Predicting Dollar Store Entries and Densities Over Space and Time - An Application of Machine Learning

Abstract

Since the early 2000s, dollar stores have proliferated throughout the United States. This proliferation has been accompanied by a largely negative narrative, associating dollar stores with poverty and decreased access to other food retailers. We assess this narrative by combining spatially and temporally dynamic data sources containing dollar store and competing food store locations with demographic, socioeconomic, and market-level geographic information. We leverage data-driven machine learning techniques to build predictive models for all dollar stores and at the chain level for three of the largest dollar store retailers. We find that, while the conventional wisdom that dollar stores tend to locate in economically-disadvantaged communities holds for the format as a whole, dollar store location strategies are more heterogeneous at the chain level than suggested by the conventional narrative. Our results provide comprehensive insight into the proliferation of dollar stores throughout the United States and contribute to the nascent literature related to the economic consequences of dollar stores in contemporary food retail.

This research was supported by the U.S. Department of Agriculture, Economic Research Service. The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy. We also gratefully acknowledge funding from Agriculture and Food Research Initiative (National Institute for Food and Agriculture, USDA).

1 Introduction

Dollar stores have grown tremendously in the United States, adding new locations at a rate of 7% per year from 2000 through 2020 (Lee, 2021; Meyersohn, 2021b). As of 2021, the three major chains (Dollar General, Family Dollar, and Dollar Tree), which operate the majority of dollar store locations, accounted for over 33,970 locations, a figure that dwarfs the total of all other food retailers. No retailer has reached store counts of this magnitude in the United States since A&P in the early-to-mid 20$^{\text{th}}$ century (Hals, 2015). While dollar stores are viewed as a remarkable business success story, the narrative surrounding them is often highly negative. In particular, popular press, policymakers, and activists associate dollar stores with poverty, inequality, poor health outcomes, crime, and a lack of alternative food retail options, both in small towns and densely-populated, majority Black and Hispanic communities (Donahue, 2018; Donahue & Bonestroo, 2019; MacGillis, 2020; S. Mitchell & Donahue, 2018; Sainato, 2019; Siegel, 2019). The negative perception surrounding dollar stores has motivated the city councils of Toledo, Oklahoma City, Tulsa, Fort Worth, Mesquite, Cleveland, Birmingham, and Dekalb County in Atlanta to restrict dollar store growth, including outright bans on new locations (Birmingham City Council, 2019; Canfield, 2018; Capelouto, 2020; FOX 4 News Dallas-Fort Worth, 2018; Hart, 2019; Higgs, 2020; Howard & Fleming, 2019; Jimenez, 2019; Toledo City Plan Commisssion, 2020; Williams, 2021). Despite growing policy interest, researchers have yet to examine comprehensively whether the narrative surrounding dollar stores is supported by data, or the product of anecdotal evidence.

We measure the growth and expansion of dollar stores at the national scale using data-driven machine learning methods and evaluate the extent to which the market-level characteristics of dollar store entry and densities conform to the conventional wisdom of dollar store locations. We analyze dollar store growth beginning in the year 2000, when dollar stores were predominantly regional retailers concentrated in the South and Midwest, through the

year 2020, at which point the leading dollar store chains operated at the national scale. We construct a spatially and temporally dynamic database containing retail store location information from NielsenIQ's TDLinx and U.S. Census demographic, socioeconomic, and market geography features to predict dollar store locations. We estimate penalized elastic net regression models and non-parametric random forest and gradient boosting machine algorithms for dollar stores as a sector and individually for three major dollar store chains, which we identify generally as Chain A, Chain B, and Chain C. We predict store locations in the neighborhood of census tract households for urban and small-town/rural markets, allowing us to better understand the patterns of dollar store growth and the urban-rural dynamics of dollar store expansion.[1]

Our results show that while elements of the conventional wisdom around dollar stores hold true, the complete story is more nuanced. The effects of demographic and socioeconomic predictors from models of all dollar stores and chain-specific models for Chain A and Chain B generally support the conventional narrative that dollar stores locate in economically disadvantaged communities. But the association between income and dollar store densities flipped over time from negative to positive for Chain B locations in urban areas, while the association between poverty and Chain A densities in small towns and rural areas changed from positive to negative. In addition, the population share that is Black and Hispanic are strong positive predictors of Chain B stores in all geographic markets, while neighborhood race composition is a weaker predictor of Chain A. Demographics and socioeconomic features are relatively less predictive of Chain C.

The relationship between features of market geography and dollar store locations similarly varies by chain. Even within small towns and rural areas, Chain A is drawn to areas with high centrality, as indicated by the increasing positive associations between the probability of Chain A entry and store densities and the number of nearby public schools, places

[1]Our use of proprietary data sources prohibit us from disclosing the names of retailers. Further, our primary goal in this paper is to highlight the heterogeneity of dollar stores relative to the conventional wisdom of dollar store locations, as opposed to understanding the location strategies of specific chains.

of worship, and mobile home parks. These features are less important predictors for Chain B and Chain C. For dollar stores as a whole, we see that the association of population with dollar store presence and density is positive and increasing over time in each geographic market. However, when predicting the binary presence of Chain A in small towns and rural areas, increasing area-wide population decreases the likelihood of observing the chain store, reflecting the chain's lower population threshold compared to other dollar stores.

When examining how dollar store entry and density are associated with measures of retail competition, we see that grocery stores tend to be negatively associated with Chain A but positively predict Chain B and Chain C, suggesting that Chain A may be the most likely to see grocery stores as competitors. This also suggests that, relative to other dollar stores, Chain A may play a more important role in increasing access to food and household goods in under-served communities. The presence of larger retail formats (e.g., supercenters and wholesale club stores) are the strongest predictors of Chain C. Co-locating with big-box stores could allow the chain to benefit from demand externalities via retailer agglomeration. At the same time, Chain A and Chain B have gradually distanced their stores from big-box mass merchandisers and supercenters.

We add to the existing literature by investigating dollar store growth, both for the dollar store sector and for the three largest dollar store chains. Our data and methods are cutting edge in the sense that we create a customized, spatially-detailed database containing rich information on demographics, socioeconomics, market geography, and retail activity and apply it to predictive models of dollar store locations using machine learning algorithms, which are designed to provide high predictive accuracy and a data-driven approach to determining the most important predictors of the outcome variable. Our use of multiple machine learning algorithms allows us to draw robust conclusions concerning the nature of dollar store location strategies. Rather than lumping all dollar store chains in with a simple narrative, our findings highlight the importance of considering the dynamics of dollar store location strategies by chain and geographic space in order to comprehensively assess dollar store growth pat-

terns over the last 20 years. The heterogeneity of dollar store profiles across chains suggests that the impact of dollar store growth on food access, consumer spending, and retail competition may also be heterogeneous. Finally, while retail activity has become increasingly concentrated in urban centers and large chain stores, our model results yield insight into how specific dollar store chains may have capitalized on this macroeconomic trend in retail by gradually increasing their store footprint in the low-income urban, small-town, and rural markets that conventional stores cannot viably operate.

In what follows, section 2 summarizes the dollar store business model and the historical and geographical context in which dollar stores emerged. Section 3 outlines the data and methods employed in our empirical analyses. In section 4, we assess and compare the results from the machine learning predictive models. We conclude in section 5 with a focus on implications for future research.

2 Background

2.1 The Dollar Store Business Model

Contemporary dollar store chains, such as Dollar General, Family Dollar, and Dollar Tree, aim to provide their customers with everyday low-price merchandise in convenient locations and easy-to-navigate store formats, enabling customers to satisfy many of their household shopping needs (Dollar General Corporation, 2021; Dollar Tree Inc., 2021).[2] Dollar stores originated in the southeastern United States with Dollar General. The early success of the dollar store format attracted entrepreneurs to Dollar General retail locations to study the one-dollar pricing policy, which fomented the creation of several regional dollar store chains in the southern states, including Family Dollar and Fred's (Fred's Inc., 2019; Junior

[2]We use reports from popular press, dollar store company annual reports, and documentation of company histories to synthesize the dollar store business model and the historical and geographical origins of the dollar store sector.

Achievement of the Carolinas Inc., 2003; Turner Jr., 2018)

Dollar stores use a small-box store model that results in a low cost of opening and operating a new store, allowing dollar stores to feasibly locate in low-income and small-population markets in which conventional retail chains often cannot operate. With small store formats, dollar stores may also more flexibly select new retail sites with favorable lease, tax, or utility incentives from local governments (Cooper, 2016; McGreal, 2018; Misra, 2018).

Dollar stores are typically categorized as general merchandise and discount variety stores (Thomas, 2021). General merchandise dollar stores are differentiated relative to discount variety dollar stores in that consumable products (e.g., health, home cleaning, paper, pet supply and perishable food) are three-quarters of sales for the former and about half of sales for the latter (Dollar General Corporation, 2021; Dollar Tree Inc., 2021). Dollar General and Family Dollar describe their stores as neighborhood general merchandisers, situated within three- to five- miles of their customers, a characteristic that may give them a locational competitive advantage relative to their big-box competitors, which rely on customers traveling longer distances to patronize their stores (Dollar General Corporation, 2015; Dollar Tree Inc., 2020). Dollar Tree, in contrast, brands its company as a discount variety store in which a majority of products have a $1.25 price point (Dollar Tree Inc., 2022). The company's low-price business strategy and preference to locate in strip malls implies that their core customers are also localized to a small shopping radius, allowing them to operate stores at a high density without cannibalizing profits from nearby locations (Dollar Tree Inc., 2020).

2.2 Dollar Store Growth Nationwide

Dollar store growth is often described in terms of expansion during and after the Great Recession (2007-2009), the widening of their low-income customer base, and their shift in product mix to everyday household products, food, and tobacco items (Nassauer, 2017; Piller & Strong, 2015). While dollar stores have targeted low-income consumers since their

inception, beginning only in the early- and mid-2000s did they gradually include food as an important item in their product mix. Reviewing company annual reports over the past 20 years, there is a temporal relationship between the period of the Great Recession, increased sales of household items and food, including perishables, the installation of coolers and freezers at store locations, and the ability to accept Supplemental Nutrition Assistance Program (SNAP) benefits as payment from their customers (Dollar General Corporation, 2010; Dollar Tree Inc., 2010; Family Dollar Stores Inc., 2010).

Unlike many retail formats, dollar store chains typically exhibit growth during times of economic uncertainty and even downturns (Nassauer, 2022; Turner Jr., 2018; Wahba, 2020). During the Great Recession, the three leading dollar store chains reported that the adverse economic impacts increased store sales, as their low- and middle-income customer base aimed to save on everyday household items by shopping more frequently at the discount dollar store (Dollar General Corporation, 2010; Dollar Tree Inc., 2010; Family Dollar Stores Inc., 2010). Food products tend to have low profit margins, but they provided a stimulus to average customer spending per trip, shopping frequency, and expanded their customer base to higher-income groups during the recession. As consumers economized on food spending, they may have reallocated budgets away from convenience and national brands and in favor of private labels and nontraditional retailers, such as dollar stores (Kumcu & Kaufman, 2011). While many retail formats faltered, dollar stores continued to grow during the COVID-19 pandemic, including the development of private label food items and new store formats with increased healthy food options, such as fresh produce (Cain, 2021; Dollar General Corporation, 2021; Meyersohn, 2021a; Wahba, 2020).

Figure 1 shows the total number of stores in each U.S. region from 2000 to 2020 for the dollar store retail channel, supercenters, wholesale-club stores, and the two largest retail food store formats in terms of store counts, superettes (i.e., small grocery stores) and supermarkets.[3] Nationally, dollar stores exceed the growth of the conventional food retail stores

[3]The figure uses store location data from NielsenIQ's TDLinx. The number of supermarkets includes

(supermarkets and superettes), and particularly following the Great Recession (2007-2009). Relative to the year 2000, the number of dollar stores had grown by 250% as of 2020.

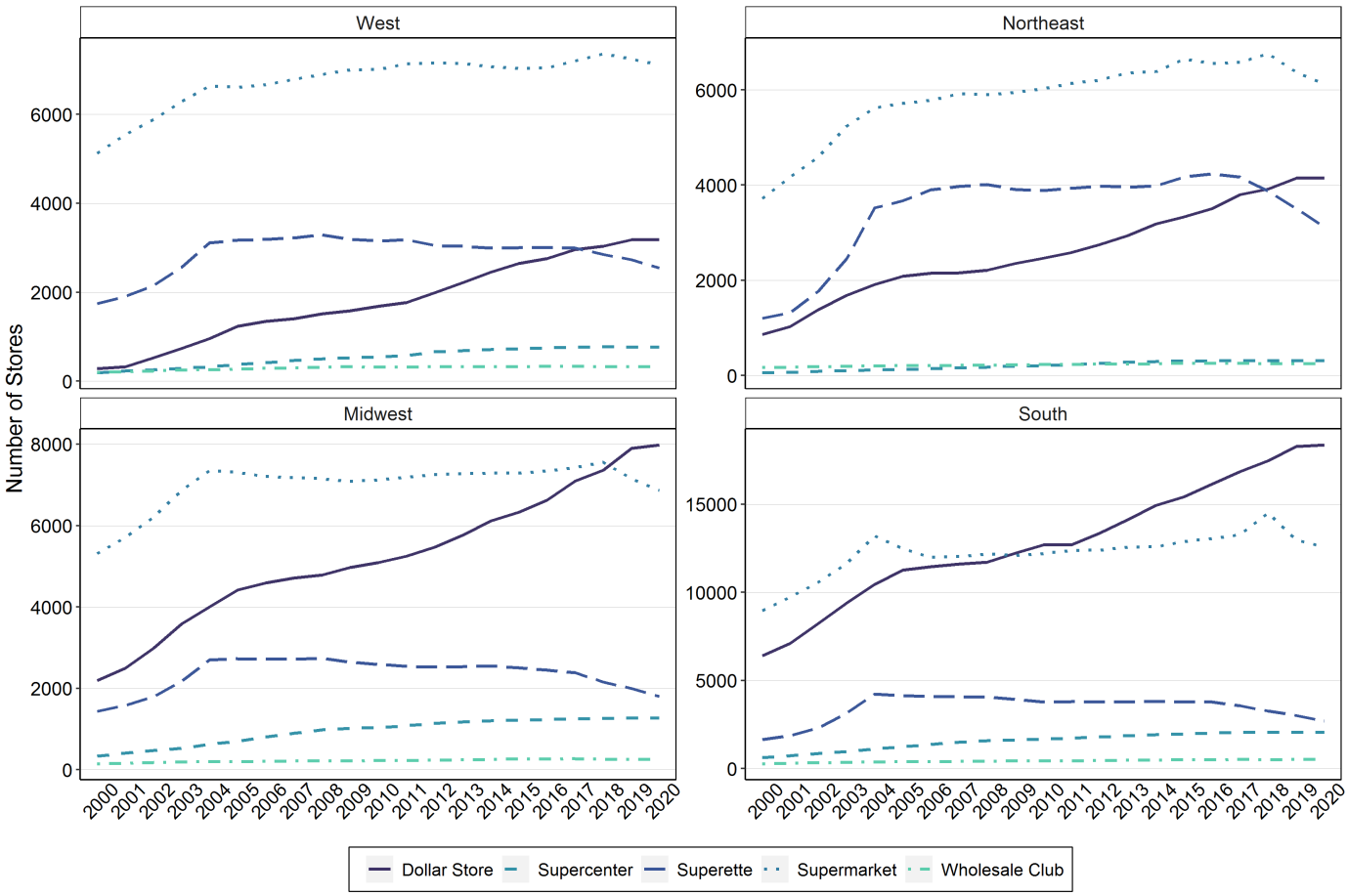

Figure 1: Number of Stores by Retail Channel and Region (2000-2020)

Figure 2 shows the share of the U.S. population whose nearest food retailer is a dollar store in each U.S. region for urbanized, urban-cluster, and rural areas.[4] [5] By 2020, approximately 77 and 66% of rural- and small-town communities in the South are most proximate

stores classified as conventional and limited assortment supermarkets and natural/gourmet food stores in TDLinx. Regions use U.S. Census Bureau definitions: `https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf`

[4]We use the 2010 U.S. Census definition and geographic delineation of urban areas to indicate whether census tracts belong to urbanized, urban-cluster, or rural areas. In our analyses, we designate urban-cluster areas as small-town markets, while we call places located outside of urbanized and urban-cluster areas rural markets: `https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html`.

[5]We compute the distance between census-tract households and the nearest food store for each TDLinx subchannel. We then found the number of times dollar stores were the closest food store option.

to a dollar store, while just over (under) 50% of rural- and small-town communities in the Midwest (Northeast) live nearest to a dollar store.

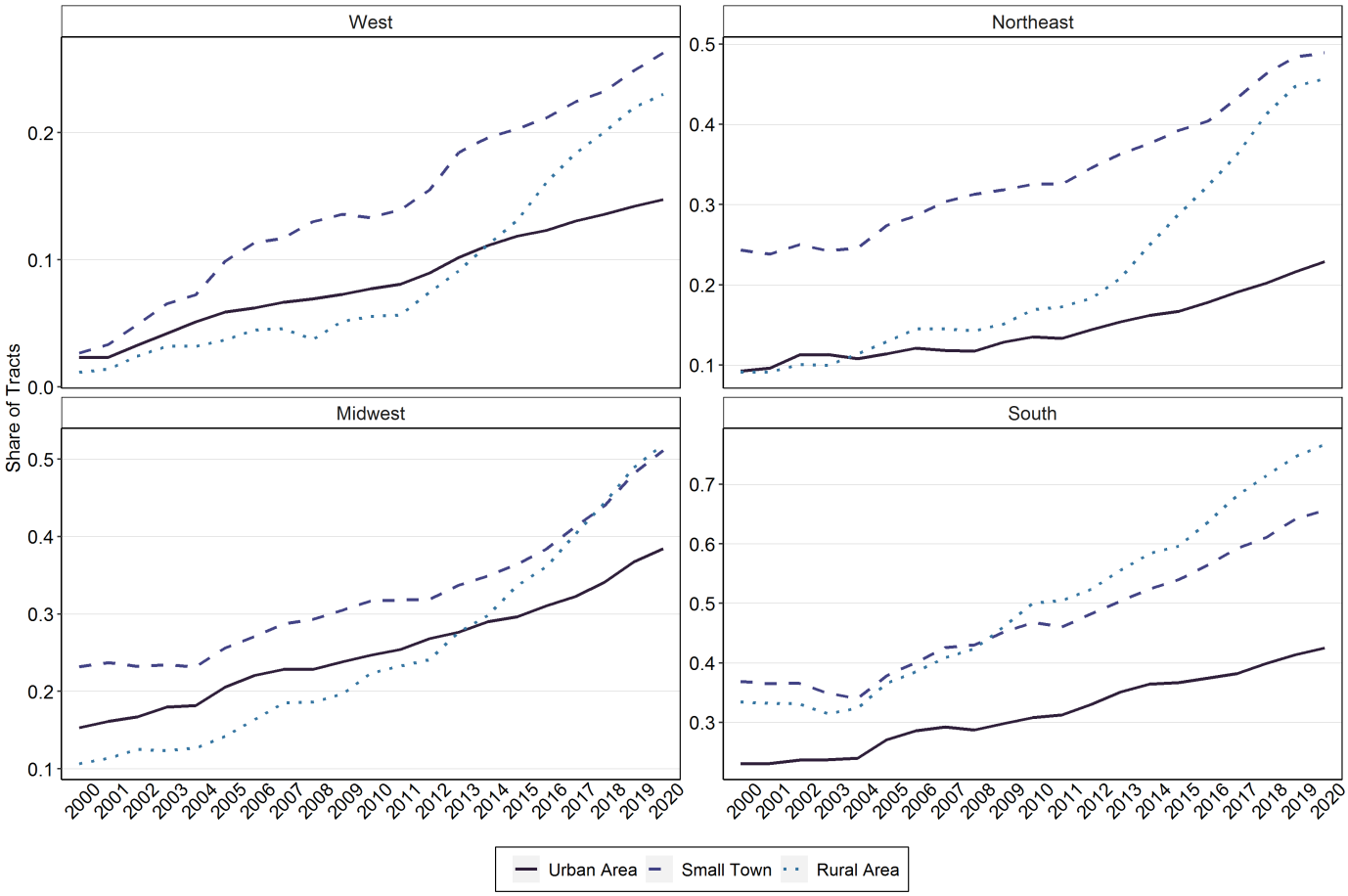

Figure 2: Share of Census Tract Households whose Nearest Food Store is a Dollar Store

2.3 The Economic Impacts of Dollar Stores

A catalyst to our study is the lack of applied research on the causes and consequences of dollar store growth. Little is known about the economic drivers of dollar store entry and densities, or the impacts of dollar stores on competitors, consumers, and local economies. We aim to take an important step towards informing these issues by measuring the factors associated with dollar store growth temporally and spatially.

Dollar store research to date has primarily focused on the sociological and public health

dimensions of the retail format. A few studies have examined the potential associations between dollar stores and demographics, such as race and income. Shrestha (2016) compared the dollar store segment to the early 20th century five-and-dime store formats and argued that dollar stores have flourished alongside rising economic inequality and intermittent economic recessions in the United States. Shannon (2020) studied the correlation between dollar store proximity and urban-area census-tract racial composition and socioeconomic variables from 2008 to 2015. He found that, controlling for income and other factors, the distance between dollar stores and census tracts with relatively higher levels of Black and Latino populations fell over time. Drichoutis, Nayga, Rouse, and Thomsen (2015) studied the link between dollar store access and childhood obesity and found no causal link, despite evidence that dollar stores offer fewer healthful food items than supermarkets.

Other research has investigated the link between dollar stores and food access. Racine, Batada, Solomon, and Story (2016) used the USDA's 2014 SNAP retailer locator to assess and qualify healthy food availability at 269 Dollar General, Family Dollar, and Dollar Tree stores in 16 counties located in the southern and western regions of North Carolina, highlighting that the three leading dollar store chains exhibited little differences in food offerings across urban and rural counties. Though they found some differences in healthier product offerings across chains, none of the three offered fresh produce. Volpe, Kuhns, and Jaenicke (2017) used household scanner data to measure negative correlations between the dietary quality of shopping baskets and the dollar store format. Chenarides, Cho, Nayga, and Thomsen (2021) investigated the propensity of dollar store entry into food deserts (i.e., areas with low-income and low-access to supermarkets) from 2000 to 2017. They showed that for the most part, dollar stores are not more likely to enter census block groups designated as food deserts. Moreover, the authors found that, conditional on dollar store entry into food deserts, exit rates of dollar stores are lower in block groups that remained a food desert throughout their study period.

