REGIONAL PREDICTORS OF THE ESTABLISHMENT, CLOSURE, AND RELOCATION OF FOOD RETAILERS IN THE LONG RUN

Syed Badruddoza[1], Modhurima Dey Amin[1], Jill McCluskey[2], and Wilson Sinclair[3]

Abstract

We used machine learning models to analyze the factors influencing the location decisions of food retailers across the United States from 2011 to 2019. We examined both short-run and long-run establishments of 13 different food store types and considered 35 socioeconomic predictors at the census tract level. Our findings revealed that the growth rates of food store establishments varied depending on the store type and time period. Convenience stores, fast food restaurants, and dollar stores were more likely to experience higher growth in areas with a higher proportion of black population, while grocery stores showed lower establishment rates in these areas. Additionally, areas with higher levels of petty crime had fewer overall store establishments, but convenience stores, dollar stores, and fast food restaurants had higher growth rates in high crime areas compared to grocery stores. Tracts with higher median household income witnessed greater expansion across all store types. Key predictors for food store establishments included the presence of full-service restaurants, percentage of rural population, presence of specialty food stores, and demographic factors. These findings emphasize the concentration effect and provide valuable insights for decision-making and resource allocation to meet the needs of communities.

Keywords: Food retailers, location decisions, machine learning, nutritional inequality, United States

JEL codes: I14; Q18; L81.

Disclaimer: The findings and conclusions in this manuscript are those of the authors and should not be construed to represent any official USDA or US Government determination or policy. This study was supported in part by the US Department of Agriculture, Economic Research Service. The views expressed here cannot be attributed to the Coleridge Initiative.

[1] Texas Tech University. Syed.Badruddoza@ttu.edu
[2] Washington State University.
[3] USDA Economic Research Service.

REGIONAL PREDICTORS OF THE ESTABLISHMENT, CLOSURE, AND RELOCATION OF FOOD RETAILERS IN THE LONG RUN

Residents of socioeconomically disadvantaged neighborhoods, especially non-white individuals, 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; Franco et al. 2008; Moore and Roux 2006; Powell et al. 2007; Laska et al. 2010). In socioeconomically deprived areas, the availability of healthful food choices decreases, leading to increased prices (Crockett et al. 1992; Larsen and Gilliand 2009; Horowitz et al. 2004). The causality between access to and demand for healthful food can operate in either direction. Nutritious foods often come at a premium price that low-income households cannot afford, even if they had access to such foods (e.g., Allcott et al. 2019). This creates a disincentive for healthful food retailers to locate in low-income neighborhoods. On the other hand, the absence of healthful food retailers in the area limits the choices available to residents and puts upward pressure on prices. In either case, the combination of limited income and restricted access to healthful food makes marginalized communities more vulnerable to poor dietary choices, inadequate nutrition, diseases, and unexpected health shocks, as evidenced during the pandemic.

The economics literature has primarily focused on exploring the correlation between food access and health-related outcomes. While the lack of access to healthful foods may contribute to health problems, there may also exist unobserved factors that jointly influence both access to healthful food and poor health outcomes. A selection problem arises because the available data alone cannot reveal the health outcomes of individuals who lack access to healthful foods, as it does not account for the potential health improvements that would occur if access were available.

A number of studies have employed machine learning techniques to understand the food retailing industry. For example, Amin et al. (2020) utilized machine learning models to demonstrate that demographic features can predict access to healthful food with 72% out-of-sample accuracy. The study found that the Hispanic and African American populations were the most important predictors of limited access to healthful food, especially in urban areas. While this study provides insights into the overall patterns of the retail food environment, more research is needed to gain a deeper understanding of the specific types and characteristics of retailers and their location decisions. The existing literature does not provide detailed insights into the dynamics of business decisions and market adjustments in the long run, particularly in urban areas where less healthful retailers substantially outnumber their healthful counterparts.

This paper utilizes census tract-level information to investigate the long-term establishment, closure, and relocation patterns of US food retailers between 2011 and 2019. The study combines data from the National Establishment Time-Series (NETS), socio-demographic information at the census tract level (including crime rates and neighborhood amenities), and data on predictors from 2010 to avoid simultaneity issues. To improve predictions and address functional form misspecification, the study employs four high-performing machine learning models: Random Forests, eXtreme Gradient Boosting, Neural Networks, Boosted OLS, and Boosted logit. Additionally, the presence of food retailers in 2010 is controlled for to account for the initial condition problem.

The outcome variables considered are the establishment, closure, and relocation of food retailers at the census tract level, categorized into four types: large retailers, small retailers, fast food restaurants, and full-service restaurants to ensure brand anonymity. The study finds that the location decisions of food retailers are significantly influenced by the presence and density of the African American population and other store types, particularly complementary ones. This effect is particularly pronounced for small retailers, fast food restaurants, and full-service restaurants. Walkability, crime rates, and road density also play a role in the establishment of food stores.

The primary objective of this study is to explore the factors that shape the location choices of food retailers at the national level. It specifically examines socioeconomic and regional factors, such as race-based inequality, income disparities, and characteristics of rural populations. By analyzing these determinants, the study aims to shed light on the underlying factors that drive the selection of locations for food retailers across the country. The research questions addressed in this study revolve around understanding how these factors influence the establishment, closure, and relocation of food retailers.

The current version of the paper focuses on analyzing the establishment of food retailers in both the short run and long run, while ongoing research is dedicated to examining the closure and relocation aspects. The findings of this study may stimulate discussions on the heterogeneous patterns of food demand based on regional characteristics and underscore the importance of tailored local solutions to address nutritional inequality.

APPROACH

In this study, we propose a model to examine the relationship between socio-demographic features of census tracts and the establishment of food retailers. We aim to understand the impact of these features on the availability and accessibility of food retail options within different communities. Socioeconomic features include variables such as population density, income levels, racial distribution, educational attainment, crime rates, and neighborhood amenities including the number of existing food stores by category. The functional form is unknown, and it captures the complex relationship between these socio-demographic factors and the food retailer establishment. Using predictors from 2010, and analyzing the establishment of food stores in 2011 and 2011-19 allows us to analyze how different socio-demographic characteristics influence the location decision and dynamics of food retailers in a given area. This will provide valuable insights into the underlying mechanisms shaping the food environment and its impact on public health.

