_j). A zero SHAP indicates no deviation from the overall mean prediction. Shapley value is the contribution of a predictor to the difference between the actual prediction and the mean prediction. The greater an absolute SHAP, the more important the respective predictor for the model. Shapley values are computationally expensive for considering every possible combination among predictors and are obtained using XGB for faster calculations.

3. Data

3.1. Predictors of the access

We utilize U.S. census tracts as the units of observation for our analysis. Census tracts generally have a population between 1200 and 8000 and offer more intra-location homogeneity because they are smaller than counties and have consistent tract boundaries with visible and identifiable features. Each tract has a unique four-digit code assigned to it, which helps us to merge the data with other sources.

In order to predict the response variables, we started with 281 variables from the American Community Survey 2010 (U.S. Census Bureau, 2019b). The data include community and housing surveys, commodity flow surveys, surveys of government, and economic censuses. Selected variables contain a range of tract-level averages of individual information, e.g., sex, race, age, household size and relationships, citizenship, geographical mobility, employment status, transportation to work, vehicle availability, education, income, occupancy characteristics, the value of living place, poverty and inequality indices, and financial characteristics. There are about 74 thousand tracts in Census 2010, of which we retain 50,212 tracts after dropping the missing observations.

The challenge is to avoid repetition in predictors and obtain interpretable results. A large number of variables are almost perfectly correlated, e.g., mean household income and mean family income. Some other variables correlate moderately, e.g., population above 25 years old with a Bachelor's degree and population above 30 years old with a Bachelor's degree. We begin with all 281 predictors to predict mRFEI in an XGB model and shortlist the top 50 predictors based on their Shapley values. Greater predictability, interpretability, and consistency with the literature serve as tiebreakers for variables offering similar information. This process leads to 20 variables that demonstrate considerable

importance in predicting mRFEI. Table 2 presents their summary statistics. Predictors mainly include income, poverty, inequality, education, race, and transportation. We omit the percentage of white (Caucasian) population to avoid high collinearity with other race variables.

3.2. Stylized facts

We begin with testing the mean difference for each variable by the level of access to healthful food retailers. Table 3 presents the means of variables grouped by binary variables (food desert, food swamp, and tracts with good access to healthful food) and results of mean-difference t-tests with unequal variances. A statistically significant difference in mean values indicates the predictor's association with access to healthful food. By construction, a typical food desert has 0% of the retailers with healthful food options, which is 6.07% in food swamps and 18.81% in healthful tracts. Mean differences indicate that food deserts and food swamps, respectively, have 18.81% points and 12.75% points fewer healthful food retailers than healthful tracts. Almost all t-statistics for predictor variables are statistically significant at the 1% level—indicating that food deserts and food swamps are different in their demographics compared to healthful tracts.

Food deserts are predominantly large, rural tracts with low population density, have fewer residents with college educations, and fewer households who receive supplemental nutrition assistance program (SNAP) benefits compared to tracts with good access to healthful food retailers (see Table 3). People living in food deserts lack public transportation and need a car for their day-to-day activities. This might also deter SNAP beneficiaries from living in a food desert. Native populations live in these tracts relatively more than other races.

Food swamps, on the other hand, are compact land areas, urban tracts, with high population densities, have lower household income, greater poverty, unemployment and inequality, low access to vehicles, and more SNAP recipients compared to other tracts. In these urban settings, food swamp tracts have greater usage of public transportation, higher property values, and have more black, Hispanic, and Asian populations. The stark contrast between food deserts and food swamps is clear from the last column of Table 3. Land area, rural population, vehicle ownership, and household income are considerably lower in food swamps compared to food deserts, whereas the poverty rate, inequality, unemployment, and non-white percentage of population are higher.

Tracts with good access to healthful food are characterized by medium land area—39 square miles compared to 104 for food deserts and 5.6 for food swamps—and moderate population density with higher incomes, college education, and employment. The black population in tracts with good access to healthful food is 3.26% points lower than food deserts and 8.11% points lower than food swamps. The average property value is $253,930—which is below food swamps ($263,480) but above food deserts ($194,450). It appears that tracts with good access to healthful food have some rural population (16.7%), which indicates that some of the residents live away from the tract centroid.

4. Training the models

This section discusses the pre-processing of the data and model training. Heterogeneous scales of predictors and high correlations among them may generate incomparable or convoluted importance of variables that are difficult to interpret. Although multicollinearity should not matter in prediction, especially in non-linear models, it might complicate the calculation of variable importance. We standardize the predictors to ensure that the scale of a variable does not influence its importance.[1] No two predictors have an absolute correlation coefficient close to one. In fact, all absolute correlations are below 0.74.

We then randomly split the data into training (70%) and independent-test (30%) samples. Training of the model consists of fine-tuning the hyperparameters until the model is optimized. We use ten-fold cross-validation during training, which means the chosen 70% of the sample is again randomly partitioned into ten equal-sized subsamples, then nine subsamples are used as training data and the remaining one is used for validation. The cross-validation process is repeated ten times (called folds) such that each of the subsamples is used only once for validation. The hyperparameters are updated based on the average of ten results. The learning completes when hyperparameters are optimal—producing minimum prediction errors in the shortest time. We then test the predictive performance of the final model using the untouched 30% test sample. This provides an unbiased estimate of the performance of the final model (e.g., Bajari et al., 2015).

A forest constructed using more and fully-grown trees results in higher accuracy. Using small depth trees may lead to poor performance and under-fitting. In RF, for instance, the researcher picks the number of trees to be grown (T) and the number of variables to be randomly considered in each split (q). Here T, q are hyperparameters, and they tend to have diminishing returns on predictive performance. Increasing q can improve the predictive performance for more predictors being available at each node of the tree, but more predictors and trees make the construction of a model computationally intensive and hence slower. The optimal number of trees and predictors depends on the prediction error and calculation time. Training XGB requires more parameter tuning than RF for its additional task of growing extra trees on residuals.

During model training, we conduct a grid search over a range of values of hyperparameters, e.g., $q \in \{3, \cdots, \sqrt{p}\}, T \in \{150, \cdots, 1000\}$, and chose the parameters that produce minimum prediction errors. Figs. A2-A4 in the appendix show the performance of the model for a range of major parameters used for RF, XGB, and LASSO. Detail grid search results are available from the authors upon request. R software version 3.5.3 was used for the analysis.

5. Prediction results

Table 4 presents the statistics on the prediction performance in test data. Performance statistics in training data are reported in the appendix for comparison (Table A1). All three methods predict food deserts and food swamps accurately around 72% of the time in the test data—indicating the robustness of the model. Given the cross-sectional, socioeconomic, and out-of-sample nature of the data, Kappa statistics show that the models perform considerably well and are not random (McHugh, 2012). True positive rates (sensitivity) exceed true negative rates (specificity) for all binary predictions, which means the models perform better at detecting the target than rejecting the benchmark. In the desert versus healthful model, for instance, XGB correctly recognizes a food desert in 91.6% cases, but a healthful tract in only 26.5% cases. That means, XGB does not detect a food desert (false negative rate) in $100 \times (1 - 0.916) = 8.4\%$ cases and does not detect a healthful tract (false positive rate) in $100 \times (1 - 0.265) = 73.5\%$ cases.

High sensitivity and low specificity can be an indicator of class imbalance problem in machine learning. We argue that our results are not due to class imbalance for the following reasons. First, we use median mRFEI cut-off that makes 23% of the tracts food deserts, 24% food swamps, and the remaining 53% healthful—none of which is a rare

[1] Fig. A1 in the appendix shows pairwise correlations among the predictors.

category. Second, we make the under-representative class as the target class for each model, which lowers the chance of mechanical result. Third, we use an Adaptive Synthetic Sampling Approach (He et al., 2008) to increase the minor class and balance the training data. The first two models, desert vs. healthful (swamp excluded) and swamp vs. healthful (desert excluded) from Table 4 were reproduced. The third model desert vs. swamp (healthful excluded) was not reproduced because deserts and swamps are almost equal in number. Table A2 in the appendix shows that sensitivity increases at the cost of specificity for both reproduced models in the test data. Given the trade-off between sensitivity and specificity (e.g., Buderer 1996), greater sensitivity is a desired outcome for the current study. Because our objective is to correctly flag every desert and swamp and not generate many false-negative results—a rule often followed in disease diagnostics (Chu, 1999).

Predictions of XGB and LASSO are consistent between training and test data due to additional regularization of parameters. A similar pattern appears in NRMSE for the continuous response mRFEI. Prediction performance with a continuous dependent variable tends to be weaker than that with binary dependent variables for its variation. Fig. A5 in the appendix plots the predicted values of mRFEI against its actual values. A 45-degree line implies perfect prediction. RF performs better in the training sample, but the XGB shows overall consistent performance across samples. Both RF and the XGB are robust to outliers—hence predictions are concentrated around the mode of observed mRFEI, and values far from the mode are predicted poorly.