The limited work to date on dollar stores broadly supports aspects of the media and

regulatory narrative surrounding dollar stores in the United States. Most empirical work suggests that dollar stores are more likely to proliferate in economically declining, Black and Hispanic communities. Relative to conventional food retailers (e.g., supermarkets), however, dollar stores do not necessarily improve food access by locating more frequently in low-income areas lacking local grocers (i.e., food deserts), though they may still positively impact food access in certain communities. While dollar store food offerings are shown to be limited relative to supermarkets and less consistent with healthy eating recommendations, such as the Dietary Guidelines for Americans, there is no clear causal evidence, of which we are aware, linking dollar stores to decreased dietary quality or adverse health outcomes. Current research mostly analyzes dollar stores as a sector, obscuring potentially important heterogeneous characteristics and impacts across dollar store chains. Further, extant research tends to have limited geographic and temporal scale, missing the urban-rural dynamics of dollar store growth since the early 2000s.

3 Data and Methods

In an effort to comprehensively understand the patterns of dollar store growth in the United States, we construct a database to which we apply machine learning techniques to predict dollar store entry and densities. We describe outcome and predictor variables in groups, which we categorize as dollar store and competing retailer densities, demographic, socioeconomic, housing, mobility, and spatial features related to market geography. The primary variable group of interest includes our outcome variables - dollar store presence and densities - and predictors related to retail competition, which impact dollar store entry decision-making. The demographic, socioeconomic, and household mobility predictors are conceptualized as demand-side factors that influence retailer decisions to enter the market, while the housing and market geography features are supply-side controls capturing the feasibility of opening stores in markets due to real estate values, distribution capacity, roadway accessibility,

distance to urban markets, and otherwise unobserved neighborhood characteristics.

3.1 Dollar Store and Competing Retailer Densities and Distances

Information on dollar store and competing retail store locations uses NielsenIQ's TDLinx database from 2000 through 2020, one of the most comprehensive data sources of retail food store locations and characteristics in the United States (Cho, W McLaughlin, Zeballos, Kent, & Dicken, 2019; Levin et al., 2018). TDLinx provides detailed store-level information of each food retailer's opening date and location, including their geocoded address, state, county, zip code, and street address. Individual stores have unique store identifiers as well as corporate-level codes indicating to which parent company the store belongs. Nielsen categorizes stores in TDLinx using primary channel and sub-channel definitions employed in the food retailer industry or developed by Nielsen (Nielsen, 2022).

Using the TDLinx data, we compute retail store densities and proximity to consumer populations for dollar and competing retail stores. In order to obtain a fine-grained measure of store densities around and distances to households, rather than use the count of stores within a given geographic area (i.e., within the block group, census tract, or county), we compute census-tract population-weighted average densities and distances to retailers using the population-weighted block-group centroids, letting the block-group centroids represent demand (Alviola IV, Nayga Jr, Thomsen, & Wang, 2013; Shannon, 2020; Wilde, Llobrera, & Ver Ploeg, 2014). This allows for a more realistic measure of store access, as consumers may not shop within the neighborhood in which they reside, defined by U.S. Census Bureau delineations.

In computing food retailer densities, we first find the number of stores of each retailer type within a given radial distance of the population-weighted block-group centroids. We aggregate these counts up to the census-tract level by calculating the population-weighted average of stores within the specified radial distance. We carry out a similar exercise in esti-

mating store proximity. We compute the distance from the population-weighted block-group centroids to the nearest store, for each retailer type. Using these distances, we calculate the population-weighted average distance between the nearest retailer and census tract households.[6] We compute weighted-average store densities for three-, five-, and ten-mile distance bands and nearest-store distances for each year of the TDLinx data (2000-2020) and retailer type. In classifying retailers by type, we leverage the sub-channel definitions from the retail trade channel categories provided by TDLinx. Table 1 shows the thirteen sub-channel store types, including the dollar store sector, for which we compute retail store densities and distances.[7] [8] We further subset the dollar store sector and compute store densities for three dollar store chains, which we label Chain A, Chain B, and Chain C. Finally, we create an "All Other" category consisting of independent and regional dollar stores.

Using the retail store densities within three, five, and ten miles, we compute inner-ring store densities for the three-to-five and five-to-ten mile distance bands for each retail channel (Seim, 2006). The three- and three-to-five mile store densities are used in our analyses for urbanized census tracts, while the five- and five-to-ten mile densities are used in urban-cluster (i.e., small-town) and rural-area census tracts. These distance bands are cited as the market areas of Dollar General and Family Dollar in their annual reports (Dollar General Corporation, 1999, 2015; Family Dollar Stores Inc., 2007, 2014). The inner-ring densities account for the fact that dollar stores may locate between high-density consumer populations

[6] Population-weighted census-tract store densities and distances are computed using the formula: $\sum_{g}^{N(i)} \left(\frac{pop_{gi}}{\sum_{g}^{N(i)} pop_{gi}} * x_{gi} \right)$, where for store densities, x_{gi} is the number of retail stores located within the specified radial distance of the population-weighted block-group centroids. For proximity, x_{gi} indicates the distance, in miles, from the population-weighted block-group centroid to the nearest retail store. The subscripts indicate the block group, g, nested in census tract, i. The summations are over the number of population-weighted block-group centroids in each census tract, N.

[7] In all analyses involving the Drug Store channel, we combine the Conventional Drug Store and Rx Only and Small Independent Drug Store into a single Drug store category, which leaves us with twelve distinct retail sub-channels. We also note that, in this paper, the word "channel" is often used interchangeably for TDLinx sub-channels.

[8] In the predictive models, we group the Conventional Supermarket, Limited Assortment Supermarket, Natural/Gourmet, Superette, and Warehouse Grocery store channels into a single "Grocery Store" variable.

and the nearest supercenter or supermarket, as well as cases in which retail agglomeration occurs just outside the three- or five-mile inner-ring thresholds.

3.2 Census Data

We pair our retail location data from TDLinx with census-tract level demographic, socioeconomic, housing, and household mobility data collected from the year 2000 U.S. Decennial Census (DC) and five-year American Community Survey (ACS) estimates. We use the census tract as our unit of analysis, as it is commonly used to proxy "neighborhoods" (Moore & Diez Roux, 2006; Morland, Wing, Roux, & Poole, 2002; Sampson, 2016; Sharkey & Faber, 2014), and therefore, is closely related to the terminology employed by Dollar General and Family Dollar in describing their operations as neighborhood stores in their annual reports.

The DC and ACS five-year estimates present two challenges in our study context. First, while our store-level data on retail locations is annual, the ACS five-year estimates are collected over a 60-month period and are pooled by the Census Bureau to compute estimates, creating temporal imprecision in pairing yearly store location data with the pooled census data estimates provided by the ACS. To address the temporal imprecision of the data, we use the middle year of each ACS period estimate as the reference year to which we match our annual store location data. Using the middle year as the reference year follows other research investigating the relationships between retail food store locations over time and the demographic, socioeconomic, housing, and mobility features using the ACS (Alviola IV et al., 2013; Shannon, 2020).[9]

Second, there is a gap in data coverage between the 2000 DC and the ACS 2005-2009. We use a similar approach as Holmes (2011) by taking convex combinations of year 2000

[9]Similarly, the IPUMS National Historical Geographic Information System (NHGIS) recommends using the 2008-2012 five-year ACS period estimates, whereby the middle year corresponds to 2010, to study changes in census estimates between the years 2000 and 2010. `https://www.nhgis.org/support/faq#compare_ACS_and_decennial`

DC and ACS 2005-2009 period estimates to interpolate the census data variables from 2001 to 2007 (e.g., to estimate census variables in the year 2001, we give $\frac{6}{7}$ weight to the 2000 DC variables and $\frac{1}{7}$ weight to the ACS 2005-2009 variables.). This approach is facilitated by the fact that the 2000 DC and ACS 2005-2009 estimates use the same year 2000 census geographies, allowing us to cleanly link census tracts over time.

3.3 Spatial Feature Creation

Recognizing that dollar stores, as well as other food retailers, do not base their location decisions only on the census-tract characteristics in which the store is located, but instead on the characteristics of the surrounding area, we create spatial features that capture the neighborhood demographic, socioeconomic, housing, and household mobility conditions that could influence dollar store location decisions. The variables used in this study were chosen based on the hypothesized economic relationships between dollar store locations and the historical characteristics of dollar store customers. We use the annual reports of dollar store chains (e.g., Dollar General, Family Dollar, Dollar Tree, Fred's), popular press (Bendix, 2018; Donahue, 2018; Misra, 2018; S. Mitchell & Donahue, 2018), economic modeling of retail market location and firm entry (Ellickson & Grieco, 2013; Holmes, 2011; Seim, 2006; Zhu & Singh, 2009), and the food environment literature (Alviola IV et al., 2013; Bonanno, Chenarides, & Goetz, 2012; Chenarides & Jaenicke, 2019; Dutko, Ver Ploeg, & Farrigan, 2012; Goetz & Swaminathan, 2006; Powell, Slater, Mirtcheva, Bao, & Chaloupka, 2007; Racine et al., 2016; Rhone, Ver Ploeg, Dicken, Williams, & Breneman, 2017; Shannon, Shannon, Adams, & Lee, 2016; Shrestha, 2016) to inform our selection of variables.

For each census tract, we find the population-weighted census-tract centroids within three, five, and ten miles to compute spatial sums or window averages of census-tract variables. We compute spatial sums of the total population and total tract size within each distance band. We find the spatial window averages for the other demographic, socioeco-

nomic, housing, and mobility variables.[10]

These spatial features coincide geographically with our measures of retail store densities measured for the three-, five-, and ten-mile distance bands. In our predictive models, we include the census-tract neighborhood spatial features within three miles for urban-area census tracts, while for census tracts in small-town and rural areas, we use the five-mile radius. Table 2 lists the complete set of demographic, socioeconomic, housing and household mobility predictors used in our predictive models of dollar stores, as well as the retail density and market geography features, discussed in the following section.

3.4 Market Geography Features

In addition to the store-level density measures and census-tract variables, we create several other spatial features as inputs to the predictive models of dollar store locations. In this section, we describe market geography features created to account for urban-rural spatial relationships, distribution and transportation infrastructure, and other spatial features that may influence dollar store location decision-making.

Demand threshold analyses for retail have employed distance to the urban core as a predictor for the number of retail establishments (Wensley & Stabler, 1998), while popular press cites dollar stores for locating in remote urban and rural communities whereby residents have fewer nearby shopping options, suggesting a positive relationship between distance and the presence of dollar stores. To account for these urban-rural market relationships in our modeling, we find the distance between each population-weighted census-tract centroid and the nearest urbanized- or urban-cluster area geographic centroid.

We control for the relationship between distribution center proximity and variation in dollar store density by collecting distribution center location information for several of the

[10]The spatial sums can be expressed as $\sum_i^{N \in B} x_i$, where x_i is a variable (e.g., population) in census-tract i, B represents the distance band (e.g., three, five, or ten miles) and N is the total number of census tracts in the neighborhood of census tract i. The spatial window averages divide the sum by N.

largest dollar store chains during our study period: Dollar General, Family Dollar, Dollar Tree, and Fred's. For each year, we compute the distance from each of the four chain's nearest distribution centers to each population-weighted census-tract centroid. In our predictive models, we use the minimum distance between census-tract population-weighted centroids and the nearest distribution facility. As the proximity between census-tract households and the nearest dollar store distribution facility decreases, we expect dollar store densities to rise, holding other factors constant. We control for census tract accessibility to road infrastructure by including the total road mileage of interstate, U.S. highway, and state highway roadway, weighted by tract area. Similar to fast-food outlets, dollar stores may co-locate along roadways to capture commuter traffic between rural, small-town, and urban areas (Nilsson & Smirnov, 2016).

We create additional predictors that help indicate the relative centrality of a given census tract. Dollar stores are cited for aiming to locate stores near central public places, including churches, post offices, and schools (Debter, 2022; Drichoutis et al., 2015; Nassauer, 2017; Sainato, 2019). We leverage three national data sources from the Homeland Infrastructure Foundation-Level Data (HIFLD) website of the Department of Homeland Security (U.S. Department of Homeland Security, 2022). Specifically, we use geo-referenced data for all places of worship, which includes point-location information for a range of religious sites (e.g., churches, temples, mosques) (N = 254,742), public schools (N = 102,334) and mobile home parks (N = 45,642).

For each census tract, we count the number of times its population-weighted centroid serves as the nearest population center of each place of worship, school, and mobile home park. If a census tract is not the nearest neighbor to a given place, its count is zero. Finally, we include state-level fixed effects to control for state-wide variation in dollar store densities as they grow over space and over time.

3.5 Machine Learning Algorithms

We use statistical machine learning (ML) algorithms to build and select optimal predictive models. In our application, the advantage of ML over conventional methods is that, given a host of market characteristics, the model development and selection processes are data-driven whereby only the variables with greatest predictive power of dollar stores are included in the final models (Athey, 2019; Breiman, 2001b). The interest in and application of ML-methods has increased in the field of economics (Athey, 2019; Blumenstock, Cadamuro, & On, 2015; Coad & Srhoj, 2019; Glaeser, Kominers, Luca, & Naik, 2018; Jean et al., 2016), as well as in studies predicting neighborhood access to healthy-food retail stores (Amin, Badruddoza, & McCluskey, 2020).

We estimate a suite of classification and regression models predicting the presence and densities of dollar stores in the neighborhood of census-tract households from 2000 to 2020, assessing each ML-model's predictive performance and each predictor's association with the outcome variables. We build models nationally for urban- and small-town/rural census tracts using the 2010 U.S. Census Bureau's urban-area definition and geographic delineations to indicate whether census tracts are located in urban, small-town, or rural markets. U.S. Census Bureau urban areas consist of urbanized areas and urban clusters. Urbanized areas are geographic territories containing at least 50,000 people, while urban clusters contain between 2,500 to 50,000 people. In our models, we define urban-area census tracts as those in or partially within urbanized areas, while census tracts intersecting urban-cluster areas are defined as census tracts in small towns. Census tracts outside of urbanized and urban-cluster areas are specified as rural-area census tracts belonging to rural counties.

3.6 Data Preparation and Spatial Sampling Tasks

In machine learning, evaluating and comparing each model's predictive performance involves applying the chosen optimal models to data not used in the model-building process. In most

machine learning applications, the data are randomly split into independent training and test subsets, depending on the availability of data (Hastie, Tibshirani, & Friedman, 2009). Variable selection and hyperparameter tuning are performed on the training data, while the model's predictive performance is assessed on the test data. Hyperparameters control model complexity. More complex models reduce bias but tend to increase the variability of out-of-sample predictions. Less complex models, on the other hand, have greater bias but less variability in out-of-sample predictions. By assumption, the observations in the training and test data are independent, which allows for an unbiased estimate of model accuracy when the trained ML-algorithm is applied to the test data to make predictions. However, randomly partitioning spatially autocorrelated data into training and test subsets can result in test error rates that overstate model accuracy (Karasiak, Dejoux, Monteil, & Sheeren, 2021; Meyer, Reudenbach, Wöllauer, & Nauss, 2019; Roberts et al., 2017; Schratz, Muenchow, Iturritxa, Richter, & Brenning, 2019).

As opposed to random sampling, spatial sampling methods that preserve the spatial structure of neighboring observations and that allow for the spatial independence of more distant observations provide unbiased estimates of the machine learning algorithm's predictive performance (Brenning, 2012; Karasiak et al., 2021; Valavi, Elith, Lahoz-Monfort, & Guillera-Arroita, 2018). In building predictive models of dollar store locations, we implement a block-based spatial sampling scheme that enables us to sample blocks of census tracts at the market level according to the urbanized and urban-cluster areas or rural counties to which census tracts belong. We train our machine learning models to predict dollar store presence and densities using census tracts within markets, and we evaluate model performance using the test subsets containing census tracts from spatially distinct geographic markets. Following this spatial sampling framework, we ensure that census tracts within the same geographic market (e.g., urbanized or urban-cluster area and rural county) are not included in the training and test data simultaneously.

We stratify each year's population of U.S. census tracts by their urbanized- and non-

urbanized area status. Each year of our data contains 481 unique urbanized markets. Non-urbanized areas include census tracts within urban clusters (i.e., small towns), and census tracts in rural counties outside of both urbanized and urban-clusters. From 2000 to 2020, there are approximately 5,476 small-town and rural county markets. Graphical examples of the spatial sampling scheme are given in Appendix A. To create the training and test data subsets for national-scale models, we assign 90% of the urbanized-area (small-town/rural) markets to the training data and the remaining 10% to the test data used exclusively to assess the models' predictive performance.

Using the training data, we conduct five-fold cross-validation (C.V.), a data re-sampling and model validation technique used to find the optimal combination of an ML-model's parameters during model selection (Kim, 2009; Kuhn & Johnson, 2013; Molinaro, Simon, & Pfeiffer, 2005). In five-fold C.V., an index ranging from 1 to 5 is assigned to each observation in the training data. All observations from a given market, as described above, are assigned the same index value. Given a set of hyperparameter values of an ML algorithm, we fit the model using observations from four of the folds. We evaluate the model's performance using the held-out, "validation" fold and record the model's accuracy with respect to a measure of prediction error. We repeat this process until each fold is included as the validation fold. Combining the estimated model errors for the five fitted models, we calculate the average error rate across all five folds for the the given set of model hyperparameters.

For each class of ML algorithm, we compute the C.V. mean squared error (MSE_{CV}) for the regression models as:

$$MSE_{CV} = \frac{1}{5}\sum_{k=1}^{5}\left(\frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2\right)_k \tag{1}$$

where i denotes the census tract and k indicates the holdout fold. For classification models, we use classification error (Err_{CV}):

$$Err_{CV} = \frac{1}{5}\sum_{k=1}^{5}\left(\frac{1}{n}\sum_{i}^{n} I(y_i \neq \hat{y}_i)\right)_k \tag{2}$$

For each unique combination in a grid of candidate hyperparameter values, we conduct five-fold C.V. The optimal model contains the set of tuning parameter values that minimizes the average C.V. error rate. Using the optimal set of parameter values, we re-estimate the algorithm using the full set of training data. The final model is applied to the test data to measure our model's predictive performance. The estimated generalization error for regression is calculated as (Hastie et al., 2009):

$$R^2_{test} = 1 - \frac{\sum_{i}^{n}(y_i^{test} - \hat{y}_i)^2}{\sum_{i}^{n}(y_i^{test} - \bar{y}^{test})^2} \tag{3}$$

The classification error is calculated as:

$$Err_{test} = \frac{1}{n}\sum_{i}^{n} I(y_i^{test} \neq \hat{y}_i) \tag{4}$$

3.7 Model Specifications

We estimate classification and regression models for all dollar stores and for each dollar store chain type - Chain A, Chain B, and Chain C - from 2000 to 2020. Our models predicting the density or entry of dollar stores can generally be expressed as:

$$\underbrace{y_{it}}_{\text{DS density/entry}} = \underbrace{\sum_{s}\sum_{b}\beta_{sb}d_{it}^{sb}}_{\text{Retail Competition}} + \underbrace{\sum_{k}\theta_k x_{kit}}_{\text{Demographics/Socioeconomics}} + \underbrace{\sum_{l}\tau_l g_{lit}}_{\text{Market Geography}} + \underbrace{\varepsilon_{it}}_{\text{Error}}$$

The machine learning algorithm presented in the main text is elastic net regression, a penalized regression method in which the regression coefficients are obtained by minimizing

the loss function subject to a penalty term that constrains model complexity so as to improve out-of-sample predictive performance. In a linear regression model, the elastic net objective function is specified as:

$$\underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij} \beta_j \right)^2 + \lambda \left(\frac{(1-\alpha)}{2} \|\beta\|_2^2 + \alpha \, |\beta| \right) \right\}. \tag{5}$$

where all variables are normalized to the same scale, and λ and α are hyperparameters chosen by cross validation (J. Friedman, Hastie, & Tibshirani, 2010). Greater values of λ further constrain the least-squares solution relative to ordinary least squares, shrinking coefficients toward zero. Note that for the elastic net penalty, $0 \leq \alpha \leq 1$, so that the elastic net constraint is a weighted combination of two terms: the sum of squared coefficients (i.e., $\|\beta\|_2^2$, the ℓ_2-norm squared) and the sum of the absolute value of the coefficients (i.e., $|\beta|$, the ℓ_1-norm). Larger values of α will result in "sparse" models with few non-zero coefficients, as in the lasso (Tibshirani, 1996), while smaller values of α will yield models with many small coefficient estimates, as in ridge regression (Hoerl & Kennard, 1970). When predictor variables are moderately or highly correlated, using a data-driven combination of these two penalty terms may yield superior performance relative to using only a single penalty term, because the elastic net allows for both variable selection and shrinkage of coefficients that have strong pairwise correlations (Zou & Hastie, 2005). Regularization of the regression coefficients helps reduce the variance of model predictions and yields models with strong predictive performance (Ahrens, Hansen, & Schaffer, 2020).

The elastic net algorithm can be used with any likelihood function from the linear exponential family (e.g., Logistic or Poisson regression) (Tay, Narasimhan, & Hastie, 2021).[11] In section 4, we present coefficient estimates from the post-selection elastic net models, which is a regression of the outcome variable on the subset of most predictive explanatory variables,

[11]The elastic net linear regression routine is outlined in Algorithm 1 of Appendix C. However, it easily accommodates logistic models predicting the entry of dollar stores using the penalized negative binomial log-likelihood as the objective function.

with explanatory variables in their original units (Belloni & Chernozhukov, 2013; Belloni, Chernozhukov, & Wei, 2016).

In Appendices D and E, we discuss two "ensemble" machine learning methods, i.e., methods that make predictions by averaging over a set of models: random forest and gradient boosting. As shown in Appendix B, random forest and gradient boosting generally outperform the elastic net in terms of predictive performance. Nevertheless, we focus on results from the elastic net because point estimates and standard errors of model coefficients easily convey the sign, magnitude, and precision of associations between predictors and our outcome variables. The same information is much more challenging to communicate with random forest and gradient boosting. Appendix F presents findings from the the random forest and gradient boosting algorithms that identify the most important predictors of the outcome variables. The conclusions from these alternative methods support the results of the elastic net.

4 Results

In this section, we discuss the most salient findings from the machine learning analysis of dollar store entry and density for the years 2000 through 2020. Each estimation features between 85 and 96 predictors, and in each case we present and summarize the findings. We separate our results into four categories - demographics, socioeconomics, market geography, and retail competition.

4.1 Demographics: Income, Poverty, and Race

The extant literature, as well as the coverage of dollar stores in the media, provides expectations for the associations linking dollar store growth to multiple demographics. Figure 3 shows the coefficient paths from each dollar store model for neighborhood median household

income, area-wide poverty rate, and four demographic predictors of race in the urban- and small-town/rural area logistic and OLS elastic net models[12] Consistent with the conventional wisdom that dollar stores target low-income and high-poverty households, in both urban- and small-town/rural areas, the impact of a unit increase in median household income (poverty rate) mostly has a negative (positive) association with the probability of observing a dollar store and on the density of dollar stores, holding constant other factors (i.e., Dollar Stores models in the 1st columns of each model panel of Figure 3). However, the relationships between income, poverty, and dollar stores are not identical across chains.