We adopt a machine learning (ML) approach, as it has demonstrated high predictive performance in out-of-sample scenarios (Bajari et al., 2015). ML models are data-driven, reducing reliance on prior assumptions regarding the functional form of the relationship. Moreover, these models offer flexibility by optimally selecting parameters through grid search. Given the unknown exact functional relationship between food access and its predictors, the ML approach is well-suited for our analysis.

The learning aspect of ML involves tuning the hyperparameters based on error rates in repeated subsamples. We divide the data from into training (80%) and testing (20%) samples. The training sample is used for model training with ten-fold cross-validation. This entails randomly partitioning the training sample into ten equal-sized subsamples, using nine for training and one for validation. This process is repeated ten times, each time with a different partition, and hyperparameters are updated to minimize prediction errors in the validation set. The learning process concludes when optimal hyperparameters are achieved. Finally, the trained model is evaluated on the held-out testing sample (20%).

We employ the following models: (1) Boosted Ordinary Least Squares (OLS), (2) Boosted Logit (BLogit), (3) Random Forest (RF), (4) eXtreme Gradient Boosting (XGB), and (5) Artificial Neural Networks (ANN). Boosted OLS, RF, XGB, and ANN are utilized for predicting the continuous response variable, such as the number of food retailers opened in 2011-19, while Boosted Logit, RF, XGB, and ANN are employed for predicting binary responses (experience-based measures). RF and XGB are utilized in regression form for the former and classification format for the latter. These models are chosen for their superior predictive performance compared to other models such as tree models or support vector machine (Chen and Guestrin, 2016). A non-technical description of these models is provided below.

Boosted OLS is a modeling technique that combines ordinary least squares regression with boosting, an iterative algorithm that builds a strong predictive model. It is effective for handling complex datasets with non-linear relationships. Weak learners, such as simple regression models, are sequentially added to create an ensemble model. Boosted OLS captures non-linear relationships and interactions while maintaining interpretability. It is commonly used in various fields for predictive modeling and identifying important predictors in high-dimensional datasets. Similarly, boosted Logit differs from conventional logit regression by iteratively fine-tuning coefficient estimates during cross-validation.

RF and XGB are ensemble tree-based models. In ML, decision trees select a regressor and split the sample into two parts, choosing the split that minimizes prediction error for the response variable (Quinlan, 1993). However, tree models without pruning and randomization may overfit the data and yield weak predictions if the trees are highly correlated (Hastie et al., 2009). Random Forest addresses this issue by growing a large number of decision trees during training with repeated subsampling and randomly chosen predictors. Each resampled portion of the training data is used independently to decorrelate the trees. The overall prediction is generated by averaging predictions from all trees in regression tasks and through majority voting in classification tasks.

The XGBoost algorithm, on the other hand, is a variation of gradient boosting that incorporates a set of decision trees. Unlike traditional gradient boosting, each new tree in XGBoost adjusts the weights on predictors based on the errors from fitting the previous tree. This enhancement is achieved through parallel processing and regularization techniques, which help mitigate overfitting and bias (Chen and Guestrin, 2016).

In contrast, Artificial Neural Networks (ANN) do not rely on a tree-based framework. Instead, they simulate the behavior of the human brain, where interconnected neurons learn from experience. ANN assigns weights to predictors and calculates errors in predicting the response variable. Linear regression and logistic regression can be seen as special cases of simple ANN models, featuring input and output layers without any hidden layers. When a hidden layer is introduced, predictors are initially assigned random weights, and the hidden layer nodes make predictions using an activation function (often linear for regression and sigmoidal for classification tasks). The output layer is then predicted

based on the hidden layer's output using the same activation function. Through iterations, such as forward feeding or backpropagation, prediction errors are minimized, and predictor weights are adjusted at each iteration. More complex ANN architectures can include additional hidden layers (Herbrich et al., 1999) for enhanced modeling capabilities.

Machine learning models are not immune to certain issues. Firstly, the presence of highly correlated predictors can complicate the assessment of their relative importance. To address this, we follow existing literature and interpretability considerations and exclude predictors that exceed a correlation threshold of 0.95 with other predictors. Secondly, class imbalance issues may arise when there are fewer target cases (e.g., food insecure households) compared to benchmark cases. To mitigate this, we employ the Adaptive Synthetic Sampling Approach (He et al., 2008), which randomly oversamples the target cases to balance the representation of the target and benchmark classes. Thirdly, the importance of continuous predictors may appear skewed if not standardized. To ensure fair comparison, we standardize all continuous predictors within the model. Finally, one limitation of machine learning models is their limited ability to provide clear interpretations of predictor weights. Importance factors for a predictor are typically calculated by assessing the change in prediction errors when including or excluding the predictor. However, this calculation is sensitive to the order of entry and combination of predictors.

To address this limitation, we employ Shapley values (SHAP) to determine the relative contribution of each regressor to the dependent variable (Chen et al., 2021; Shapley, 1951). SHAP values are computed as the weighted sum of prediction differences with and without each predictor, where the weights are derived from all possible combinations of predictors with different orderings. Each observation is assigned a SHAP value, resulting in a SHAP matrix of the same dimension as the predictor matrix. The sum of SHAP values across columns represents the variation beyond the overall mean of the predicted variable for each observation. By multiplying SHAP values with the sign of the correlation coefficient between SHAP and a predictor, we can ascertain the direction of the association. Thus, larger SHAP values indicate greater relative importance of the predictor, while smaller values indicate the opposite. This approach allows us to gain insights into the relative importance of predictors in a more meaningful and interpretable manner.