How does the reported prediction accuracy compare to that of existing literature? ML may predict purely scientific models up to 100% cases in test data. Economic studies implementing ML for cross-sectional out-of-sample predictions are rare, given the level of heterogeneity and interconnectivity of socioeconomic variables. Hossain et al. (2019) use 323 observations and test the performance of RF and XGB (with 70% training, 30% validation, 0% test data) and obtain over 90% accuracy. However, model validation is used for refining the hyperparameters to improve the model and do not indicate out-of-sample performance. A more appropriate test of prediction is to employ the model in an independent test sample, as commonly practiced in computer science. The current study randomly divides data into two parts 70% and 30%, trains the model on 70%, and then tests the trained model on the held-out 30%. It can be possible to raise predictive accuracy by increasing the share of the training sample, e.g., Bajari et al (2015) and Racca et al. (2016) use 75%-25% split. Given the size of our held-out samples with at least seven thousand observations, and the heterogeneous nature of cross-sectional and socioeconomic data, the model used in this study arguably demonstrates a robust predictive power.

6. Results from feature extraction

6.1. Features of deserts and swamps

Fig. A6 in the appendix plots the Shapley values. The SHAP values and the predictors are placed respectively on the horizontal and vertical axis. Each dot represents a census tract; hence, the number of dots against each predictor equates the sample size. The figures on the vertical axis are the mean of SHAP values that indicate the average contribution of the corresponding variable in prediction. A positive (negative) SHAP value represents an increase (decrease) in the predicted variable across all possible coalitions of the predictors. In Fig. A6(A) for example, the predictor "rural" increases the predicted values (SHAP > 0) for most observations when included in the coalition. Darker shades imply greater values of the predictor, e.g., greater values of rural population are observed where SHAP > 0. This implies that the rural variable is positively associated with a tract being desert. Thus, the correlation between the predictor and SHAP values indicate the direction of association, whereas the mean SHAP values show the predictor's contribution.

Fig. 1 summarizes the Shapley values for better insights. Mean SHAP values are shown against the predictors on a bar diagram, colored according to the correlations between the predictor and its SHAP values. Rural population and black population have high positive associations with food deserts, whereas population density, no vehicle, and property value have negative associations. Interestingly, deserts seem to have higher household incomes. This contradiction is resolved in the next subsection where we find that the greater household income is primarily a feature of urban deserts—possibly caused by upscale residential areas. It is possible, however, that low-income consumers in food deserts without vehicle access may have migrated to food swamps or healthful tracts for survival.

The variable land area appears to be an important contributor to the predictions of the desert, but has a low correlation with SHAP (-0.13). This means inclusion of land area in the model increases $\hat{y}$ for some tracts and reduces $\hat{y}$ for many others (also evident from Fig. A6-A). Intuitively, greater land area is not always a feature of food deserts, but keeping the variable in the model is essential for other predictors, and omitting it would lower the predictability.

Food swamps are negatively related to the rural population (Fig. 1B). Swamps are further characterized by higher population density, black and Hispanic population, and lack of vehicle access compared to healthful tracts. Poverty and unemployment rates are greater in swamps, whereas property values and household income are lower. Fig. 1C captures the main differences between deserts and swamps. Compared to swamps, deserts have more rural population and household income, but less population density, non-white population, property value, public transport, inequality, and households with no vehicles. Education variables receive less importance in predicting the difference; however, swamps seem to have more college educated than deserts. Fig. 1(A-C) provide evidence that food deserts and food swamps are structurally different and may need distinct policy interventions.

Table A3 presents the LASSO estimates for robustness. Standard errors for LASSO estimates are omitted because they are redundant for feature extraction and do not have a straightforward formula (e.g., Lockhart et al., 2014). The LASSO reduced dimension for the Swamp vs. Healthful and Desert vs. Swamp models by pushing some of the coefficients to zero (Table A3). The remaining predictors are consistent with the ones declared important by Shapley values in Fig. 1. Noticeably, black population is assigned a zero coefficient by LASSO in desert versus swamp model—pointing at the fact that black population is a common feature to both food deserts and swamps.

6.2. Urban and rural subsamples

In this subsection, we repeat the above tasks separately on urban and rural samples. The motivation is to understand the difference between urban food deserts (swamps) and rural food deserts (swamps). An urban area comprises a densely settled core of census tracts/blocks that have at least 2,500 people and meet some other criteria on the adjacent territory (U.S. Census Bureau, 2019b). About 80% of U.S. Census tracts are urban by this definition. Table A4 in the appendix shows the performance statistics in the test data and Fig. 2 plots the summarized Shapley values.

In the desert vs. healthful model using the urban-only sample, property values (with a negative correlation) and the percentage of

black population (with a positive correlation) gain importance relative to the full sample in predicting food deserts. While in the rural-only sample, population density is the top predictor of food deserts. In the rural-only sample, the percentage of Asian and Hispanic residents gains importance with negative associations relative to the full sample and the urban sample.

In the swamp vs. healthful model, the results are largely similar to the full data set with the percentage of black population as the most important predictor with a positive correlation. With the urban-only data, the percentage of no-vehicle households gains importance with positive association relative to the full sample. In the same model with rural-only data, land area gains in importance as a predictor with a positive association.

In the desert vs. swamp model with urban-only data, the results are consistent with the full data set with minor relative changes in relative importance. In the same model with the rural-only data, the percentage of black residents becomes more important with a negative association.

7. Policy implications

Machine learning can provide insights at a relatively low cost without imposing parametric assumptions on measures of access to healthful foods, its predictors, and the functional form. One can use the model above to predict the access in regions where mRFEI data are not available yet, and obtain a reliable approximation. For instance, we present three random counties in Fig. 3 where mRFEI data are not available for some of the tracts. The left column of Fig. 3 shows actual mRFEI values in Schuyler county (New York), Pine county (Minnesota), and Reno county (Kansas); and the right column shows the predicted values from the mRFEI model. Darker shades represent greater values of mRFEI. Grey shaded tracts on the left column are missing the mRFEI data. The actual and predicted shades of mRFEI appear to be almost similar for tracts whose data are available. Given the predictive performance of machine learning, it is highly likely for the predicted tracts to reflect the reality. Applying our model to 2020 census data or any recent data where the aforementioned demographic variables are available will help update the mRFEI. Tracts with low predicted mRFEI can then be scrutinized for policy intervention.

Combining the results from machine learning and mean difference t-tests, we observe that food deserts—tracts that have no healthful retailers—are lightly populated tracts with more rural population and low ethnic diversity. Conversely, food swamps—tracts that have disproportionally low healthful retailers—are densely populated urban tracts that suffer from low income, employment, vehicle access, and have greater representation of non-whites. Overall, the major predictors of the retail food environment are population density, black and Hispanic population, property value, vehicle access, and income. This justifies USDA's definition of food desert based on income, vehicle access, and location. However, the presence of minorities in the area appears to be a critical determinant and can be weighted in a multidimensional measure of access. For example, greater black population is a common feature for both deserts and swamps relative to the healthful tracts—more so in urban areas.

We found a clear association between food swamps and black and Hispanic populations. If the difference in healthful diet is mainly driven by the differences in supply (e.g., Feather, 2003; Olsho et al., 2016; Richardson et al., 2017), our results suggest that more healthful retailers should be encouraged (e.g., with tax credits or other policies) in places where healthful retailers are inundated by unhealthful retailers—especially in black and Hispanic majority areas. Another option can be restricting less healthful retailers. On the other hand, if the difference in healthful diet is mainly driven by the demand, then introducing heathful retailers will not change the demand (e.g., Bartlett et al., 2014; Stern et al., 2016; Vaughan et al., 2017; Allcott et al., 2019; Sharpe et al., 2020). These studies argue that the lack of healthful food may be caused by the lack of demand for it, and policy intervention is required to induce the residents towards healthful consumption (e.g., Allcott et al., 2019). Then our results suggest that black and Hispanic majority areas are more deprived of healthful foods and can be prioritized.

The findings are consistent with Rose et al. (2009), Hager et al. (2016), and Cooksey-Stowers et al. (2017) that food swamps are intrinsically different from deserts and deserve separate policy attention. Goodman et al. (2020), for example, discuss separate policies for deserts and swamps. A food desert calls for incentives to build grocery stores in unserved areas, whereas a food swamp can be benefitted from zoning or limiting less healthful establishments (Goodman et al., 2020). This is intuitive because we found swamps having higher poverty, inequality, and unemployment compared to other tracts—particularly in urban areas. Not surprisingly, recent news discussed how black and Hispanic populations were disproportionately affected by COVID-19 in New York City because living in food swamps worsened their comorbidity (Adams, 2020). Besides, residents in food swamps lack vehicle ownership and depend on public transportation in both urban and rural areas more than deserts. Dependence on public transportation adds another restriction in reaching the desired retailers for swamp residents. For example, a typical consumer in a food swamp may prefer a fast food restaurant to a supercenter if the restaurant is on the bus route, even if the supercenter has a shorter drive time.

8. Concluding remarks

The current study utilizes machine learning to predict access to healthful food. Using data from 50,212 U.S. census tracts, we derive the optimal models with at most twenty demographic variables that can predict access to healthful food with high accuracy. This study is first of its kind to employ machine learning in food policy with appropriate test data. Instead of using multidimensional measures of food deserts and food swamps, we extracted their salient features using an index of healthful food retailers. Major distinguishing characteristics indicate that food deserts are characterized by lightly populated rural areas, whereas food swamps are characterized by densely populated poverty-prone urban areas with more non-white population.