The Dollar Stores logistic and OLS urban- and small-town/rural area coefficient paths for the neighborhood median household income and poverty rate tend to resemble the coefficient trajectories of Chain A and Chain B. These patterns are explained by the large store presence of Chain B in urban areas and Chain A in small towns and rural areas. The poverty rate coefficient magnitudes are especially large in the Dollar Stores, Chain A, and Chain B store density urban-area OLS models (bottom-left panel of Figure 3).

The associations linking income and poverty with Chain A and Chain B locations are also not constant over time. In the urban-area Chain B logistic and OLS models (top- and bottom-left panels of Figure 3), the coefficient sizes associated with median household income increase from 2000-2020 but are always negative until 2019 and 2016, respectively, where they switch to become positive. In the small-town/rural Chain A OLS models (bottom-right panel of Figure 3), the poverty-rate coefficient sign switches from positive to negative beginning in 2015. These patterns suggest that, as these chains have expanded their store network, they have also broadened their customer base to higher income populations.

In the Chain C logistic and OLS models, the neighborhood poverty rate is selected less frequently. Conditional on selection by the elastic net algorithm, the coefficient is often statistically insignificant or the sign is predominantly negative, implying that rising poverty

[12]As detailed in 3.2, the demographic and socioeconomic variables are measured within a three- and five-mile radius for urban and small-town/rural models, respectively.

decreases the likelihood of Chain C entry and store densities, illustrating their contrasting location profile relative to the other two dollar store chains. For example, contradicting conventional wisdom, the logistic urban models in the early 2000s often yield negative and statistically significant associations between the neighborhood poverty rate and Chain C, before the elastic net algorithm shrinks the coefficient to zero.

The relationship between race and dollar stores similarly varies by chain and geographic market. In the Dollar Stores logistic and OLS models, in both urban and small-town/rural areas, the share of a neighborhood's population that is Black and Hispanic increases both the likelihood of observing a dollar store and the density of stores, confirming the conventional wisdom that dollar stores tend to locate in majority Black and Hispanic areas, particularly in urban markets. Observing only the coefficient paths of the Dollar Stores models, however, obscures the heterogeneity of neighborhood location preferences of the three largest dollar store chains.

The composition of the neighborhood's race is an important predictor of Chain B in both urban and small-town/rural geographic areas, while race is more predictive of Chain A in small-town/rural areas. Conditional on other market-level covariates, the share of the population that is Black or Hispanic is strongly associated with the likelihood of Chain B entry and store densities in urban-area models; this is expected given the strong urban presence of Chain B, but these associations carry over to small-town/rural area models as well. The effects of race on Chain A vary for models predicting the presence of at least one store location and densities. The Chain A logistic small-town/rural area model (top-right panel of Figure 3) supports a narrative that the chain is more likely to locate in majority-white, small-town and rural communities, and that this association has grown stronger over time. But the logistic models in both urban and small-town/rural areas (top two panels of Figure 3) also show that, controlling for other market-level characteristics, the association between the share of the population that is Black and the likelihood of observing at least one Chain A store switched from negative to positive in urban and small-town/rural areas,

suggesting that Chain A has expanded over time into markets with more heterogeneous demographic profiles. On the other hand, when predicting multiple stores, the OLS Chain A models (bottom two panels of Figure 3) show that the share of the population that is Black and Hispanic mostly decreases Chain A store densities. The neighborhood's race is less important for Chain C, although the share of the population that is Hispanic does have a strong effect in the OLS and logistic models in some years, likely reflecting the strong presence of Chain C in western states.

4.2 Socioeconomics

Figure 4 displays the coefficient paths of socioeconomic variables related to economic, education, housing, and mobility conditions in the census-tract neighborhood. It is evident that socioeconomic variables are selected less frequently for Chain C. Conditional on selection, the estimates tend to have smaller magnitudes relative to the two other dollar store chains. On the other hand, the coefficient paths for the socioeconomic variables in the Chain A and Chain B models are quite consistent with the conventional narrative that dollar stores locate in socioeconomically disadvantaged communities with poor housing conditions.

Chain B has the strongest preference for communities with high shares of the population unemployed, with less than a high-school education, low vehicle access and greater reliance on governmental public assistance, as reflected in the large magnitudes of predictor estimates. Still, the social capital of the communities in which Chain A enters is clearly low, as increases in the share of the population with at least a bachelor's degree decreases the likelihood of Chain A entry and store density in each geographic market and model. Further, the coefficient paths of Chain A indicate that, holding constant other predictors, rising unemployment tends to increase the presence and density of Chain A in both geographic areas, and particularly around the period of the Great Recession (2007-2009), consistent with reports in popular press regarding dollar store growth (Corkery, 2021; Nassauer, 2017).

Chain B may locate to a greater extent in communities where households have larger transportation constraints, as reflected by the stronger relative associations with the two mobility predictors. In the logistic, and to a lesser extent, OLS, urban-area models, the association between Chain A and Chain B entry and lack of vehicle ownership has grown stronger over time. These results suggest that dollar store impact on increasing neighborhood food access in transportation-constrained communities may actually be positive, and is therefore likely more complex than that suggested in the media (Sainato, 2019), varying substantially by dollar store chain.

4.3 Market geography

Popular press reports that the low-cost business model of dollar stores allows them to locate in under-served, small-town and rural communities in which conventional supermarkets cannot feasibly operate (W. Brown, 2022; Debter, 2022; Del Mastro, 2021; McGreal, 2018; Morris, 2017). We assess this dimension of the conventional wisdom of dollar store locations by studying the associations between dollar stores and market geography indicators of census-tract remoteness, accessibility, and centrality. Figure 5 displays the coefficient paths for several market geography predictors from the logistic and OLS small-town/rural area models, including the total population within a five-mile radius, several proxies of census-tract centrality (i.e., mobile home parks, places of worship, and public schools), and measures of accessibility, specifically the census-tract's distance to the nearest urban core and distribution center.

Figure 5 shows that for dollar stores as a whole and in two of the three dollar store chain models, a ceteris paribus rise in population increases both the likelihood of observing at least one store and the density of dollar stores within a five-mile radius. However, the relationship between population and Chain A varies with the density of the chain's store network. In the logistic small-town/rural model (left panel of Figure 5), increases in area-wide population

are negatively associated with the presence of a single Chain A store. On the other hand, the OLS model implies that increasing Chain A store densities in small-towns and rural areas requires larger populations to support multiple stores. Based on Figure 5, relative to the other two dollar store chains, Chain A is more likely to operate in markets for which population is not necessarily the most critical factor to economic viability, while Chain B and Chain C tend to have larger population thresholds for operating in small-towns and rural communities.[13]

Coefficient paths for indicators of census-tract centrality suggest that Chain A still selects locations with a high degree of centrality, despite often locating in less populous rural areas. The number of mobile home parks, places of worship, and public schools all show consistently strong positive effects on the presence and density of Chain A stores, and the coefficient magnitudes have grown over time. These associations tend to be weaker for Chain B and mostly negative for Chain C.

The urban-cluster area binary variable is included in the small-town/rural area models to control for differences in dollar store levels across small towns (i.e., urban clusters) and rural areas. The urban-cluster indicator is equal to one if a tract is in an urban cluster and zero otherwise. The coefficient path for the logistic and OLS small-town/rural models indicate that relative to more populous small towns, dollar stores are less likely to locate in rural census tracts. However, the urban-cluster area coefficients from the logistic model are trending downwards in all dollar store models, suggesting that these three dollar store chains have grown in less populous rural areas over time.

As expected, holding other factors constant, an increase in a census tract's distance to a dollar store distribution center tends to decrease both the likelihood of observing a store and the density of stores. But the association between distance to the nearest distribution

[13]Mirroring these population estimates, Chain A is the only chain for which increasing distance between census-tract populations and the nearest urban core increases store density, controlling for other factors (right panel of Figure 5).

center and dollar store locations changes over time, becoming less negative in the logistic and OLS models. In some cases, the coefficient sign switches from negative to positive. Initially each of the dollar store chains may aim to establish their store network within a given radius of distribution centers. As distribution efficiencies improve for servicing the most proximate stores, each retail chain can more easily expand its store network by taking advantage of economies of density, locating stores further from the nearest distribution center (Holmes, 2011). This also raises questions about how dollar stores are acquiring groceries and the potential role that wholesalers and distributors play in supplying dollar store chains. Finally, the patterns described for small-town/rural area models are largely replicated for urban areas in Figure 6, where measures of centrality again appear to play a larger role for Chain A relative to Chain B and Chain C.

4.4 Retail Competition

Figure 7 displays the effects of dollar store chain densities on the presence and densities of competing dollar store chains within a three-mile (urban) and five-mile (small-town/rural) radius of census-tract households for the logistic and OLS post-selection elastic net models. In the urban-area models, the three dollar store chains tend to co-locate within census-tract neighborhoods in which the other dollar store chain operates. Communities that have imposed bans on dollar store densities are predominantly situated in urban areas (Donahue & Smith, 2022), and the results in Figure 7 provide some credence to the notion that dollar stores cluster in urban markets. The pattern of proliferation may also be taking hold in small-towns and rural areas. Chain A and Chain B are positively correlated with one another outside of urban areas, and the association between Chain A and Chain C in the OLS small-town/rural area model switches from negative to positive in 2019 and 2020.

Figure 8 suggests heterogeneous relationships between dollar store and grocery store

densities that vary by chain and geographic space.[14] Whereas the OLS urban and small-town/rural Dollar Stores models imply positive and increasing associations between dollar store and grocery store densities, the relationship does not hold for all chains. For the same pair of models, the marginal effects of three-mile (urban areas) and five-mile (small-town/rural areas) grocery store densities are mostly negative in the Chain A models, while the effects are positive and increasing in the Chain C and Chain B models.

The coefficient paths of the other competing retailer densities indicate that Chain A and Chain B tend to locate in neighborhoods distinct from those in which large-store formats operate. On the other hand, Chain C clearly prefers markets with retail agglomeration. Figure 8 shows that predictors associated with big-box mass merchandisers, supercenters, and wholesale club stores are consistently positive and increasing in magnitude for Chain C. Despite the clear statistically significant relationships revealed in our models, conventional wisdom says little about the importance of larger store formats in determining dollar store locations.

In contrast, the associations between the same big-box store channels, Chain A, and to a lesser extent, Chain B, are consistently negative or decline over time, switching from positive to negative. Along with dollar stores, the supercenter format experienced one of the largest growth rates in terms of store counts from 2000 to 2020. However, the fact that the coefficient magnitude associated with supercenters declines over time in the Chain A models suggests that the chain is increasingly locating in markets in which supercenters (e.g., Wal-Mart) cannot viably operate. While the result aligns with media reports citing that communities distant from big-box competition are more profitable for dollar stores (Nassauer, 2017), we show that the location strategy has actually evolved over time and varies by dollar store chain. In the year 2000, the coefficient estimates of supercenter densities in the Chain A

[14]Note that we do not include the drug and convenience store channels in the urban-area models because they are almost perfectly correlated with population and grocery store densities. To facilitate comparisons between urban- and small-town/rural models, we do not include the convenience and drug store estimates in the small-town/rural area models for Figure 8.

models are positive in each geographic area. By the year 2020, the estimates for the same predictor are negative.

Note that for the Dollar Stores models, (1^{st} columns in each panel of Figure 8), the predictor estimates and coefficient paths mostly contradict the conditional relationships observed for individual dollar store chain models. For example, in the OLS small-town/rural area model (bottom-right panel of Figure 8), the coefficient path of five-mile supercenter density mirrors that of Chain A and Chain B but clearly misrepresents the relationship with Chain C, underscoring again that the nature of dollar store growth is deeply nuanced by chain and geographic market.

4.5 Summary and Discussion

To summarize the elastic net regularization and post-selection regression patterns across predictors and predictor categories, we generate bar plots of the counts and shares of the number of times predictors are selected over time for each dollar store type for the urban and small-town/rural area models. In the elastic net algorithm, predictors that are not sufficiently correlated with the outcome conditional on already-included predictors can be eliminated from the selected model. The variable selection feature of the elastic net provides us with valuable information regarding the relative importance of predictors for each dollar store chain type.

Figures 9 and 10 plot the number of times variables are selected by the logistic and OLS elastic net algorithms for each dollar store type in urban and small-town/rural models, respectively.[15] Each bar corresponding to a model predictor provides information about the number of times the variable is selected and whether the coefficient sign is positive or negative, summarizing the relationship of the variable with the outcome variable. Bars with a single color suggest a concise, one-way relation, while bars with a balance of color imply

[15]The summary figures include all model predictors except for the state-level fixed effects.

that the predictor-outcome variable relation may vary over time. Predictors are organized by category, which we define in terms of demographic, socioeconomic, market geography, and retail features.

In Figure 11, for each dollar store model and geographic area, we select the predictor(s) from each of the four variable categories with the highest share of statistically significant post-selection coefficients across the 21 model years (2000-2020). Specifically, we choose the predictors whose share of statistically significant post-selection estimates at the $\alpha <= 0.001$ confidence level are in the 90^{th} percentile of statistically significant shares for each feature category. The top row of bar charts shows predictors for urban-area models, while the bottom shows predictors for small-town/rural areas.

The figures highlight several of the most important characteristics that determine the location strategies of each dollar store chain. Figures 9 and 10 indicate that, relative to Chain A and Chain B, the demographic, socioeconomic, and market geography predictors are less frequently selected in the Chain C logistic and OLS elastic net models. Several predictors in these feature categories are selected less than half of the total 21 years. The retail predictors, on the other hand, are selected with more frequency in Chain C models, especially for predictors corresponding to large-store formats.[16] The demographic, socioeconomic, and market geography features in the Chain A and Chain B models are more frequently selected as highly predictive. For Chain A, the socioeconomic and market geography predictors, such as the population share with at least a bachelor's degree, average worker commuting travel time, and distance to the most proximate distribution center, are usually selected and enter the logistic and OLS models in both urban and small-town/rural areas with a negative sign in each model year. On the other hand, the neighborhood poverty rate and share of the population that is Black and Hispanic are some of the most selected socioeconomic and

[16]In several instances of the Chain A and Chain B models, the wholesale club format three-to-five and five-to-ten mile densities are selected more frequently than the three- and five-mile densities and have positive coefficients, suggesting that, as opposed to co-locating with the big-box store format, Chain A and Chain B dollar stores often locate in-between larger stores to shorten the distance that consumers must travel to obtain everyday household products.

demographic variables for Chain B, reflecting their preference for locating in multi-racial communities with poor economic conditions.

Figure 11 reveals that there is very little overlap between the most selected predictors in Chain C models and in models of the other two chains. Big-box stores, including supercenters, mass merchandisers, and wholesale club stores are the retail competition predictors with the highest share of statistically significant post-selection estimates for Chain C. The most selected socioeconomic predictors of Chain A and Chain B sometimes overlap, reflecting their tendency to both locate in economically distressed communities. But, Chain B displays a clear difference relative to Chain A in that the population shares that are Black and Hispanic are more often Chain B's top demographic predictors.

We draw seven generalized findings from the results, each of which informs the literature on dollar stores, corroborates or refutes the conventional wisdom of the format, and suggests future research on the topic.

1. The relationship between dollar store entry and densities and neighborhood income and poverty is more nuanced than the narrative in popular press and most research to date. The coefficient paths in the Dollar Stores models tend to reflect the coefficient paths of specific chains, which have changed over time. Since 2015, the neighborhood poverty rate is negatively associated with the density of Chain A in small-town and rural areas, while the association between median household income and Chain B store entry and densities in urban areas becomes moderately positive beginning in 2016.

2. The predictive power of race varies importantly across dollar store chains. Contrary to many media reports that dollar stores are mostly concentrated in majority Black and Hispanic areas, our data-driven models reveal that the effect of neighborhood race on predicting Chain A store locations is mixed, varying significantly by model (logistic or OLS) and geographic market (urban or small-town/rural), while the likelihood of Chain B entry and store densities increase most with the population share that is Black

and Hispanic. The neighborhood's race composition is less predictive of Chain C.

3. For Chain A and Chain B, socioeconomic factors such as education and unemployment levels, neighborhood housing environment, and household mobility largely conform to the conventional narrative that dollar stores are located in resource-constrained communities, though the set of socioeconomic features most important in each model varies for the two chains. The predictive power of socioeconomic predictors in Chain C models is mixed.

4. Dollar stores are expanding at increasing distances from their distribution centers. As they expand their store network over time, each dollar store chain may be leveraging economies of density in distribution. This may also suggest that these chains are receiving food and beverage products from wholesalers or distributors, and that the role of dollar stores in determining food access or food insecurity may not be explained by chain infrastructure.

5. While still maintaining strong growth in rural areas, dollar stores tend to locate in more populous areas over time, though we note some exceptions in the case of Chain A. Distinct from the other two dollar store chains, measures of centrality, including public schools, places of worship, and mobile home parks, are consistently strong, positive predictors of Chain A store locations. Even in small-town and rural markets, Chain A is attracted to areas with increased centrality.

6. The associations between dollar store locations and the density of competing retailers is heterogeneous, varying substantially by chain. Grocery store densities are mostly negatively correlated with Chain A but tend to be positive for Chain B and Chain C, holding other factors constant. This suggests that, of the three largest dollar store chains, Chain A is the most likely to view supermarkets as direct competitors and that the chain may have the largest impact on increasing food access in communities with few alternative store options.

7. Based on the negative and/or declining estimates associated with mass merchandisers, supercenters, and wholesale club store densities, Chain A and Chain B increasingly locate in areas more distant from large-store formats. The two dollar store chains may seek to appeal to more proximate price-conscious households with economic, transportation, and retail access constraints. Chain C, on the other hand, consistently prefers co-locating with big-box stores to attract nearby customers.

5 Conclusion

Popular press, policymakers, and activists aver that dollar stores tend to concentrate in economically distressed, majority Black and Hispanic, densely populated cities and sparsely populated small towns and rural communities (Debter, 2022; Donahue, 2018; Sainato, 2019). Dollar store research, however, has mostly been limited to a handful of cities or use a single narrative, missing or obscuring the heterogeneous characteristics of individual chains across geographic markets (Chenarides et al., 2021; Donahue & Bonestroo, 2019; Shannon, 2020). In this paper, we comprehensively study the entry and density of dollar stores in the United States from 2000 to 2020, as a sector and by chain, applying data-driven machine learning techniques to a novel, spatially and temporally dynamic database of dollar stores, competing retailers, and census-tract level information on neighborhood demographics, socioeconomics, and market geography features. We describe the expansion of dollar stores from small regional discount retailers to a national store format and evaluate the extent to which the conventional narrative surrounding dollar store locations is supported by the data. We draw several generalized facts about dollar store growth in the United States.

We find that demographics, socioeconomics, market geography, and measures of retail competition all have predictive power for dollar store entry and density, but the set of relevant predictors varies significantly across chains. The predictive power of demographic and socioeconomic features in models using all dollar stores as the outcome variable are mostly

consistent with the conventional narrative of dollar store entry and densities. However, dollar store sector-level models tend to reflect a combination of the model estimates from specific chains, obfuscating the nuance of location strategies across dollar store chains and geographic markets.

Our results reveal that there are important differences in terms of the growth patterns and significant predictors for each retailer. Conditioned on geographic market (i.e., urban or small-town/rural areas), no machine learning model yields the same predictive insights for all three chains. Moreover, we show that many of the predictors of dollar store entry and densities have changed over time. In some cases, predictor point estimates even switch signs, suggesting that dollar store location strategies are evolving as these stores compete for sales of household goods and food. In particular, a salient finding from our models is that, since the year 2000, some dollar stores (i.e., Chain A and Chain B) have gradually distanced their stores from markets in which big-box supercenters are present, better positioning their stores as the local discount retailers.

Each machine learning algorithm indicates that distance to the nearest distribution center is one of the most predictive features of dollar store entry and densities. The elastic net estimates show that the impact of distance gradually declines each year. As they locate further from distribution centers over time, dollar store chains likely have developed more efficient distribution networks. At this stage, very little is known about how dollar stores leverage economies of density or manage their logistics for food and other household goods. The extent to which dollar stores invest in refrigerated storage both in their stores and throughout their supply chain (e.g., distribution centers and reefer transportation) or are partnering with established food wholesalers and manufacturing companies, will be crucial in explaining the capacity for growth in household food expenditures in this format (Redman, 2022; Repko, 2021). Studies using mixed methods can help answer these questions and, in turn, better understand the growth and impact of dollar store chains in the food system.

Since the early- and mid-1990s, researchers and policymakers have noted the increasing influence that non-traditional food retailers have in food marketing (e.g., supercenters and wholesale club stores) (Bonanno & Goetz, 2012; Crowley & Stainback, 2019; Hortaçsu & Syverson, 2015; Kaufman, 1998; Newton, 1993). Perhaps due to their small-store formats, tendency to operate in under-served communities, and more recent transition to food sales at the national scale, little empirical work has investigated the economic impacts of dollar stores on food retail. Fruitful avenues for research include the study of the effects of dollar stores on food access, food assistance benefits, food expenditures, dietary quality, retail competition, food prices, and local economic development. As dollar store research accelerates with the proliferation of dollar stores throughout the United States, our analyses suggest that dollar store impacts identified in studies may also be heterogeneous, varying by chain and geographic market, in urban cities, small-towns, and rural communities.

References

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2020). lassopack: Model selection and prediction with regularized regression in Stata. *The Stata Journal*, *20*(1), 176–235.

Alviola IV, P. A., Nayga Jr, R. M., Thomsen, M. R., & Wang, Z. (2013). Determinants of food deserts. *American Journal of Agricultural Economics*, *95*(5), 1259–1265.

Amin, M. D., Badruddoza, S., & McCluskey, J. J. (2020). Predicting access to healthful food retailers with machine learning. *Food Policy*, 101985.

Athey, S. (2019). The Impact of Machine Learning on Economics. In *The economics of artificial intelligence* (pp. 507–552). University of Chicago Press.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*(2), 521–547.

Belloni, A., Chernozhukov, V., & Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, *34*(4),

606–619.

Bendix, A. (2018, 12). *Dollar stores are feeding more Americans than Whole Foods, and it's leading some communities into crisis* (US edition ed.). New York. Retrieved from https://login.lp.hscl.ufl.edu/login?url=https://www.proquest.com/newspapers/dollar-stores-are-feeding-more-americans-than/docview/2406926025/se-2?accountid=10920http://resolver.ebscohost.com/openurl?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rfr_id=i

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967.

Birmingham City Council. (2019). *Mayor Woodfin secures successful passage of amendment to help reduce food deserts in Birmingham* (Tech. Rep.). Birmingham, AL. Retrieved from https://www.birminghamal.gov/2019/07/10/mayor-woodfin-secures-successful-passage-of-amendment-to-help-reduce-food-deserts-in-birmingham/

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076.

Bonanno, A., Chenarides, L., & Goetz, S. J. (2012). *Limited Food Access as an Equilibrium Outcome: An Empirical Analysis* (Tech. Rep.).

Bonanno, A., & Goetz, S. J. (2012). WalMart and local economic development: A survey. *Economic Development Quarterly*, *26*(4), 285–297.