To evaluate the performance of machine learning (ML) models, we adopt the evaluation metrics recommended by Amin et al. (2021). For binary target variables, we utilize accuracy. In the case of continuous response variables, we employ the normalized root mean squared error (NRMSE). Later version of the current paper will include more robust metrics of the performance. The accuracy metric is computed as the sum of the true positive and true negative predictions divided by the sum of all predictions in the confusion matrix. It provides an overall measure of the model's correctness in predicting both positive and negative cases. The NRMSE represents the root mean squared error (RMSE) normalized by the range of the response variable. It provides a standardized measure of the model's prediction accuracy.

It is important to note that our approach focuses on the predictive capacity of the models rather than establishing causal relationships. By utilizing machine learning techniques, we can construct flexible and data-driven models that minimize the out-of-sample prediction errors. This allows us to account for the varying factors not captured in the data and obtain accurate predictions for our target variables.

The dataset used in our analysis consists of three type of response variables (y variables) related to the establishment of food retailers. The first response variable type represents the number of food retailers opened in 2011, providing insights into the short-run scenario. The second type reflects the number of food retailers established between 2011 and 2019, offering a perspective on the long-run scenario. Finally, the third response variable type is a binary variable indicating whether at least one food retailer was established in the census tract during the 2011-2019 period, allowing us to examine the presence or absence of food retailers. These variables are categorized into 13 different store types, capturing various aspects of the food retail industry.

The 13 store types encompass a wide range of food retail categories. These include groceries, convenience stores, supercenters (which combine a supermarket with a discount department store), department stores, other general retailers, limited-service (fast food) restaurants, full-service restaurants, cafeterias, snack bars, drinking places, beer retailers, specialty food stores, and dollar stores. Each store type represents a distinct aspect of the food retail industry, offering a comprehensive view of the establishments and their presence within the census tracts. We discuss data in more detail below.

It is important to note that, we use density-based measure, e.g., number of store established in a census tract, instead of distance-based measure, such as meters from household to the nearest grocery. The use of distance as a measure for proximity to grocery stores presents both advantages and disadvantages. However, it is important to consider the limitations of this measure for this exercise. First, with the increasing prevalence of online shopping, the relevance of distance to physical stores may be diminishing. Online orders allow consumers to access groceries without the need to

consider geographical proximity. Second, when measuring distance from the centroid of a tract, it can result in a substantial aggregation of data across households, leading to imprecision in assessing individual households' accessibility. Third, relying solely on distance fails to account for the complex dynamics of substitute and complement effects in shopping behavior. For instance, a consumer may opt not to visit a drug store if a dollar store is within reach, even if it is slightly farther away. Fourth, measuring distance individually for each household is a costly and impractical endeavor, particularly when considering the entire United States. Lastly, using driving distance or meters as a proxy for distance is imperfect, as the mode of transportation can vary significantly among households. Overall, we argue that density-based measure is more suitable for understanding the long-run trend in food store establishments across the country.

DATA

We collect information from multiple sources to provide comprehensive insights. We utilize data from the US census 2010 for our analysis, with the US census tracts serving as our units of observation. The data provides comprehensive social, economic, housing, and demographic information at the geographical level, offering data in the form of population counts, percentages, and aggregate measures such as means and medians. Our choice of predictor variables is guided by Amin et al. (2021), who employ machine learning models to predict access to healthful food retailers. They identify a set of 281 predictors, shortlist 50 variables based on Shapley values, and ultimately select 20 variables that best predict the modified Retail Food Environment Index while remaining interpretable and consistent with existing literature. We anticipate that these same variables will be relevant for predicting the establishment of food retailers. These predictors encompass various characteristics at the census tract level, such as age, race and ethnicity, unemployment, education, mode of transportation to work, median household income, poverty rate, and vehicle availability. We further add crime rate, road density, and walkability index. The walkability index is collected from the Environmental Protection Agency.

In our dataset, we initially have a total of 72,538 census tracts. After removing tracts with missing information, we are left with 67,618 tracts, covering approximately 97% of the US population and over 95% of the land area in the contiguous states. Table 1 provides a summary of the selected predictors, including variables such as median household income, poverty rate, the proportion of housing units receiving Supplemental Nutrition Assistance Program (SNAP) benefits, inequality as measured by the Gini Index, unemployment rate, educational attainment levels, property values, public transportation usage, vehicle availability, land area, population density, and the percentages of various racial and ethnic populations.

The National Establishment Time Series (NETS) data is used to record the annual establishment, employment, and location of food retailers identified by the North American Industry Classification System (NAICS) codes. We use the starting year information to capture the establishment of the food store types, and NAICS code to identify the food retailer categories. The food retailer categories include:

1. Supermarket and other grocery retailers: These establishments primarily engage in retailing a general line of food products, such as canned and frozen foods, fresh fruits and vegetables, and fresh and prepared meats, fish, and poultry.

2. Warehouse Clubs and Supercenters: These establishments are known as supermarkets and other grocery retailers that offer a general line of food products, similar to the above category.

3. Convenience stores: These establishments are primarily engaged in retailing a limited line of goods that typically includes items like milk, bread, soda, and snacks. Some convenience stores may also sell automotive fuels in combination with groceries.

4. Specialized food stores: This category includes establishments primarily engaged in retailing specific types of food products. Examples include stores selling fresh, frozen, or cured meats and poultry, fresh fruits and vegetables, baked goods, candy and confections, and other specialty foods.

In addition to food retailers, the study also considers different types of restaurants:

1. Limited Service Restaurants: These establishments provide food services where patrons order or select items and pay before eating. Food and drink can be consumed on the premises, taken out, or delivered. This category

includes chain vendors (national and regional), takeout vendors (Asian and other cuisines), and other limited-service vendors.

2. Full-Service Restaurants: These establishments provide food services to patrons who order and are served while seated. They may offer waiter/waitress service and the option to consume food and drink on the premises. This category includes casual dining and fine dining restaurants.

3. Cafeterias, Grill Buffets, and Buffets: These establishments prepare and serve meals for immediate consumption using cafeteria-style or buffet serving equipment. Patrons can select from food and drink items displayed in a continuous cafeteria line or buffet stations.