Given the complexity of working with socioeconomic data, machine learning models can be a promising alternative to traditional ways of analyzing the food sector. Our model provides a convenient starting place to identify specific geographic areas of states that could benefit from further policy interventions aimed at improving the retail food environment. Future studies will focus on capturing the non-demographic predictors of access to healthful foods.

CRediT authorship contribution statement

Modhurima Dey Amin: Conceptualization, Data curation, Formal analysis, Software. **Syed Badruddoza:** Conceptualization, Data curation. **Jill J. McCluskey:** Conceptualization, Supervision.

Appendix A

See Figs. A1–A6. and Tables A1–A4.

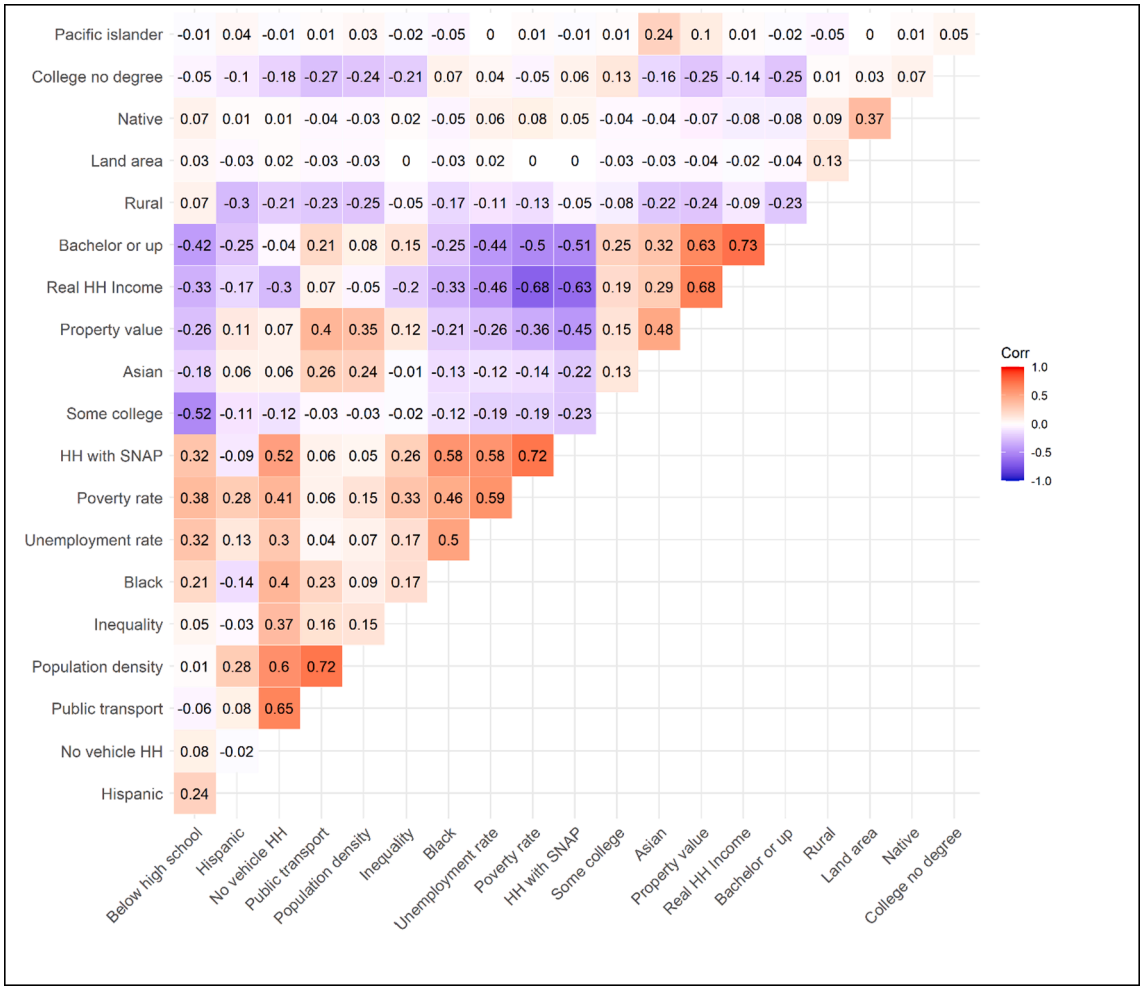

Fig. A1. Pearson correlation coefficients of the predictors. Number of predictors = 20, observations = 50,212. Darker shades represent greater positive (red) or negative (blue) correlation.

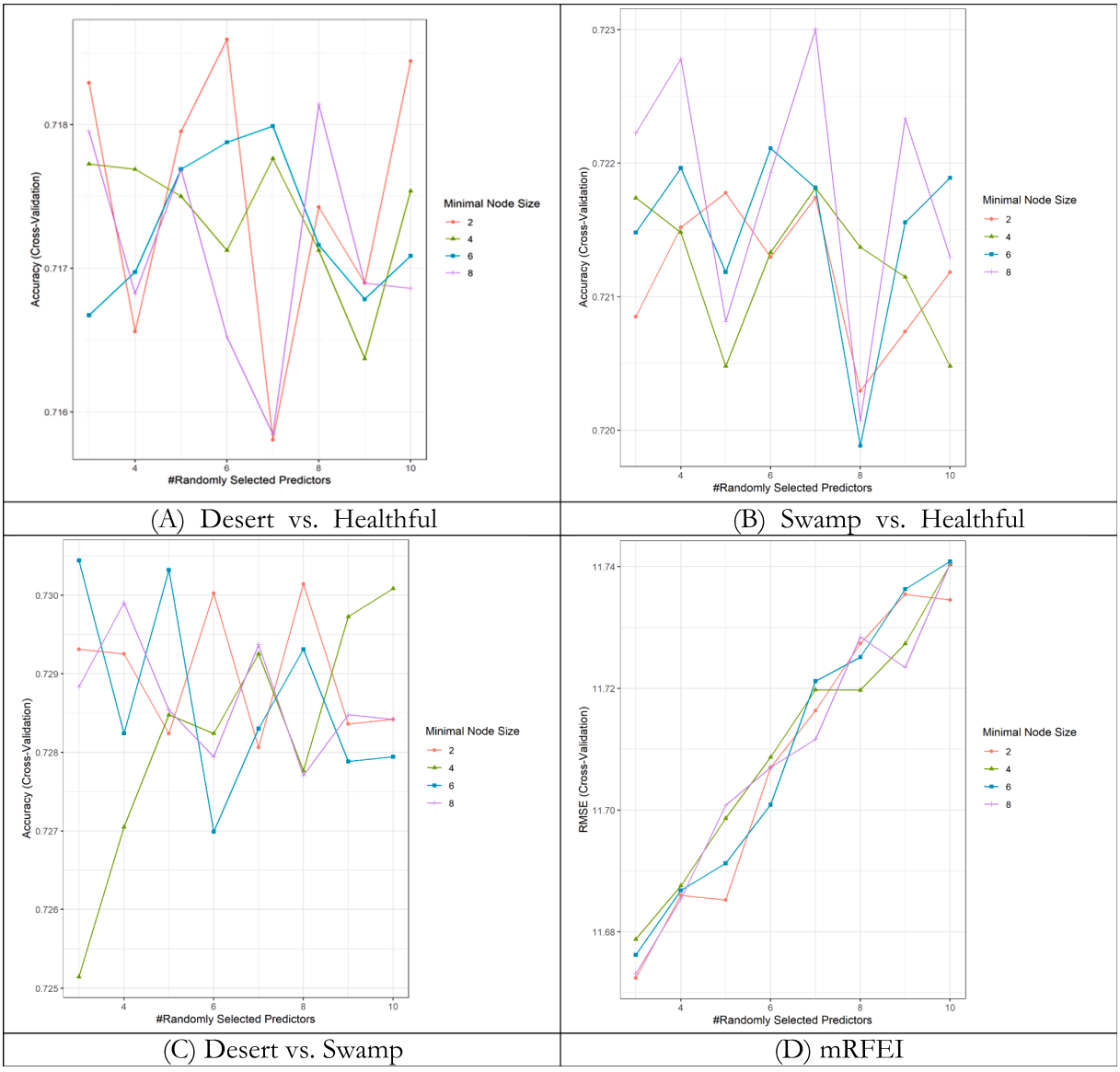

Fig. A2. Tuning hyperparameters for Random Forests. Minimal Node Size is the minimum number of cases required splitting a tree node (usual range 1 to 10). The number of predictors randomly chosen (q) at each node usually ranges from 3 to square root of the number of predictors. A, B, C show binary, and D shows the continuous response variable. Definitions: (A) Target = desert, benchmark = healthful, swamp tracts are not included in the sample, (B) target = swamp, benchmark = healthful, desert tracts are not included in the sample, and (C) target = desert, benchmark = swamp, healthful tracts are not included in the sample. Final model selection criteria: Accuracy = the higher the better, RMSE = the lower the better.