Breiman, L. (2001a). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199–231.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In *2012 ieee international geoscience*

and remote sensing symposium (pp. 5372–5375). IEEE.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453. Retrieved from https://www.sciencedirect.com/science/article/pii/S095741741101342X doi: https://doi.org/10.1016/j.eswa.2011.09.033

Brown, W. (2022, 4). *As Dollar Stores Proliferate, Some Communities Push Back.* Retrieved from https://civileats.com/2022/04/13/dollar-stores/

Cain, A. (2021). *Dollar General CEO details the dollar chain's modernization strategy as it looks to compete with Walmart in fresh grocery and as a one-stop shop.* Retrieved from https://www.businessinsider.com/dollar-general-ceo-todd-vasos-talks-grocery-future-walmart-2021-2#:~:text=InsiderspoketoCEOTodd,ofInsiderformorestories.

Canfield, K. (2018). *City Council approves restrictions on dollar stores in north Tulsa.* Retrieved from https://tulsaworld.com/news/local/city-council-approves-restrictions-on-dollar-stores-in-north-tulsa/article_994a90f3-0609-51db-9a07-6278e7abe3e0.html

Capelouto, J. (2020). *DeKalb County again extends temporary ban on dollar stores.* Retrieved from https://www.ajc.com/news/local/dekalb-county-extends-temporary-ban-dollar-stores/4JXRx9YxV90qndrHYKRprL/

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

Chenarides, L., Cho, C., Nayga, R. M., & Thomsen, M. R. (2021). Dollar stores and food deserts. *Applied Geography*, *134*, 102497. Retrieved from https://www.sciencedirect.com/science/article/pii/S0143622821001132 doi: https://doi.org/10.1016/j.apgeog.2021.102497

Chenarides, L., & Jaenicke, E. C. (2019). Documenting the Link Between Poor Food Access

and Less Healthy Product Assortment Across the U.S. *Applied Economic Perspectives and Policy*, *41*(3), 434–474. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1093/aepp/ppy018 doi: 10.1093/aepp/ppy018

Cho, C., W McLaughlin, P., Zeballos, E., Kent, J., & Dicken, C. (2019). *Capturing the Complete Food Environment With Commercial Data: A Comparison of TDLinx, ReCount, and NETS Databases* (Tech. Rep.).

Coad, A., & Srhoj, S. (2019). Catching Gazelles with a Lasso: Big data techniques for the prediction of high-growth firms. *Small Business Economics*, 1–25.

Cooper, R. K. (2016). *Dollar General poised to receive $11 million in tax breaks.* Retrieved from https://www.bizjournals.com/albany/news/2016/01/11/dollar-general-poised-to-receive-11-million-in-tax.html

Corkery, M. (2021). *'Everything Going the Wrong Way': Dollar Stores Hit a Pandemic Downturn.* Retrieved from https://www.nytimes.com/2021/09/30/business/dollar-stores-struggling-pandemic.html

Crowley, M., & Stainback, K. (2019). Retail Sector Concentration, Local Economic Structure, and Community Well-Being. *Annual Review of Sociology*, *45*(1), 321–343. Retrieved from https://doi.org/10.1146/annurev-soc-073018-022449 doi: 10.1146/annurev-soc-073018-022449

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In *Ensemble machine learning* (pp. 157–175). Springer.

Debter, L. (2022). *How Dollar General Is Spreading Like Hot Gossip In Tiny Towns Across The Country.* Retrieved from https://www.forbes.com/sites/laurendebter/2022/05/20/dollar-general-opening-new-stores-across-small-town-america/?sh=44dcb72ce5f9

Del Mastro, A. (2021). Dollar Stores And Reinvention Of Rural Retail. *The American Conservative*. Retrieved from https://www.theamericanconservative.com/urbs/dollar-stores-and-reinvention-of-rural-retail/

Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, *7*(1), 1–13.

Dollar General Corporation. (1999). *Dollar General, Form 10-K, Annual Report.* Retrieved from https://sec.report/Document/0000914317-00-000307/

Dollar General Corporation. (2010). *Dollar General, Form 10-K, Annual Report.* Retrieved from https://sec.report/Document/0001047469-10-003012/

Dollar General Corporation. (2015). *Dollar General, Form 10-K, Annual Report.* Retrieved from https://sec.report/Document/0001047469-15-002540/

Dollar General Corporation. (2021). *Dollar General, Form 10-K, Annual Report.* Retrieved from https://sec.report/Document/0001558370-21-003245/

Dollar Tree Inc. (2010). *Dollar Tree, Form 10-K, Annual Report.* Retrieved from https://sec.report/Document/0000935703-10-000010/

Dollar Tree Inc. (2020). *Dollar Tree, Form 10-K, Annual Report.* Washington, DC.. Retrieved from https://sec.report/Document/0000935703-20-000006/

Dollar Tree Inc. (2021). *Dollar Tree, Form 10-K, Annual Report.* Retrieved from https://sec.report/Document/0000935703-21-000014/

Dollar Tree Inc. (2022). *Dollar Tree, Form 10-K, Annual Report.* Retrieved from https://sec.report/Document/0000935703-22-000020/

Donahue, M. (2018). *The Impact of Dollar Stores and How Communities Can Fight Back (Fact Sheet).* Retrieved from https://ilsr.org/dollar-stores-factsheet/

Donahue, M., & Bonestroo, H. (2019). *Maps Show Alarming Pattern of Dollar Stores' Spread in U.S. Cities.* Washington, DC.. Retrieved from https://ilsr.org/new-maps-dollar-stores-spread/

Donahue, M., & Smith, K. (2022). *Dollar Store Restrictions* (Tech. Rep.). Institute for Local Self-Reliance. Retrieved from https://ilsr.org/rule/dollar-store-dispersal-restrictions/

Drichoutis, A. C., Nayga, R. M., Rouse, H. L., & Thomsen, M. R. (2015). Food environment

and childhood obesity: the effect of dollar stores. *Health economics review*, *5*(1), 37.

Dutko, P., Ver Ploeg, M., & Farrigan, T. (2012). *Characteristics and influential factors of food deserts* (Tech. Rep.).

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of animal ecology*, *77*(4), 802–813.

Ellickson, P. B., & Grieco, P. L. (2013). Wal-Mart and the geography of grocery retailing. *Journal of Urban Economics*, *75*, 1–14. Retrieved from https://www.sciencedirect.com/science/article/pii/S0094119012000630 doi: 10.1016/J.JUE.2012.09.005

Family Dollar Stores Inc. (2007). *Family Dollar, Form 10-K, Annual Report.* Retrieved from https://sec.report/Document/0001193125-07-229693/

Family Dollar Stores Inc. (2010). *Family Dollar, Form 10-K, Annual Report.* Retrieved from https://sec.report/Document/0001193125-10-236638/

Family Dollar Stores Inc. (2014). *Family Dollar, Form 10-K, Annual Report.* Washington, DC.. Retrieved from https://sec.report/Document/0000034408-14-000010/

Ferreira, A. J., & Figueiredo, M. A. T. (2012). Boosting algorithms: A review of methods, theory, and applications. *Ensemble machine learning*, 35–85.

FOX 4 News Dallas-Fort Worth. (2018). *New Mesquite ordinance will limit number of dollar stores.* Retrieved from https://www.fox4news.com/news/new-mesquite-ordinance-will-limit-number-of-dollar-stores

Fred's Inc. (2019). *Fred's Form 10-K, Annual Report.* Retrieved from https://sec.report/Document/0001564590-19-015740/fred-10k_20190202.htm

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, *14*(771-780), 1612.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of statistics*, *28*(2), 337–407.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear

models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367–378.

Genuer, R., Poggi, J.-M., & Tuleau, C. (2008). Random Forests: some methodological insights. *arXiv preprint arXiv:0811.3619*.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0167865510000954` doi: https://doi.org/10.1016/j.patrec.2010.03.014

Glaeser, E. L., Kominers, S. D., Luca, M., & Naik, N. (2018, 1). BIG DATA AND BIG CITIES: THE PROMISES AND LIMITATIONS OF IMPROVED MEASURES OF URBAN LIFE. *Economic Inquiry*, *56*(1), 114–137. Retrieved from `https://doi.org/10.1111/ecin.12364` doi: https://doi.org/10.1111/ecin.12364

Goetz, S. J., & Swaminathan, H. (2006). Wal-Mart and county-wide poverty. *Social Science Quarterly*, *87*(2), 211–226.

Greenwell, B. M., & Boehmke, B. C. (2019). *Variable Importance Plots: An Introduction to Vip.*

Hals, S. S. Y., Tom; Kurane. (2015). *Grocery store chain A&P files for bankruptcy again.* Retrieved from `https://www.reuters.com/article/greatatlantic-bankruptcy/grocery-store-chain-ap-files-for-bankruptcy-again-idUSL1N1010EH20150721`

Hart, A. (2019). *In A Fight For Healthier Food, Fort Worth Is Fending Off Dollar Stores.* Retrieved from `https://www.kut.org/texas/2019-12-19/in-a-fight-for-healthier-food-fort-worth-is-fending-off-dollar-stores`

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data*

mining, inference, and prediction. Springer Science & Business Media.

Higgs, R. (2020). *Cleveland Mayor Frank Jackson OKs ban on new dollar stores while city considers new regulations.* Retrieved from https://www.cleveland.com/cityhall/2020/06/cleveland-mayor-frank-jackson-oks-ban-on-new-dollar-stores-while-city-considers-new-regulations.html

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Holmes, T. J. (2011). The diffusion of Wal-Mart and economies of density. *Econometrica*, *79*(1), 253–302.

Hortaçsu, A., & Syverson, C. (2015). The ongoing evolution of US retail: A format tug-of-war. *Journal of Economic Perspectives*, *29*(4), 89–112.

Howard, K., & Fleming, M. (2019). *Moratorium Passed To Fight 'Food Desert' In Northeast Oklahoma City.* Retrieved from https://www.kgou.org/post/moratorium-passed-fight-food-desert-northeast-oklahoma-city

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794.

Jimenez, C. (2019). *Fort Worth passes regulations to limit dollar stores.* Retrieved from https://www.star-telegram.com/news/local/fort-worth/article238036189.html

Junior Achievement of the Carolinas Inc. (2003). *Laureate: Leon Levine.* Retrieved from https://www.historync.org/laureate-LeonLevine.htm

Karasiak, N., Dejoux, J.-F., Monteil, C., & Sheeren, D. (2021). Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing. *Machine Learning*. Retrieved from https://doi.org/10.1007/s10994-021-05972-1 doi: 10.1007/s10994-021-05972-1

Kaufman, P. R. (1998). Nontraditional retailers are challenging traditional grocery stores.

Food Review/National Food Review, *21*(1482-2016-121513), 31–33.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53*(11), 3735–3745. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167947309001601 doi: https://doi.org/10.1016/j.csda.2009.04.009

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Kumcu, A., & Kaufman, P. R. (2011). *Food spending adjustments during recessionary times* (Tech. Rep.).

Lee, J. (2021). *Dollar Stores Deserve a Break; Dollar General gave muted guidance for 2021, but there are plenty of tailwinds that should help its main customer base.* New York, N.Y.. Retrieved from https://login.lp.hscl.ufl.edu/login?url=https://www.proquest.com/newspapers/dollar-stores-deserve-break-general-gave-muted/docview/2502300879/se-2?accountid=10920http://resolver.ebscohost.com/openurl?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rfr_id=

Levin, D., Noriega, D., Dicken, C., Okrent, A. M., Harding, M., & Lovenheim, M. (2018). *Examining food store scanner data: A comparison of the IRI InfoScan data with other data sets, 2008–2012* (Tech. Rep.).

MacGillis, A. (2020). *How Dollar Stores Became Magnets for Crime and Killing* (Tech. Rep.). ProPublica. Retrieved from https://www.propublica.org/article/how-dollar-stores-became-magnets-for-crime-and-killing

McGreal, C. (2018). *Where even Walmart won't go: how Dollar General took over rural America.* Retrieved from https://infoweb.newsbank.com/apps/news/document-view?p=WORLDNEWS&docref=news/16DC84C8A10FCE60

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction. *Ecological Modelling*, *411*, 108815.

Meyersohn, N. (2021a). *Dollar stores are starting to offer fresh food after years*

of criticism. Retrieved from https://www.cnn.com/2021/06/02/business/dollar-general-family-dollar-fresh-food/index.html

Meyersohn, N. (2021b). *Nearly 1 in 3 new stores opening in the US is a Dollar General.* Retrieved from https://www.cnn.com/2021/05/06/business/dollar-store-openings-retail/index.html

Misra, T. (2018). The Dollar Store Backlash Has Begun. *Bloomberg CityLab*. Retrieved from https://www.bloomberg.com/news/articles/2018-12-20/when-the-closest-grocery-store-is-a-dollar-store

Mitchell, M. W. (2011). Bias of the Random Forest out-of-bag (OOB) error for certain input parameters. *Open Journal of Statistics*, *1*(03), 205.

Mitchell, S., & Donahue, M. (2018). *Dollar Stores Are Targeting Struggling Urban Neighborhoods and Small Towns. One Community Is Showing How to Fight Back.* Retrieved from https://ilsr.org/dollar-stores-target-cities-towns-one-fights-back/

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005, 8). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, *21*(15), 3301–3307. Retrieved from https://doi.org/10.1093/bioinformatics/bti499 doi: 10.1093/bioinformatics/bti499

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Moore, L. V., & Diez Roux, A. V. (2006, 2). Associations of Neighborhood Characteristics With the Location and Type of Food Stores. *American Journal of Public Health*, *96*(2), 325–331. Retrieved from https://doi.org/10.2105/AJPH.2004.058040 doi: 10.2105/AJPH.2004.058040

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

Morris, F. (2017). *How Dollar General Is Transforming Rural America.* Na-

tional Public Radio, Inc. (NPR). Retrieved from http://lp.hscl.ufl.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=edsggo&AN=edsgcl.518440768&site=eds-live

Nassauer, S. (2017). *How Dollar General Became Rural America's Store of Choice.* New York, N.Y.. Retrieved from https://login.lp.hscl.ufl.edu/login?url=https://search.proquest.com/docview/1972105377?accountid=10920http://resolver.ebscohost.com/openurl?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rfr_id=info:sid/ProQ%3Aabiglobal&rft_val_fmt=info:ofi/fmt:kev:mtx:j

Nassauer, S. (2022). *Dollar General, Dollar Tree Stores Show Strength in Face of Inflation.* Retrieved from https://www.wsj.com/articles/dollar-stores-earnings-show-surprising-strength-in-face-of-inflation-11653567874

Natekin, A., & Knoll, A. (2013, 12). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21. Retrieved from https://pubmed.ncbi.nlm.nih.gov/24409142https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3885826/ doi: 10.3389/fnbot.2013.00021

Newton, D. J. (1993). Nontraditional retailers challenge the supermarket industry. *Food Review/National Food Review*, *16*(1482-2017-3314), 2–7.

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC bioinformatics*, *11*(1), 1–13.

Nielsen. (2022). *Retail Trade Channel and Sub-Channel Definitions.* Retrieved from https://environicsanalytics.com/docs/default-source/us---data-product-support-documents/tdlinx-retail-channel-sub-channel-definitions.pdf

Nilsson, I. M., & Smirnov, O. A. (2016). Measuring the effect of transportation infrastructure on retail firm co-location patterns. *Journal of Transport Geography*, *51*, 110–118. Retrieved from https://www.sciencedirect.com/science/article/pii/S0966692315002227 doi: https://doi.org/10.1016/j.jtrangeo.2015.12.002

Piller, J., & Strong, J. S. (2015). *The High Price of Dollar Stores: Dollar Tree and Dollar General Battle for Family Dollar* (Tech. Rep.). Babson Park, MA: Raymond A. Mason School of Business Foundation Board. Retrieved from https://www.babson.edu/academics/executive-education/babson-insight/strategy-and-innovation/the-high-price-of-dollar-stores/#

Powell, L. M., Slater, S., Mirtcheva, D., Bao, Y., & Chaloupka, F. J. (2007). Food store availability and neighborhood characteristics in the United States. *Preventive medicine*, *44*(3), 189–195.

Probst, P., & Boulesteix, A.-L. (2017). To tune or not to tune the number of trees in random forest. *J. Mach. Learn. Res.*, *18*(1), 6673–6690.

Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research*, *20*(1), 1934–1965.

Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1301.

Racine, E. F., Batada, A., Solomon, C. A., & Story, M. (2016). Availability of Foods and Beverages in Supplemental Nutrition Assistance Program: Authorized Dollar Stores in a Region of North Carolina. *Journal of the Academy of Nutrition and Dietetics*, *116*(10), 1613–1620.

Redman, R. (2022). *Dollar General pushes ahead with brick-and-mortar expansion.* Retrieved from https://www.supermarketnews.com/retail-financial/dollar-general-pushes-ahead-brick-and-mortar-expansion

Repko, M. (2021). *Dollar stores got aggressive as the rest of retail hunkered down, and Wall Street likes the strategy.* Retrieved from https://www.cnbc.com/2021/01/04/dollar-stores-got-aggressive-as-the-rest-of-retail-hunkered-down.html

Rhone, A., Ver Ploeg, M., Dicken, C., Williams, R., & Breneman, V. (2017). *Low-income*

and low-supermarket-access census tracts, 2010-2015 (Tech. Rep.).

Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package. *Update*, *1*(1), 2007.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... Dormann, C. F. (2017, 8). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. Retrieved from https://doi.org/10.1111/ecog.02881 doi: https://doi.org/10.1111/ecog.02881

Sainato, M. (2019). Dollar Stores Prey on the Poor: Communities are doing what they can to impose restrictions. *The Progressive VO - 83*(5), 32. Retrieved from http://lp.hscl.ufl.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=edsgov&AN=edsgcl.604716412&site=eds-live

Sampson, R. J. (2016). Individual and community economic mobility in the Great Recession era: The spatial foundations of persistent inequality. *Economic mobility: Research and ideas on strengthening families, communities and the economy*, 261–287.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109–120. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304380019302145 doi: https://doi.org/10.1016/j.ecolmodel.2019.06.002

Seim, K. (2006). An empirical model of firm entry with endogenous product-type choices. *The RAND Journal of Economics*, *37*(3), 619–640.

Shannon, J. (2020). Dollar Stores, Retailer Redlining, and the Metropolitan Geographies of Precarious Consumption. *Annals of the American Association of Geographers*, 1–19.

Shannon, J., Shannon, S., Adams, G. B., & Lee, J. S. (2016). Growth in SNAP retailers was associated with increased client enrollment in Georgia during the Great Recession. *Health Affairs*, *35*(11), 2100–2108.

Sharkey, P., & Faber, J. W. (2014). Where, when, why, and for whom do residential contexts

matter? Moving away from the dichotomous understanding of neighborhood effects. *Annual review of sociology*, *40*, 559–579.

Shrestha, S. (2016). Dollars to dimes: Disparity, uncertainty, and marketing to the poor at US dollar stores. *International Journal of Cultural Studies*, *19*(4), 373–390. Retrieved from https://doi.org/10.1177/1367877913515869 doi: 10.1177/1367877913515869

Siegel, R. (2019, 2). *As dollar stores move into cities, residents see a steep downside.* Retrieved from https://www.washingtonpost.com/business/economy/as-dollar-stores-move-into-cities-residents-see-a-steep-downside/2019/02/15/b3676cbe-2f09-11e9-8ad3-9a5b113ecd3c_story.html

Sutton, C. D. (2005). 11 - Classification and Regression Trees, Bagging, and Boosting. In C. R. Rao, E. J. Wegman, & J. L. B. T. H. o. S. Solka (Eds.), *Data mining and data visualization* (Vol. 24, pp. 303–329). Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169716104240111 doi: https://doi.org/10.1016/S0169-7161(04)24011-1

Tay, J. K., Narasimhan, B., & Hastie, T. (2021). Elastic Net Regularization Paths for All Generalized Linear Models. *arXiv preprint arXiv:2103.03475*.

Thomas, B. (2021). *Dollar & Variety Stores in the US, Industry Report 45299* (Tech. Rep.). IBISWorld.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Toledo City Plan Commisssion. (2020). *Study of Small Box Discount Sotres, per Ord. 166-20* (Tech. Rep.). Toledo, Ohio: Toledo-Lucas County Plan Commissions. Retrieved from https://toledo.legistar.com/LegislationDetail.aspx?ID=4742265&GUID=B9AF9025-345E-4FE1-82F4-74D1A0C2E595&G=E5BDDE21-C978-430C-804D-EE97005A5E73&Options=&Search=

Turner Jr., C. (2018). *My Father's Business*. Hachette Book Group, Inc.

U.S. Department of Homeland Security. (2022). *Homeland Infrastructure Foundation-Level Data (HIFLD).* Retrieved from https://hifld-geoplatform.opendata.arcgis.com/

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2018). blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *bioRxiv*, 357798.

Volpe, R., Kuhns, A., & Jaenicke, T. (2017). *Store formats and patterns in household grocery purchases* (Tech. Rep.).

Wahba, P. (2020). Why Dollar General thinks coronavirus can help business. *Fortune.com*. Retrieved from http://lp.hscl.ufl.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=buh&AN=142203598&site=eds-live

Wensley, M. R. D., & Stabler, J. C. (1998). Demand-Threshold Estimation for Business Activities in Rural Saskatchewan. *Journal of Regional Science*, *38*(1), 155–177.

Wilde, P., Llobrera, J., & Ver Ploeg, M. (2014). Population density, poverty, and food retail access in the United States: an empirical approach. *International Food and Agribusiness Management Review*, *17*(1030-2016-82991), 171–186.

Williams, C. (2021). *Toledo city council considers new regulations for dollar stores.* Retrieved from https://www.13abc.com/2021/01/05/toledo-city-council-considers-new-regulations-for-dollar-stores/

Wright, M. N., & Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409*.

Yeturu, K. (2020). Chapter 3 - Machine learning algorithms, applications, and practices in data science. In A. S. R. Srinivasa Rao & C. R. B. T. H. o. S. Rao (Eds.), *Principles and methods for data science* (Vol. 43, pp. 81–206). Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169716120300225 doi: https://doi.org/10.1016/bs.host.2020.01.002

Zhang, C., Liu, C., Zhang, X., & Almpanidis, G. (2017). An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*, *82*, 128–150. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417417302397 doi: https://doi.org/10.1016/j.eswa.2017.04.003

Zhu, T., & Singh, V. (2009). Spatial competition with endogenous location choices: An application to discount retailing. *QME*, *7*(1), 1–35.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. Retrieved from https://doi.org/10.1111/j.1467-9868.2005.00503.x doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x

6 Tables

6.1 Variable Tables

Table 1: Retail Store Densities Using the Retail Channel-Subchannel Classifications of TDLinx

Trade Channel	Sub-channel	Examples
Grocery	Conventional Supermarket	Kroger, Food Lion, IGA, Cub Foods
	Limited Assortment	Aldi Food Store, Save-A-Lot, Tesco
	Natural/Gourmet	Trader Joes, Whole Foods, Dean & DeLuca
	Supercenter	Wal-Mart, Meijer Supermarket
	Superette	IGA, Maynards, La Michoacana, Select Markets
	Cash and Carry Warehouse Stores	Cash and Carry, Smart and Final
Drug	Conventional Drug Store	CVS, Walgreen's
	Rx Only and Small Independent Drug Store	United Drug, Medicine Shoppe
Mass, General Merchandiser, and Dollar Store	Mass Merchandiser	Wal-Mart, K-Mart, Target
	General Merchandiser	Trader Horn, Pic-N-Save
	Dollar Store	Dollar General, Dollar Tree, Family Dollar
Wholesale Club	Wholesale Club	Sam's Club, Costco
Convenience Store	Conventional	7-Eleven

Data Sources: NielsenIQ TDLinx retail stores (2000-2020).
Retail store densities are computed for each subchannel in the table using three-, five-, and ten-mile distance bands from population-weighted centroids at the block-group level and aggregated up to the census tract as population-weighted averages.
Inner-ring densities are also created for all retail store channels at the three-to-five and five-to-ten mile distance bands.