4. Snack and Nonalcoholic Beverage Bars: These establishments primarily focus on serving specialty snacks or nonalcoholic beverages for consumption on or near the premises. Examples include dessert/bakery vendors, cafes, juice/tea bars, and other snack and beverage vendors.

5. Drinking Places (Alcoholic Beverages): These establishments primarily prepare and serve alcoholic beverages for immediate consumption. They may also provide limited food services. Examples include bars, taverns, nightclubs, and drinking places.

The comprehensive data on these various types of food retailers and restaurants enables us to analyze the establishment, closure, and relocation patterns in the context of regional predictors and demographic factors.

RESULTS

Descriptive statistics

Table 1 presents summary statistics in terms of the mean and standard deviations of the variables in our dataset. We have a total of 26 response variables and 35 predictors. These response variables correspond to 13 food store types, with each type having two categories: short-run establishments (2011) and long-run establishments (2011-2019). Thus, we have a total of 26 response variables resulting from the combination of the food store types and establishment periods. The dataset comprises information from 67,618 census tracts.

As expected, the values of the response variables are generally higher for the long-run establishments (2011-2019) compared to the short-run establishments (2011) since food stores may not open in a census tract within a single year but can be established over a period of ten years. On average, there were 0.168 grocery stores per census tract in 2011, while the number increased to 1.364 in the long run. Notably, grocery stores, limited service or fast food restaurants, full-service restaurants, drinking places, and specialty stores experienced significant growth during the period 2011-2019. Particularly, full-service restaurants had a remarkable growth rate, with an average of 5.176 new establishments per census tract in the last decade (2011-2019).

Among the predictor variables, we have included the existing number of stores in 2010 for each store type. On average, there were 4.8 full-service restaurants per census tract, with a standard deviation of 6.3. The second most common food store type was specialty stores, followed by grocery stores.

In addition to the store-related variables, we also incorporated neighborhood socioeconomic factors such as median income, poverty rate, crime rate, etc., which were collected from the US Census Bureau, specifically from the 2010 census. The walkability index, on the other hand, was obtained from the Environmental Production Agency. We selected these predictors based on relevant literature, including the work of Amin et al. (2021).

Figure 1 illustrates the relationship between the growth rates of selected store types (grocery, convenience stores, fast food or limited service, and dollar stores) and various socioeconomic variables. The growth rates, expressed as percentage change, are plotted against four major socioeconomic indicators: the percentage of the black population, median household income, and rural population. In Figure 1, the logarithm of median income is categorized to enhance the visibility of patterns. The choice of these predictor variables is based on the literature highlighting the disparities in food access, particularly in areas with higher proportions of black population, lower income levels, higher crime rates, and rural populations.

Several key observations can be made from Figure 1. Firstly, the establishment of new grocery stores is less prevalent in census tracts with higher black population percentages. Conversely, convenience stores, fast food retailers, and dollar stores exhibit more balanced growth across different population demographics, with a notable presence in tracts with higher black population percentages. This suggests a distinct long-term pattern in the establishment of grocery stores in black neighborhoods.

A similar pattern emerges when examining the relationship between food store establishments and petty crime rates. Tracts with higher levels of petty crime tend to have fewer store establishments in the 2011-2019 period. Most stores, however, were established in areas with lower crime rates. Notably, convenience stores, dollar stores, and fast food restaurants experienced higher growth rates in high crime areas compared to grocery stores. This indicates that factors such as shoplifting may play a role in the location preferences of larger grocery stores that offer a wider variety of fresh produce options.

Tracts with higher median household income demonstrate greater expansion of grocery stores, convenience stores, fast food restaurants, and dollar stores. However, the distribution of dollar stores across income categories is more uniform, indicating substantial growth in both low and high-income neighborhoods compared to grocery stores. Convenience stores also exhibit a more even distribution of growth. Hence, income emerges as a critical determinant in the location choices of these types of stores, with convenience and dollar stores being relatively less likely to exclusively locate in high-income neighborhoods.

Lastly, Figure 1 highlights that all types of stores primarily experienced growth in urban tracts during the 2011-2019 period. The establishment of new grocery stores in rural areas was relatively uncommon, whereas convenience stores, fast food restaurants, and dollar stores showed higher growth rates in rural areas. Particularly, the establishment of new fast food and dollar stores was more likely in rural areas during the specified period.

Overall, Figure 1 provides valuable insights into the spatial patterns of store growth rates and their associations with socioeconomic variables. The findings suggest disparities in the establishment of different store types based on factors such as population demographics, crime rates, income levels, and urban-rural distinctions.

Results from machine learning

Table 2 presents the out-of-sample predictive performance of four models: Boosted OLS, Random Forests, XGBoost, and ANN. The evaluation is based on the normalized root mean squared errors (NRMSE) for both the short-run store openings (food stores that opened only in 2011) and the long-run store openings (food stores that opened in 2011-2019). The NRMSE values, which indicate the accuracy of the predictions, are generally low, indicating a high level of predictive accuracy for our models.

Among the four models, XGBoost demonstrates the best performance overall, with lower NRMSE values compared to the other models. This superior performance may be attributed to the ability of XGBoost to tune a large number of hyperparameters and its robustness against overfitting. Boosted OLS performs particularly well in the case of short-run store openings in 2011. On the other hand, Random Forests outperforms Boosted OLS in many cases for the long-run store openings in 2011-2019. This suggests that while a linear model like Boosted OLS may be effective in predicting the location of food retailers in the short run, a more complex and nonlinear model, such as Random Forests, is required for better predictability in the long run.

The results highlight the importance of using machine learning approaches, such as Random Forests and XGBoost, when dealing with complex and dynamic phenomena like the opening of food stores over an extended period. These models can capture the nonlinear relationships and account for various factors that influence store openings, leading to improved predictive performance.