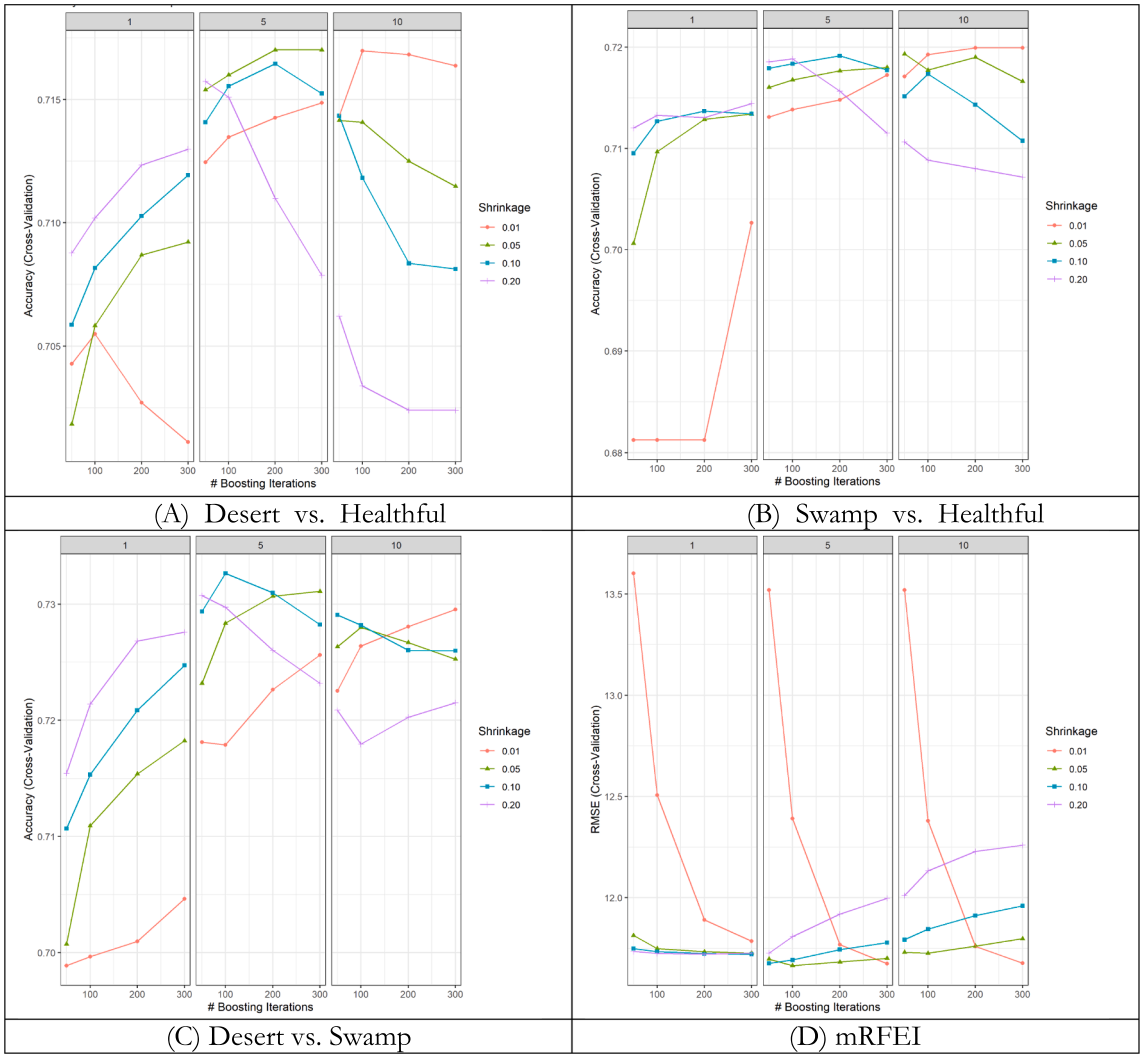

Fig. A3. Tuning hyperparameters for eXtreme Gradient Boost by maximum tree depth={1,5,10}. Maximum tree depth limits the detailed growth of a tree (default value = 6). Deeper trees are more complex and consume more CPU power. Boosting iterations are the number of iterations used for adjusting the weights on the predictors. Shrinkage is the step size used in the update to prevent overfitting. Greater shrinkage makes weighting more conservative (range is 0 to 1). A, B, C show binary, and D shows the continuous dependent variable. Definitions: (A) Target = desert, benchmark = healthful, swamp tracts are not included in the sample, (B) target = swamp, benchmark = healthful, desert tracts are not included in the sample, and (C) target = desert, benchmark = swamp, healthful tracts are not included in the sample. Final model selection criteria: Accuracy = the higher the better, RMSE = the lower the better.

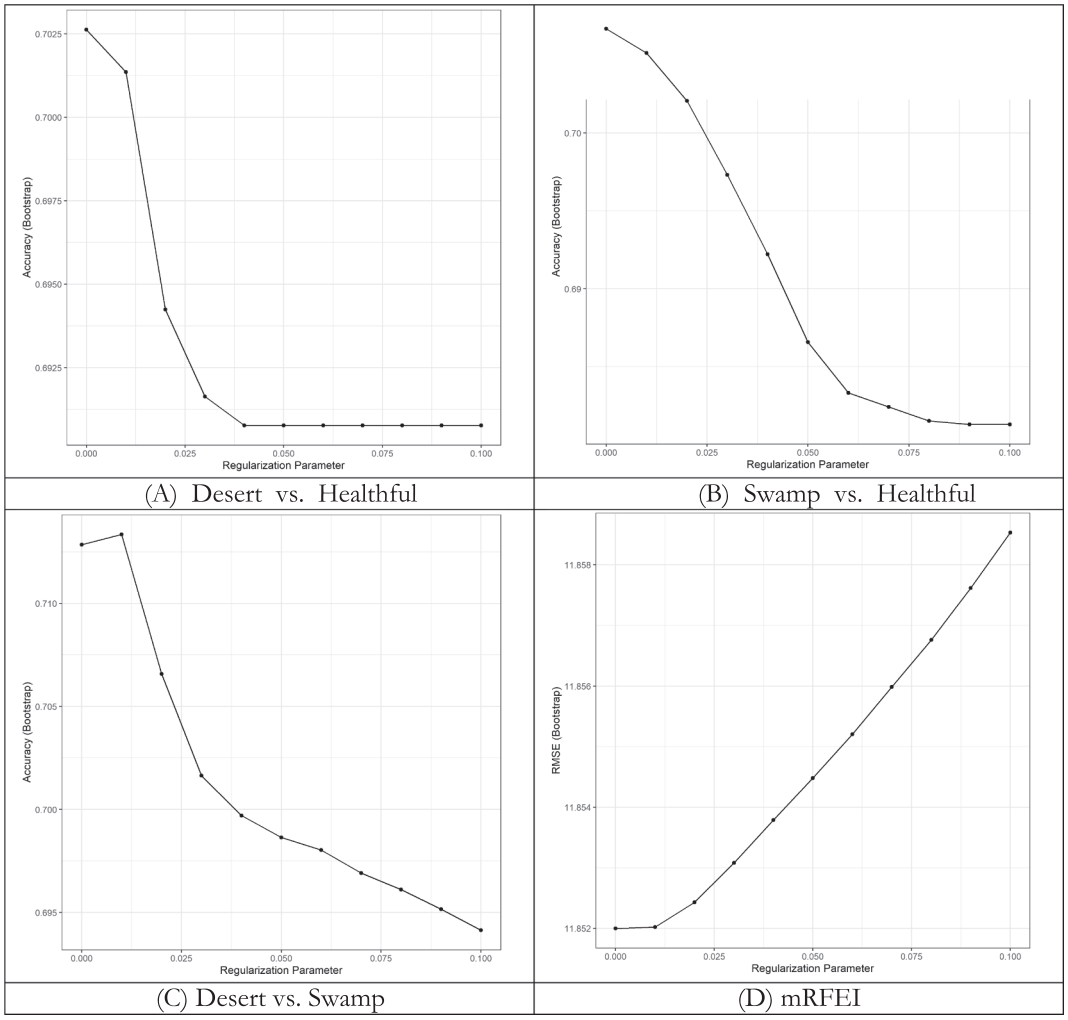

Fig. A4. Tuning hyperparameters for LASSO. The regularization parameter (λ) represents the penalty term on coefficient (zero or above). Zero value means all predictors are important so the LASSO leads to OLS regression. A, B, C show binary, and D shows the continuous dependent variable. Definitions: (A) Target = desert, benchmark = healthful, swamp tracts are not included in the sample, (B) target = swamp, benchmark = healthful, desert tracts are not included in the sample, and (C) target = desert, benchmark = swamp, healthful tracts are not included in the sample. Final model selection criteria: Accuracy = the higher the better, RMSE = the lower the better.