Table 2: Demographic, Socioeconomic, Housing, Mobility, Market Geography, and Retail Variables

Demographics	Socioeconomics	Housing and Mobility	Market Geography	Retail Densities
Population (100,000s)	Median household income ($10,000s)	Median house values ($100,000s)	Places of worship	Grocery Stores
Population Shares:	*Population Shares*:	Worker average commute time (Hrs.)	Public schools	General Merchandisers
White	Poverty rate	*Population Shares*:	Mobile home parks	Mass Merchandisers
Black	Receiving public assistance	Vacant housing units	Urban-cluster Area (1/0)	Supercenters
Hispanic	Unemployed	Population without vehicle	Distance to nearest urban area (miles)	Wholesale Club Stores
Asian	Education less than high school diploma		Distance to nearest distribution center (100s of miles)	Convenience Stores
Age 18-34	Bachelor's degree or higher		Interstate miles	Drug Stores
Age 35-65			U.S. highway miles	Dollar Stores
Age 65 and over			State highway miles	Chain A
			Tract area (mi^2)	Chain B
			State Fixed Effects	Chain C
				All Other

Data Sources: U.S. Decennial Census (2000) and ACS Five-Year Period estimates 2005-2009 to 2015-2019; U.S. Census Geography Program; Nielsen TDLinx retail stores (2000-2020).
Dollar store distribution center information is collected from company annual reports and websites.
All demographic, socioeconomic, housing, and mobility variables in the table, with the exception of population, are calculated as spatial window averages within three and five miles of each census tract's population-weighted centroid. The population and tract area variables in the market geography predictor column are computed as spatial sums within three- and five-mile distance bands. Retail store densities are computed within three and five miles and three-to-five and five-to-ten miles of census-tract population-weighted centroids.
All nominal dollar values are converted to real 2019 dollars using the R-CPI-U-RS.
The variables employed in each model depend on the geographic extent (i.e., urban or small-town/rural) and the dollar store model type (all dollar stores or chain specific).

Table 3: Predictors used for National-Scale Machine Learning Models by Market Geography (Urban and Small-Town/Rural)

Urban Models			
Dollar Stores	Chain A	Chain B	Chain C
Grocery Stores	Grocery Stores	Grocery Stores	Grocery Stores
General Merchandisers	General Merchandisers	General Merchandisers	General Merchandisers
Mass Merchandisers	Mass Merchandisers	Mass Merchandisers	Mass Merchandisers
Supercenters	Supercenters	Supercenters	Supercenters
Wholesale Club Stores	Wholesale Club Stores	Wholesale Club Stores	Wholesale Club Stores
	Chain B	Chain A	Chain A
	Chain C	Chain C	Chain B
	All Other Dollar Stores	All Other Dollar Stores	All Other Dollar Stores
Demographic	Demographic	Demographic	Demographic
Socioeconomic	Socioeconomic	Socioeconomic	Socioeconomic
Housing	Housing	Housing	Housing
Mobility	Mobility	Mobility	Mobility
Market Geography*	Market Geography*	Market Geography*	Market Geography*
State Fixed Effects	State Fixed Effects	State Fixed Effects	State Fixed Effects
Total Predictors: 85	91	91	91
Small-Town/Rural Models			
Dollar Stores	Chain A	Chain B	Chain C
Grocery Stores	Grocery Stores	Grocery Stores	Grocery Stores
General Merchandisers	General Merchandisers	General Merchandisers	General Merchandisers
Mass Merchandisers	Mass Merchandisers	Mass Merchandisers	Mass Merchandisers
Supercenters	Supercenters	Supercenters	Supercenters
Wholesale Club Stores	Wholesale Club Stores	Wholesale Club Stores	Wholesale Club Stores
Convenience Stores	Convenience Stores	Convenience Stores	Convenience Stores
Drug Stores	Drug Stores	Drug Stores	Drug Stores
	Chain B	Chain A	Chain A
	Chain C	Chain C	Chain B
	All Other Dollar Stores	All Other Dollar Stores	All Other Dollar Stores
Demographic	Demographic	Demographic	Demographic
Socioeconomic	Socioeconomic	Socioeconomic	Socioeconomic
Housing	Housing	Housing	Housing
Mobility	Mobility	Mobility	Mobility
Market Geography*	Market Geography*	Market Geography*	Market Geography*
State Fixed Effects	State Fixed Effects	State Fixed Effects	State Fixed Effects
Total Predictors: 90	96	96	96

The outcome variable is either dollar store entry, a binary (1/0) variable indicating whether the dollar store locates within three (five) miles of census-tract populations in urban (small-town/rural) models, or is dollar store density, computed in the same distance bands as the entry indicator.

The store-channel predictors are computed as store densities within three and three-to-five miles of census-tract households for the urban-area models and five and five-to-ten miles for the small-town/rural area models.

The urban-area models do not include the convenience and drug store channel densities, as their respective correlation with grocery store channel densities and neighborhood population were approximately 0.95 in urban areas. The small-town/rural area models include the densities for both convenience and drug stores.

*In the small-town/rural area models, we control for differences in dollar store densities in small-town and rural census tracts using the urban-cluster area (1/0) dummy variable.

We use Connecticut and North Carolina as the base variable in the state-fixed effects for the urban- and rural-models, respectively.

7 Figures

7.1 Elastic Net Coefficient Plots

7.1.1 Income, Poverty, and Race

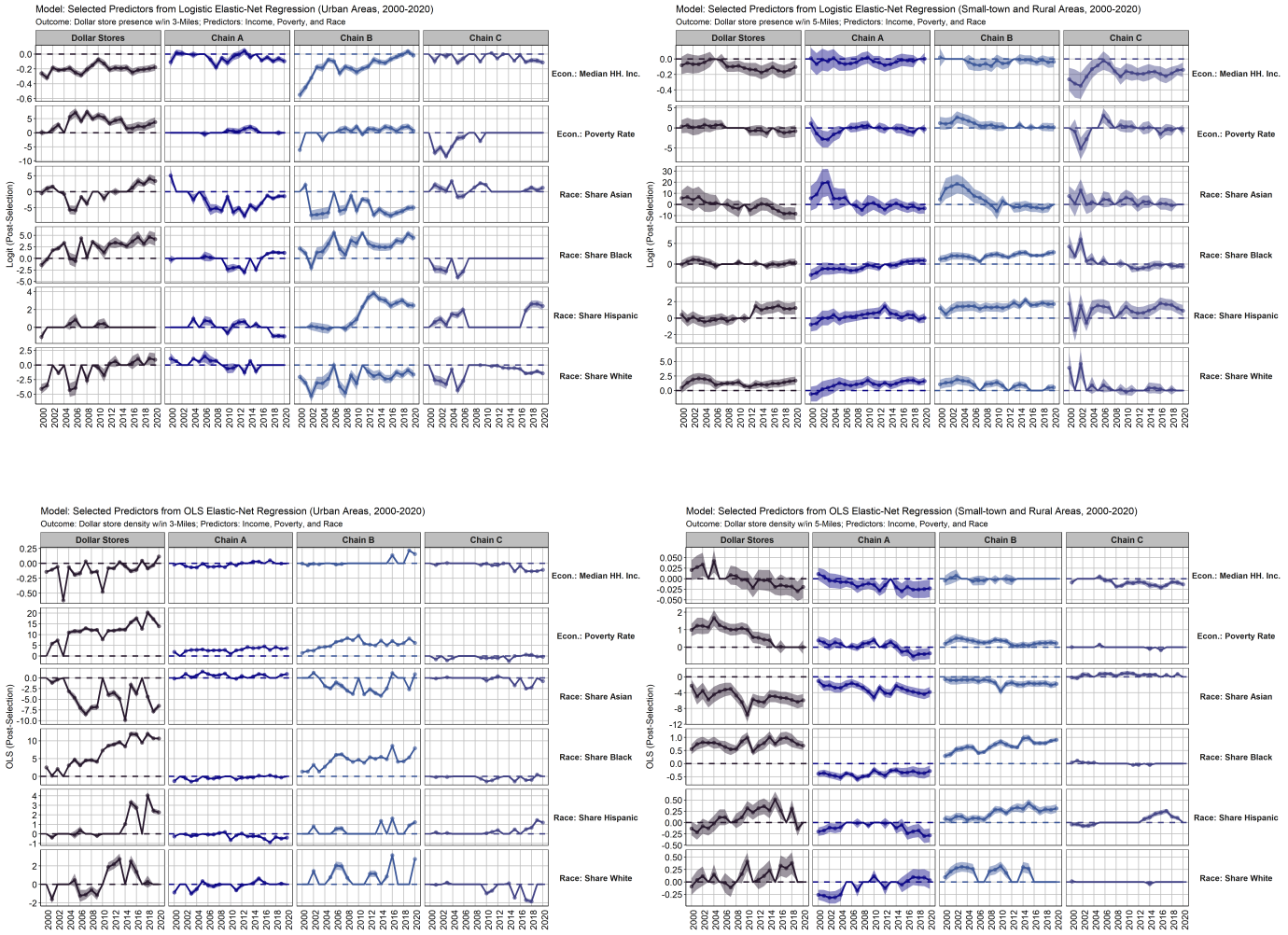

Figure 3: Income, Poverty, and Race Post-Selection Regression Coefficient Estimates for Logistic-Urban (top-left), Logistic-Small Town/Rural (top-right), OLS-Urban (bottom-left) and OLS-Small Town/Rural (bottom-right)

7.1.2 Socioeconomics

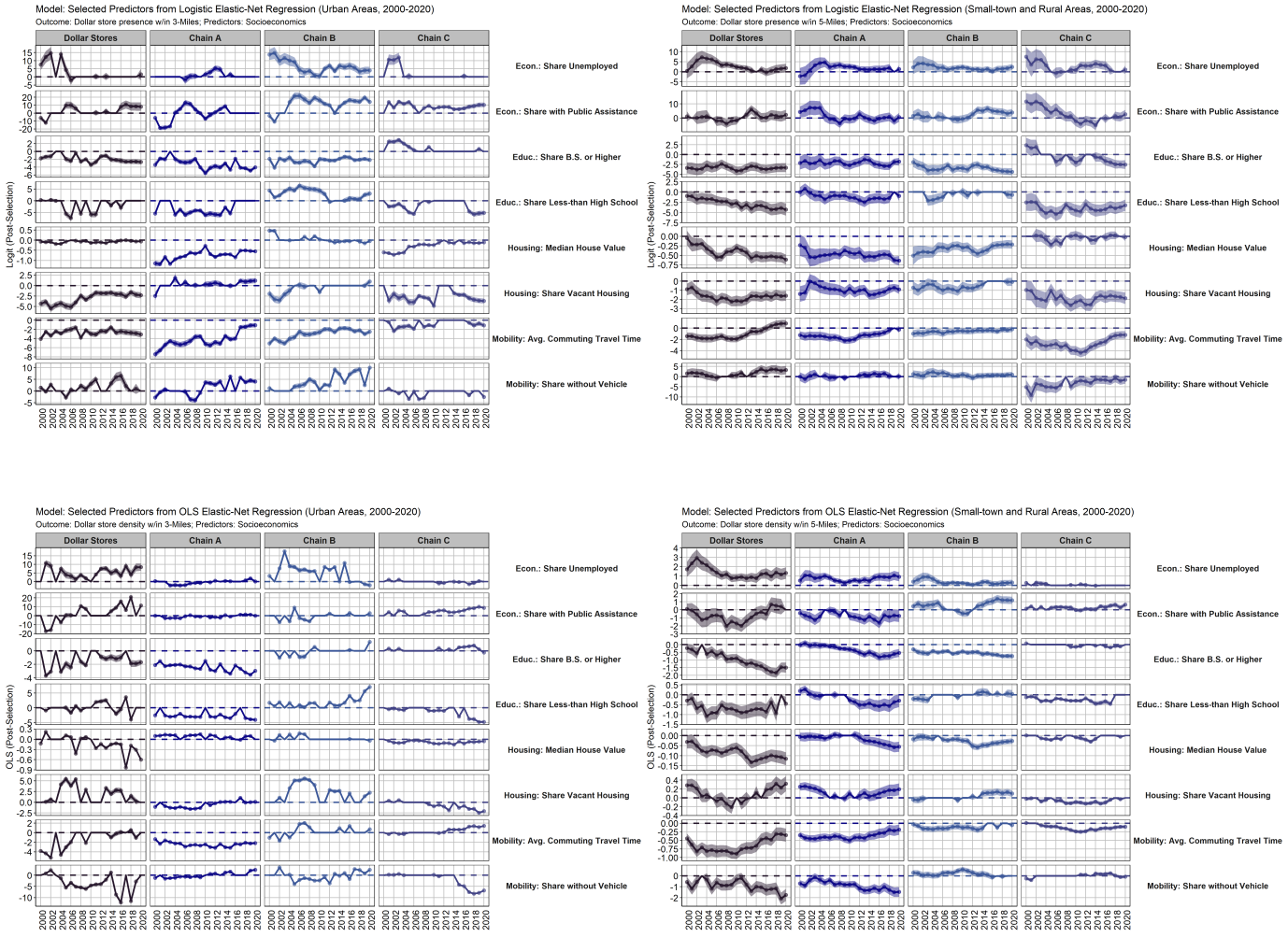

Figure 4: Socioeconomic Post-Selection Regression Coefficient Estimates for Logistic-Urban (top-left), Logistic-Small Town/Rural (top-right), OLS-Urban (bottom-left) and OLS-Small Town/Rural (bottom-right)

7.1.3 Market Geography

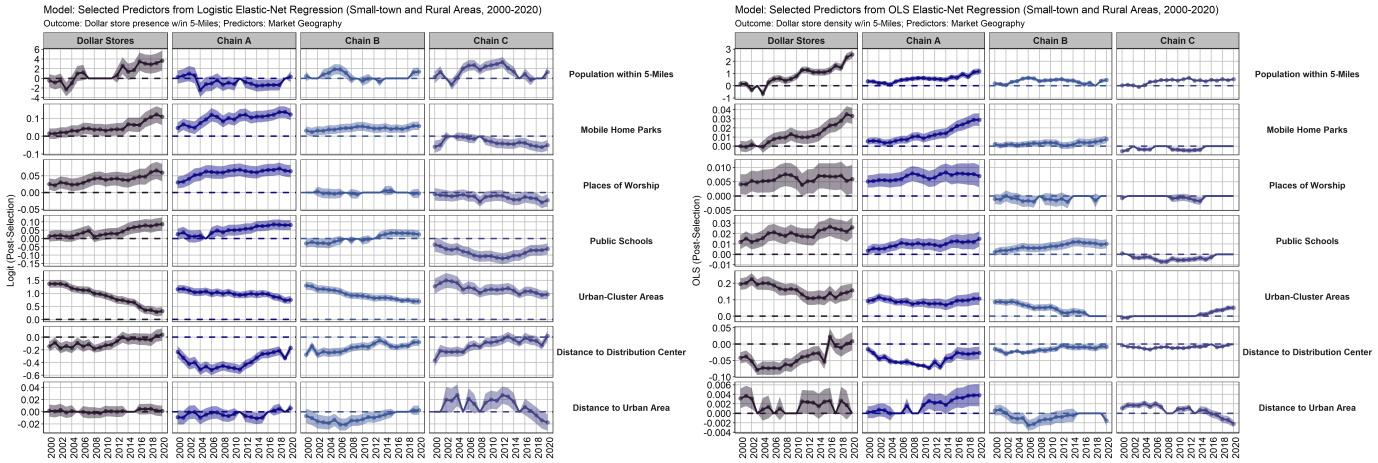

Figure 5: Market Geography Post-Selection Regression Coefficient Estimates for Logistic (left) and OLS (right) Small-Town/Rural Models

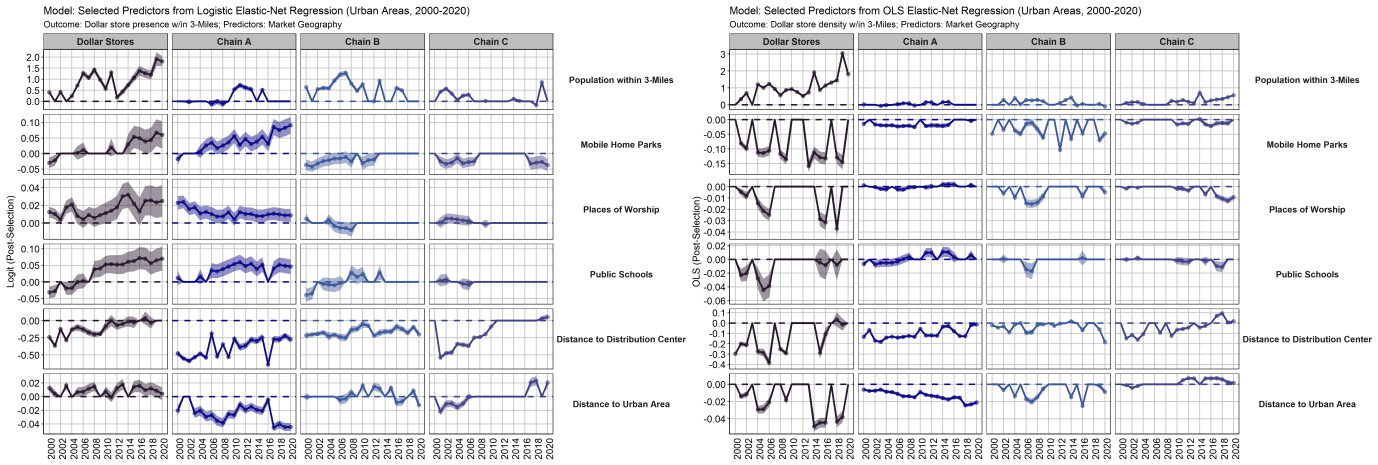

Figure 6: Market Geography Post-Selection Regression Coefficient Estimates for Logistic (left) and OLS (right) Urban Models

7.1.4 Retail Competition

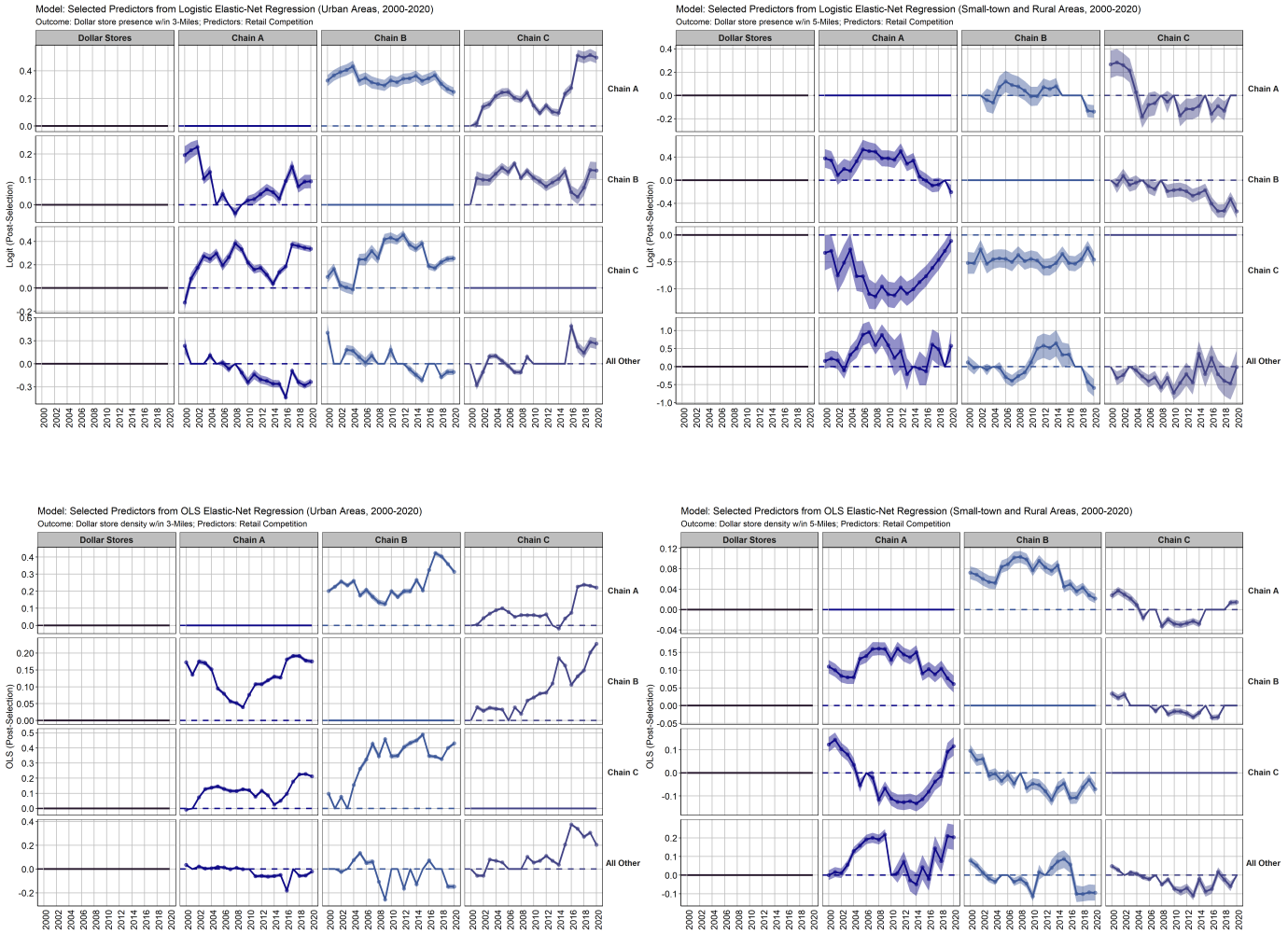

Figure 7: Dollar Store Competition Post-Selection Regression Coefficient Estimates for Logistic-Urban (top-left), Logistic-Small Town/Rural (top-right), OLS-Urban (bottom-left) and OLS-Small Town/Rural (bottom-right)