Table 3 displays the accuracy scores obtained from four models: Boosted Logit, Random Forests, XGBoost, and ANN. In this analysis, the response variables are binary, indicating whether at least one store of the respective category opened in the census tract during the period of 2011-2019. A value of 1 signifies the presence of a store, while a value of 0 indicates no store opening.

The table presents the out-of-sample predictive accuracy for each model, with a focus on 20% of the data. Overall, all models achieve accuracy scores around 76%. However, XGBoost demonstrates superior performance compared to the other models, exhibiting higher average accuracy. In most cases, the accuracy of XGBoost reaches

approximately 80%. Notably, certain store types, such as dollar stores and supercenters, drinking places, and specialty stores, show a stronger association with socioeconomic predictors, resulting in more accurate predictions.

It is important to interpret raw accuracy scores cautiously, as they can be influenced by the randomness of the data. As part of our ongoing research, we are actively considering additional performance indicators such as sensitivity, specificity, and Cohen's Kappa to provide a more robust assessment of model performance. These measures will further enhance our understanding of the models' predictive capabilities and their ability to capture the complexities of store openings in the census tracts.

The high predictive accuracy of the xgboost model suggests its suitability for feature extraction in determining the establishment of various food store types within a census tract. Given the 35 predictor variables, machine learning classifiers prove to be more robust as they do not assume any linear functional form. Consequently, we employ the xgboost model to predict the presence of at least one food store in each category during the 2011-2019 period. This approach offers insights into the factors that contribute to the establishment of different food store types within a census tract.

Figure 2 presents the relative importance of each predictor in predicting the establishment of 12 food store types. These response variables represent the presence of at least one respective food store type in the 2011-2019 period, providing valuable information on the determinants of long-term food store establishment. The relative importance factors are calculated using median Shapley values multiplied by the correlation between the predictor and response variable. Positive importance factors indicate a positive association between the predictor and the response, while negative importance factors indicate a negative association. To facilitate interpretation, Figure 2 includes dashed red lines at the horizontal axis value of 0. A data point to the right of the red line indicates a positive association, while a data point to the left indicates a negative association. The distance between a data point and the red line represents the magnitude of the predictor's relative importance.

The predictor variables are arranged on the y-axis based on the mean absolute value of their relative importance across all plots. Notably, the presence of full-service restaurants exhibits the highest average relative importance in predicting the establishment of various food store types. Following that, the percentage of rural population and the presence of specialty food stores rank as important predictors. Among demographic variables, the black population emerges as a significant predictor, displaying a negative association with the establishment of most food retailer types, particularly supercenters and snack bars. Areas with higher Asian population proportions are less likely to have supercenters. Walkability and college education level also hold predictive value, positively influencing the presence of all food store types. Intriguingly, median income shows a large negative association with the establishment of supercenters when accounting for other predictors. This implies that areas with lower median income have experienced at least one supercenter establishment in the 2011-2019 period, controlling for other factors. It is worth noting that we controlled for poverty rate, households with SNAP benefits, and inequality, all of which display a negative relationship with the presence of supercenters.

Additionally, there is evidence of a concentrating effect for each store category, whereby the likelihood of another store locating in a census tract during the 2011-2019 period increases when similar stores already existed in 2010. Similarly, the presence of a supercenter in the long run is more likely in tracts where another supercenter was already established in 2010. This suggests that the existing composition of stores contributes to market efficiency in the long run, rather than intense competition among similar store types. Notably, the presence of department stores exhibits a negative association with supercenters, indicating competition between these two types of stores.

Finally, regarding limited service or fast food restaurants, convenience stores, and dollar stores, we observe distinct patterns in their location preferences. Dollar stores are more likely to be situated in close proximity to other dollar stores, indicating a concentration effect. However, they are less likely to be found in high-income areas characterized by high property values and road density. Interestingly, dollar stores exhibit a lower likelihood of locating near supercenters but a higher likelihood of being situated near grocery stores.

Fast food restaurants demonstrate a greater propensity for locating in rural areas, particularly in regions where fast food restaurants and department stores are already established. This suggests a preference for areas with existing commercial infrastructure. On the other hand, convenience stores display a lower likelihood of being located in regions with high property values and a more educated customer base. These stores are more commonly found in rural areas, indicating their significance in catering to the needs of residents in these settings.

CONCLUDING REMARKS

This paper examines the establishment, closure, and relocation patterns of US food retailers from 2011 to 2019 using census tract-level data. It combines the National Establishment Time-Series (NETS) dataset with socio-demographic information to predict and analyze these patterns. Advanced machine learning models, including Random Forests, eXtreme Gradient Boosting, Neural Networks, Boosted OLS, and Boosted logit, are employed to improve prediction accuracy. The presence of food retailers in 2010 is controlled for to address the initial condition problem.

In our preliminary analysis, we found that the growth rates of food store establishments varied across store types and time periods. Tracts with higher proportions of black population experienced growth in convenience stores, fast food restaurants, and dollar stores, while grocery stores showed lower establishment rates in these areas. Higher levels of petty crime were associated with fewer store establishments overall, but convenience stores, dollar stores, and fast food restaurants exhibited higher growth rates in high crime areas compared to grocery stores. Tracts with higher median household income demonstrated greater expansion across all store types. XGBoost consistently outperformed other models in terms of predictive accuracy, indicating its suitability for feature extraction.

Additionally, our analysis revealed important predictors in determining the establishment of food stores. Full-service restaurants, the percentage of rural population, the presence of specialty food stores, and demographics were key factors influencing the growth of food store establishments. Demographic variables such as the black population and Asian population played significant roles in explaining the establishment of supercenters in 2011-19. Furthermore, the analysis highlighted the concentration effect, where the likelihood of another store locating in a census tract increased when similar stores already existed. These findings provide valuable insights into spatial dynamics and factors influencing the location choices and growth patterns of food stores, enabling stakeholders to make informed decisions and allocate resources effectively to meet the diverse needs of communities.