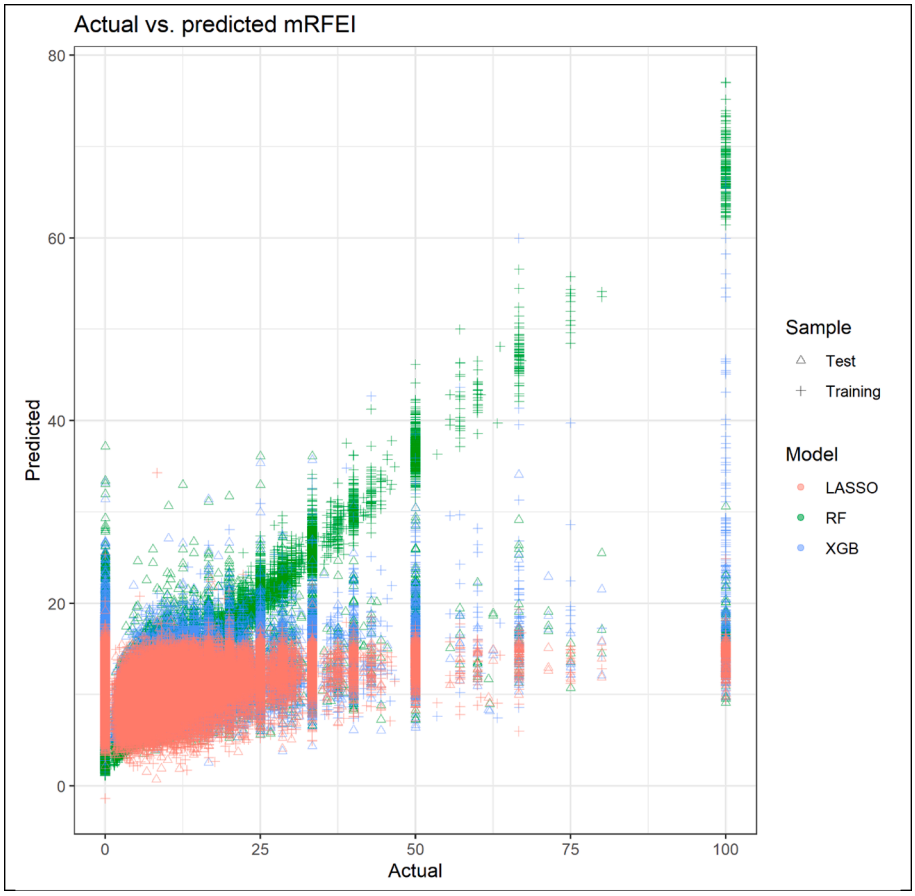

Fig. A5. The modified Food Retail Food Environment Index (mRFEI) predicted by RF, XGB, and LASSO. The total sample is 50,212 census tracts. The sample is randomly split into training (70%) and test (30%) data. Training implies choosing the best hyperparameters from a grid of hyperparameters with 10-fold cross-validation. Test refers to the independent sample held-out during training.

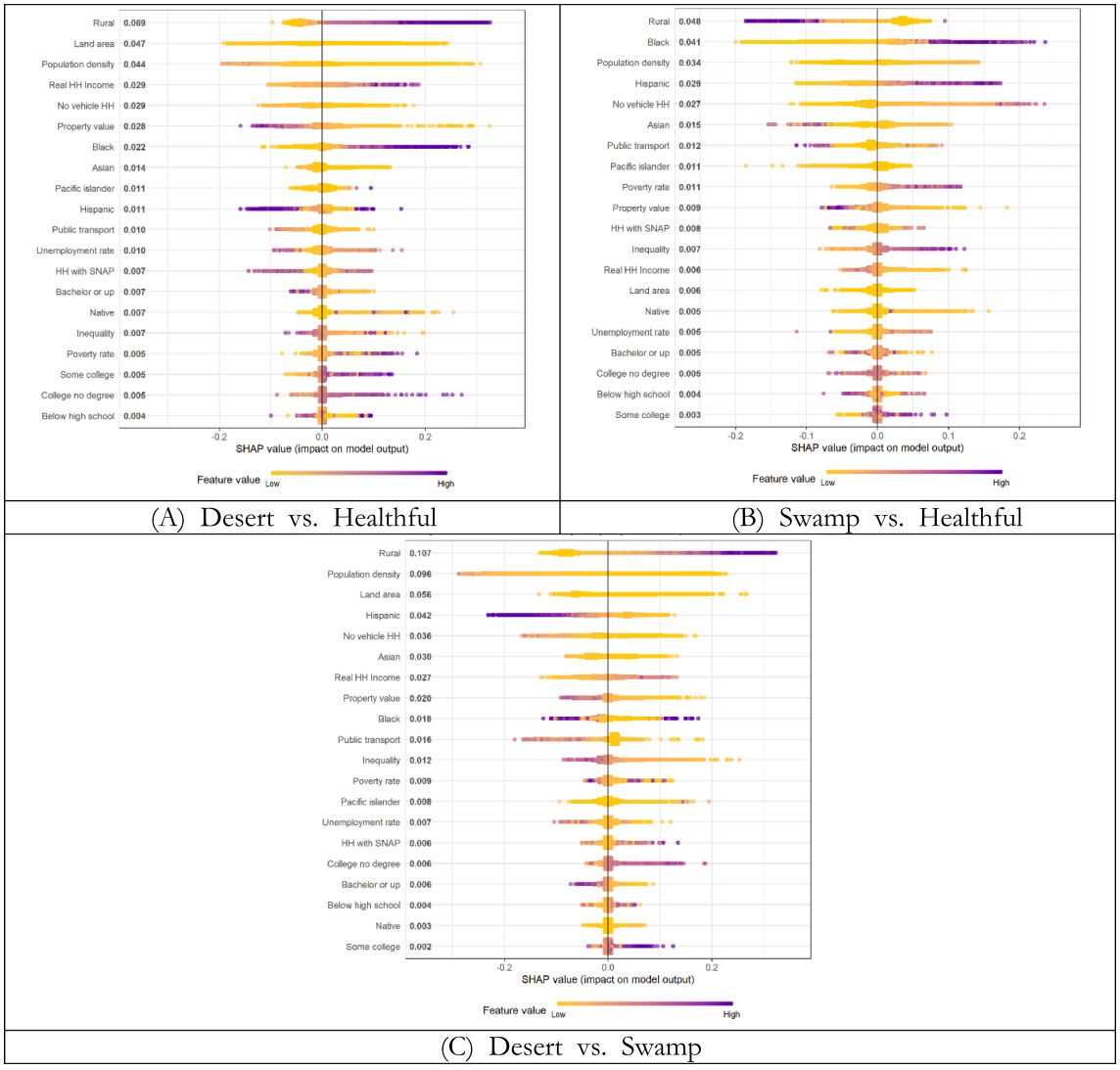

Fig. A6. Shapley values against the predictors. Each dot represents a census tract; hence, the number of dots against each predictor equates the sample size. A positive (negative) SHAP value represents an increase (decrease) in the predicted variable across all possible coalitions of the predictors. The mean of SHAP values indicate the average contribution of the variable in prediction and is provided on the vertical axis. Darker shades imply greater values of the predictor. Definitions: (A) Target = desert, benchmark = healthful, swamp tracts are not included in the sample, (B) target = swamp, benchmark = healthful, desert tracts are not included in the sample, and (C) target = desert, benchmark = swamp, healthful tracts are not included in the sample.

Table A1
Model performance statistics (training data).

Model	Statistic	RF	GB	LASSO
Target = Desert,	Accuracy	1.000	0.772	0.703
Benchmark = Healthful	95% CI	(1.000,1.000)	(0.767,0.777)	(0.698,0.709)
(Swamps not included	Sensitivity	1.000	0.956	0.921
in this sample, N = 26,563)	Specificity	1.000	0.361	0.216
	Kappa	1.000	0.372	0.165
Target = Swamp,	Accuracy	0.998	0.840	0.708
Benchmark = Healthful	95% CI	(0.998,0.999)	(0.836,0.844)	(0.702,0.713)
(Deserts not included	Sensitivity	1.000	0.984	0.927
in this sample, N = 26,939)	Specificity	0.994	0.534	0.240
	Kappa	0.996	0.584	0.199
Target = Desert,	Accuracy	0.998	0.791	0.711
Benchmark = Swamp	95% CI	(0.998,0.999)	(0.785,0.797)	(0.704,0.718)
(Healthful not included	Sensitivity	0.997	0.868	0.884
in this sample, N = 16,798)	Specificity	1.000	0.711	0.530
	Kappa	0.997	0.580	0.417
mRFEI(full training sample N = 35,151)	NRMSE	37.1	92.2	98.1

Note: N = respective observations, CI = confidence intervals, Sensitivity = True Positive Rate, Specificity = True Negative Rate, Kappa is the rate of agreement between actual and predicted classes beyond what is expected by chance. NRMSE = Normalized Root-Mean-Squared Error. All accuracy estimates are significant at p-values < 0.01. The total sample is (11,730 Desert + 12,266 Swamp + 26,216 Healthful) 50,212 tracts, all of which are used in the prediction of mRFEI. In each case, the sample is randomly split into training (70%) and test (30%) data. Training implies choosing the best hyperparameters from a grid of hyperparameters with 10-fold cross-validation. Test refers to the independent sample held out during training. Performance in the training data only shows how the models are trained. Although a good performance is expected in training data, it does not guarantee the model performance in test data.

Table A2
Reproduction of Table 4 with synthetically balanced training data.