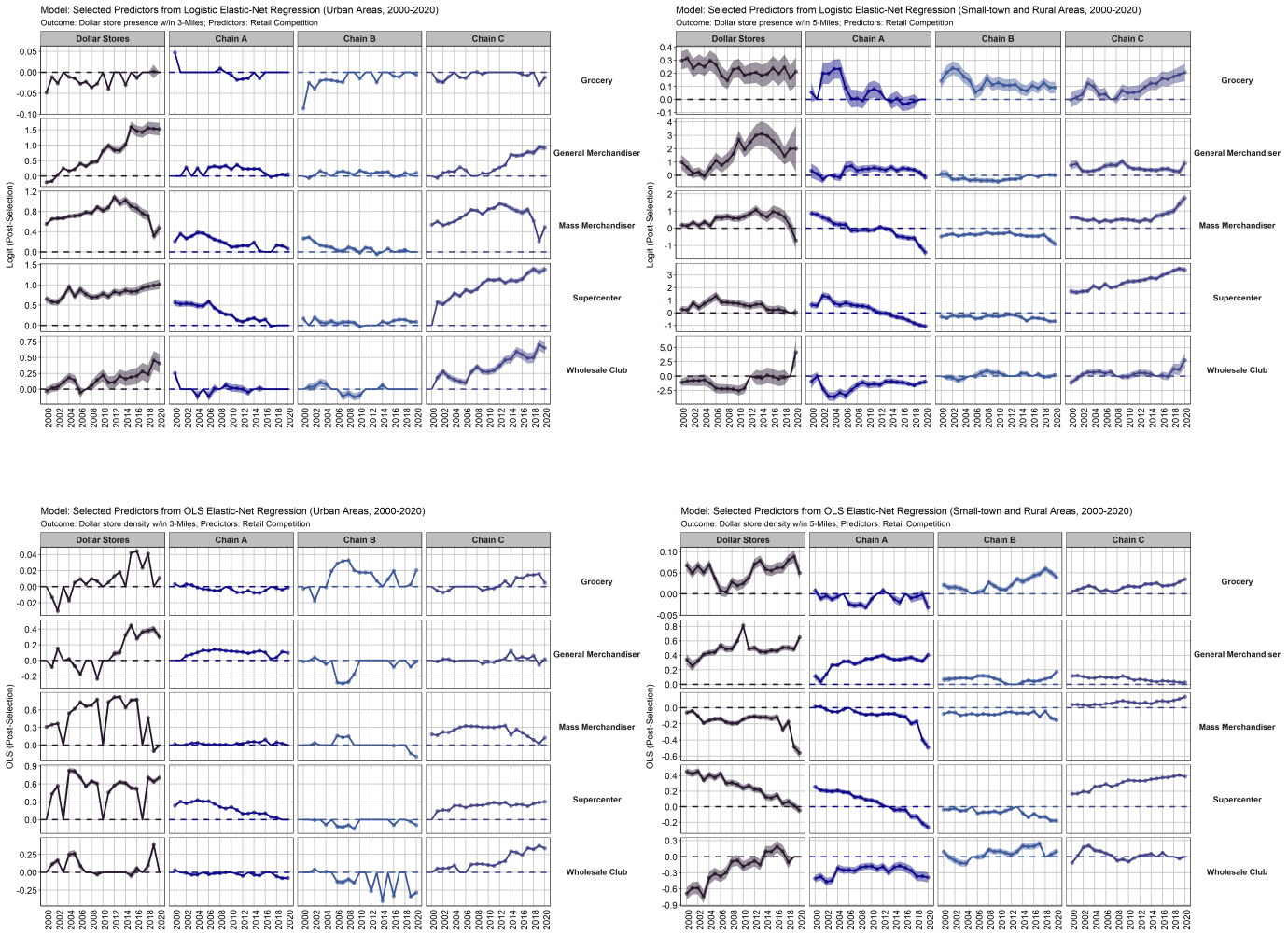

Figure 8: Grocery and Big-Box Store Competition Post-Selection Regression Coefficient Estimates for Logistic-Urban (top-left), Logistic-Small Town/Rural (top-right), OLS-Urban (bottom-left) and OLS-Small Town/Rural (bottom-right)

7.2 Elastic Net Variable Selection Summaries

7.2.1 Urban Areas

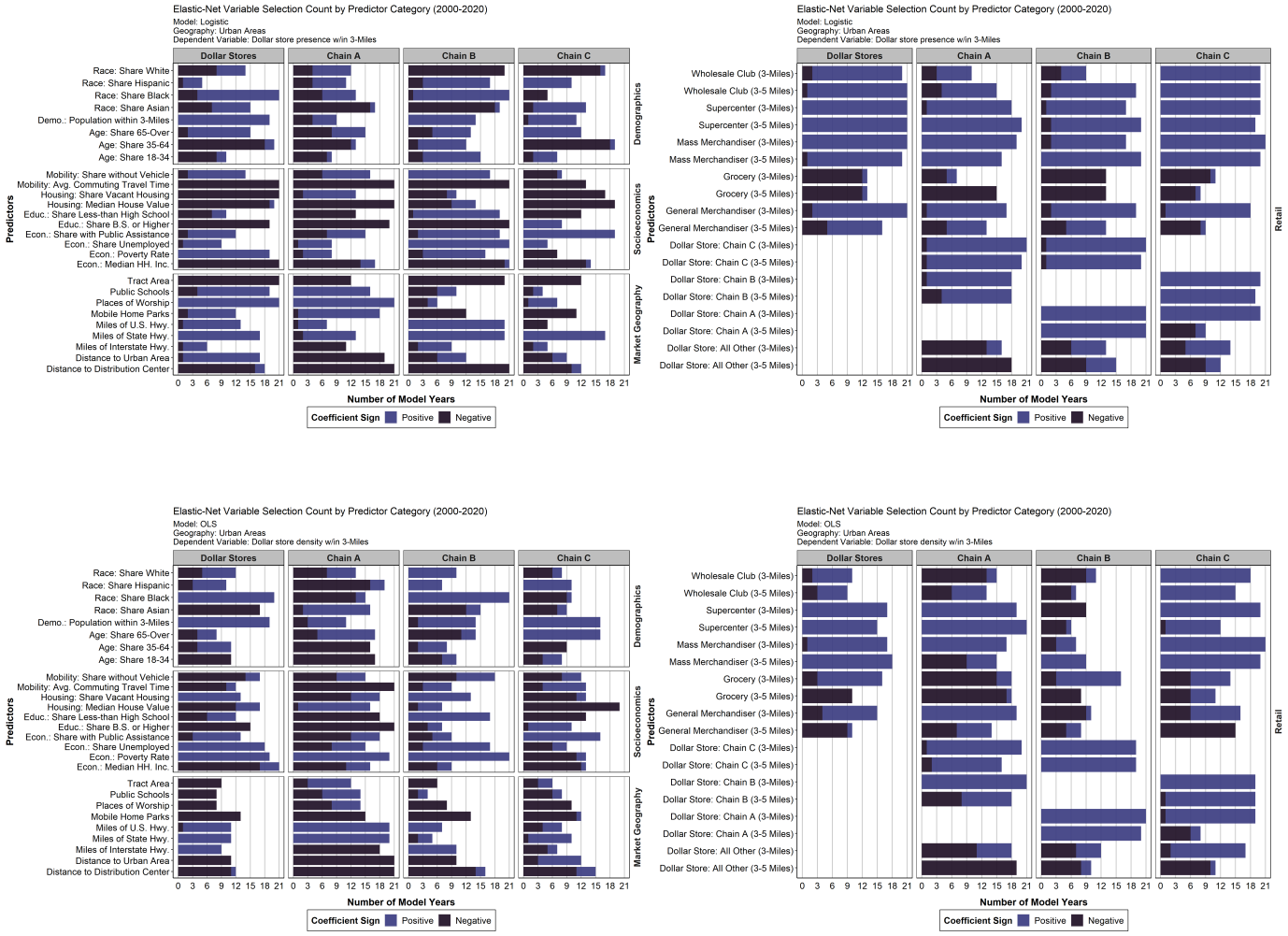

Figure 9: Summary of Urban-Area elastic net Regression Model Variable Selection in Logistic (top) and OLS Models (bottom)

7.2.2 Small-town/Rural Areas

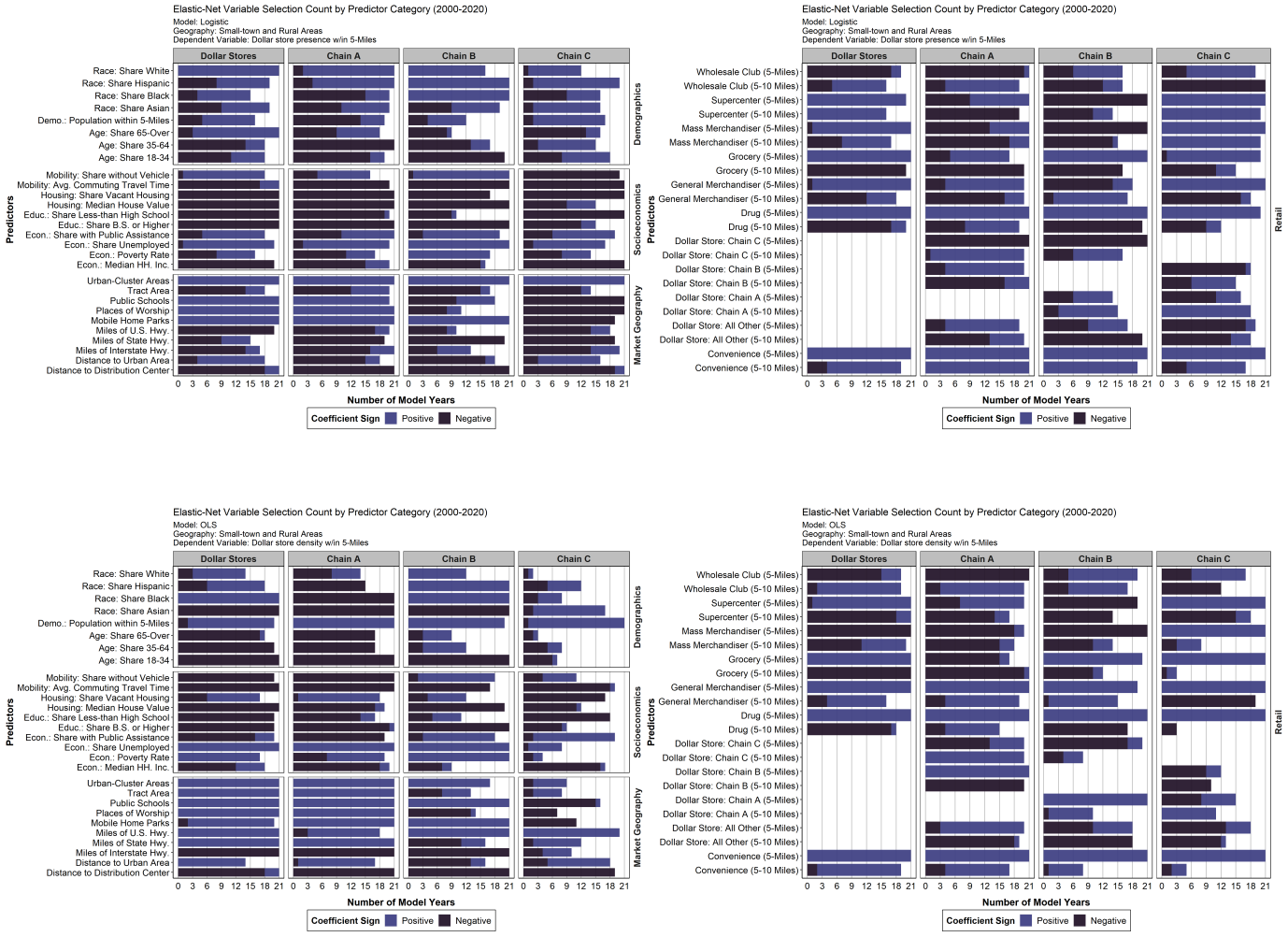

Figure 10: Summary of Small-town/Rural elastic net Regression Model Variable Selection in Logistic (top) and OLS Models (bottom)

7.2.3 Variable Selection and Statistical Significance Summary by Predictor

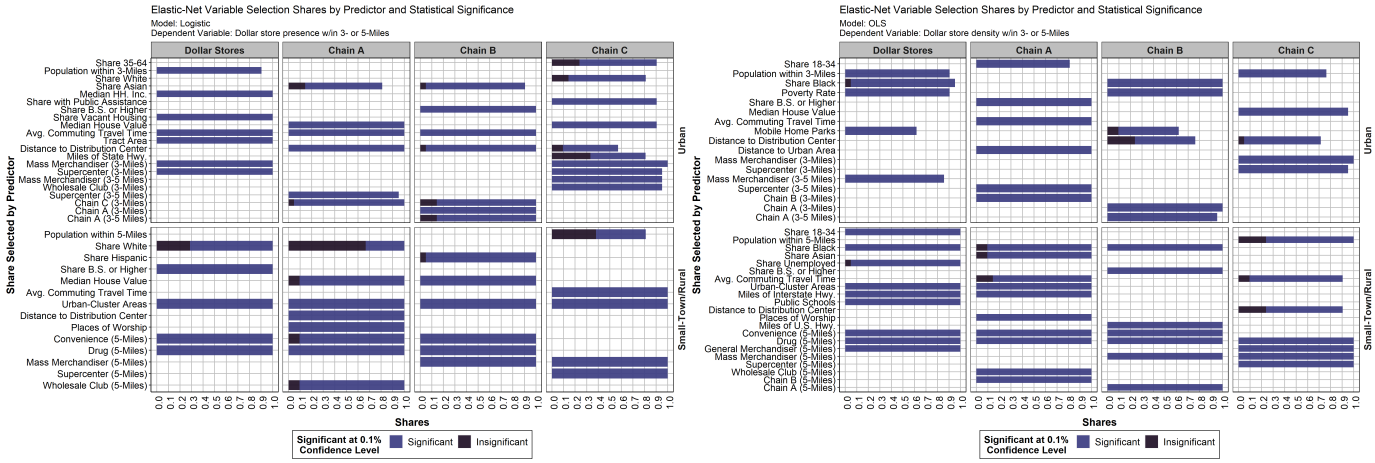

Figure 11: Summary of Logistic (left) and OLS (right) elastic net Variable Selection and Statistical Significance by Predictor

A Training, Validation, and Test Folds

Figures 12, 13, 14 and 15 display the block-based spatial sampling scheme for the national-scale urban-area models using year 2020 training, validation, and test identifiers. For graphing, we subset the markets by region. While their geographic delineations are not visible in the figures, each urban-area market contains individual census tracts. Census tracts within a given market that are included in the training (test) data cannot have census tracts from the same market in the test (training) data. The individual markets are color-coded by their fold identification, which are used to subset the data during cross validation. Census tracts not included in the training data are assigned to the test data.

Figure 16 shows a detailed view of the block-based spatial sampling cross-validation and model assessment method for small-town/rural markets for the state of Louisiana.[17] The colored triangles represent the rural markets and correspond spatially to the geographic centroids of counties. When a rural county is randomly selected for inclusion in the training,

[17]Louisiana was the randomly sampled state chosen from the 48 contiguous states in the United States to be included in the figure.

validation, or test data, all of the census tracts belonging to the county are selected so that that out-of-sample predictions provide biased-reduced estimates.

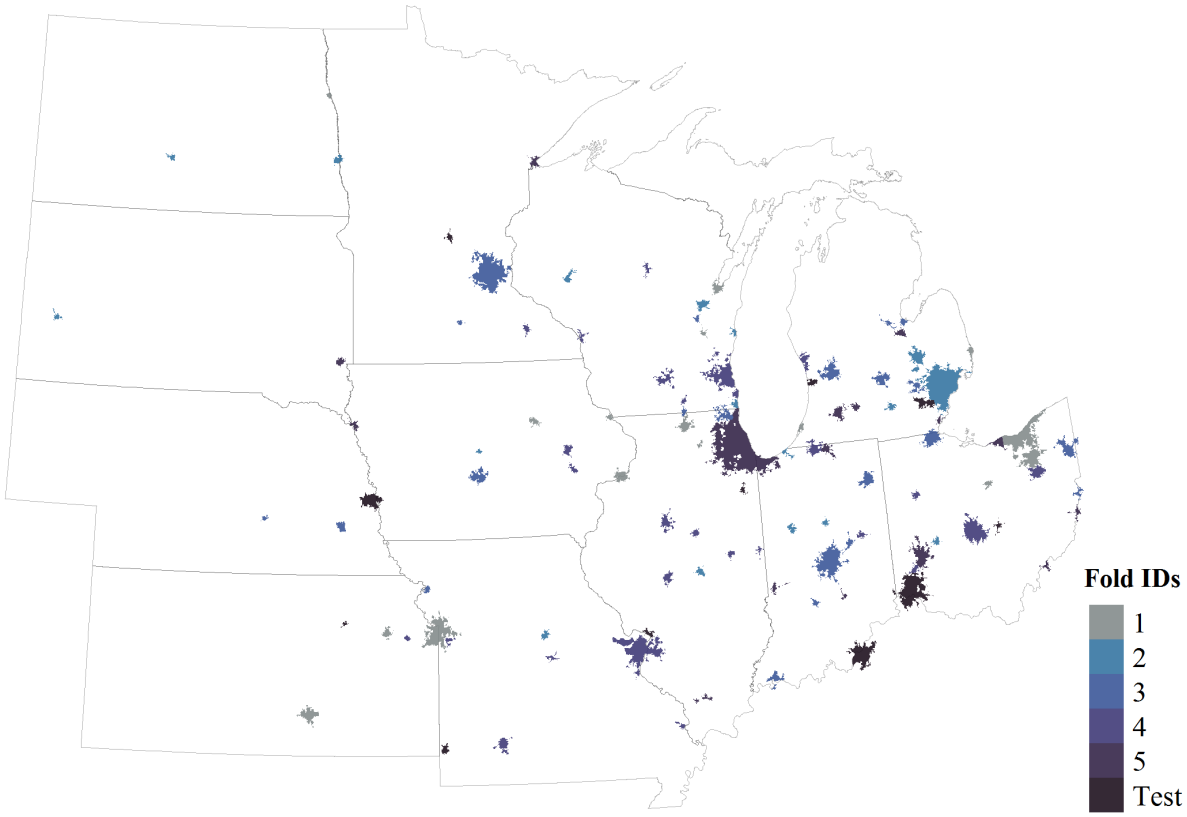

Figure 12: Block-Spatial Sampling for Cross-Validation and Test Data - Urban Markets (Midwest)

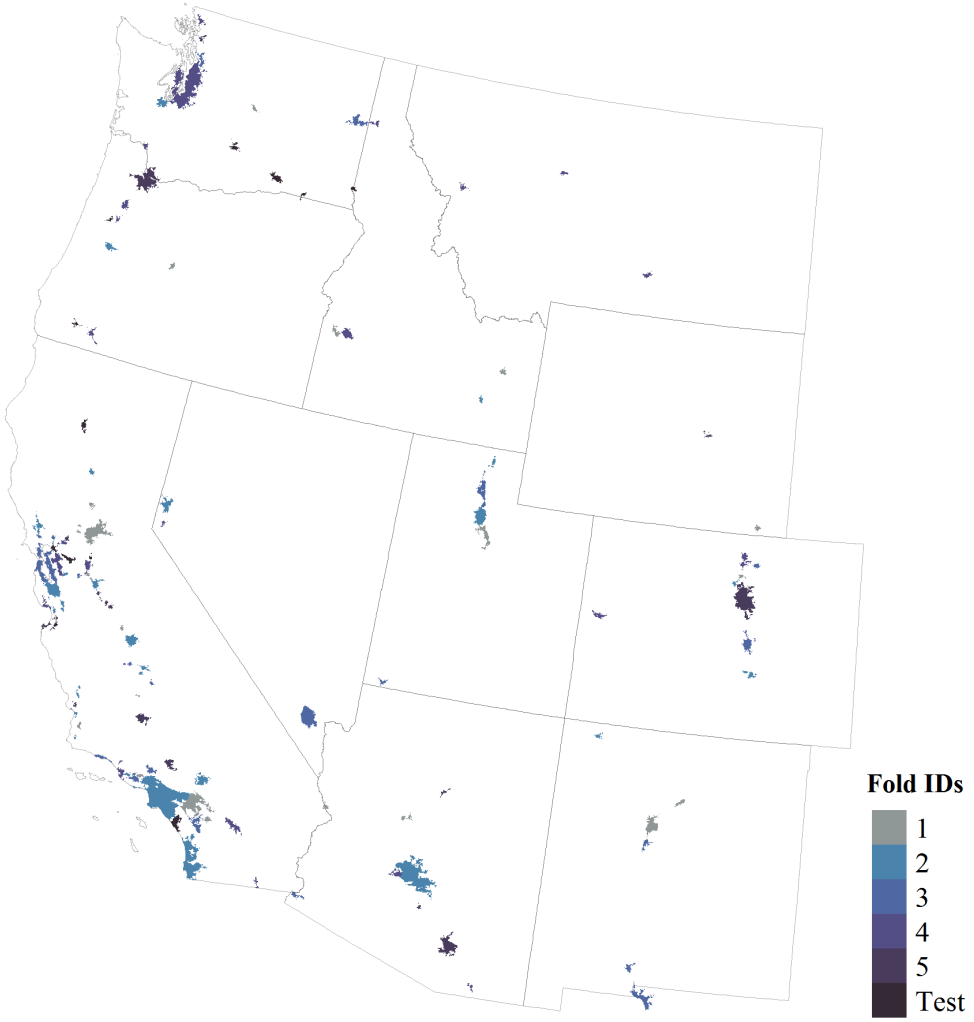

Figure 13: Block-Spatial Sampling for Cross-Validation and Test Data - Urban Markets (West)