REFERENCES

Allcott, H., Diamond, R., Dubé, J.P., Handbury, J., Rahkovsky, I. and Schnell, M., 2019. Food deserts and the causes of nutritional inequality. *The Quarterly Journal of Economics*, *134*(4), pp.1793-1844.

Alwitt, L.F. and Donley, T.D., 1997. Retail stores in poor urban neighborhoods. Journal of consumer affairs, 31(1), pp.139-164.

Amin, M.D., Badruddoza, S. and McCluskey, J.J., 2021. Predicting access to healthful food retailers with machine learning. *Food Policy*, *99*, p.101985.

Bajari, P., Nekipelov, D., Ryan, S.P. and Yang, M., 2015. Machine learning methods for demand estimation. *American Economic Review*, *105*(5), pp.481-485.

Baker, E.A., Schootman, M., Barnidge, E. and Kelly, C., 2006. Peer reviewed: The role of race and poverty in access to foods that enable individuals to adhere to dietary guidelines. Preventing chronic disease, 3(3).

Chen, H., Lundberg, S. and Lee, S.I., 2021. Explaining models by propagating Shapley values of local components. *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability*, pp.261-270.

Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Chung, C. and Myers Jr, S.L., 1999. Do the poor pay more for food? An analysis of grocery store availability and food price disparities. Journal of consumer affairs, 33(2), pp.276-296.

Crockett, S.J. and Sims, L.S., 1995. Environmental influences on children's eating. *Journal of Nutrition Education*, *27*(5), pp.235-249.

Franco, M., Roux, A.V.D., Glass, T.A., Caballero, B. and Brancati, F.L., 2008. Neighborhood characteristics and availability of healthy foods in Baltimore. American journal of preventive medicine, 35(6), pp.561-567.

Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H., 2009. *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

He, H., Bai, Y., Garcia, E.A. and Li, S., 2008, June. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). IEEE.

Herbrich, R., Keilbach, M., Graepel, T., Bollmann-Sdorra, P. and Obermayer, K., 1999. Neural networks in economics: Background, applications and new developments. *Computational techniques for modelling learning in economics*, pp.169-196.

Horowitz, C.R., Colson, K.A., Hebert, P.L. and Lancaster, K., 2004. Barriers to buying healthy foods for people with diabetes: evidence of environmental disparities. *American journal of public health*, *94*(9), pp.1549-1554.

Larsen, K. and Gilliland, J., 2009. A farmers' market in a food desert: Evaluating impacts on the price and availability of healthy food. *Health & place*, *15*(4), pp.1158-1162.

Laska, M.N., Hearst, M.O., Forsyth, A., Pasch, K.E. and Lytle, L., 2010. Neighbourhood food environments: are they associated with adolescent dietary intake, food purchases and weight status?. *Public health nutrition*, *13*(11), pp.1757-1763.

Moore, L.V. and Diez Roux, A.V., 2006. Associations of neighborhood characteristics with the location and type of food stores. American journal of public health, 96(2), pp.325-331.

Morland, K., Wing, S., Roux, A.D. and Poole, C., 2002. Neighborhood characteristics associated with the location of food stores and food service places. American journal of preventive medicine, 22(1), pp.23-29.

Powell, L.M., Auld, M.C., Chaloupka, F.J., O'Malley, P.M. and Johnston, L.D., 2007. Associations between access to food stores and adolescent body mass index. *American journal of preventive medicine*, *33*(4), pp.S301-S307.

Quinlan, J.R., 1993, June. Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on machine learning* (pp. 236-243).

Shapley, Lloyd S. (1951). "Notes on the n-Person Game -- II: The Value of an n-Person Game" (PDF). Santa Monica, Calif.: RAND Corporation.

Zenk, S.N., Schulz, A.J., Israel, B.A., James, S.A., Bao, S. and Wilson, M.L., 2006. Fruit and vegetable access differs by community racial composition and socioeconomic position in Detroit, Michigan. Ethnicity & disease, 16(1), pp.275-280.

TABLES

Table 1. Summary statistics

Response variables

	Mean	SD
Stores opened in 2011		
1. Grocery	0.168	0.455
2. Convenience	0.062	0.256
3. Supercenter	0.002	0.041
4. Department	0.040	0.211
5. Other general retailers	0.041	0.207
6. Limited service	0.150	0.436
7. Full service	0.362	0.775
8. Cafeterias	0.007	0.083
9. Snack bars	0.031	0.183
10. Drinking places	0.129	0.415
11. Beer retailers	0.022	0.151
12. Specialty	0.276	0.594
13. Dollar	0.022	0.149
Stores opened in 2011-19		
14. Grocery	1.364	1.833
15. Convenience	0.619	0.915
16. Supercenter	0.009	0.099
17. Department	0.199	0.553
18. Other general retailers	0.557	0.890
19. Limited service	1.348	2.023
20. Full service	5.176	7.005
21. Cafeterias	0.065	0.299
22. Snack bars	0.638	1.117
23. Drinking places	1.286	2.184
24. Beer retailers	0.286	0.681
25. Specialty	2.956	3.009
26. Dollar	0.141	0.408

Predictor variables

	Mean	SD
Existing number in 2010		
1. Grocery	2.324	2.445
2. Convenience	1.514	1.567
3. Supercenter	0.029	0.186
4. Department	0.407	1.039
5. Other general retailers	0.646	0.982
6. Limited service	1.983	2.625
7. Full service	4.852	6.301
8. Cafeterias	0.098	0.369
9. Snack bars	0.583	1.065
10. Drinking places	1.288	2.122
11. Beer retailers	0.621	0.982
12. Specialty	2.776	2.776
13. Dollar	0.183	0.464
Neighborhood predictors		
14. Median income	56,029.060	27,036.040
15. Poverty rate	564.414	494.898
16. HH with SNAP	79.918	96.007
17. Inequality	0.408	0.064

18.	Unemployment	8.331	5.300
19.	Below high school	27.185	23.877
20.	College no degree	7.428	3.318
21.	Some college	17.082	10.270
22.	Bachelors	10.057	9.145
23.	Property	237,181.000	182,739.000
24.	Public transport	26.453	7.479
25.	No vehicle	133.562	210.799
26.	Black	515.604	891.393
27.	Hispanic	668.630	1,098.729
28.	Asian	5.045	27.444
29.	Native	32.724	163.875
30.	Pacific islander	19.678	40.632
31.	Rural population	552.414	908.809
32.	Tract population	4,255.253	1,853.881
33.	Petty crime	24.377	10.621
34.	Walkability	9.614	3.965
35.	Road density	15.474	9.619

N=67,618 census tracts. See data section for details.