	After balancing the training data with ADASYN			
Model	Statistic	RF	XGB	LASSO
Target = Desert,	Accuracy	0.677	0.670	0.602
Benchmark = Healthful	95% CI	(0.669, 0.686)	(0.661, 0.678)	(0.593, 0.611)
(Swamps not included	Sensitivity	0.765	0.779	0.586
in this sample, N = 11,383)	Specificity	0.482	0.425	0.639
	Kappa	0.246	0.209	0.194
Target = Swamp,	Accuracy	0.673	0.682	0.590
Benchmark = Healthful	95% CI	(0.664, 0.681)	(0.674, 0.691)	(0.581, 0.599)
(Deserts not included	Sensitivity	0.726	0.765	0.499
in this sample, N = 11,543)	Specificity	0.558	0.506	0.784
	Kappa	0.274	0.270	0.231

Note: N = respective observations, CI = confidence intervals, Sensitivity = True Positive Rate, Specificity = True Negative Rate, Kappa is the rate of agreement between actual and predicted classes beyond what is expected by chance. NRMSE = Normalized Root-Mean-Squared Error. All accuracy estimates are significant at p-values < 0.01. The total sample is (11,730 Desert + 12,266 Swamp + 26,216 Healthful) 50,212 tracts, all of which are used in the prediction of mRFEI. In each case, the sample is randomly split into training (70%) and test (30%) data. Training implies choosing the best hyperparameters from a grid of hyperparameters with 10-fold cross-validation. Test refers to the independent sample held out during training. Training sample was balanced using the Adaptive Synthetic Sampling Approach (ADASYN) that creates observations in the under-representative class weighted by their difficulty in learning that improves upon simple over- or under-sampling methods based on nearest neighbors (See He et al., 2008 for an elaborate discussion). The third model, desert vs. swamp was not reproduced with ADASYN because desert and swamp tracts are close in numbers: 11,730 (49%) and 12,266 (51%) respectively.

Table A3
LASSO Estimates (Test Data).

	Dependent variable		
Predictor	(A) Desert vs. Healthful	(B) Swamp vs. Healthful	(C) Desert vs. Swamp
Real HH Income	0.426	−0.101	0.128
Poverty rate	0.127	0.041	0
HH with SNAP	0.012	−0.009	0
Inequality	−0.013	0.071	−0.083
Unemployment rate	0.092	0.019	0.011
Below high school	−0.018	−0.055	0
College no degree	0.020	0.003	0
Some college	0.025	−0.013	0
Bachelor or up	−0.052	0.124	−0.044
Property value	−0.350	−0.093	−0.067
Public transport	−0.001	0.078	−0.140
No vehicle HH	0.005	0.150	−0.029
Land area	0.026	−0.107	0
Population density	−0.118	0.065	−0.409
Black	0.188	0.244	0
Hispanic	−0.044	0.288	−0.250
Asian	−0.029	0.017	−0.059
Native	0.006	0	0
Pacific islander	0.018	−0.009	0
Rural	0.429	−0.646	1.019
(Intercept)	−0.878	−0.934	0.044
Obs.	11,383	11,543	7,198

Note: LASSO estimates above are obtained on the test sample using the model described in the paper. The total sample is (11,730 Desert + 12,266 Swamp + 26,216 Healthful) 50,212 tracts, all of which are used in the prediction of mRFEI. However, the columns represent pairwise combinations of binary responses. Definitions: (A) Target = desert, benchmark = healthful, swamp tracts are not included in the sample, (B) target = swamp, benchmark = healthful, desert tracts are not included in the sample, and (C) target = desert, benchmark = swamp, healthful tracts are not included in the sample. In each case, the sample is randomly split into training (70%) and test (30%) data. Training implies choosing the best hyperparameters from a grid of hyperparameters with 10-fold cross-validation. Test refers to the independent sample held out during training. Standard errors for LASSO estimates are omitted because they are redundant for feature extraction and do not have a straightforward formula. LASSO reduced dimension for the Swamp vs. Healthful and Desert vs. Swamp models by pushing some of the coefficients to zero. The non-zero estimates are better suited for feature extraction than making inference because they can be biased and inconsistent.

Table A4
Model performance statistics: rural-urban comparison (Test data).

			Urban			Rural	
Model	Statistic	RF	GB	LASSO	RF	GB	LASSO
(A)	Accuracy	0.751	0.750	0.746	0.613	0.613	0.590
Desert vs.	95% CI	(0.742, 0.761)	(0.741, 0.759)	(0.737, 0.756)	(0.595, 0.630)	(0.595, 0.630)	(0.573, 0.608)
Healthful	Sensitivity	0.970	0.974	1.000	0.676	0.697	0.665
	Specificity	0.107	0.093	0.000	0.540	0.515	0.504
	Kappa	0.105	0.091	0.000	0.217	0.214	0.170
	N	8372	8372	8372	3010	3010	3010
(B)	Accuracy	0.687	0.688	0.665	0.912	0.910	0.912
Swamp vs.	95% CI	(0.678, 0.697)	(0.678, 0.697)	(0.656, 0.675)	(0.898, 0.925)	(0.896, 0.923)	(0.898, 0.925)
Healthful	Sensitivity	0.905	0.910	0.906	1.000	0.998	1.000
	Specificity	0.301	0.292	0.239	0.006	0.006	0.000
	Kappa	0.233	0.230	0.166	0.012	0.007	0.000
	N	9771	9771	9771	1771	1771	1771
(C)	Accuracy	0.678	0.680	0.658	0.903	0.903	0.899
Desert vs.	95% CI	(0.666, 0.690)	(0.668, 0.692)	(0.645, 0.670)	(0.887, 0.918)	(0.887, 0.917)	(0.883, 0.914)
Swamp	Sensitivity	0.838	0.842	0.884	0.147	0.167	0.000
	Specificity	0.412	0.411	0.282	0.988	0.985	1.000
	Kappa	0.267	0.271	0.186	0.202	0.220	0.000
	N	5645	5645	5645	1551	1551	1551
mRFEI	NRMSE	96.6	96.9	98.0	99.1	98.4	98.5
	N	11,894	11,894	11,894	3167	3167	3167

Note: N = respective observations, CI = confidence intervals, Sensitivity = True Positive Rate, Specificity = True Negative Rate, Kappa is the rate of agreement between actual and predicted classes beyond what is expected by chance. NRMSE = Normalized Root-Mean-Squared Error. All accuracy estimates are significant at p-values < 0.01. The total sample is (11,730 Desert + 12,266 Swamp + 26,216 Healthful) 50,212 tracts, all of which are used in the prediction of mRFEI. However, the models represents three binary responses: (A) target = desert, benchmark = healthful, swamp tracts are not included in the sample, (B) target = swamp, benchmark = healthful, desert tracts are not included in the sample, and (C) target = desert, benchmark = swamp, healthful tracts are not included in the sample. mRFEI is a continuous response. In each case, the sample is randomly split into training (70%) and test (30%) data. Training implies choosing the best hyperparameters from a grid of hyperparameters with 10-fold cross-validation. Test refers to the independent sample held out during training.

References

Adams, E., 2020. Our unjust food system is a driver of racial disparities in COVID-19 deaths. King County Politics. May 13, 2020. Retrieved from https://www.kingscountypolitics.com/op-ed-our-unjust-food-system-is-a-driver-of-racial-disparities-in-covid-19-deaths/.

Allcott, H., Diamond, R., Dubé, J., Handbury, J., Rahkovsky, I., Schnell, M., 2019. Food deserts and the causes of nutritional inequality. Qtrly. J. Econ. 134 (4), 1793–1844.

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26 (10), 1340–1347.

Alviola, P.A., Nayga, R.M., Thomsen, M.R., Wang, Z., 2013. Determinants of food deserts. Am. J. Agric. Econ. 95 (5), 1259–1265.

Alwitt, L.F., Donley, T.D., 1997. Retail stores in poor urban neighborhoods. J. Consumer Affairs 31 (1), 139–164.

Apparicio, P., Cloutier, M.S., Shearmur, R., 2007. The case of Montreal's missing food deserts: evaluation of accessibility to food supermarkets. Int. J. Health Geographics 6 (1), 4–10.

Athey, S., Imbens, G.W., 2019. Machine learning methods that economists should know about. arXiv no. 1903.10075.

Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. Annals Stat. 47 (2), 1148–1178.

Bajari, P., Nekipelov, D., Ryan, S.P., Yang, M., 2015. Machine learning methods for demand estimation. Am. Econ. Rev. 105 (5), 481–485.

Baker, E., Schootman, M., Barnidge, E., Kelly, C., 2006. Access to foods that enable individuals to adhere to dietary guidelines: the role of race and poverty. Preventing Chronic Disease 3 (3), 1–11.

Ball, K., Timperio, A., Crawford, D., 2009. Neighbourhood socioeconomic inequalities in food access and affordability. Health and Place 15 (2), 578–585.

Bartlett, S., Klerman, J., Olsho, L., Logan, C., Blocklin, M., Beauregard, M., Enver, A., Wilde, P., Owens, C., Melhem, M., 2014. Evaluation of the Healthy Incentives Pilot: Final Report, Prepared by Abt Associates for the U.S. Department of Agriculture, Food and Nutrition Service.

Berg, N., Murdoch, J., 2008. Access to grocery stores in dallas. Int. J. Behav. and Healthcare Res. 1 (1), 22–37.

Berkowitz, S.A., Karter, A.J., Corbie-Smith, G., Seligman, H.K., Ackroyd, S.A., Barnard, L. S., Atlas, S.J., Wexler, D.J., 2018. Food insecurity, food "deserts", and glycemic control in patients with diabetes: a longitudinal analysis. Diabetes Care 41 (6), 1188–1195.