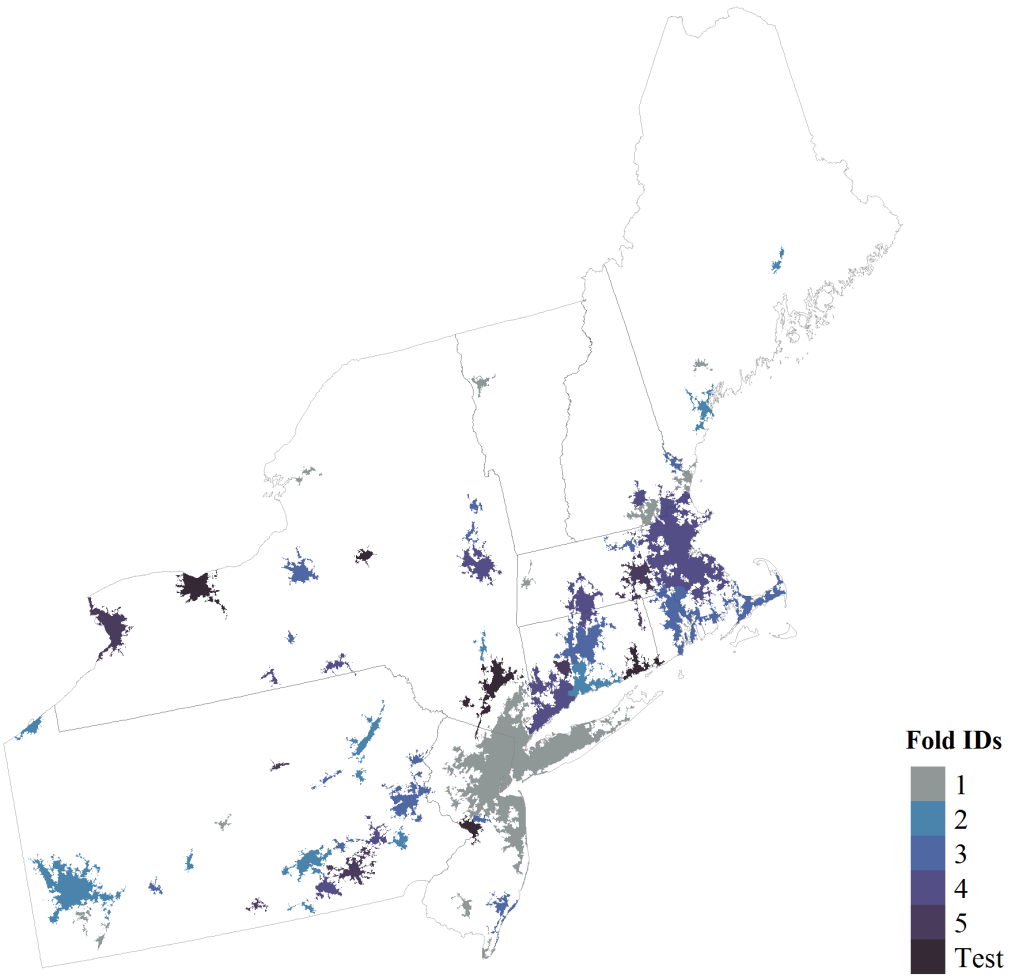

Figure 14: Block-Spatial Sampling for Cross-Validation and Test Data - Urban Markets (Northeast)

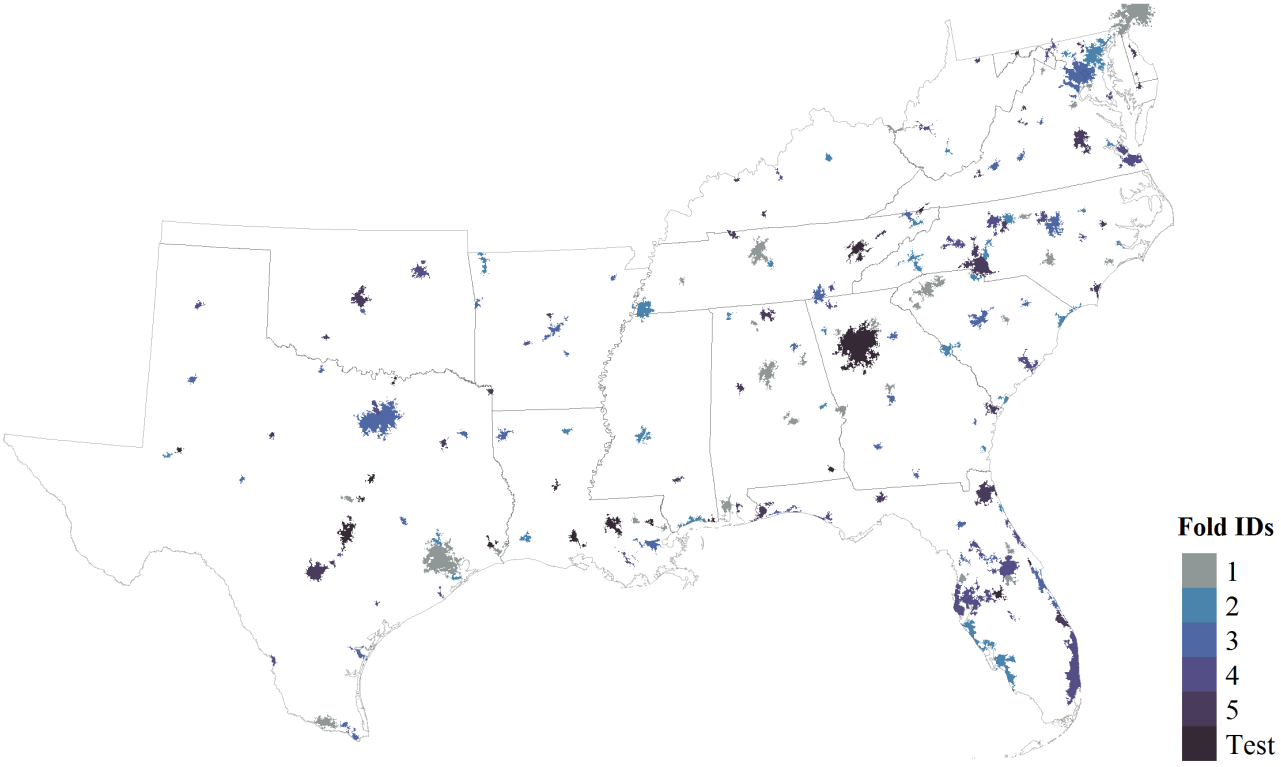

Figure 15: Block-Spatial Sampling for Cross-Validation and Test Data - Urban Markets (South)

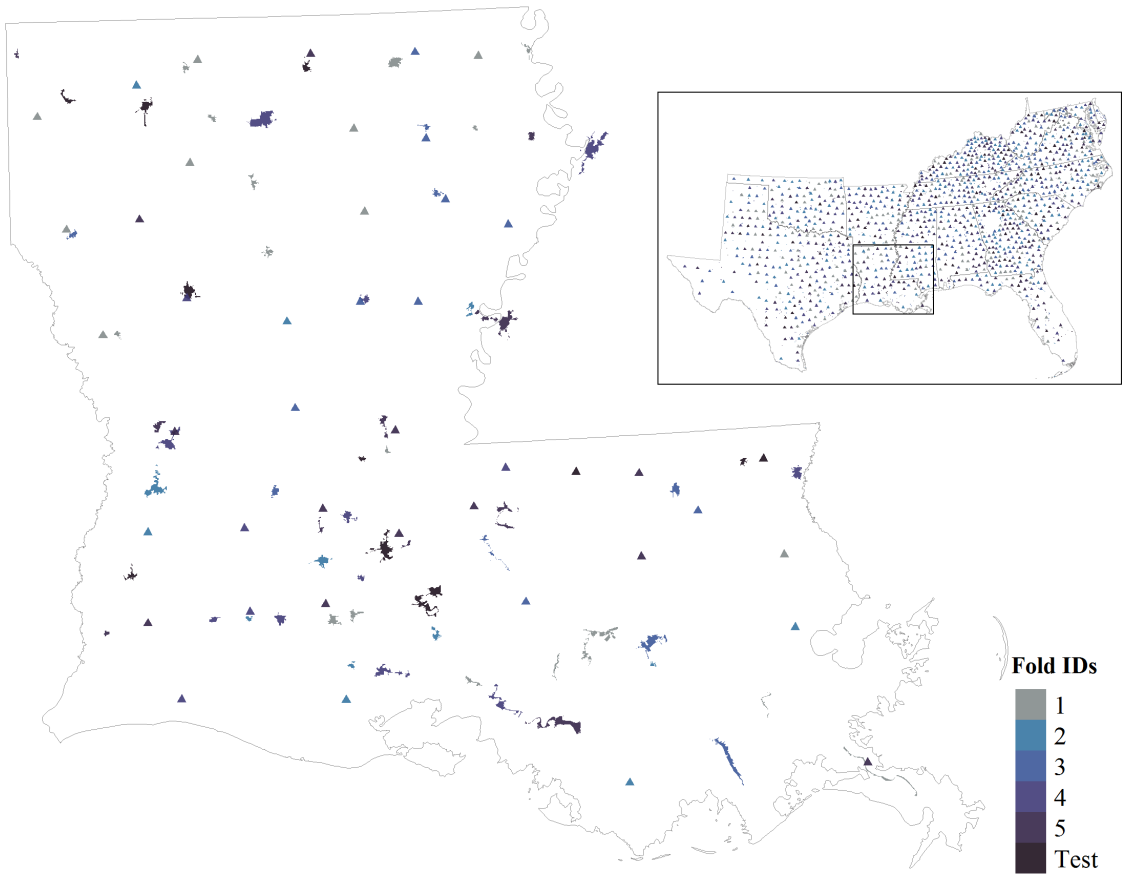

Figure 16: Block-Spatial Sampling for Cross-Validation and Test Data - Small-Town and Rural Markets in Louisiana

B Model Performance

We compare the results of the three algorithms using the out-of-sample error rates in the test data (i.e., R-squared and classification error). Figures 17, 18, 19, and 20 compare the test classification error and R-squared of the elastic net, random forest (RF), and XGBoost (XGB) algorithms for national-scale models in urban- and small-town/rural areas. In nearly all model types and years, RF and XGB consistently have lower out-of-sample test error compared to the elastic net regression models. The RF and XGB models perform similarly

with respect to accurately predicting dollar store entry and densities.[18]

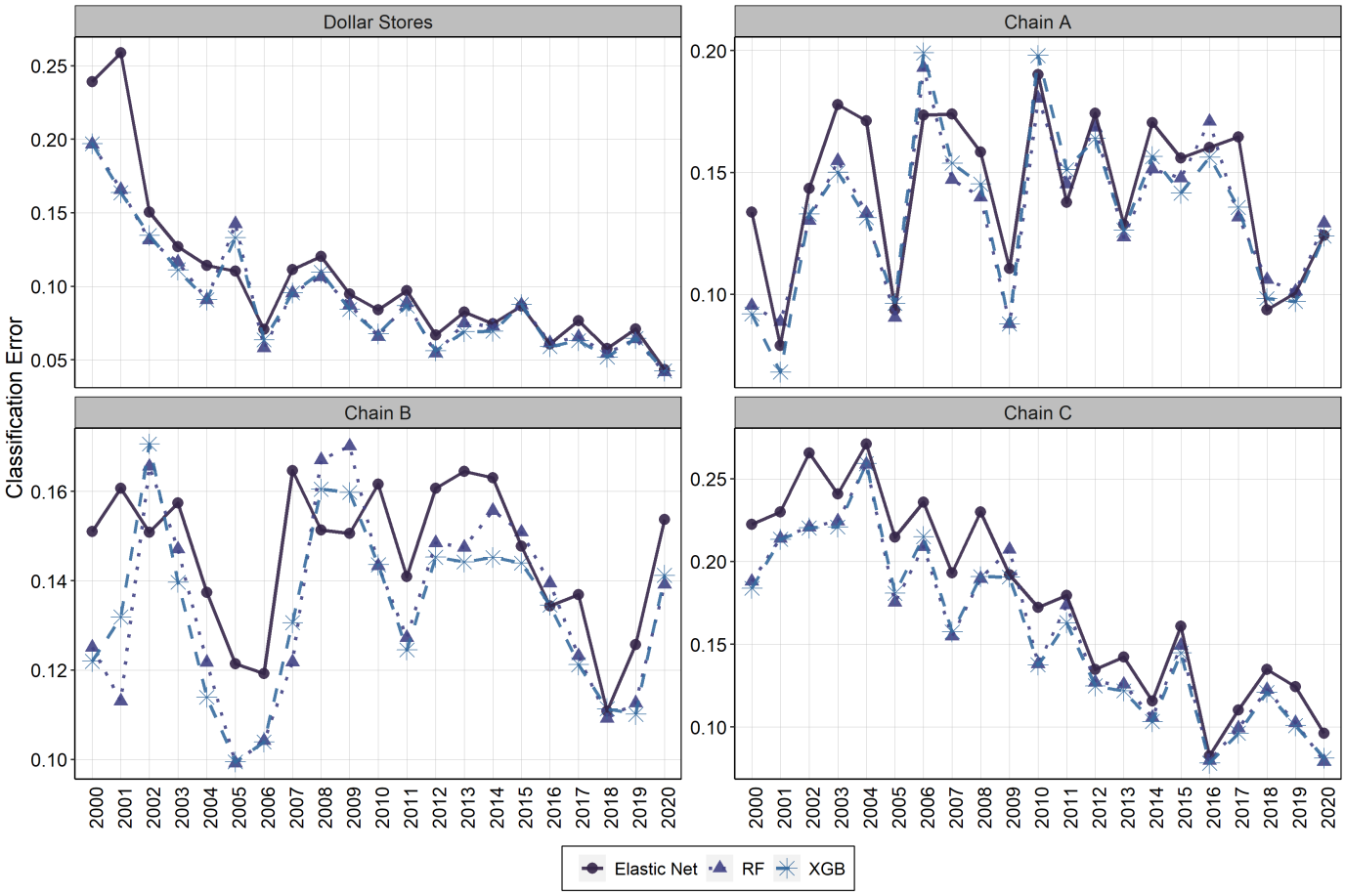

Figure 17: Classification Error by Dollar Store Type and ML Model - Urban Areas (2000-2020)

[18]For select years in the urban-area models, census tracts of large urban markets, specifically, New York–Newark, NY–NJ–CT and Los Angeles–Long Beach–Anaheim, CA, are randomly placed in the test data. In these years, the out-of-sample prediction error increases notably for each model type. Relative to elastic net, however, RF and XGB still tend to have more robust predictive performance. For comparable estimates of out-of-sample error over time, the test error computed in Figures 17, 18, 19, and 20 do not include census tracts from the above-mentioned markets.

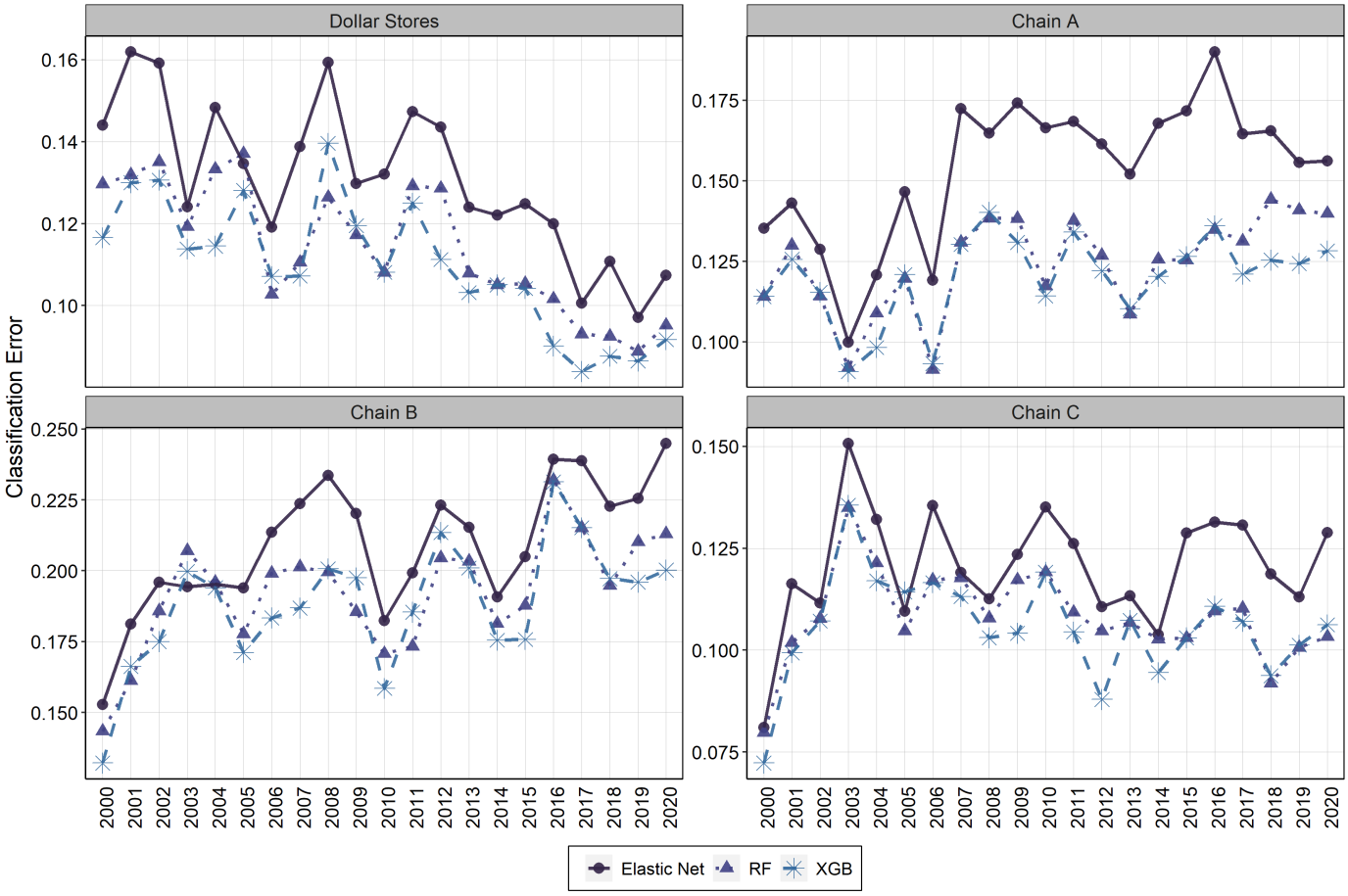

Figure 18: Classification Error by Dollar Store Type and ML Model - Small-Town/Rural Areas (2000-2020)

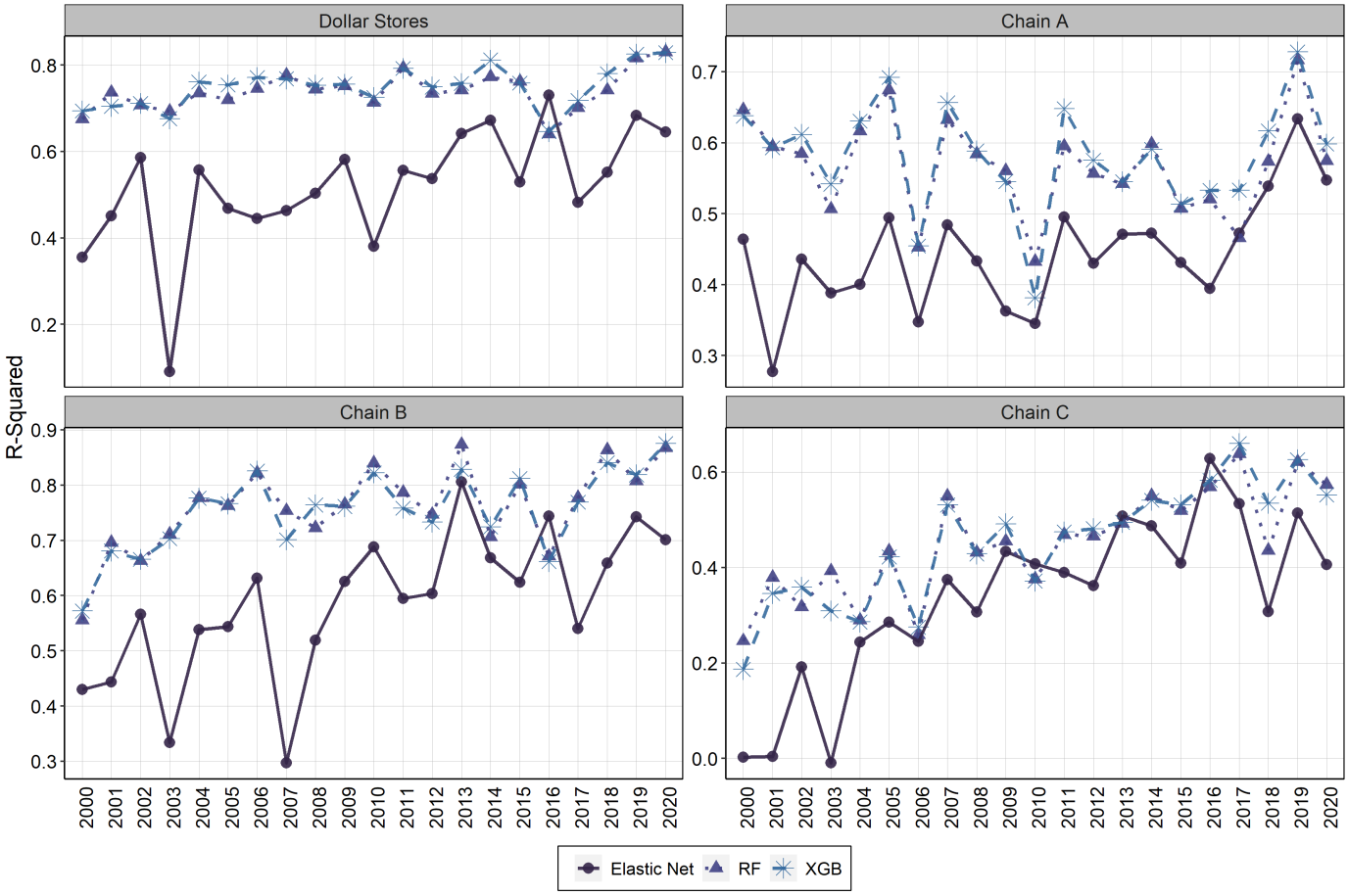

Figure 19: R-Squared by Dollar Store Type and ML Model - Urban Areas (2000-2020)

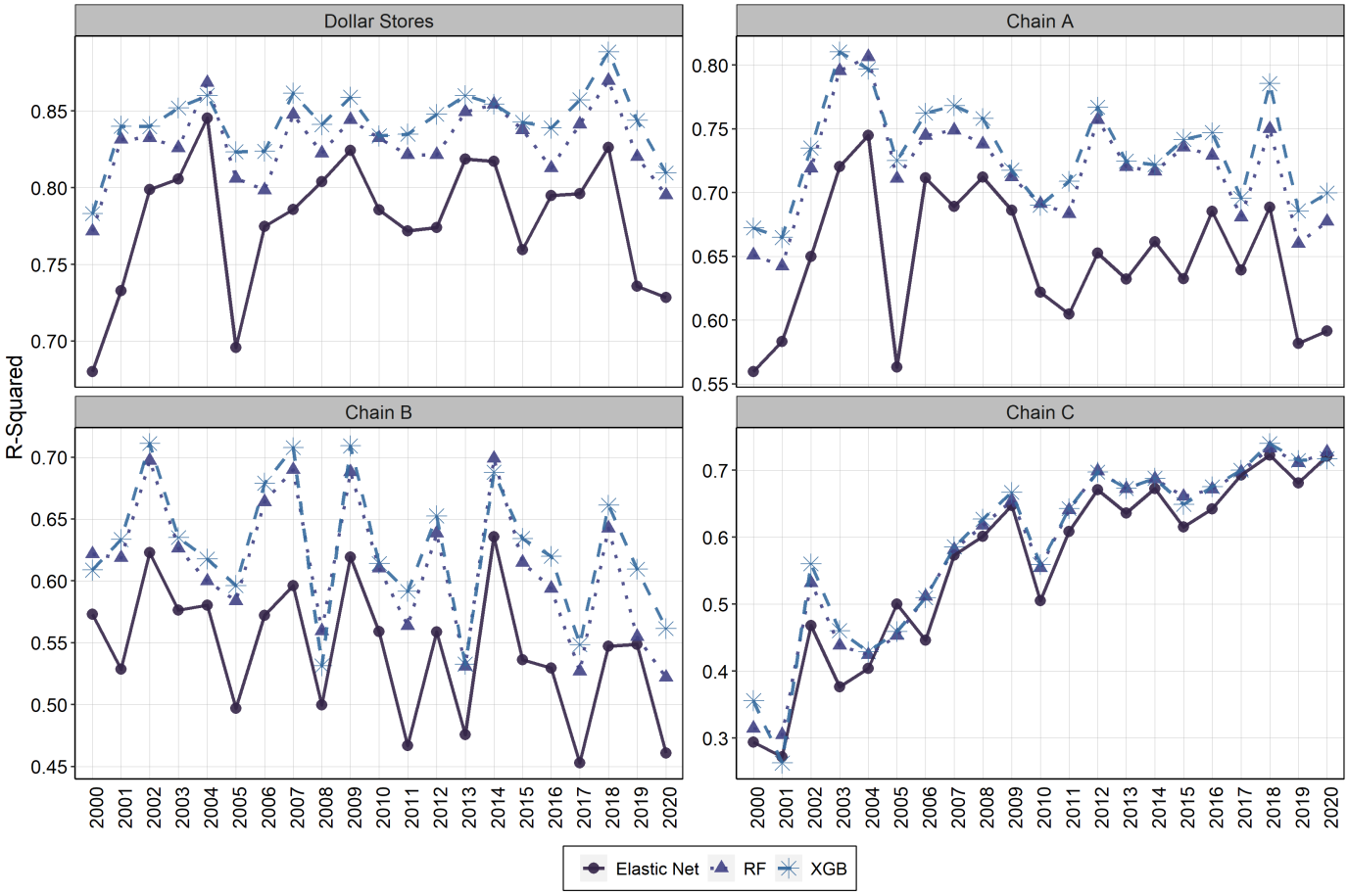

Figure 20: R-Squared by Dollar Store Type and ML Model - Small-Town/Rural Areas (2000-2020)

C Elastic Net

The elastic net linear regression model is presented in Algorithm 1. For the logistic model, the penalized negative binomial log-likelihood can be substituted for the optimization function (5) and the classification error, Err_{test}, is used in place of R^2_{test}. Lines 6 to 7 of the algorithm indicate that after finding the optimal penalized elastic net regression model by cross-validation, we output a final post-selection model, regressing the outcome variable on the subset of most predictive explanatory variables using the full set of census-tract observations. In the case of lasso (i.e., $\alpha = 1$), post-regularization regression coefficient parameter

estimates have lower bias and similar convergence rates relative to the lasso (Belloni & Chernozhukov, 2013; Belloni et al., 2016). The elastic net models are estimated using the R statistical software package, glment (J. Friedman et al., 2010).