Table 2. Out of sample predictive performance of models
Response: Number of stores opened in the respective category

Response variables	Boosted OLS	Random Forests	XGBoost	ANN
Stores opened in 2011				
1. Grocery	0.419	0.431	0.414	0.441
2. Convenience	0.249	0.256	0.244	0.253
3. Supercenter	0.045	0.046	0.040	0.048
4. Department	0.207	0.211	0.202	0.210
5. Other general retailers	0.208	0.212	0.203	0.212
6. Limited service	0.397	0.406	0.392	0.417
7. Full service	0.639	0.644	0.634	0.664
8. Cafeterias	0.088	0.089	0.083	0.089
9. Snack bars	0.179	0.183	0.174	0.183
10. Drinking places	0.380	0.386	0.375	0.393
11. Beer retailers	0.146	0.150	0.141	0.149
12. Specialty	0.533	0.543	0.528	0.559
13. Dollar	0.148	0.152	0.143	0.151
Stores opened in 2011-19				
14. Grocery	1.318	1.307	1.313	1.327
15. Convenience	0.795	0.801	0.790	0.821
16. Supercenter	0.475	0.486	0.470	0.494
17. Department	0.796	0.801	0.791	0.821
18. Other general retailers	1.380	1.360	1.375	1.405
19. Limited service	3.502	3.355	3.497	3.262
20. Full service	0.285	0.282	0.280	0.289
21. Cafeterias	0.879	0.883	0.874	0.909
22. Snack bars	1.488	1.425	1.483	1.435
23. Drinking places	0.587	0.583	0.582	0.603
24. Beer retailers	1.936	1.884	1.931	1.898
25. Specialty	0.091	0.095	0.086	0.093
26. Dollar	0.386	0.390	0.381	0.399

N=67,618 census tracts. The lower RMSE the better. Response variables are number of food stores.

Table 3. Out of sample predictive performance of models
Response: At least one store opened in the respective category (2011-19)

Food store type	Boosted Logistic	Random Forests	XGB	ANN
1. Grocery	0.767	0.770	0.780	0.757
2. Convenience	0.869	0.869	0.879	0.859
3. Supercenter	0.969	0.969	0.979	0.959
4. Department	0.886	0.887	0.897	0.876
5. Other general retailers	0.790	0.792	0.802	0.780
6. Limited service	0.798	0.801	0.811	0.788
7. Full service	0.993	0.993	0.893	0.983
8. Cafeterias	0.877	0.877	0.887	0.867
9. Snack bars	0.783	0.785	0.795	0.773
10. Drinking places	0.952	0.952	0.962	0.942
11. Beer retailers	0.767	0.768	0.778	0.757
12. Specialty	0.900	0.890	0.910	0.970
13. Dollar	0.987	0.987	0.987	0.977

N=67,618 census tracts. The greater accuracy scores the better. Response variables are binary.