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25 (2), 197–227.

Block, D., Chavez, N., Birgin, J., 2008. Finding Food in Chicago and the Suburbs: Report of the Northeastern Illinois Community Food Security Assessment, Report to the Public, Chicago State Univ. Frederick Blum Neighborhood Assistance Center and Univ. of Illinois-Chicago School of Public Health, Div. Community Health Sci.

Block, D., Kouba, J., 2006. A comparison of the availability and affordability of a market basket in two communities in the Chicago area. Pub. Health Nutrition 9 (7), 837–845.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breyer, B., Voss-Andreae, A., 2013. Food mirages: geographic and economic barriers to healthful food access in Portland, Oregon. Health and Place 24, 131–139.

Buderer, N.M.F., 1996. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. Academic Emerg. Med. 3 (9), 895–900.

Carroll, K.A., Samek, A., 2018. Field experiments on food choice in grocery stores: A 'how-to' guide. Food Policy 79, 331–340.

Centers for Disease Control and Prevention (CDC), 2009. State indicator report on fruits and vegetables. Atlanta: CDC.

CDC, 2011a. Children's food environment state indicator report. CDC, Atlanta.

CDC, 2011b. Census Tract Level State Maps of the Modified Retail Food Environment Index (mRFEI). Retrieved from https://www.cdc.gov/obesity/downloads/census-tract-level-state-maps-mrfei_TAG508.pdf.

Chen T., Guestrin C., 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining, pp. 785-794.

Chen, X., Clark, J., 2013. Interactive three-dimensional geovisualization of space–time access to food. Appl. Geog. 43, 81–86.

Chen, X., Yang, X., 2014. Does food environment influence food choices? A geographical analysis through tweets. Appl. Geog. 51, 82–89.

Chu, K., 1999. An introduction to sensitivity, specificity, predictive values, and likelihood ratios. Emergency Medicine. 11 (3), 175–181.

Chung, C., Myers Jr., S.L., 1999. Do the poor pay more for food? An analysis of grocery store availability and food price disparities. J. Cons. Affairs 33 (2), 276–296.

Cohen, J.A., 1960. Coefficient of agreement for nominal scales. Ed. and Psy. Measurement 1960, 37–46.

Cooksey-Stowers, K., Schwartz, M.B., Brownell, K.D., 2017. Food swamps predict obesity rates better than food deserts in the United States. Int. J. Env. Res. and Pub. Health 14 (1366): 1-20.

Courtemanche, C., Carden, A., Zhou, X., Ndirangu, M., 2019. Do Walmart supercenters improve food security? Appl. Econ. Perspectives and Policy 41 (2), 177–198.

Crockett, E.G., Clancy, K.L., Bowering, J., 1992. Comparing the cost of a thrifty food plan market basket in three areas of New York state. J. Nutr. Ed. 24 (1), 71S–78S.

Davis, J., Heller, S.B., 2017. Using causal forests to predict treatment heterogeneity: An application to summer jobs. Am. Econ. Rev. 107 (5), 546–550.

Drewnowski, A., Barratt-Fornell, A., 2004. Do healthier diets cost more? Nutr. Today 39 (4), 161–168.

Drewnowski, A., Specter, S.E., 2004. Poverty and obesity: the role of energy density and energy costs. Am. J. Clin. Nutr. 79 (1), 6–16.

Dutko, P., Ver Ploeg, M., Farrigan, T., 2012. Characteristics and influential factors of food deserts, No. 1477–2017-3995. USDA, Washington DC.

Feather, P.M., 2003. Valuing food store access: Policy implications for the food stamp program. Am. J. Agric. Econ. 85 (1), 162–172.

Fitzpatrick, K., Greenhalgh-Stanley, N., Ver Ploeg, M., 2015. The impact of food deserts on food insufficiency and SNAP participation among the elderly. Am. J. Agric. Econ. 98 (1), 19–40.

Fitzpatrick, K., Greenhalgh-Stanley, N., Ver Ploeg, M., 2019. Food deserts and diet-related health outcomes of the elderly. Food Policy 87, 1–8.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Statist. Software 33 (1), 1–22.

Goodman, M., Thomson, J., Landry, A., 2020. Food environment in the lower mississippi delta: food deserts, food swamps and hot spots. Int. J. Env. Res. and Public Health 17 (10), 1–13.

Grimm, K.A., Moore, L.V., Scanlon, K.S., 2013. Access to healthier food retailers—United States, 2011. CDC Health Disparities and Inequalities Report—United States 62 (3): 1-20.

Gustafson, A.A., Lewis, S., Wilson, C., Jilcott-Pitts, S., 2012. Validation of food store environment secondary data source and the role of neighborhood deprivation in Appalachia. Kentucky. BMC Pub. Health 12 (688), 1–12.

Hager, E.R., Cockerham, A., O'Reilly, N., Harrington, D., Harding, J., Hurley, K.M., Black, M.M., 2016. Food swamps and food deserts in Baltimore City, MD, USA: association with dietary behaviours among urban adolescent girls. Pub. Health Nut. 20 (14), 2598–2607.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd ed. Springer-Verlag, New York.

He, H., Bai, Y., Garcia, E.A., Li, S.A., 2008. Adaptive synthetic sampling approach for imbalanced learning. IEEE Int. Joint Conference Neural Networks 2008, 1322–1328.

Horowitz, C.R., Colson, K.A., Hebert, P.L., Lancaster, K., 2004. Barriers to buying healthy foods for people with diabetes: evidence of environmental disparities. Am. J. Pub. Health 94 (9), 1549–1554.

Hossain, M., Mullally, C., Asadullah, M.N., 2019. Alternatives to calorie-based indicators of food security: An application of machine learning methods. Food Pol. 84, 77–91.

Keenan, N.L., Rosendorf, K.A., 2011. Prevalence of hypertension and controlled hypertension–United States, 2005-2008. Morbidity Mortality Weekly Rep. 60, 94–97.

Kestens, Y., Lebel, A., Daniel, M., Thériault, M., Pampalon, R., 2010. Using experienced activity spaces to measure foodscape exposure. Health and Place 16 (6), 1094–1103.

Laraia, B.A., Siega-Riz, A.M., Kaufman, J.S., Jones, S.J., 2004. Proximity of supermarkets is positively associated with diet quality index for pregnancy. Preventive Med. 39 (5), 869–875.

Larsen, K., Gilliland, J., 2008. Mapping the evolution of 'food deserts' in a Canadian city: Supermarket accessibility in London, Ontario, 1961–2005. Int. J. Health Geog. 7 (1), 1–16.

Larson, N.I., Story, M.T., Nelson, M.C., 2009. Neighborhood environments: disparities in access to healthy foods in the U.S. Am. J. Prev. Med. 36 (1), 74–81.

Laska, M.N., Hearst, M.O., Forsyth, A., Pasch, K.E., Lytle, L., 2010. Neighbourhood food environments: are they associated with adolescent dietary intake, food purchases and weight status? Pub. Health Nutr. 13 (11), 1757–1763.

LeDoux, T.F., Vojnovic, I., 2013. Going outside the neighborhood: The shopping patterns and adaptations of disadvantaged consumers living in the lower eastside neighborhoods of Detroit, Michigan. Health and Place 19, 1–14.

Liu, S., Manson, J.E., Lee, I., Cole, S.R., Hennekens, C.H., Willett, W.C., Buring, J.E., 2000. Fruit and vegetable intake and risk of cardiovascular disease: the women's health study. Am. J. Clin. Nutr. 72 (4), 922–928.

Lockhart, R., Taylor, J., Tibshirani, R.J., Tibshirani, R., 2014. A significance test for the LASSO. Ann. Stat. 42 (2), 413–468.

Luan, H., Law, J., Quick, M., 2015. Identifying food deserts and swamps based on relative healthy food access: a spatio-temporal Bayesian approach. Int. J. Health Geog. 14 (1), 37.

MacDonald, J.M., Nelson Jr., P.E., 1991. Do the poor still pay more? Food price variations in large metropolitan areas. J. Urban Econ. 30 (3), 344–359.

Maguire, E.R., Burgoine, T., Monsivais, P., 2015. Area deprivation and the food environment over time: A repeated cross-sectional study on takeaway outlet density and supermarket presence in Norfolk, UK, 1990–2008. Health and Place 33, 142–147.

Mason, K.E., Bentley, R.J., Kavanagh, A.M., 2013. Fruit and vegetable purchasing and the relative density of healthy and unhealthy food stores: evidence from an Australian multilevel study. J. Epid. and Community Health 67 (3), 231–236.

McEntee, J., Agyeman, J., 2010. Towards the development of a GIS method for identifying rural food deserts: Geographic access in Vermont, USA. Appl. Geography 30 (1), 165–176.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochemia Medica 22 (3), 276–282.

McKinnon, R.A., Reedy, J., Morrissette, M.A., Lytle, L.A., Yaroch, A.L., 2009. Measures of the food environment: a compilation of the literature, 1990–2007. Am. J. Prev. Med. 36 (4), S124–S133.