Algorithm 1 elastic net Algorithm for Linear Regression Models

1: Estimate (5) using the training data and cross-validation folds to find the optimal tuning parameter pair (α_{min}, λ_{min}).
2: Estimate a final model from the full set of training data to obtain $\left(\hat{\boldsymbol{\beta}}^{enet}, \alpha_{min}, \lambda_{min}\right)$
3: Using the matrix of test data, $\boldsymbol{X}_{test}$, and elastic net parameters $\left(\hat{\boldsymbol{\beta}}^{enet}, \alpha_{min}, \lambda_{min}\right)$, make predictions, $\hat{\boldsymbol{y}}$, for the holdout test observations, $\boldsymbol{y}^{test}$.
4: Estimate $R^2_{test} = 1 - \frac{\sum_i^n \left(y_i^{test} - \hat{y}_i\right)^2}{\sum_i^n \left(y_i^{test} - \bar{y}^{test}\right)^2}$
5: Using the complete data set, estimate the post-selection model:
6: $\hat{\boldsymbol{\beta}}_{post} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2 \right\}$
7: such that $\beta_j = 0$ when $\hat{\beta}_j^{enet} = 0$.

D Random Forest

Random forest is an "ensemble" classification and regression tree (CART) method that makes predictions by averaging the predictions of independently-grown decision trees (Breiman, 2001b; Breiman, Friedman, Stone, & Olshen, 1984). Aggregating predictions over multiple trees improves prediction accuracy. A decision tree is created by searching over all possible splits of the predictors and dividing the sample at the split point that minimizes a measure of total prediction error (e.g., MSE). The algorithm repeatedly splits the data using this same process until reaching a stopping rule. When a decision tree is complete, predictions are made for each observation by assigning the average value of the outcome, or majority vote in the case of classification, for the set of observations in a terminal-node or -leaf, where a leaf is a group of observations created by the sequence of variable splits.

In random forests, each decision tree is grown using a new bootstrap sample or subsample,

where the tree sample size is $n < N$. At each node in the tree, a randomly-selected subset of predictors is used as candidate split variables, $m < M$, where M is the full set of predictors. Randomizing the selection of observations and candidate variables in building the trees reduces the correlation of predictions between trees, stabilizing prediction variance (Hastie et al., 2009).

Random forests have several tuning parameters that control the amount of randomization and structure of individual trees and the overall forest (Probst, Wright, & Boulesteix, 2019). In our random forest models, we tune the number of randomly chosen candidate split variables, m (specified by $mtry$), because this hyperparameter tends to generate the largest gain in prediction accuracy (M. W. Mitchell, 2011; Probst, Boulesteix, & Bischl, 2019). We perform five-fold cross-validation following the spatial cross-validation approach described in 3.6 to find the hyperparameter value that minimizes the cross-validation prediction error. For other random forest hyperparameters, we use optimal values cited in the literature.

For each tree grown in the random forest, we subsample, without replacement, 70% of the training observations, controlled by $sample.fraction$ and $replace$ (Nicodemus, Malley, Strobl, & Ziegler, 2010; Probst & Boulesteix, 2017). The individual tree and forest structures are determined by the tree depth and the number of independent trees grown in the random forest, which are respectively controlled by the minimum number of observations required for a terminal node in each tree ($min.node.size$) and the total number of trees, T ($ntree$). We set $min.node.size = 1$ and $min.node.size = 5$ for classification and regression problems (Díaz-Uriarte & De Andres, 2006). We set $ntree = 500$, which is shown to provide stable prediction accuracy (Genuer, Poggi, & Tuleau, 2008; Genuer, Poggi, & Tuleau-Malot, 2010). We use the gini index for classification and the variance for regression as the node split criterion for each binary split during tree construction ($splitrule$).[19] Table 4 summarizes the parameter values.

[19]The gini index for node m is $\sum_{k=1}^{K} \hat{p}_k (1 - \hat{p}_k) = 1 - \sum_{k=1}^{K} (\hat{p}_k)^2$, where $\hat{p}_k$ is the proportion of observations in class k and K is the total number of classes. In our application predicting the presence of dollar stores using a 1/0 indicator variable, $K = 2$. Smaller values of the gini index imply increased purity of node splits.

Table 4: Random Forest Parameter Values for Model Training

Parameter	**Classification**	**Regression**
ntree	500	500
min.node.size	1	5
sample.fraction	0.7	0.7
replace	FALSE	FALSE
splitrule	Gini index	Variance
mtry*	$\{\frac{1}{2}x, \frac{1}{3}x, \frac{1}{6}x\sqrt{x}\}$	$\{\frac{1}{2}x, \frac{1}{3}x, \frac{1}{6}x, \frac{1}{9}x\}$

*We implement five-fold cross-validation to find the optimal values of *mtry* that minimize the cross-validated prediction error.

Algorithm 2 provides the random forest routine adapted from Cutler, Cutler, and Stevens (2012) and Hastie et al. (2009). We fit a final random forest model using the full set of training data and the optimal value of *mtry*. Regression and classification model prediction errors are computed using the test data and equations 3 and 4. The random forest models are implemented using the R statistical software package, ranger (Wright & Ziegler, 2015).

Algorithm 2 Random Forest Algorithm for Regression and Classification Models

1: **for** $b = 1$ **to** B **do**
2: Subsample, without replacement, 70% of training data $\boldsymbol{X}^*_{train}$.
3: Grow a random forest decision tree, T_b, using $\boldsymbol{X}^*_{train}$ by recursively repeating
4: the following steps for each terminal node of the tree,
5: until the minimum node size, *min.node.size*, is reached.

i Randomly select m of the M predictors as candidate split variables.
ii Of the m predictors, find the variable-split pair that minimizes the prediction error.
iii Split the node into two descendent nodes.

6: **end for**

7: Output the ensemble of trees $\{T_b\}_1^B$

To make a prediction at a new point $\boldsymbol{x}$:

Regression: $\hat{f}^B_{\text{rf}}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} T_b(\boldsymbol{x})$
Classification: $\hat{f}^B_{\text{rf}}(\boldsymbol{x}) = \arg\max_y \sum_{b=1}^{B} I\left(T_b(\boldsymbol{x}) = y\right)$

where $T_b(\boldsymbol{x})$ is the prediction of the outcome variable at $\boldsymbol{x}$ using the b^{th} tree.

E Gradient Boosting

Boosting is an ensemble machine learning method that combines a series of base learners (i.e., models slightly better than random guessing) in a stagewise fashion (Ferreira & Figueiredo, 2012; Freund, Schapire, & Abe, 1999). In gradient boosting (J. Friedman, Hastie, & Tibshirani, 2000; J. H. Friedman, 2001; Ridgeway, 2007), models are successively fit to the residuals generated from the previous model's predictions. Whereas random forest combines multiple full-sized regression trees to make predictions, reducing the model variance, boosting reduces model bias by successively adding simple models to increase prediction accuracy (Sutton, 2005; Yeturu, 2020). While gradient boosting can be implemented using various types of base learners, the most common base-learners in practice are decision trees (Natekin & Knoll, 2013). In gradient boosting decision trees, predictions are formed with the objective of minimizing a loss function, $\sum_{i=1}^{N} L\left(y_i, f\left(x_i\right)\right)$ (J. H. Friedman, 2002). Common loss functions include the squared-error loss in the case of regression and log loss in the case of binary classification (Hastie et al., 2009; Natekin & Knoll, 2013).

The gradient boosting algorithm is initiated by an initial prediction.

$$f_0(x) = \arg\min_{\alpha} \sum_{i=1}^{N} L\left(y_i, \alpha\right)$$

In the case of gradient boosted decision trees that use squared-error loss, the initial prediction is the average of the response variable in the data. For each iteration, b, the negative gradient of the loss function is computed with respect to the model's previous predictions.

$$r_{ib} = -\left[\frac{\partial L\left(y_i, f\left(x_i\right)\right)}{\partial f\left(x_i\right)}\right]_{f=f_{b-1}}$$

The values r_{ib} are the pseudo-residuals (J. H. Friedman, 2002) for observation i at the

b^{th} iteration of the boosting algorithm. A small regression tree is fit to the residuals (i.e., the loss function's negative gradient). Regression tree predictions in each terminal node minimize:

$$\hat{f}_{bk} = \arg\min_{f_b(x)} \sum_{x_i \in S_{kb}} L\left(y_i, f_{b-1}(x_i) + f_b\left(x_i\right)\right)$$

where S_{kb} is the set of observations in terminal node k at boosting iteration b (Hastie et al., 2009; Ridgeway, 2007).

Predictions are updated by adding the new decision tree, or the fitted values of the residuals, to the model. After B iterations, the boosted model yields predictions for observation i as:

$$\hat{y}_i = \hat{f}(x_i) = f_0(x) + \sum_{b=1}^{B} \eta \hat{f}_b(x_i)$$

Using the squared-error loss function, gradient boosting is equivalent to iteratively fitting regression trees to the updated model residuals (J. H. Friedman, 2002; Ridgeway, 2007). Each boosted model is multiplied by a small constant, $\eta = (0, 1]$, which shrinks the contribution of each boosted model and controls the model learning rate (Elith, Leathwick, & Hastie, 2008). Smaller values of η (e.g., 0.01) imply that a larger number of boosted iterations, B, are required to achieve strong prediction accuracy. However, higher B increases model complexity, which can lower the predictive performance on test data (Hastie et al., 2009). In practice, optimal values of η and B are found via cross-validation.

We implement the XGBoost version of gradient boosting (Chen & Guestrin, 2016), which is optimized for computational speed and includes additional model tuning parameters and functionality to allow for greater customization of the boosting procedure.[20] We tune several parameters that impact the size, number, and contribution of boosted trees using cross validation. To guide our tuning strategy and selection of other XGBoost parameter settings,

[20]We use the xgboost package in R (Chen & Guestrin, 2016).

we reference the recommended hyperparameter values from several machine learning model-comparison studies that include gradient boosting (Bentéjac, Csörgő, & Martínez-Muñoz, 2021; I. Brown & Mues, 2012; Probst, Boulesteix, & Bischl, 2019; Zhang, Liu, Zhang, & Almpanidis, 2017).

We tune each model to find the optimal learning rate, *eta* (η), number of boosting iterations, *nrounds*, depth of the regression trees in each boosting iteration, *max_depth*, and *gamma* (γ), a regularization term that controls tree complexity. Higher values of γ reduce model complexity by requiring that the reduction in model error from internal node splits is larger, whereas smaller values of γ allow for additional splits from small improvements in model performance. Table 5 displays the hyperparameter specifications for training the XGBoost models.

As in random forest, the predictive accuracy of gradient boosting can be further enhanced by adding randomness to each boosted iteration (J. H. Friedman, 2002). We set $subsample = 0.7$ for both classification and regression models, which controls the fraction of randomly selected (without replacement) training data observations to be used on each boosting iteration. We set $colsample\_bytree = 0.75$ and $colsample\_bylevel = 0.6$, which specify the fraction of predictors randomly selected for each boosted tree and as split-candidate variables at each tree level.

In both classification and regression models, we find the optimal hyperparameter combination by implementing the spatial cross-validation approach described in 3.6 for a grid of 12 tuning parameter vectors. For each unique set of hyperparameters, we use five-fold cross validation to find the optimal number of boosting iterations, *nrounds*. We fit a final boosted model to the complete set of training data using the globally optimal tuning parameter combination and *nrounds* that minimize the cross-validation error. Finally, we assess the predictive accuracy of the optimal model using the test data and equations 3 and 4.

Table 5: XGBoost Hyperparameter Values for Model Training

Parameter	Values
eta* (η)	$[0.025, 0.05, 0.1]$
gamma* (γ)	$[0.2, 0.3]$
max_depth*	$[6, 10]$
subsample	0.7
colsample_bytree	0.75
colsample_bylevel	0.6
nrounds	500

*We implement five-fold cross-validation and grid search to find the optimal tuning parameter combination that minimizes the cross-validated prediction error.
To prevent over-fitting, we stop the cross-validation procedure if the prediction error does not decrease after 10 boosting iterations.

F Random Forest and XGBoost: Permutation Feature Importance

We supplement the elastic net model results by analyzing and comparing the permutation feature importance estimates from the random forest and XGBoost algorithms (Breiman, 2001a). Permutation feature importance measures the influence of each predictor on the outcome by computing the change in model accuracy after randomly permuting the k^{th} explanatory variable. Randomly shuffling the predictor removes the correlation structure between x_k and the outcome variable, as well as interaction effects between x_k and other correlated predictors.

For each model year, we compute the permutation feature importance of predictor x_k by taking the ratio of the permuted and original model errors using the test data (Greenwell & Boehmke, 2019; Molnar, 2020)[21] Permutation feature importance estimates greater than one indicate that the prediction error with x_k permuted increases the test error relative to

[21]Given our spatial sampling approach of test markets, the feature importance estimates provide insight into how variables contribute to the prediction of dollar store entry and densities in spatially independent markets.

the original model test error. Using the random forest and XGBoost permutation feature importance estimates from each dollar store model and year, we scale the values such that feature importance estimates range from 0 to 1.[22] We compute the mean-scaled permutation feature importance for each predictor by averaging the estimates over the 21 model years (2000 to 2020).

F.1 Mean Permutation Feature Importance (2000-2020)

Figures 21 and 22 display the predictors from the random forest (RF) and XGBoost (XGB) classification and regression models whose mean-scaled permutation feature importance scores are in the 90^{th} percentile for the given dollar store model. The top row of bar charts shows predictors from urban-area models, while the bottom row shows predictors from small-town/rural areas.

The patterns of most predictive features in the RF and XGB models are consistent with the elastic net results, supporting our conclusions that dollar store entry and densities are more heterogeneous than suggested by conventional wisdom. In addition, the RF and XGB algorithms both yield quite similar results for each prediction task, further validating our findings. The most predictive features in the Dollar Stores models (1^{st} columns in each of Figures 21 and 22) largely reflect a combination of important features for each of the dollar store chains. Yet, the most predictive features vary considerably across the three dollar store retailers.

The mean permutation feature importance estimates suggest that, while both Chain A and Chain B locate in areas associated with poor economic conditions, the neighborhood's race composition, particularly the share of the population that is Black, is more predictive of Chain B. With the exception of area-wide population, Figures 21 and 22 indicate that

[22] For a given dollar store model and year, we scale feature importance estimates by $\frac{I_k - I_{\min}}{I_{\max} - I_{\min}}$, where I_k is the unscaled feature importance of predictor k, and $I_{\min}$ and $I_{\max}$ are the predictors whose unscaled feature importance estimates are respectively the minimum and maximum of the predictor set.

demographic variables (e.g., race) are rarely in the 90th percentile of features predicting Chain A entry and densities. Neighborhood median house values and distance to the nearest distribution center are the most important socioeconomic and market geography predictors across all Chain A models and markets (i.e., urban- and small-town/rural areas). The share of neighborhood vacant housing and the population share with less than a high-school education are top socioeconomic predictors in several of Chain A's RF models.

The most predictive features of Chain B entry and densities consistently include broader sets of demographic and socioeconomic variables associated with race and economically disadvantaged neighborhoods, in addition to distribution center proximity. The share of the population that is Black is located in all but one panel of Figures 21 and 22. Factors related to socioeconomic status, such as the neighborhood poverty rate, median household income, median house values, and the share of the population without a vehicle tend to be predictors ranked in the 90th percentile of permutation feature importance estimates for Chain B models.

The RF and XGB models suggest that the densities of the other dollar store chain and grocery stores are important retail competition predictors of Chain A and Chain B. The elastic net models similarly indicate that Chain A and Chain B tend to co-locate, and that, while Chain A mostly distances itself from conventional grocers, Chain B entry and densities tend to increase with the number of grocery stores in the area.

Finally, the RF and XGB models indicate that Chain C is highly differentiated from Chain A and Chain B by the relative importance of predictors associated with retail competition, and specifically the densities of big-box supercenters and mass merchandisers. In urban-area RF and XGB classification and regression models, the density of mass merchandisers is the first or second highest average permutation feature importance estimate, while in small-town/rural area models, supercenter density is the most important predictor of Chain C entry and densities. Area-wide population and grocery store densities are also con-

sistently ranked in the 90$^{\text{th}}$ percentile of predictors. The elastic net model results similarly signal Chain C's preference for markets characterized by large populations and high levels of retail agglomeration.

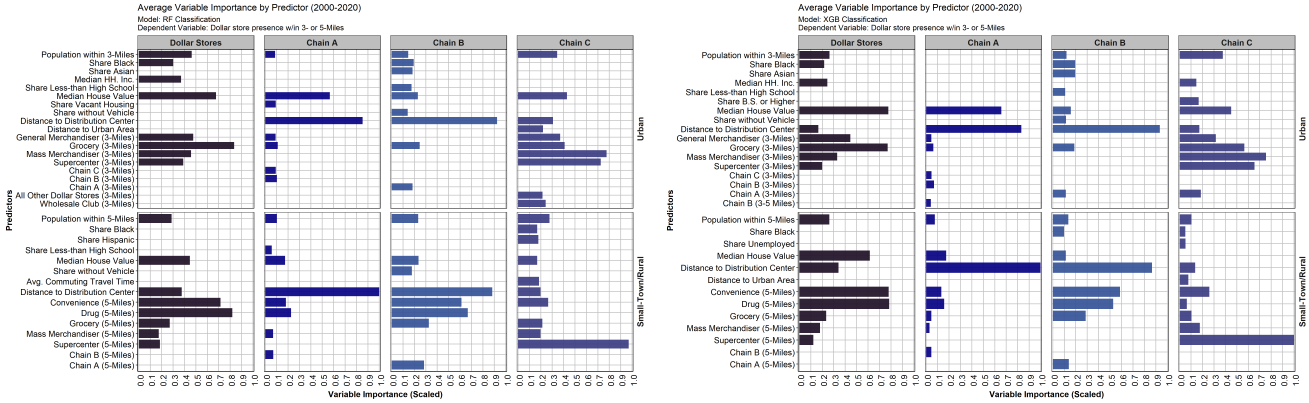

Figure 21: Mean-Scaled Permutation Feature Importance from Random Forest (left) and XGBoost (right) Classification (90$^{\text{th}}$ percentile; 2000-2020)

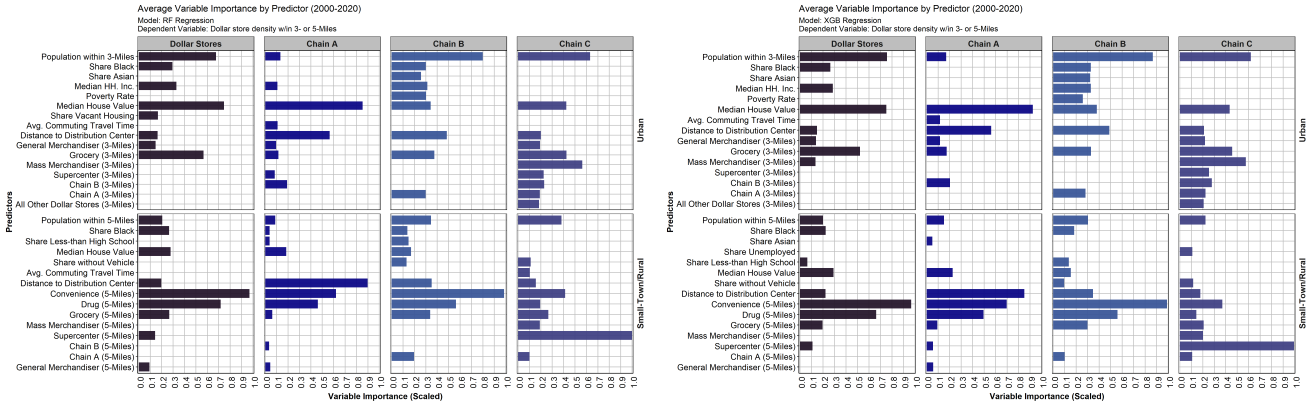

Figure 22: Mean-Scaled Permutation Feature Importance from Random Forest (left) and XGBoost (right) Regression Models (90$^{\text{th}}$ percentile; 2000-2020)