FIGURES

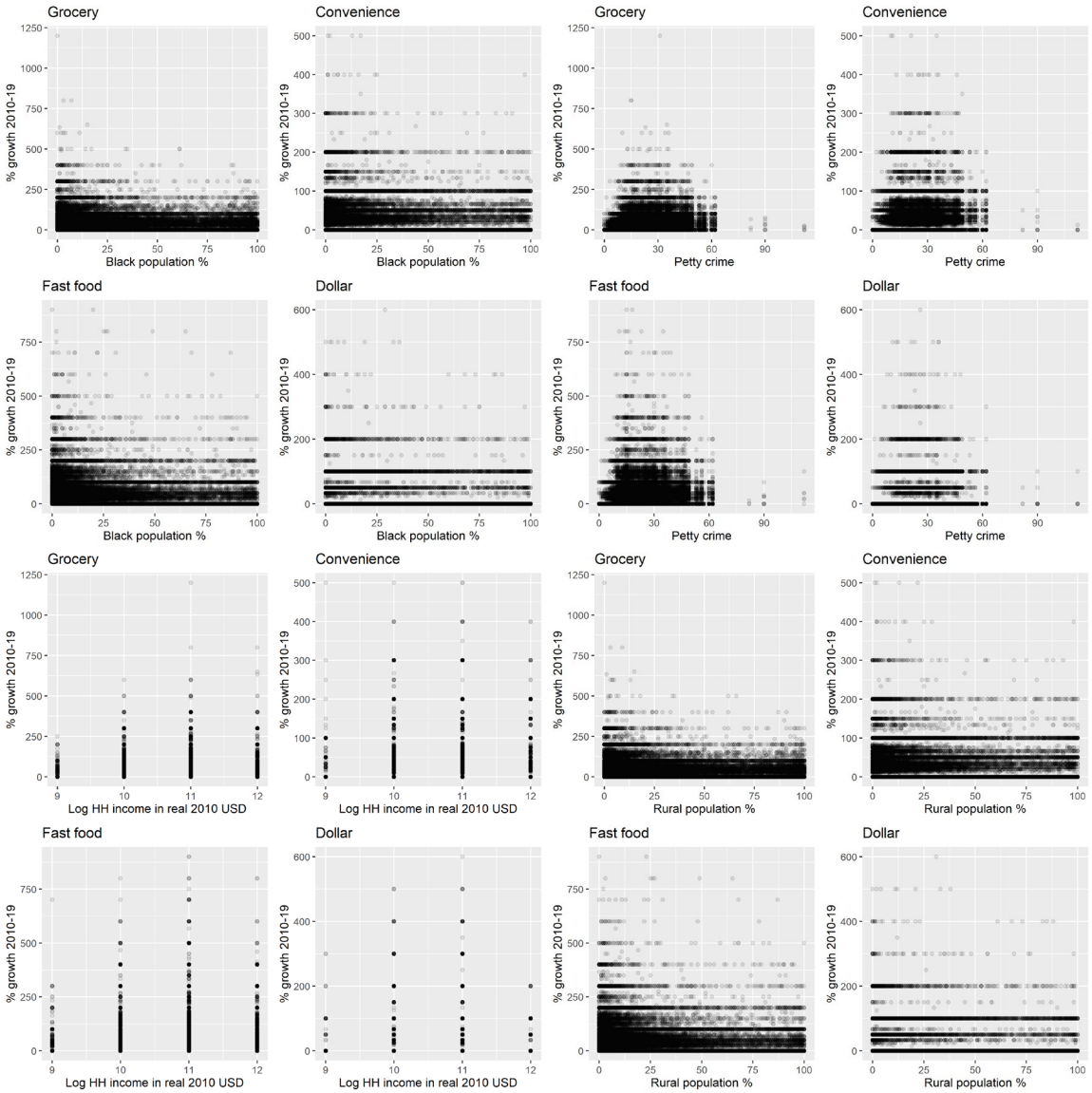

Fig 1. Percentage change in the number of retailers in 2011-19 compared to 2010. The figure shows four common store types, grocery, convenience, fast food or limited service, and dollar stores. Their % change was plotted against four major socio-economic variables: black population %, petty crime index, household median income, and rural population. Log of median income is categorized in this graph for clear visibility of pattern. See data section for details.

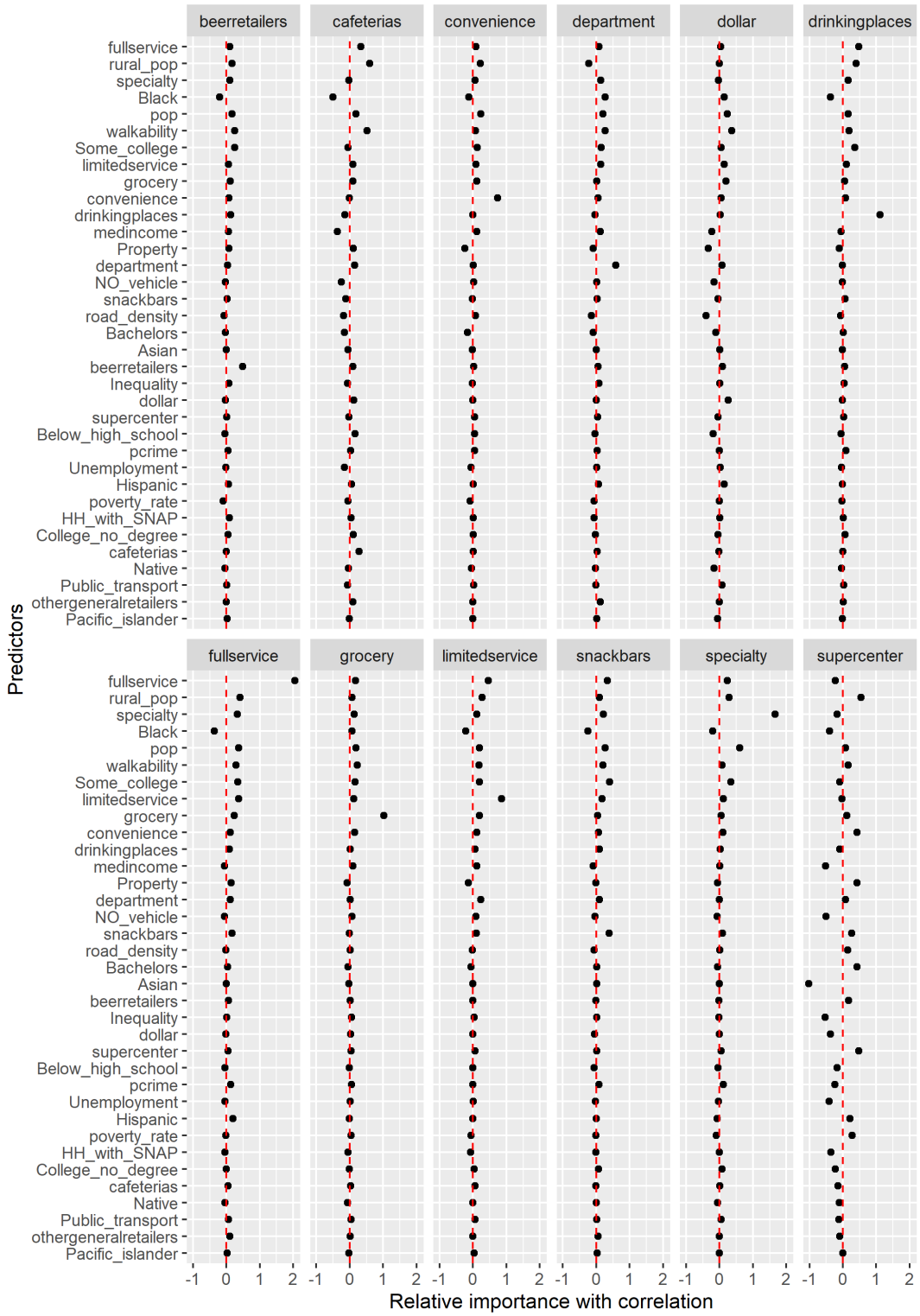

Fig 2. Relative importance of predictors in predicting the establishment of at least one store type in 2011-19. Machine learning SHAP values generated from XGBoost and multiplied with correlation between the predictor and predicted response. The predictors are arranged by the mean absolute value of their relative importance for all plots. One of the response variables, "other general retailers" was not included in the figure for brevity. See the method section for a discussion on SHAP values.