Mooney, C., 1990. Cost and availability of healthy food choices in a London health district. J. Human Nutr. and Dietetics 3 (2), 111–120.

Moore, L.V., Roux, A.V.D., 2006. Associations of neighborhood characteristics with the location and type of food stores. Am. J. Pub. health 96 (2), 325–331.

Moore, L.V., Roux, A.V.D., Nettleton, J.A., Jacobs Jr., D.R., 2008. Associations of the local food environment with diet quality—a comparison of assessments based on surveys and geographic information systems: the multi-ethnic study of atherosclerosis. Am. J. Epidem. 167 (8), 917–924.

Morland, K., Wing, S., Roux, A.V.D., 2002. The contextual effect of the local food environment on residents' diets: the atherosclerosis risk in communities study. Am. J. of Pub. Health 92 (11), 1761–1768.
North American Industry Classification System. 2007. Retrieved from census.gov/eos/www/naics.
Olsho, L.E.W., Klerman, J.A., Wilde, P.E., Bartlett, S., 2016. Financial incentives increase fruit and vegetable intake among supplemental nutrition assistance program participants: a randomized controlled trial of the USDA Healthy Incentives Pilot. Am. J. Clinical Nutrition 104 (2), 423–435.
Opfer, P.R., 2010. Using GIS technology to identify and analyze "food deserts" on the Southern Oregon coast. Oregon State University, June, Master's Essay.
Patel, D., Cogswell, M.E., John, K., Creel, S., Ayala, C., 2017. Knowledge, attitudes, and behaviors related to sodium intake and reduction among adult consumers in the United States. Am. J. Health Promotion 31 (1), 68–75.
Powell, L.M., Slater, S., Mirtcheva, D., Bao, Y., Chaloupka, F.J., 2007. Food store availability and neighborhood characteristics in the United States. Preventive Med. 44 (3), 189–195.
Racca, P., Casarin, R., Squazzoni, F., Dondio, P., 2016. Resilience of an online financial community to market uncertainty shocks during the recent financial crisis. J. Computational Sci. 16, 190–199.
Rhone, A., Ver Ploeg, M., Dicken, C., Williams, R., Breneman, V., 2017. Low-income and low-supermarket-access census tracts, 2010–2015. USDA Econ. Res, Service EIB, p. 165.
Riboli, E., Norat, T., 2003. Epidemiologic evidence of the protective effect of fruit and vegetables on cancer risk. Am. J. Clinical Nutrition 78 (3), 559S–569S.
Richardson, A.S., Ghosh-Dastidar, M., Beckman, R., Flórez, K.R., DeSantis, A., Collins, R. L., Dubowitz., T., 2017. Can the introduction of a full-service supermarket in a food desert improve residents' economic status and health?. Ann. Epidem. 27 (12): 771–776.
Rose, D., Bodor, J.N., Swalm, C.M., Rice, J.C., Farley, T.A., Hutchinson, P.L., 2009. Deserts in New Orleans? Illustrations of urban food access and implications for policy. National Poverty Center Working Paper. Ann Arbor: Univ. of Michigan.
Rose, D., Richards, R., 2004. Food store access and household fruit and vegetable use among participants in the US Food Stamp Program. Public Health Nutr. 7 (8), 1081–1088.
Sage, J.L., McCracken, V.A., Sage, R.A., 2013. Bridging the gap: Do farmers' markets help alleviate impacts of food deserts? Am. J. Agric. Econ. 95 (5), 1273–1279.
Shapley, L.S., 1953. A value for n-person games. Ann. Mathematics, Study No. 28, in Kuhn, H.W. and Tucker, A.W. (eds.) Contributions to the Theory of Games, v. II, Princeton University Press. Princeton, NJ. Pp. 307-318.
Sharkey, J.R., Horel, S., 2008. Neighborhood socioeconomic deprivation and minority composition are associated with better potential spatial access to the ground-truthed food environment in a large rural area. J. Nutrition 138, 620–627.
Sharpe, P.A., Bell, B.A., Liese, A.D., Wilcox, S., Stucker, J., Hutto, B.E., 2020. Effects of a food hub initiative in a disadvantaged community: a quasi-experimental evaluation. Health & Place 63 (102341), 1–13.
Smoyer-Tomic, K.E., Spence, J.C., Amrhein, C., 2006. Food deserts in the prairies? Supermarket accessibility and neighborhood need in Edmonton Canada. Prof. Geographer 58 (3), 307–326.
Sooman, A., Macintyre, S., Anderson, A., 1993. Scotland's health–a more difficult challenge for some? The price and availability of healthy foods in socially contrasting localities in the west of Scotland. Health Bull. 51 (5), 276–284.
Sparks, A., Bania, N., Leete, L., 2009. Finding food deserts: methodology and measurement of food access in Portland. National Poverty Center retrieved from www.npc.umich.edu/news/events/food-access/index.php, Oregon.
Stern, D., Poti, J.M., Ng, S.W., Robinson, W.R., Gordon-Larsen, P., Popkin, B.M., 2016. Where people shop is not associated with the nutrient quality of packaged foods for any racial-ethnic group in the United States. Am. J. Clin. Nutr. 103, 1125–1134.
Testa, A., Jackson, D.B., 2019. Food insecurity, food deserts, and waist-to-height ratio: Variation by sex and race/ethnicity. J. Community Health 44 (3), 444–450.
Thomsen, M.R., Nayga, R.M., Alviola, P.A., Rouse, H.L., 2015. The effect of food deserts on the body mass index of elementary schoolchildren. Am. J. Agric. Econ. 98 (1), 1–18.
Travers, K.D., Cogdon, A., McDonald, W., Wright, C., Anderson, B., MacLean, D.R., 1997. Availability and cost of heart healthy dietary changes in Nova Scotia. J. Can. Dietetic Assoc. 58, 176–183.
U.S. Census Bureau. 2019a. Geographic Terms and Concepts –Census Tract. Retrieved from https://www2.census.gov/geo/pdfs/reference/GTC_10.pdf.
U.S. Census Bureau. 2019b. American Community Survey 2010. Retrieved from https://www2.census.gov/acs2010_1yr/.
U.S. Dept. Agric. 2009. Access to affordable and nutritious food: Measuring and understanding food deserts and their consequences—Report to Congress, AP-036, USDA, Econ. Res. Service. Retrieved from http://www.ers.usda.gov/Publications/AP/AP036/.
U.S. Depart. Agric., Econ. Res. Service. 2019. Food Access Research Atlas. Retrieved from https://www.ers.usda.gov/data-products/food-access-research-atlas/.
U.S. Dept. of Health and Human Services and U.S. Dept. of Agric. 2015. Dietary Guidelines for Americans 2015-20. 8th Ed. Retrieved from http://health.gov/dietaryguidelines/2015/guidelines/.
Varian, H.R., 2014. Big data: New tricks for econometrics. J. Econ. Perspectives 28 (2), 3–28.
Vaughan, C.A., Cohen, D.A., Ghosh-Dastidar, M., Hunter, G.P., Dubowitz, T., 2017. Where do food desert residents buy most of their junk food? Supermarkets. Pub. Health Nutr. 20, 2608–2616.
Ver Ploeg, M., 2010. Access to affordable, nutritious food is limited in 'food deserts'. Amber Waves 8 (1), 20–27.
Ver Ploeg, M., Breneman, V., Farrigan, T., Hamrick, K., Hopkins, D., Kaufman, P., Lin, B. H., Nord, M., Smith, T.A., Williams, R. Kinnison, K., 2009. Access to affordable and nutritious food: measuring and understanding food deserts and their consequences: report to congress (No. 2238-2019-2924). Washington DC: USDA.
Ver Ploeg, M., Wilde, P.E., 2018. How do food retail choices vary within and between food retail environments? Food Pol. 79, 300–308.
Volpe, R., Okrent, A., Leibtag, E., 2013. The effect of supercenter-format stores on the healthfulness of consumers' grocery purchases. Am. J. Agric. Econ. 95 (3), 568–589.
Walker, R.E., Keane, C.R., Burke, J.G., 2010. Disparities and access to healthy food in the United States: A review of food deserts literature. Health and Place 16 (5), 876–884.
Walker, R.E., Fryer, C.S., Butler, J., Keane, C.R., Kriska, A., Burke, J.G., 2011. Factors influencing food buying practices in residents of a low-income food desert and a low-income food oasis. J. Mixed Methods Res. 5 (3), 247–267.
Walker, R.E., Block, J., Kawachi, I., 2012. Do residents of food deserts express different food buying preferences compared to residents of food oases? A mixed-methods analysis. Int. J. Behav. Nutr. Phys. Activity 9 (1), 41–48.
Wang, H., Tao, L., Qiu, F., Lu, W., 2016. The role of socio-economic status and spatial effects on fresh food access: two case studies in Canada. Appl. Geography 67, 27–38.
Yang, M., Wang, H., Qiu, F., 2020. Neighbourhood food environments revisited: When food deserts meet food swamps. Can. Geographer 64 (1), 135–154.
Zenk, S.N., Schulz, A.J., Israel, B.A., James, S.A., Bao, S., Wilson, M.L., 2006. Fruit and vegetable access differs by community racial composition and socioeconomic position in Detroit. Michigan. Ethnicity and Disease 16 (1), 275–280.